# Modeling and Valuation of Energy Structures

Daniel Mahoney

Analytics, Econometrics, and Numerics

Modeling and Valuation of Energy Structures

**Applied Quantitative Finance series**

*Applied Quantitative Finance* is a new series developed to bring readers the very latest market tested tools, techniques and developments in quantitative finance. Written for practitioners who need to understand how things work "on the floor", the series will deliver the most cutting-edge applications in areas such as asset pricing, risk management and financial derivatives. Although written with practitioners in mind, this series will also appeal to researchers and students who want to see how quantitative finance is applied in practice.

*Also available*

Oliver Brockhaus
EQUITY DERIVATIVES AND HYBRIDS
*Markets, Models and Methods*

Enrico Edoli, Stefano Fiorenzani and Tiziano Vargiolu
OPTIMIZATION METHODS FOR GAS AND POWER MARKETS
*Theory and Cases*

Roland Lichters, Roland Stamm and Donal Gallagher
MODERN DERIVATIVES PRICING AND CREDIT EXPOSURE ANALYSIS
*Theory and Practice of CSA and XVA Pricing, Exposure Simulation and Backtesting*

Zareer Dadachanji
FX BARRIER OPTIONS
*A Comprehensive Guide for Industry Quants*

Ignacio Ruiz
XVA DESKS: A NEW ERA FOR RISK MANAGEMENT
*Understanding, Building and Managing Counterparty and Funding Risk*

Christian Crispoldi, Peter Larkin and Gérald Wigger
SABR AND SABR LIBOR MARKET MODEL IN PRACTICE
*With Examples Implemented in Python*

Adil Reghai
QUANTITATIVE FINANCE
*Back to Basic Principles*

Chris Kenyon and Roland Stamm
DISCOUNTING, LIBOR, CVA AND FUNDING
*Interest Rate and Credit Pricing*

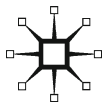
Marc Henrard
INTEREST RATE MODELLING IN THE MULTI-CURVE FRAMEWORK
*Foundations, Evolution and Implementation*

# Modeling and Valuation of Energy Structures

## Analytics, Econometrics, and Numerics

Daniel Mahoney
*Director of Quantitative Analysis, Citigroup, USA*

palgrave
macmillan





Softcover reprint of the hardcover 1st edition 2016 978-1-137-56014-8



First published 2016 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States, the United Kingdom, Europe and other countries.

ISBN 978-1-349-56688-4 ISBN 978-1-137-56015-5 (eBook)
DOI 10.1007/978-1-137-56015-5

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

*To Cathy, Maddie, and Jack*

This page intentionally left blank

# Contents

# List of Figures

# List of Tables

# Preface

Energy markets (and commodity markets in general) present a number of challenges for quantitative modeling. High volatilities, small sample sizes, structural market changes, and operational complexity all make it very difficult to straightforwardly apply standard methods to the valuation and hedging of products that are commonly encountered in energy markets. It cannot be denied that there is an unfortunate tendency to apply, with little skeptical thought, methods widely used in financial (*e.g.*, bond or equity) markets to problems in the energy sector. Generally, there is insufficient appreciation for the trade-off between theoretical sophistication and practical performance. (This problem is compounded by the temptation to resort to, in the face of multiple drivers and physical constraints, computational machinations that give the illusion of information creation through ease of scenario generation *i.e.*, simulation.) The primary challenge of energy modeling is to correctly adapt what *is* correct about these familiar techniques while remaining fully cognizant of their limitations that become particularly acute in energy markets. The present volume is an attempt to perform this task, and consists of both general and specialized facets.

First, it is necessary to say what this book is *not*. We do not attempt to provide a detailed discussion of any energy markets or their commonly transacted products. There exist many other excellent books for this purpose, some of which we note in the text. For completeness and context, we provide a very high-level overview of such markets and products, at least as they appear in the United States for natural gas and electricity. However, we assume that the reader has sufficient experience in this industry to understand the basics of the prevailing market structures. (If you think a toll is just a fee you pay when you drive on the highway, this is probably not the right book for you.) Furthermore, this is not a book for people, regardless of existing technical ability, who are unfamiliar with the basics of financial mathematics, including stochastic calculus and option pricing. Again, to facilitate exposition such concepts will be introduced and summarized as needed. However, it is assumed that the reader has a reasonable grasp of such necessary tools that are commonly presented in, say, first-year computational finance courses. (If your first thought when someone says "Hull" is convex hull, then you probably have not done sufficient background work.)

So, who *is* this book for? In truth, it is aimed at a relatively diverse audience, and we have attempted to structure the book accordingly. The book is aimed at readers with a reasonably advanced technical background who have a good familiarity with

energy trading. Assuming this is not particularly helpful, let us elaborate. Quantitative analysts ("quants") who work on energy-trading desks in support of trading, structuring, and origination and whose job requires modeling, pricing, and hedging natural gas and electricity structures should have interest. Such readers should have the necessary industry background as well as familiarity with mathematical concepts such as stochastic control. In addition, they will be reasonably expected to have analyzed actual data at some point. They presumably have little trepidation in rolling up their sleeves to work out problems or code up algorithms (indeed, they should be eager to do so). For them, this book will (hopefully) present useful approaches that they can use in their jobs, both for statistical work and model development. (As well, risk control analysts and quantitatively oriented traders who must understand, at least at a high level, valuation methodologies can also benefit, at least to a lesser extent.)

Another category of the target audience is students who wish not only to understand more advanced techniques than they are likely to have seen in their introductory coursework, but also to get an introduction to actual traded products and issues associated with their analysis. (More broadly, academics who have the necessary technical expertise but want to see applications in energy markets can also be included here.) These readers will understand such foundational concepts as stochastic calculus, (some) measure theory, and option pricing through replication, as well as knowing how to run a regression if asked. Such readers (at least at the student level) will benefit from seeing advanced material that is not normally collected in one volume (*e.g.*, affine jump diffusions, cointegration, Lévy copulas). They will also receive some context on how these methods should (and should not) be applied to examples actually encountered in the energy industry.

Note that these two broad categories are not necessarily mutually exclusive. There are of course practitioners at different levels of development, and some quants who know enough about tolling or storage, say, to operate or maintain models may want to gain some extra technical competency to understand these models (and their limitations) better. Similarly, experienced students may require little technical tutoring but need to become acquainted with approaches to actual structured products. There can definitely be overlap across classes of readership.

The structure of the book attempts to broadly satisfy these two groups. We divide the exposition into the standard blocks of theory and application; however, we reverse the usual order of presentation and begin with applications before going into more theoretical matters. While this may seem curious at first, there is a method to the madness (and in fact our dichotomy between practice and theory is rather soft, there is overlap throughout). As stated in the opening paragraph, we wish to retain what is correct about most quantitative modeling while avoiding those aspects that are especially ill-suited for energy (and commodity) applications. Broadly speaking, we present valuation of structured products as a replication/decomposition problem, in conjunction with robust estimation (that is, estimation that is not

overly sensitive to the particular sample). We essentially view valuation as a portfolio problem entailing representations in terms of statistical properties (such as variance) that are comparatively stable as opposed to those which are not (such as mean-reversion rates or jump probabilities). By discussing the core econometric and analytical issues first, we can more seamlessly proceed to an overview of valuation of some more popular structures in the industry.

In Part I the reader can thus get an understanding for how and why we choose our particular approaches, as well as see how the approaches manifests themselves. Then, in Part II the more theoretical issues can be investigated with the proper context in mind. (Of course, there is cross-referencing in the text so that the reader can consult certain ideas before returning to the main flow.) Although we advise against unthinkingly applying popular sophisticated methods for their own sake, it is unquestionably important to understand these techniques so as to better grasp why they can break down. Cointegration, for example, is an important and interesting idea, but its practical utility is limited (as are many econometric techniques) by the difficulty of separating signal from noise in small samples. Nonetheless, we show that cointegration has a relationship to variance scaling laws, which *can* be robustly implemented. We thus hope to draw the reader's attention to such connections, as well as provide the means for solving energy market problems.

The organization is as follows. We begin Part I with a (very) brief overview of energy markets (specifically in the United States) and the more common structured products therein. We then discuss the critical econometric issue of time scaling and how it relates to the conventional dichotomy stationarity/non-stationarity and variance accumulation. Next, we present valuation as a portfolio construction problem that is critically dependent on the prevailing market structure (via the availability of hedging instruments). We demonstrate that the gain from trying to represent valuation in terms of the actual *qualitative* properties of the underlying stochastic drivers is typically not enough to offset the costs. Finally we present some valuation examples of the aforementioned structured products.

Part II, as already noted, contains more theoretical material. In a sense, it fills in some of the details that are omitted in Part I. It can (hopefully) be read more profitably with that context already provided. However, large parts of it can also serve as a stand-alone exposition of certain topics (primarily the non-econometric sections). We begin this part with a discussion of (stochastic) process modeling, not for the purposes of valuation as such, but rather to provide a conceptual framework for being able to address the question of *which* qualitative features should be retained (and which features should be ignored) for the purposes of robust valuation. Next we continue with econometric issues, with an eye toward demonstrating that many standard techniques (such as filtering) can easily break down in practice and should be used with great caution (if at all). Then, numerical methods are discussed. The obvious rationale for this topic is that at some point in any problem, actual computations must be carried out, and we go over techniques particularly relevant

for energy problems (*e.g.*, stochastic control and high-dimensional quadrature). Finally, given the key role joint dependencies play in energy markets, we present some relevant ideas (copulas being chief among these).

We should point out that many of the ideas to be presented here are more generally applicable to commodity markets as such, and not simply the subset of energy markets that will be our focus. Ultimately, commodity markets are driven by final (physical) consumption, so many of the characteristics exhibited by energy prices that are crucial for proper valuation of energy structures will be shared by the broader class of commodities (namely, supply-demand constraints and geographical concentration, small samples/high volatilities, and most critically, volatility scaling). We will not provide any specific examples in, say, agriculture or metals, except to note when certain concepts are more widely valid. We will also employ the term "commodity" in a generic, plain language sense. (So, reader beware!)

# Acknowledgments

I would like to thank Alexander Eydeland and an anonymous referee for their helpful comments on earlier drafts of this book. They have helped make this a much-improved product; any remaining flaws and errors are entirely mine. I would also like to thank Piotr Grzywacz, Mike Oddy, Vish Krishnamoorthy, Marcel Stäheli, and Wilson Huynh for many fruitful and spirited discussions on quantitative analysis. I must also express a special intellectual and personal debt to Krzysztof Wolyniec. This book arose from a number of projects we have collaborated on over the years, and could not have come into being without his input and insights. His influence on my thinking about quantitative modeling simply cannot be understated. I would also like to thank Swiss Re for their support, and SNL for their permission to use their historical data.

# 1 Synopsis of Selected Energy Markets and Structures

## 1.1 Challenges of modeling in energy markets

Although it is more than ten years old at the time of this writing, Eydeland and Wolyniec (2003, hereafter denoted by EW) remains unparalleled in its presentation of both practical and theoretical techniques for commodity modeling, as well as its coverage of the core structured products in energy markets.[1] We will defer much discussion of the specifics of these markets to EW, as our focus here is on modeling techniques. However, it will still be useful to highlight some central features of energy markets, to provide the proper context for the subsequent analysis.[2]

### 1.1.1 High volatilities/jumps

Energy markets are characterized by much higher volatilities than those seen in financial or equity markets. Figure 1.1 provides an illustration.

It is worth noting that the general pattern (of higher commodity volatility) has persisted even in the post-crisis era of collapsing volatilities across markets. In large part, this situation reflects the time scales associated with the (physical) supply and demand factors that drive the dynamics of price formation in energy markets. These factors require that certain operational balances be maintained over relatively small time horizons, and that the arrival of new information propagates relatively quickly. Demand is a reflection of overall economic growth as well as stable (so to speak[3]) drivers such as weather. Supply is impacted by the marginal cost of those factors used in the production of the commodity in question. A familiar example is the generation stack in power markets, where very hot or very cold weather can increase demand to sufficiently high levels that very inefficient (expensive) units must be brought online.[4] See Figure 1.2. for a typical example.

The presence of high volatilities makes the problem of extracting useful information from available data much more challenging, as it becomes harder to distinguish signal from noise (in a sample of a given size). This situation is further exacerbated by the fact that, in comparison to other markets, we often do not have much data to analyze in the first place.

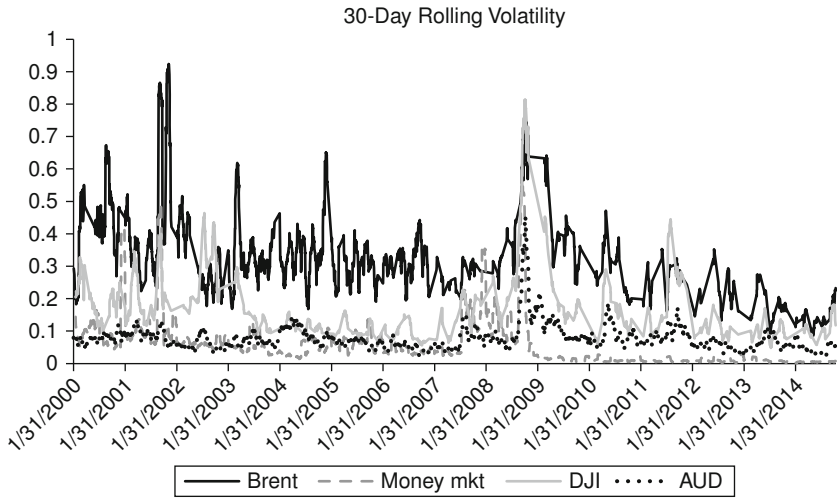



**Figure 1.1** Comparison of volatilities across asset classes. Resp. Brent crude oil (spot), Federal funds rate, Dow Jones industrial average, and Australian dollar/US dollar exchange rate.
*Source*: quandl.com.

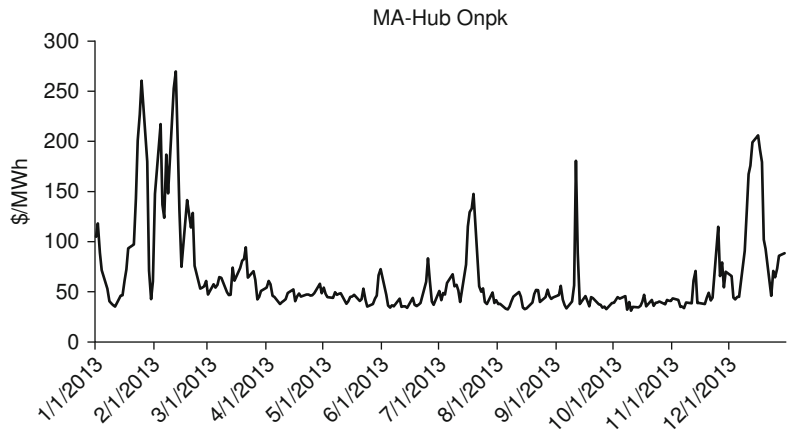



**Figure 1.2** Spot electricity prices.
*Source*: New England ISO (www.iso-ne.com).

### 1.1.2 Small samples

The amount of data, both in terms of size and relevance, available for statistical and econometric analysis in energy markets is much smaller than that which exists in other markets. For example, some stock market and interest rate data go back to the early part of the 20th century. Useful energy data may only go back to the 1980s at best.[5] This situation is due to a number of factors.

Commodity markets in general (and especially energy markets) have traditionally been heavily regulated (if not outright monopolized) entities (*e.g.*, utilities) and have only relatively recently become sufficiently open where useful price histories and time series can be collected.[6] In addition (and related to prevailing and historical regulatory structures), energy markets are characterized by geographical particularities that are generally absent from financial or equity markets. A typical energy deal does not entail exposure to natural gas (say) as such, but rather exposure to natural gas in a specific physical location, *e.g.* the Rockies or the U.S. Northeast.[7] Certain locations possess longer price series than others.

Finally, and perhaps most importantly, we must make a distinction between spot and futures/forward[8] prices. Since spot commodities are not traded as such (physical possession must be taken), trading strategies (which, as we will see, form the backbone of valuation) must be done in terms of futures. The typical situation we face in energy markets is that for most locations of interest, there is either much less futures data than spot, or there is no futures data at all. The latter case is invariably associated with illiquid physical locations that do not trade on a forward basis. These include many natural gas basis locations or nodes in the electricity generation system. However, even for the liquidly traded locations (such as Henry Hub natural gas or PJM-W power), there is usually a good deal more spot data than futures data, especially for longer times-to-maturity.

### 1.1.3 Structural change

Along with the relatively recent opening up of energy markets (in comparison to say, equity markets), has come comparatively faster structural change in these markets. It is well beyond the scope of this book to cover these developments in any kind of detail. We will simply note some of the more prominent ones to illustrate the point:

- the construction of the Rockies Express (REX) natural gas pipeline, bringing Rockies gas into the Midwest and Eastern United States (2007–09)
- the so-called shale revolution in extracting both crude oil and natural gas (associated with North Dakota [Bakken] and Marcellus, respectively; 2010–present)
- the transition of western (CAISO) and Texas (ERCOT) power markets from bilateral/zonal markets to LMP/nodal markets (as prevail in the East; 2009–2010).

These developments have all had major impacts on price formation and dynamics and, as a result, on volatility. In addition, although not falling under the category of structural change as such, macro events such as the financial crisis of 2008 (leading to a collapse in commodity volatility and demand destruction) and regulatory/political factors such as Dodd-Frank (implemented after the Enron scandal in the early 2000s and affecting various kinds of market participants) have amounted

to kinds of regime shifts (so to speak) in their own right. The overall situation has had the effect of exacerbating the aforementioned data sparseness issues. The (relatively) small data that we have is often effectively truncated even more (if not rendered somewhat useless) by structural changes that preclude the past from providing any kind of guidance to the future.

### 1.1.4 Physical/operational constraints

Finally, we note that many (if not most) of the structures of interest in energy markets are heavily impacted by certain physical and operational constraints. Some of these are fairly simple, such as fuel losses associated with flowing natural gas from a production region to a consumer region, or into and out of storage. Others are far more complex, such as the operation of a power plant, with dispatch schedules that depend on fuel costs from (potentially) multiple fuel sources, response curves (heat rates) that are in general a function of the level of generation, and fixed (start-up) costs whose avoidance may require running the plant during unprofitable periods.[9,10] Some involve the importance of time scales (a central theme of our subsequent discussion), which impact how we project risk factors of interest (such as how far industrial load can move against us over the time horizon in question).[11]

In general, these constraints require optimization over a very complex set of operational states, while taking into account the equally complex (to say nothing of unknown!) stochastic dynamics of multiple drivers. A large part of the challenge of valuing such structures is determining how much operational flexibility must be accounted for. Put differently, which details can be ignored for purposes of valuation? This amounts to understanding the *incremental* contribution to value made by a particular operational facet. In other words, there is a balance to be struck between how much detail is captured, and how much value can be reasonably expected to be gained. It is better to have approximations that are robust given the data available, than to have precise models which depend on information we cannot realistically expect to extract.

## 1.2 Characteristic structured products

Here we will provide brief (but adequately detailed) descriptions of some of the more popular structured products encountered in energy markets. Again, EW should be consulted for greater details.

### 1.2.1 Tolling arrangements

Tolling deals are, in essence, associated with the spread between power prices and fuel prices. The embedded optionality in such deals is the ability to run the plant

(say, either starting up or shutting down) only when profitable. The very simplest form a tolling agreement takes is a so-called spark spread option, with payoff given by

$$(P_T - H \cdot G_T - K)^+ \tag{1.1}$$

with the obvious interpretation of $P$ as a power price and $G$ as a gas price (and of course $x^+ \equiv \max(x, 0)$. The parameters $H$ and $K$ can be thought of as corresponding to certain operational costs, specifically a heat rate and variable operation and maintenance (VOM), respectively[12] The parameter $T$ represents an expiration or exercise time. (All of the deals we will consider have a critical time horizon component.)

Of course, tolling agreements usually possess far greater operational detail than reflected in (1.1). A power plant typically entails a volume-independent cost for starting up (that is, the cost is denominated in dollars, and not dollars per unit of generation),[13] and possibly such a cost for shutting down. Such (fixed) costs have an important impact on operational decisions; it may be preferable to leave the plant on during uneconomic periods (*e.g.*, overnight) so as to avoid start-up costs during profitable periods (*e.g.*, weekdays during business hours). In general, the pattern of power prices differs by temporal block, *e.g.*, on-peak vs. off-peak. In fact, dispatch decisions can be made at an hourly resolution, a level at which no market instruments settle (a situation we will see also prevails for load following deals). There are other complications. Once up, a plant may be required to operate at some (minimum) level of generation. The rate at which fuel is converted to electricity will in general be dependent on generation level (as well as a host of other factors that are typically ignored). Some plants can also operate using multiple fuel types. There may also be limits on how many hours in a period the unit can run, or how many start-ups it can incur. Finally, the very real possibility that a unit may fail to start or fail to operate at full capacity (outages and derates, resp.) must be accounted for.

The operational complexity of a tolling agreement can be quite large, even when the contract is tailored for financial settlement. It remains the case, however, that the primary driver of value is the codependence of power and fuel and basic spread structures such as (1.1). The challenge we face in valuing tolling deals (or really any other deal with much physical optionality) is integrating this operational flexibility with available market instruments that, by their nature, do not align perfectly with this flexibility. We will see examples in later chapters, but our general theme will always be that it is better to find robust approximations that bound the value from below,[14] than to try to perform a full optimization of the problem, which imposes enormous informational requirements that simply cannot be met in practice. Put differently, we ask: how much operational structure must we include in order to represent value in terms of both market information and entities (such as realized volatility or correlation) that can be robustly estimated? Part of our objective here is to answer this question.

### 1.2.2 Gas transport

The characteristic feature of natural gas logistics is flow from regions where gas is produced to regions where it is consumed. For example, in the United States this could entail flow from the Rockies to California or from the Gulf Coast to the Northeast. The associated optionality is the ability to turn off the flow when the spread between delivery and receipt points is negative. There are, in general, (variable) commodity charges (on both the receipt and delivery ends), as well as fuel losses along the pipe. The payoff function in this case can be written

$$\left(D_T - \frac{1}{1-f}R_T - K\right)^+ \tag{1.2}$$

where $R$ and $D$ denote receipt and delivery prices respectively, $K$ is the (net) commodity charge, and $f$ is the fuel loss (typically small, in the 1–3% range).[15] Although transport is by far the simplest[16] structure we will come across in this book, there are some subtleties worth pointing out.

In U.S. natural gas markets, most gas locations trade as an offset (either positive or negative) to a primary (backbone or hub) point (NYMEX Henry Hub). This offset is referred to as the basis. In other words, a leg (so to speak) price $L$ can be written as $L = N + B$ where $N$ is the hub price and $B$ is the basis price. Thus, transacting (forward) basis locks in exposure relative to the hub; locking in total exposure requires transacting the hub, as well. Note that (1.2) can be written in terms of basis as

$$\left(B_T^D - \frac{1}{1-f}B_T^R - \frac{f}{1-f}N_T - K\right)^+ \tag{1.3}$$

Thus, if there are no fuel losses ($f = 0$), the transport option has *no* hub dependence. Hence, the transport spread can be locked in by trading in basis points *only*. Alternatively, (1.3) can be written as

$$\left(B_T^D - B_T^R - K - \frac{f}{1-f}R_T\right)^+ \approx (B_T^D - B_T^R - K)^+ - \frac{f}{1-f}R_T \cdot H(B_T^D - B_T^R - K) \tag{1.4}$$

We thus see that transport options are essentially options on a basis spread, and not a price spread as such. (Mathematically, we might say that a Gaussian model is more appropriate than a lognormal model.) Decomposing the payoff structure as in (1.4) we see that the optionality consists of both a regular option and a digital option, as well. We emphasize these points because they illustrate another basic theme here: market structure is critical for proper valuation of a product. Looking at leg prices can be misleading because in general (depending on the time horizon) the hub is far more volatile than basis. Variability in the leg often simply reflects variability in the hub. This is of course a manifestation of differences in liquidity,

which as we will see is a critical factor in valuation. For transport deals with no (or small) fuel costs, hedging (which is central to valuation through replication) will be conducted purely through basis, and care must be taken to not attribute value to hub variability.[17] These points are illustrated in Figure 1.3.[18] The implications here concern not simply modeling but (more importantly) the identification of the relevant exposure that arises from hedging and trading around such structures.

### 1.2.3 Gas storage

Another common gas-dependent structure is storage. Due to seasonal (weather-driven) demand patterns, it is economically feasible to buy gas in the summer (when it is relatively cheap), physically store it, and sell it in the winter (when it is relatively expensive). The embedded optionality of storage is thus a seasonal spread option:

$$\left((1-f_{wdr})P_T^{wdr} - \frac{1}{1-f_{inj}}P_T^{inj} - K\right)^+ \tag{1.5}$$

As with transport, there are typically fuel losses (on both injection and withdrawal), as well as (variable) commodity charges (on both ends, aggregated as $K$ in (1.5). However, unlike transport, there is no common backbone or hub involved in the spread in (1.5), and the underlying variability is between leg prices (for different temporal flows[19]).

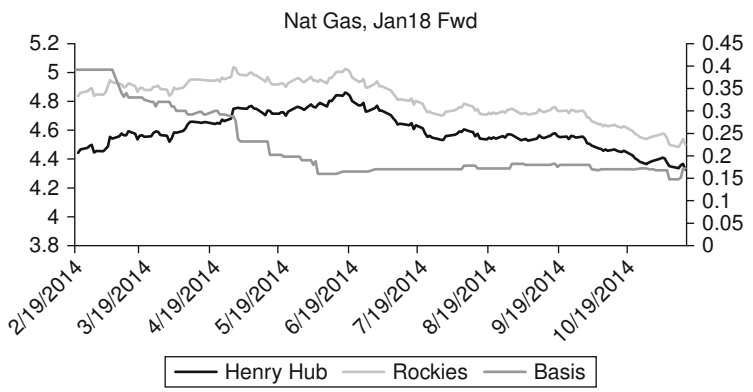



**Figure 1.3** Comparison of basis, leg, and backbone. The (long-dated) Rockies (all-in) leg price clearly covaries with the benchmark Henry Hub price, but in fact Rockies (like most U.S. natural gas points) trades as an offset (basis) to Henry Hub. This basis is typically less liquid than the hub (esp. for longer times-to-maturity), hence the co-movement of Rockies with hub is due largely to the hub moving, and *not* because of the presence of a common driver (stochastic or otherwise). *Source*: quandl.com.

One may think of the expression in (1.5) as generically representing the seasonal structure of storage. More abstractly, storage embodies a so-called stochastic control problem, where valuation amounts to (optimally) choosing how to flow gas in and out of the facility over time:

$$-\int_t^T q_s(\mathfrak{f}(q_s, Q_s)S_s + \mathfrak{c}(q_s, Q_s))d_s, \dot{Q} = q \tag{1.6}$$

where $q$ denotes a flow rate (negative for withdrawals, positive for injections), $Q$ is the inventory level, $S$ is a spot price, and $\mathfrak{f}$ and $\mathfrak{c}$ are (action- and state-dependent) fuel and commodity costs, respectively. A natural question arises. The formulations of the payoffs in (1.5) and (1.6) appear to be very different; do they in fact represent very different approaches to valuation, or are they somehow related? As we will see in the course of our discussion, there is in fact a connection. The formulation in (1.5) can best be understood in terms of traded (monthly) contracts that can be used to lock in value through seasonal spreads, and in fact more generally through monthly optionality that can be captured as positions are rebalanced in light of changing price spreads (*e.g.*, a Dec–Jun spread may become more profitable than a Jan–Jul spread). In fact, once monthly volumes have been committed to, one is always free to conduct spot injections/withdrawals. We will see that the question of relating the two approaches (forward-based vs. spot-based) comes down to a question of market resolution (or more accurately the resolution of traded instruments). Put roughly, as the resolution of contracts becomes finer (*e.g.*, down to the level of specific days within a month), the closer the two paradigms will come.

As with tolling, there can be considerable operational constraints with storage that must be satisfied. The most basic form these constraints take are maximum injection and withdrawal rates. These are typically specified at the daily level, but they could apply over other periods as well, such as months. Other volumetric constraints are inventory requirements; for example, it may be required that a facility be completely full by the end of October (*i.e.*, you cannot wait until November to fill it up) or that it be at least 10% full by the end of February (*i.e.*, you cannot completely empty it before March). These kinds of constraints are actually not too hard to account for. A bit more challenging are so-called ratchets, which are volume-dependent flow rates (for injection and/or withdrawal). For example, an injection rate may be 10,000 MMBtu/day until the unit becomes half full, at which point the injection rate drops to 8,000 MMBtu/day. We will see that robust lower bound valuations can be obtained by crafting a linear programming problem in terms of spread options such as (1.5). The complications induced by ratchets effectively render the optimization problem nonlinear. As we stated with tolling, our objective will be to understand how much operational detail is necessary for robust valuation.

### 1.2.4 Load serving

The final structured product we will illustrate here differs from those we have just considered in that it does not entail explicit spread optionality. Load-serving deals (also known as full requirements deals) are, as the name suggests, agreements to serve the electricity demand (load) in a particular region for a particular period of time at some fixed price. The central feature here is volumetric risk: demand must be served at every hour of every day of the contract period, but power typically only trades in flat volumes for the on- and off-peak blocks of the constituent months. (Load does not trade at all.) Hedging with (flat) futures generally leaves one under-hedged during periods of higher demand (when prices are also generally higher) and over-hedged during periods of lower demand (when prices are also generally lower).

Of obvious interest is the cost-to-serve, which is simply price multiplied by load.[20] On an expected value basis, we have the following useful decomposition:

$$\begin{aligned} E_t L_{T'} P_{T'} &= E_t E_T (L_{T'} - E_T L_{T'} + E_T L_{T'})(P_{T'} - E_T P_{T'} + E_T P_{T'}) \\ &= E_t \left[ E_T (L_{T'} - E_T L_{T'})(P_{T'} - E_T P_{T'}) + E_T L_{T'} \cdot E_T P_{T'} \right] \end{aligned} \tag{1.7}$$

Alternatively, we can write

$$\begin{aligned} &E_t (L_{T'} - E_t L_{T'})(P_{T'} - E_t P_{T'}) \\ &\quad = E_t E_T (L_{T'} - E_T L_{T'} + E_T L_{T'} - E_t L_{T'})(P_{T'} - E_T P_{T'} + E_T P_{T'} - E_t P_{T'}) \\ &\quad = E_t \left[ E_T (L_{T'} - E_T L_{T'})(P_{T'} - E_T P_{T'}) + (E_T L_{T'} - E_t L_{T'})(E_T P_{T'} - E_t P_{T'}) \right] \end{aligned} \tag{1.8}$$

In the expressions (1.7) and (1.8), $t$ is the current time, $T'$ is a representative time within the term (say, middle of a month), and $T$ is a representative intermediate time (say, beginning of a month). These decompositions express the expected value of the cost-to-serve, conditioned on current information, in terms of expected values conditioned on intermediate information. For example, from (1.7), we see that the expected daily cost-to-serve (given current information) is the expected monthly cost-to-serve $E_t [E_T L_{T'} \cdot E_T P_{T'}]$ plus a cash covariance term $E_t [E_T (L_{T'} - E_T L_{T'})(P_{T'} - E_T P_{T'})]$. (By cash we mean intra-month [say], conditional on information prior to the start of the monthly.) This decomposition is useful because we often have market-supplied information over these separate time horizons (*e.g.*, monthly vs. cash) that can be used for both hedging and information conditioning. (A standard approach is to separate a daily volatility into monthly and cash components.)

It is helpful to see the role of the covariance terms from a portfolio perspective. Recall that the deal consists of a fixed price $P_X$ (payment received for serving the

load), and assume we put on a (flat) price hedge (with forward price $P_F$) at expected load ($\overline{L}$):

$$\Pi = -L_{T'}P_{T'} + L_{T'}P_X + \overline{L}(P_{T'} - P_F) = P_F\overline{L}\left(\frac{L_{T'}}{\overline{L}} - 1\right)\left(\frac{P_{T'}}{P_F} - 1\right) + (P_X - P_F)L_{T'} \tag{1.9}$$

Since changes in (expected) price and load and typically co-move, we see from (1.9) that the remaining risk entails both over- and under-hedging (as already noted). The larger point to be made, however, is that in many deals the (relative) covariation contribution $\left(\frac{L_{T'}}{\overline{L}} - 1\right)\left(\frac{P_{T'}}{P_F} - 1\right)$ covaries with realized price volatility.[21] (This behavior is typically seen in industrial or commercial load deals [as opposed to residential]). Thus, an option/vega hedge (*i.e.*, an instrument that depends on realized volatility) can be included in the portfolio. As such, this relative covariation is not a population entity, but rather a *pathwise* entity. The fixed price $P_X$ must then be chosen to not only finance the purchase of these option positions, but to account for residual risk, as well. Of course, this argument assumes that power volatility trades in the market in question; this is actually often not the case, as we will see shortly. However, in many situations one has load deals as part of a larger portfolio that includes tolling positions, as well, the latter of which have a "natural" vega component, so to speak. There is thus the possibility of exploiting intra-desk synergy between the two structures, and indeed, without complementary tolling positions load following by itself is not a particularly viable business (unless one has complementary physical assets such as generation [*e.g.*, as with utilities]).[22] What we see here is a theme we will continue to develop throughout this book: the notion of valuation as a portfolio construction problem.

A final point we should raise here is the question of expected load. As already noted, load does not trade, so not only can load *not* be hedged, there are no forward markets whose prices can be used as any kind of projection of load. Thus, we must always perform some estimation of load. We will begin discussing econometric issues in Chapter 2 (and further in Chapter 6), but we wish to note here two points. First, the conditional decomposition between monthly and cash projections proves quite useful for approaching the problem econometrically. We often have a good understanding of load on a monthly basis, and can then form estimates conditional on these monthly levels (*e.g.*, cash variances,[23] *etc.*). Furthermore, we may be able to reckon certain intra-month (cash) properties of load robustly by conditioning on monthly levels. Second, load is an interesting example of how certain time scales (a central theme in this work) come into play. Some loads (residential) have a distinct seasonal structure, as they are driven primarily by weather-dependent demand. After such effects are accounted for, there is a certain residual structure whose informational content is a function of the time horizon in question. Currently, high or low demand relative to "normal" levels will generally affect our projections of future levels (again, relative to normal) inversely with time horizon.[24] On the other hand,

other load types (industrial or commercial) generally do not display sharp seasonal patterns, and are dominated by the responsiveness of customers to price and switching of providers (so-called migration). These loads (perhaps not surprisingly) have statistical properties more reminiscent of financial or economic time series such as GDP.[25] The informational content of such load observations accumulates directly with time horizon. We will give more precise meaning to these notions in Chapter 2.

## 1.3 Prelude to robust valuation

In this brief overview of energy markets and structures, we have already managed to introduce a number of important concepts that will receive fuller exposition in due course. Chief among these are the following facts:

- The perfect storm of small data sets, high volatility, structural change, and operational complexity make it imperative that modeling and analysis properly balance costs and benefits.
- Structured products do not appear *ab initio* but always exist within the context of a certain (energy) market framework that only permits particular portfolios to be formed around those structures.

These points are actually not unrelated. The fact that we have only specific market instruments available to us means that we can approach a given structured product from the point of view of replication or relative valuation, which of course means a specific kind of portfolio formation. Since different portfolios create different kinds of exposure, it behooves us to identify those portfolios whose resulting exposure entails risks we are most comfortable with.

More accurately, these risks are *residual* risks, *i.e.*, the risks remaining *after* some hedges have been put on.[26] Portfolio constructs or hedging strategies that require information that cannot be reliably obtained from the available data are not particularly useful, and must be strenuously avoided. We will have much more to say about valuation as portfolio formation in Chapter 3. Before that, we will first turn to the precursor of the valuation problem, namely the identification of entities whose estimation can be robustly performed given the data constraints we inevitably face in energy markets.

# 2 Data Analysis and Statistical Issues

## 2.1 Stationary vs. non-stationary processes

### 2.1.1 Concepts

*2.1.1.1 Essential issues, as seen through proxy hedging*

Let us start with an example that is very simple, yet illustrates well both the kind of econometric problems faced in energy markets as well as the essential features the econometric analysis must address. Invariably, we are not interested in estimation as such, but only within the context of valuing some structured product/deal. Suppose we are to bid on a deal that entails taking exposure to some (non-traded) entity $y$, which we believe (for whatever reason) to stand in a relation to some (forward traded) entity $x$. We thus anticipate putting on some kind of hedge with this entity $x$. A critical aspect of this deal is that the exposure is realized at a specific time in the future, denoted by $T$. Examples of proxy hedging include:

- Hedging short-term exposure at a physical node in PJM with PJM-W
- Hedging very long-term PJM-W exposure with Henry Hub natural gas
- Hedging illiquid Iroquois gas basis with more liquid Algonquin basis.

To determine the price $K$ at which we would be willing to assume such exposure, we consider (as we have throughout) the resulting portfolio:[1]

$$\Pi = y_T - K - \Delta(x_T - F_x) = y_T - \Delta \cdot x_T + \Delta \cdot F_x - K \tag{2.1}$$

where $F_x$ is the forward price of $x$. The first thing to note from (2.1) is that we must be concerned with the *residual* exposure that results from the relationship between the exposure $y$ and the hedge $x$. So, it would be sensible to try to estimate this relationship, and a natural (and common) first attempt is to look for a linear relationship:

$$y_T = \alpha x_T + \beta + \varepsilon_T \tag{2.2}$$

where $\varepsilon_T$ is some (as yet unspecified) zero-mean random variable, commonly referred to as the relationship disturbance (or statistical error).[2] In the course of this chapter (and Chapter 6) we will discuss various techniques (and their associated assumptions/limitations) for estimating models more general than (2.2). In addition, we must have some understanding of how the estimation procedure relates to some property (or properties) of the assumed model (and hence the underlying model parameters). In other words, we must be able to relate the sample to the population. In fact, this issue is closely related to (but ultimately distinct from) the question of how strongly the statistical evidence supports the assumption of the purported relationship. We will fill in these details shortly; our concern here is with what information we require for the valuation problem at hand.

#### *2.1.1.2 The requirements: relationships and residuals*

If the estimation did provide evidence of a strong relationship between $x$ and $y$, then the next step would be clear: take the hedge to be $\Delta = \hat{\alpha}$ and $K = \hat{\alpha} F_x + \hat{\beta} + \phi_\varepsilon$ (where hats denote estimated entities), and $\phi_\varepsilon$ denotes a risk adjustment based on the estimated disturbance (*e.g.*, we may want coverage at the 25th percentile, such that we would only lose money 25% of the time[3]). This latter point (concerning risk adjustment) is critical, because we now consider the case where the econometrics does *not* provide evidence of a strong relationship. A natural question to ask is: should we still form our price in light of the estimated (weak) relationship, or disregard the econometrics and bid based on unconditional information? The answer to this question is actually not obvious. There may well be situations where we have *a priori* reasons to believe that there *is* a relationship between the exposure and the available hedging instrument, yet the formal diagnostics (meaning tests of statistical significance of the assumed relationship) prove inconclusive (or possibly offer rejection). (We note in passing that liquid futures markets are known to be efficient processors of information.) In such cases, it may well be the case that the estimated disturbances from (2.2), although (formally) weak evidentially, *do* provide useful information that can be combined with *conditional* data (namely, the current state of the market, *e.g.*, through forward prices). This issue will become much more apparent when we discuss the impact that small samples have on such assessments, because it is precisely in this all-too-common situation that formal diagnostics can be especially misleading. We must always seek a balance between what the data itself says, and what relationships can be exploited via liquid (futures) markets.

To make these points more abstractly, write (2.1) in the following way:

$$\Pi = y_T - K - \Delta(x_T - F_x) = \varepsilon_{xy}(\Delta; T) + \Delta \cdot F_x - K \tag{2.3}$$

where $\varepsilon_{xy}$ denotes the (realized) residuals from a particular hedge $\Delta$ over a particular time horizon $T$. Obviously, this residual is simply the stochastic variable $y_T - \Delta \cdot x_T$, but written in the form (2.3), we see clearly the fact that the entity

of interest, namely, an actual portfolio, does not *necessarily* depend on the existence of a relationship (linear or otherwise) between $x$ and $y$ (as in (2.2)).[4] Rather, the critical dependence is on the hedge volume *and* the time horizon in question for which the resulting exposure in (2.3) is deemed preferable to a naked position. That is to say, the econometric question of interest is not simply estimation as such, but rather what kind of (historical) information can be exploited to construct portfolios around some structured product. Now, formal econometric techniques can still be very useful in this task (and we will devote time to discussing some of these techniques), but their limitations must always be kept in mind (and subordinated to the larger goal of robust valuation).

We thus see that even in this very simple example (which is not at all contrived, it is a very common situation in energy markets that longer-term deals at illiquid, physical locations cannot be directly hedged using existing market instruments, hence some sort of proxy ("dirty") hedge *must* be employed (thus rendering even the notion of intrinsic value illusory, as some residual risk must be worn)) that a number of subtle econometric issues are involved, that are often given short shrift in many standard treatments of the subject. Much of what we have said to this point is a cautionary tale about conventional statistical techniques needing to be congruent with the primary goals of robust valuation, which itself *always* takes place in the context of an actual market environment (with a specific level of liquidity, across both time horizons and products). We will see examples where some of these techniques break down surprisingly quickly in fairly simple problems. The point we wish to focus on here, however, is the role that *time scales* play in identifying the salient features of a problem, features that *must* be adequately captured by an econometric analysis, even (especially?) if it means ignoring other, (ostensibly) more complex features.[5]

#### *2.1.1.3 Formal definitions and central principles*

To this end, we start with a description of certain general categories that serve to delineate the kinds of (stochastic) time series for which various techniques must be brought to bear.[6] The broadest category here is that of stationary vs. non-stationary time series. A stationary time series is essentially one for which the statistics of future states do not depend on the current time. More accurately, the joint distribution of future states depends only on the relative times of those states, and not absolute times, *i.e.*,

$$\Pr(x_{t+\tau_1}, x_{t+\tau_2}, \ldots x_{t+\tau_n}) = \Pr(x_{\tau_1}, x_{\tau_2}, \ldots x_{\tau_n}) \tag{2.4}$$

This time invariance means, for example, that it is meaningful to speak of *unconditional* statistics such as the mean and variance. Examples of stationary time series include white noise, physical processes such as (deseasonalized) temperature, and economic entities such as heat rates. (Mean reversion is commonly associated

with stationarity, and while they are not unrelated, they are not really the same concept either.) Not surprisingly, a non-stationary time series is one that is not stationary; in other words, its statistics are not time invariant and there is no meaningful sense in which we can refer to unconditional mean and variance (say). (Only conditional statistics can be formed.) Examples include Brownian motion and financial or economic time series such as GDP, inflation rates, and prices in general. Note, of course, that non-stationary time series can often be analyzed by looking at constituent series that are stationary, *e.g.*, (log-) returns in GBM. (Heuristically speaking, a stationary time series cannot "wander" too far off over a long enough time horizon, whereas a non-stationary time series can; we will render these common-language notions more precise when we discuss variance scaling laws in Section 2.2.)

There are various categories of stationarity (*e.g.*, covariance stationarity), the distinction between which we will not dwell on here. The critical point (a theme we have emphasized throughout) concerns the nature of information flows, *i.e.*, volatility. For a stationary time series, there are unconditional or time-invariant properties (*e.g.*, a long-term mean) which limit the extent to which *current* information (or more accurately, a *change* in current information) is relevant for projecting future states of the process in question. In other words, the incremental benefit (so to speak) of new information is of declining value as the distance to those future states increases. By contrast, there are no such restrictions on a non-stationary process and new information is of great relevance for projecting future states; the incremental value of new information tends to be uniform over the time horizon in question. To illustrate, a hot day today means that tomorrow will probably be hot, as well. (Indeed, notions of "hot" or "cold" only make sense in reference to some "normal" – *i.e.*, time-invariant – level.) However, there is little reason to think that next *month's* temperature will be very different from normal.[7] In distinction, our projection of what next quarter's or next year's GDP will be depends very largely on where GDP is right now. There is no long-term level to which we could refer in making such projections.

Note that we have introduced a very important concept in this discussion, namely the crucial notion of *time scales*.[8] In truth, what is important is not whether a process is distributionally time-invariant as such, but rather the time scales over which information relevant to that process accumulates. We already indicated this when we mentioned that, over a small enough time horizon, (changes in) current information about temperature *is* important. For prices the situation is a bit more complicated, and we do not intend to discuss in detail the issue of whether there is (formal) mean reversion in commodity markets (see EW for a fuller exposition). We will see, though, that there is definite evidence for nonuniformity of time scales for information flows in energy markets (and commodity markets in general). Indeed, there is no doubt of the existence of the so-called Samuelson effect in commodity markets, namely the term structure of volatility. While this effect is commonly

associated with mean reversion of the underlying spot process,[9] this does not necessarily have to be the driving force. We will return to these points in great detail later, but they should be kept in mind as we discuss conventional techniques that rely on a rather strict delineation between stationary and non-stationary processes.

#### *2.1.1.4 Statistical significance: why it matters (and why it does not)*

This distinction between stationary and non-stationary processes is important for a number of reasons. One reason of course is that the projection of the future value of some random variable is very different when there is a (long-term) unconditional distribution to which reference can be made, and when there is not. Another important reason is that most common econometric methods are only valid when the underlying time series to which they are applied are stationary.[10] However, most time series encountered in actual markets are, to a large degree, non-stationary.[11] When such familiar methods are applied outside their range of validity, extremely misleading results can arise. On top of this, many of these techniques only provide diagnostics in an *asymptotic* sense, *i.e.*, in the limiting case of very large samples. Even in the case of stationarity, depending on the operative time scales, these asymptotic results may be of little practical value, and may in fact generate erroneous conclusions.

It needs to be stressed that estimation *must* be accompanied by some means of carrying out diagnostics (which ultimately are a way of detecting whether the exercise is simply fitting to noise or not). By this we mean assessing the results for sensitivity to noise (that is, determining whether they are simply an artifact of the sample in question or whether they are robust to other realizations of [non-structural] noise [and hence indicative of some actual *population* property]). In standard terminology this can be thought of as statistical significance, but as will be seen we mean it in a more general sense. (It should be noted that common measures of estimation *quality*, such as goodness-of-fit [*e.g.*, *R*-squared values in ordinary least squares to be considered below], while useful for some purposes, are *not* diagnostics or tests of significance.) Recall our prime objective: valuation through replication via portfolio formation, plus appropriate accounting for exposure that cannot be replicated (residual risk). This objective requires that we put on *specific* positions (in some hedging instruments), and charge for residual risk at a *specific* level of (remaining) exposure. *Both* of these aspects of the valuation problem rely on two sides of the same econometric coin, namely the distinction between structure and noise (or more, accurately, the need to be able to distinguish between the two). We will see exactly such an example in Section 6.1.1, when spurious regressions based on unrelated, non-stationary time series are considered: there is no meaningful distinction between structure and noise, *no matter how large the sample.*

A very broadly encompassing example is finding relationships between so-called value drivers (see Chapter 3) that cannot be hedged and other variables which can

be (recall the example in (2.1)). A specific example would be a (pathwise) relationship between price and load convexity in full requirements deals[12] and realized price quadratic variation. Since the latter entity can (at least in certain markets) be vega-hedged with options, knowing a relationship between convexity (which is the relevant exposure in structured deals such as full requirements/load serving) and quadratic variation amounts to being able to hedge, and hence value, the former. Thus we must be able to determine two things from any estimation procedure:

1. Does it reveal a real relationship that we can exploit?
2. Are the resulting residuals (estimation errors) indicative of the kinds of remaining exposures/risks we face after the relationship is accounted for?

To dispel any potential misunderstanding, we should emphasize that our principal objective here is *not* hypothesis testing of models as such, although the formal language may often be employed. Rather, the concern is with the *conditional* information that can be provided by econometric analysis. In this sense, while points 1 and 2 above are both important (and related), it is really point 2 that is of greater relevance. This central point must be stressed here. For the purposes of pricing and hedging (which as we have endeavored to show throughout, are isomorphic, so to speak), we seldom care *only* about explicit estimations such as (2.2). What we also care about is residual exposure, as manifested in actual portfolios such as (2.1) or (2.3). This is precisely why formal hypothesis testing (and more generally tests of statistical significance) is tangential to our primary concern. Hypothesis testing is, ultimately, a binary procedure: either a hypothesized relationship is deemed true (at some specified level of significance), or it is not. By contrast, the formation of optimal (in the appropriate sense) portfolios is a continuous problem across hedge volumes and un-hedged risk (in terms of some distributional property, say, percentiles). It is not obvious how these two problems relate to one another (if at all), but one thing is certain: they are *not* equivalent, and the former issue must be subordinate to the latter issue for purposes of valuation.

The ultimate question is: do the estimated residuals (that arise from any econometric procedure) provide useful information about realized exposures *after* suitable hedges are implemented? Again, it is difficult to provide any kind of definitive answer to this question in light of standard econometric practices (such as hypothesis testing for formal parameter estimates). Doubtless, confirmatory econometric information is useful to have. However, it must *always* be kept in mind that our chief concern in valuation is having a good understanding of residual error, because valuation takes place in the context of a portfolio where one exposure is exchanged for another. Thus, whenever we speak of diagnostics here, we are principally concerned with whether the resulting residuals have informational content, and *not*

whether a particular estimated relationship takes one set of values or another. It remains the case that one must still understand the underlying mechanics, so at the risk of misdirecting the proper emphasis, we will continue to speak in terms of the conventional language of hypothesis testing/diagnostics and pursue the appropriate analysis. Clearly, inferring whether a relationship exists can be useful (although not always decisive) in projecting the nature of residuals that a particular portfolio is exposed to.

#### *2.1.1.5 A quick recap*

After this somewhat lengthy discourse, one may reasonably ask: what does all of that have to do with the topic of stationary vs. non-stationary time series? Precisely the point is that, as assumptions of stationarity (under which many common estimation techniques are constructed) are relaxed, familiar diagnostics may be very misleading, if not completely wrong. As we will formalize presently, estimators are maps from a realization of some random variable to an estimate of some property of that random variable. As such, estimators inherit, if only indirectly, certain statistical features of that variable. As the assumptions under which the estimator is known to relate sample entities to population entities are weakened, it correspondingly becomes less valid to use that estimator to draw conclusions. (We remind the reader again about the time dependence in (2.3), an aspect of the problem that is usually only implicitly accounted for [if at all].) In other words, potentially disastrous inferences can be drawn from common statistical tests if great care is not exercised.

It is important to understand that the categories stationary and non-stationary, even allowing for gray areas introduced by time scales, are *population* categories. As we have already noted (and will discuss in great detail in Section 2.2), the variance scaling law of a process is of critical importance. We can broadly characterize the behavior of a process in terms of how the variance accumulates over different time scales: does it eventually flatten out, or continue growing (say, linearly) without limit? Now, it is certainly useful conceptually to understand these concepts (in a rather binary sense) and their associated (theoretical) econometric analysis (hence our efforts to that end here). However, it must always be kept in mind that *any* actual econometric study takes place in the context of a data set of a particular *finite* size. In other words, we are always faced with a problem of samples, and not populations. (Or more accurately, the issue concerns how to derive population information from actual samples, without assuming that the samples at our disposal are arbitrarily large.) The challenges presented by time scaling will depend greatly on sample size. We will see an example in Section 2.1.2 of a time series that is stationary in population, but very weakly so.[13] However, standard, popular techniques applied to this series in finite samples will perform very poorly.[14] The operational challenge can be characterized as an interaction between population (variance) time scales and (actual) sample size.

Our objective in this chapter is to make clear those issues that have to be taken into account, in order to effectively conduct econometric analysis. A not-insignificant part of this task will be to examine where, precisely, standard techniques break down. To do so, we will obviously have to understand what those techniques entail. While we do not intend to be (nor can we be) an econometrics text book, we will have to provide some amount of overview.

#### *2.1.1.6 Estimators*

Now, the essential feature of any econometric procedure is to infer the characteristics of some population from a sample. That is to say, for some parameter-dependent stochastic variable $X(\theta)$ and a set of realizations $\{X_i\}$ of (finite) size $N$, we seek a map (the estimator) $\mathfrak{E} : \{X_i\} \to \hat{\theta}$ taking the sample to an estimated parameter *and* a (perhaps probabilistic) relationship $\Re_N(\theta, \hat{\theta}) = 0$ between the estimate and the true parameter value. In addition, the model underlying $X$ implies some probability distribution $\Pr(X;\theta)$. Note that the relationship is (typically) dependent on the sample size. This is actually a very important aspect of the problem. As we will see in greater detail in the course of the discussion, the relationship associated with a particular estimator is often only known (analytically) *asymptotically*, that is, for very large sample sizes. In other words, the actual relationship associated with the estimator usually takes the form $\lim_{N\to\infty} \Re_N(\theta, \hat{\theta}) = 0$. As we will see further, the question of what qualifies as "very large" is very much dependent on the problem in question. This problem is amplified in energy markets, where it is quite common to have sample data of small size relative to equity and bond markets.

#### *2.1.1.7 Ordinary least squares*

Let us lead into some of the underlying issues by considering the well-known method of ordinary (linear) least squares (OLS). This method assumes there is a linear relationship between two sets of data $x$ and $y$ with Gaussian (white) noise:

$$y = \alpha x + \beta + \varepsilon \tag{2.5}$$

with $\varepsilon \sim N(0, \sigma^2)$, and with each realization being independent of other realizations. (Compare with (2.2).) Now, this model has certain assumptions, which we will not explicitly spell out here as they are most likely familiar to the reader (see Hamilton [1994])[15]. The point we wish to make is that OLS is an estimation procedure, with estimators formally obtained from the following optimization problem:

$$(\hat{\alpha} \quad \hat{\beta})^T = \arg\min_{\alpha,\beta} \sum_i (y_i - \alpha x_i - \beta)^2 \tag{2.6}$$

The solution of (2.6) is easily found to be:[16]

$$\begin{aligned}\hat{\alpha} &= \frac{\langle xy\rangle - \frac{1}{N}\langle x\rangle\langle y\rangle}{\langle x^2\rangle - \frac{1}{N}\langle x\rangle^2} \\ \hat{\beta} &= \frac{1}{N}(\langle y\rangle - \hat{\alpha}\langle x\rangle) \\ \hat{\sigma}^2 &= \frac{1}{N-2}\langle\hat{\varepsilon}^2\rangle\end{aligned} \tag{2.7}$$

where $N$ is the number of data points in the sample, $\langle\rangle$ denotes ensemble sum, and $\hat{\varepsilon} \equiv y - \hat{\alpha}x - \hat{\beta}$ are the (realized) *residuals*. Clearly, it can be seen how (2.7) conforms to our abstract definition of an estimator (as a function of a [realized] sample).

Now, the obvious question is the following: what does the output from the recipe in (2.6) really mean? Put differently, what can we say about the relationship between the sample entities in (2.7), and the population entities of the data generating process (DGP) in (2.5)? Note that the following relationship holds:[17]

$$\hat{\alpha} = \alpha + \frac{\langle x\varepsilon\rangle - \frac{1}{N}\langle x\rangle\langle\varepsilon\rangle}{\langle x^2\rangle - \frac{1}{N}\langle x\rangle^2} \tag{2.8}$$

The seemingly uninteresting expression in (2.8) is important for a number of reasons. First, it shows that the estimator is itself a random variable, depending not only on the regressors $x$ but also on the realized (random) deviations $\varepsilon$. Second, it shows an important relationship between the estimator and the "true" parameter (*i.e.*, the entity that the estimator is estimating), namely *unbiasedness*:

$$E\hat{\alpha} = \alpha \tag{2.9}$$

Back to our abstract framework, (2.9) is an example of a relationship between estimate and true parameter value. (Note further that in this case, the estimator is unbiased *in any sample size*; this is usually not the case, *i.e.*, unbiasedness is often an asymptotic relationship.)

Finally, (2.8) gives important information about the distribution of the (random) estimator. For example, the variance can be seen to be (given the underlying assumptions)

$$E(\alpha - \hat{\alpha})^2 = \frac{\sigma^2}{\langle x^2\rangle - \frac{1}{N}\langle x\rangle^2} \tag{2.10}$$

We do not write out the specifics here (again, consult Hamilton [1994]), but we emphasize the crucial point that (2.8) allows us to conduct *inferences* about the model parameters in (2.5).[18] For example, we can ask: if the true value of $\alpha$ was zero (so that there really was no relationship between $x$ and $y$), how likely would

it be that a particular realization of the data (such as the one being analyzed) would produce the observed *numerical* estimate $\hat{\alpha}$? In other words, are the results of the estimation statistically significant? Depending on one's threshold for acceptance/rejection (conventionally, 5% is a standard criteria for deeming a particular output as unlikely), a given model may be rejected.[19,20]

#### *2.1.1.8 Maximum likelihood*

Another very common estimator, at the heart of a great many econometric techniques, is so-called *maximum likelihood* estimation (MLE). We will employ it frequently here (as well as emphasize its shortcomings), so it merits some explication here. As the name suggests, the essence of the idea is to find values of the model parameters that maximize the probability of the (realized) sample:

$$\hat{\theta} = \arg\max_{\theta} \Pr(X_1, \ldots, X_N; \theta) \tag{2.11}$$

Typically, the assumption of identical independent distribution (i.i.d.) across the sample is made, so that (2.11) can be written as

$$\hat{\theta} = \arg\max_{\theta} \prod_i \Pr(X_i; \theta) \tag{2.12}$$

We will later investigate the ramifications of weakening the i.i.d. assumptions, but for now we will simply assume its validity for the problems we examine. It is convenient to then conduct the analysis in terms of the *log-likelihood function*, defined as

$$\mathfrak{L}(X;\theta) \equiv \frac{1}{N} \sum_i \log \Pr(X_i;\theta) \tag{2.13}$$

The first order condition for optimality implies

$$\frac{\partial}{\partial \theta} \mathfrak{L}(X;\hat{\theta}) = \frac{1}{N} \sum_i \frac{\partial}{\partial \theta} \log \Pr(X_i;\hat{\theta}) = 0 \tag{2.14}$$

Now, by the law of large numbers, under the prevailing assumptions (i.i.d.), ensemble averages like that in (2.14) converge[21] in the limit to the corresponding expectation:

$$\frac{1}{N} \sum_i \frac{\partial}{\partial \theta} \log \Pr(X_i;\hat{\theta}) = 0 \leftrightarrow E \frac{\partial}{\partial \theta} \log \Pr(X;\theta^*) = 0 \tag{2.15}$$

where $\theta^*$ denotes the true parameter value. The expectation in (2.15) follows from

$$\begin{aligned}
&\int \Pr(X;\theta)dX = 1 \Rightarrow \\
&\int \frac{\partial}{\partial\theta}\Pr(X;\theta)dX = \int \Pr(X;\theta)\frac{\partial}{\partial\theta}\log\Pr(X;\theta)dX = 0 \Rightarrow \\
&E\frac{\partial}{\partial\theta}\log\Pr(X;\theta^*) = 0
\end{aligned} \tag{2.16}$$

We thus see that the intuitive (and popular) econometric technique of maximum likelihood corresponds (in the limit) to a specific population condition. We will see later how more detailed (asymptotic) information regarding the estimates (such as the covariances of the estimator[22]) can be derived. We will see further just how dangerous it can be to rely on asymptotic results when dealing with the sample sizes typical of most energy market problems. Before doing so, however, we must continue with some (appropriately focused) overview.

### 2.1.2 Basic discrete time models: AR and VAR

*2.1.2.1 AR estimator and its properties*

Consider the following popular extension of (2.5), the so-called auto-regressive AR process:

$$x_t = \phi x_{t-1} + \varepsilon_t \tag{2.17}$$

with $|\phi| < 1$ (which, we will see, implies that the process is stationary so it makes sense to speak of its long-term or unconditional means and variances). The term $\varepsilon_t \sim N(0,\sigma^2)$ is assumed to be independent of $x_{t-1}$ (*i.e,*. it is unconditionally normal). It is not difficult to show that, under MLE[23], the estimator for the coefficient $\phi$ is formally the same as the OLS result:

$$\hat{\phi} = \frac{\langle x_{t-1}x_t\rangle}{\langle x_{t-1}^2\rangle} \tag{2.18}$$

which implies the following relationship:

$$\hat{\phi} = \phi + \frac{\langle x_{t-1}\varepsilon_t\rangle}{\langle x_{t-1}^2\rangle} \tag{2.19}$$

It is worth noting here that, as one of the standard assumptions of OLS is relaxed, namely non-stochasticity of the regressors, we lose the unbiassedness property (2.9) of the estimator.[24] However, it can be shown that the estimator is still *consistent*, *i.e.* as the sample size increases the bias gets smaller. Thus (2.18) is still useful. Furthermore, in this case we can no longer appeal to exact distributional results for

the estimator, but instead must rely on large sample-size (asymptotic) results. In fact (leaving aside the question of how large the sample must be for asymptotic results to be valid [which turns out to very non-trivial]) the asymptotic distribution is (standard) normal, making diagnostics quite easy to carry out.

Specifically, we have

$$\sqrt{T}(\hat{\phi} - \phi) = \frac{\frac{1}{\sqrt{T}}\langle x_{t-1}\varepsilon_t \rangle}{\frac{1}{T}\langle x_{t-1}^2 \rangle} \tag{2.20}$$

where $T$ is the sample size. For large $T$, the denominator in (2.20) is asymptotically the long-term variance (which is easily seen to equal $\sigma^2/(1-\phi^2)$; remember the assumption of stationarity), while the numerator is asymptotically normal with variance $\sigma^4/(1-\phi^2)$). (See Hamilton [1994] for a more rigorous demonstration.) Thus

$$\sqrt{T}(\hat{\phi} - \phi) \sim N(0, 1-\phi^2) \tag{2.21}$$

demonstrating the aforementioned consistency of the estimator: as the sample size increases, the estimator clusters around the true value with standard deviation shrinking like $O\left(\frac{1}{\sqrt{T}}\right)$.

#### *2.1.2.2 Non-stationarity and the limits of asymptotics*

However, certain complications can arise. Suppose that $\phi = 1$, so that (2.17) describes a random walk (discrete time Brownian motion). In other words, (2.17) becomes a non-stationary process. The most obvious problem is that the diagnostics in (2.21) become trivial! What is going on of course is that the estimator is converging at a rate faster than $O(\frac{1}{\sqrt{T}})$, in fact with rate $O(\frac{1}{T})$. In fact, intuitively we can see (via the scaling property of Brownian motion) that

$$T(\hat{\phi} - 1) \sim \frac{\int_0^1 w_s dw_s}{\int_0^1 w_s^2 ds} = \frac{\frac{1}{2}(w_1^2 - 1)}{\int_0^1 w_s^2 ds} \tag{2.22}$$

where $w$ is a standard Brownian motion. This expression is of course the basis of the well-known Dickey-Fuller test for unit roots.[25] The entity in (2.22) has a non-standard distribution and the critical values (for acceptance/rejection of inferences) are typically obtained via simulation. The point we are trying to make here is that, in the presence of non-stationary, the relevant diagnostics can radically change, even if the underlying estimation algorithm is unchanged.

This point is worth emphasizing. The power of many of these standard tests can be very low in small samples. (The power of a test refers to its ability to detect a relationship when it really exists, *i.e.*, to reject the null hypothesis [of no relationship] when the null relationship is actually false.) This is particularly true when one relies on asymptotic results (which are often the only theoretical results available). It is not hard to see why. Consider a near-unit root process, *e.g.*,

$$x_n = 0.999x_{n-1} + \varepsilon_n \tag{2.23}$$

For a *given* sample size, it will be very hard to distinguish this stationary series from a non-stationary random walk, based on asymptotic results such as (2.22). (This is also true of non-asymptotic tests such as Dickey-Fuller, which use as the test statistic the OLS estimate of $\phi$ divided by its standard error [*i.e.*, the usual $t$-statistic from linear regression], with associated critical values obtained via simulation.) In fact, simple simulations show (again, for a given sample) that the distribution of the estimator (2.18) is much greater than the asymptotic results in (2.21) would imply for this value of $\phi$ (about 3.5 times larger, for a sample size of 500), with the distribution being decidedly non-Gaussian. (This reflects the omission of higher-order terms in approximating $\langle x_{t-1}^2 \rangle$ in the denominator of (2.19) by the long-term [that is, asymptotic] population variance.) Thus, the estimator would tend to over-suggest the presence of (strong) mean reversion (or alternatively under-suggest the presence of near non-stationarity). These results can be seen in Figures 2.1 and 2.2 (note the extreme skewness of the estimator).

Interestingly, the estimator of the disturbance variance in (2.23) *does* conform to its asymptotic distribution. The MLE estimator of the variance is given by

$$\hat{v} = \frac{1}{T}\langle (x_n - \hat{\phi} x_{n-1})^2 \rangle = \frac{1}{T}\left( \langle \varepsilon_n^2 \rangle - \frac{\langle \varepsilon_n x_{n-1} \rangle^2}{\langle x_{n-1}^2 \rangle} \right) \tag{2.24}$$

where $v \equiv \sigma^2$. Now, the expression in (2.24) should be compared with the estimator for the AR coefficient in (2.19). For the AR coefficient, the estimator consists of

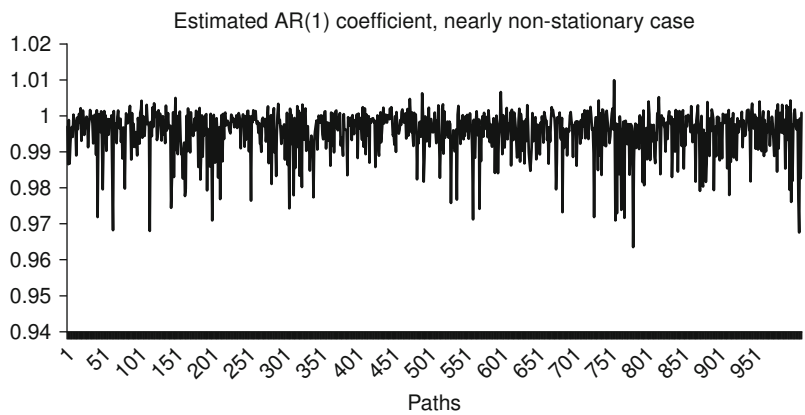



**Figure 2.1** AR(1) coefficient estimator, nearly non-stationary process

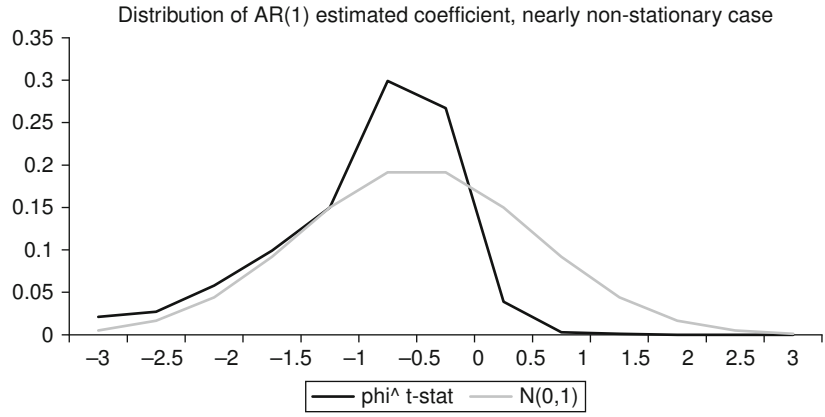



**Figure 2.2** Distribution of *t*-statistic, AR(1) coefficient, nearly non-stationary process

the exact value plus some non-standard RV. It is precisely the extreme deviations from normality of this RV that leads to a breakdown of asymptotic results in small samples. By contrast, the variance estimator consists of an average of the squared (realized but unobserved) disturbances plus another non-standard RV (similar to the one for the AR coefficient). Since the disturbances are (by assumption) i.i.d., we would expect the average of their squared sum to be a good representative of the true variance. This sub-estimator (so to speak) has its own variability, which is characterized by approximate normality with $O(T^{-1/2})$ standard deviation.[26] So, *if* this variability greatly exceeds the variability of the non-standard component of the estimator, overall estimator variability will be asymptotically normal. This condition is typically satisfied in practice, and we frequently see that variance estimators are much better behaved than mean-reversion estimators. These claims are illustrated in Figures 2.3 and 2.4. Note further that we would anticipate that the uncertainty associated with variance estimators will decrease much faster with sample size[27] compared with the uncertainty associated with mean-reversion estimators: we will typically require far fewer data to form robust estimations of variance than we would for mean-reversion rates. We will make great use of these properties in the course of our exposition.

The failure of asymptotically valid results to provide reliable diagnostics in finite samples points to the need for alternative means of assessing the output of any econometric technique. A powerful and useful approach is the so-called bootstrap methodology, which falls under the general category of resampling; we will consider this topic in greater detail in Section 6.4. Another cautionary tale about conducting econometric tests with non-stationary variables concerns so-called spurious regressions, regressions that show statistically significant results when in fact no real relationship exists. This topic will also be revisited in Section 6.1.1 when

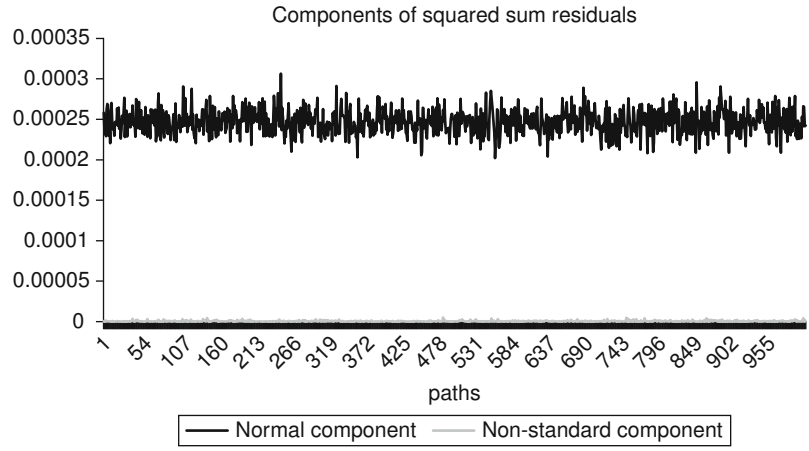



**Figure 2.3** Components of AR(1) variance estimator, nearly non-stationary process. We take $\sigma = 30\%$, daily time step

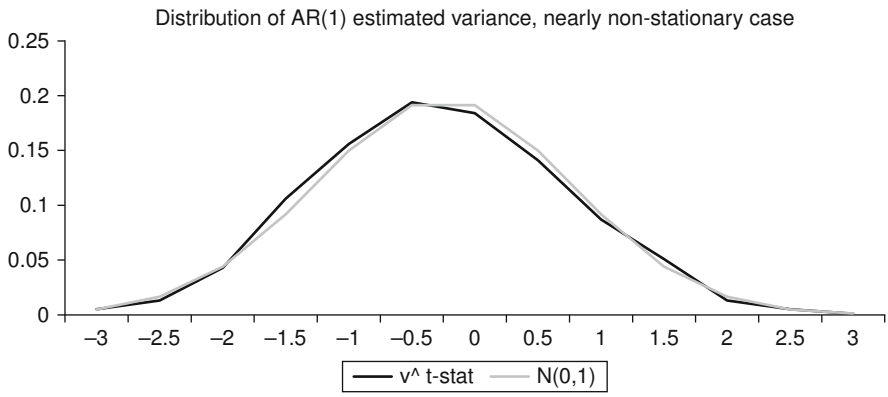



**Figure 2.4** Distribution of *t*-statistic, AR(1) variance, nearly non-stationary process

we consider cointegration, but we will first need to consider a generalization of the one-dimensional case studied thus far.

#### *2.1.2.3 Extension to higher dimensions*

To better illustrate the underlying methods, it is worthwhile to go into a bit of operational detail for a multivariate situation. Consider the vector autoregression model

$$x_t = \Pi x_{t-1} + \varepsilon_t \tag{2.25}$$

for some $N$-dimensional process, with the noise term independent between time steps but possessing a contemporaneous covariance structure denoted by $\mathbf{\Omega}$; specifically, $\varepsilon_t \sim N(0, \mathbf{\Omega})$. In addition, we assume all of the eigenvalues of the matrix $\Pi$ are within the unit circle, so that the process is stationary. With this latter assumption we can validly rewrite (2.25) as an infinite series moving average (MA) representation as

$$x_t = \varepsilon_t + \Pi\varepsilon_{t-1} + \Pi^2\varepsilon_{t-2} + \cdots \tag{2.26}$$

We will return to (2.26) shortly. After a bit of algebra, ML estimation of (2.25) yields[28]

$$\hat{\Pi} = \langle x_t x_{t-1}^T \rangle \langle x_{t-1} x_{t-1}^T \rangle^{-1} \tag{2.27}$$

It thus follows that

$$\hat{\Pi} - \Pi = \langle \varepsilon_t x_{t-1}^T \rangle \langle x_{t-1} x_{t-1}^T \rangle^{-1} \tag{2.28}$$

To determine the asymptotic diagnostics of this estimator, consider the second matrix factor in (2.28). Using (2.26), we have that

$$\frac{1}{T}\langle x_{t-1} x_{t-1}^T \rangle = \frac{1}{T} \sum_t \sum_{i,j} \Pi^i \varepsilon_{t-i} \varepsilon_{t-j}^T \Pi^{Tj} \tag{2.29}$$

In the large sample size limit, (2.29) converges in probability to (recall the condition on the eigenvalues of $\Pi$ and the non-contemporaneous independence of $\varepsilon$):

$$\sum_i \Pi^i \mathbf{\Omega} \Pi^{Ti} \tag{2.30}$$

If we further assume that the eigenvalues $\lambda$ of $\Pi$ are distinct,[29] we have the factorization $\Pi = V\Lambda V^{-1}$ where $\Lambda$ is a diagonal matrix with the eigenvalues of $\Pi$ along the diagonal, and $V$ is a matrix comprising the corresponding eigenvectors as its columns. Consequently (2.30) can be written as:

$$V \left( \sum_i \Lambda^i V^{-1} \mathbf{\Omega} V^{-T} \Lambda^i \right) V^T \tag{2.31}$$

Writing $\mathbf{\Omega}^V \equiv V^{-1}\mathbf{\Omega}V^{-T}$ and exploiting the diagonal form of $\Lambda$, the matrix in (2.31) becomes $V\mathbf{\Omega}^\lambda V^T$, where the elements of $\mathbf{\Omega}^\lambda$ are given by:

$$\mathbf{\Omega}_{ij}^\lambda = \frac{\mathbf{\Omega}_{ij}^V}{1 - \lambda_i \lambda_j} \tag{2.32}$$

Denote the matrix in (2.31) by $Q$; it is essentially the long-term covariance matrix for the (stationary) process (2.25). (Compare with the results for the one-dimensional case, where the long-term variance is easily seen to be $\sigma^2/(1-\phi^2)$). The asymptotic diagnostics thus become:

$$\sqrt{T}\left(\hat{\Pi}-\Pi\right) \sim \frac{1}{\sqrt{T}}\langle \varepsilon_t x_{t-1}^T\rangle Q^{-1} \tag{2.33}$$

It should be clear from (2.33) that the estimator, though biased, is consistent. We can say a bit more by noting the $(i,j)$ element of the estimator is distributed as:

$$\frac{1}{\sqrt{T}}\langle \varepsilon_t^i x_{t-1}^k\rangle Q_{kj}^{-1} \tag{2.34}$$

which has expectation (in population) 0 and covariance (again in population):

$$\begin{aligned}&\frac{1}{T}E\langle \varepsilon_t^i x_{t-1}^k Q_{kj}^{-1}\varepsilon_{t'}^{i'}x_{t'-1}^{k'}Q_{k'j'}^{-1}\rangle \\ &\quad=\frac{1}{T}\langle x_{t-1}^k Q_{kj}^{-1}x_{t-1}^{k'}Q_{k'j'}^{-1}\Omega_{ii'}\rangle \sim Q_{kj}^{-1}Q_{k'j'}^{-1}\Omega_{ii'}Q_{kk'}=\Omega_{ii'}Q_{jj'}^{-1}\end{aligned} \tag{2.35}$$

Thus, we anticipate that:[30]

$$\sqrt{T}(\mathrm{vec}(\hat{\Pi})-\mathrm{vec}(\Pi)) \sim N(0,\Omega\otimes Q^{-1}) \tag{2.36}$$

where $\otimes$ denotes the Kronecker product and $\mathrm{vec}(A)$ is a vector formed by stacking the rows of the matrix $A$. (See Proposition 11.1 in Hamilton [1994] for a more detailed derivation.)

#### *2.1.2.4 Limits to asymptotics revisited*

Again, apart from illustrating some operational details, we wish to highlight the importance of the stationarity assumption in deriving the necessary diagnostics. In particular, note the non-standard diagnostics (along with completely different asymptotic behavior) in the presence of non-stationarity. Intuitively, for large $T$, we would expect the process in (2.25) to behave like its continuous time limit over a scaled time horizon, *i.e.*:

$$dx = Axdt + dw \tag{2.37}$$

with $I + Adt \to \Pi$ and where $w$ represents a vector-valued Brownian motion.[31] We would thus expect the estimator in (2.28) to behave like:

$$\hat{\Pi}-\Pi \sim \frac{1}{\sqrt{T}}\int_0^1 dw_s x_{Ts}^T\left(\int_0^1 ds\cdot x_{Ts}x_{Ts}^T\right)^{-1} \tag{2.38}$$

Now, for a stationary process, the long-term limits of $x_{Ts}$ make sense and it can be seen that (2.38) recovers the result in (2.36). However, for non-stationary processes, the dynamics of $x$ will have a purely Brownian component (as can be seen from an eigenvalue analysis of the matrix $A$; we will demonstrate this fact when we consider the relationship between cointegration and scaling laws of variance later[32]). Thus, due to the scaling property of Brownian motion, it can be seen from (2.38) that, at least for some of the components, $O(\frac{1}{T})$ will be introduced. (Note that since the $Q$-matrix exhibits pathological behavior via (2.32), the asymptotic diagnostics in (2.36) break down.) It is not immediately clear if any tractable general results can be stated via this approach, but for the very simple case $\mathbf{\Omega} = I$, the relevant diagnostics are:

$$\hat{\Pi} - \Pi \sim \frac{1}{T} \int_0^1 dw_s w_s^T \left( \int_0^1 ds \cdot w_s w_s^T \right)^{-1} \tag{2.39}$$

which can be seen as a multidimensional extension of the Dickey-Fuller statistic for unit root testing. (Compare with (2.22); yet again we refer the reader to Hamilton [1994] for a more thorough discussion).

We have thus seen already with even this brief overview of standard time series results that a great many problems can arise when these methods are applied outside their theoretical setting and in the actual situations we face in energy markets, namely small samples in the presence of non-stationary effects. Indeed, strict non-stationarity is not at all necessary to create problems, as the example in (2.23) shows. (The interplay of such effects is particularly acute in energy markets, where price formation depends on *both* more-or-less stationary, fundamental demand drivers on the one hand and more-or-less non-stationary capital supply effects on the other.) It is thus imperative that we have an alternative (or more accurately, more robust) framework for extracting information from available data. The idea of variance as a measure of information accumulation over specific time horizons proves crucial here.

## 2.2 Variance scaling laws and volatility accumulation[33]

In truth, most time series of interest do not fall clearly into one category (stationary vs. non-stationary) or the other. A simple example is the standard mean-reverting process (the continuous-time[34] version of the AR process in (2.17)):

$$dz = \kappa(\theta - z)dt + \sigma\, dw \tag{2.40}$$

with $\kappa > 0$. The conditional statistics are normal, with:

$$z_T \sim N\left(ze^{-\kappa\tau} + \theta(1 - e^{-\kappa\tau}), \sigma^2 \frac{1 - \exp(-2\kappa\tau)}{2\kappa}\right) \tag{2.41}$$

with $\tau \equiv T - t$ denoting the time horizon (and $z$ the current [time $t$] value of the process). Clearly, the distribution of future states *is* dependent on the state at the current time, in terms of the technical definition the process cannot be called stationary. However, just as clearly this dependence is weak over sufficiently long time horizons (which of course depends on the strength of mean reversion). Put differently, we can ask: how much does current information affect our projections of the future? From (2.41), the expected value of the process can be written as $\theta + (z - \theta)e^{\kappa\tau}$, so on the one hand, there is a sense in which we can talk about unconditional means: namely, the (long-term) mean reversion level $\theta$. On the other hand, our projection must take into account *both* the current deviation from that (unconditional) level as well as the time horizon in question. The shorter the time horizon and/or the weaker the mean-reversion rate, the more important current information will be in forming projections. There is thus an interplay and balance between these two effects that must be properly accounted for. This point has obvious ramifications for trading strategies based on mean reversion (such as convergence plays): one may indeed identify a situation where a price (say) is above or below its long-term mean (assuming this mean is known, an important but separate issue), but the question of whether one should take a counterposition (sell if above, buy if below) depends crucially on how long one expects it to take before the position moves into profitable territory.[35]

Now, as is well known, directly estimating mean-reversion rates (*e.g.*, from regressions) is extremely non-robust. This is actually a special case of the issue discussed previously, on the low power of many econometric tests in small samples. However, there is a robust alternative to detecting the presence of mean-reverting effects (which we will improperly conflate with stationarity, as the two concepts have sufficiently common features of importance). As we have discussed previously, the variance of the process (2.40), as a function of time horizon, has a distinctly different structure from that of a non-stationary process (such as Brownian motion). Specifically, the variance of a mean-reverting process tends to flatten out over long enough time horizons, while the variance of a non-stationary process tends to grow linearly with time.[36] In fact, depending on the strength of mean reversion, the time scale over which flattening out occurs can vary. Given that variances and covariances are typically more robust to estimate than mean-reversion rates, looking at how the variance of a process behaves over different time horizons can serve as a useful diagnostic for detecting the presence of stationarity. This will be our next topic to consider.

### 2.2.1 The role of fundamentals and exogenous drivers

As we have stated, energy markets present a number of econometric challenges, chief among them comparatively small sample sizes and relatively large volatilities. It is therefore critical to exploit as many advantages as we may in analyzing such data. One ameliorating feature of energy markets is that we often have some idea of certain fundamental or exogenous[37] drivers that affect prices and volatilities, if only indirectly.[38] These drivers can be broadly characterized as reflective of supply-and-demand conditions. A bit more narrow (but still fairly inclusive) characterization would be the capital structure (*e.g.*, the generation stack in power or the pipeline network in natural gas) that produces commodities, and weather and economic conditions (*e.g.*, greater demand for electricity in the summer and gas in winter or general economic growth influencing the price of oil) that drive demand for commodities. To the extent that (at least some of) these fundamental factors may be relatively well understood (in comparison to commodities themselves), such information should be exploited as much as possible. In particular, we should endeavor to understand how the time scales influencing these factors manifest themselves in the dynamics of energy prices and, if possible, separated out from the analysis of these dynamics.

#### *2.2.1.1 A confluence of effects*

As a simple example, consider a spot price model for (log-) power: $p = g + h + l$ where

$$\begin{aligned} dg &= \mu_g dt + \sigma_g dw_g \\ dh &= \kappa_y(\theta - h)dt + \sigma_h dw_h \\ dl &= -\kappa_l l dt + j dq \end{aligned} \tag{2.42}$$

which is an augmentation of an example to be considered in Section 3.1.3. The constituent terms in (2.42) can be thought of respectively as (log-) gas (the fuel cost of generation), (log-) heat rate (of the marginal generating unit), and an outage term (signifying jumps within the stack, *i.e.*, supply shocks). For our purposes we can think of the heat rate as equivalent to temperature or load/demand. Now, a useful application of the characteristic function methods to be studied in Chapter 5 is to derive expressions for the forward prices (we will ignore any questions as to the measure being employed) and the corresponding dynamics. (We are of course very interested in forwards and their associated dynamics, since they are the primary hedging instruments in energy markets.) If $F^P_{t,T} = E_t e^{p_T}$ (with corresponding expressions for the other entities), then with these methods it is straight-forward to

show that

$$\log F^P_{t,T} = \log F^G_{t,T} + \log F^H_{t,T} + \log F^L_{t,T} + \gamma_{T-t}$$
$$\frac{dF^P_{t,T}}{F^P_{t,T}} = -\lambda_l E(\exp(je^{-\kappa_l(T-t)}) - 1)dt + \sigma_g dw_g + \sigma_h e^{-\kappa_h(T-t)} dw_h + (\exp(je^{-\kappa_l(T-t)}) - 1)dq \quad (2.43)$$

(In (2.43), $\gamma$ is some [deterministic] function of time-to-maturity whose precise form is unimportant here.)

*2.2.1.2 Impact on information flows*

Now, each of the stochastic drivers in the dynamics in (2.43) has a ready interpretation in terms of time scales and the different horizons over which certain kinds of information are important (and over which other kinds of information are unimportant). First, gas contributes an essentially flat term structure to the flow of information. This is not surprising, as in this model gas is purely non-stationary, so a change in today's gas price affects expectations of future power prices equally across time horizons/times-to-maturity. Put differently, the incremental contribution of information flow (as measured by cumulative variance) is constant for gas (linear variance scaling law). Second, the heat rate (or more broadly considered, demand) contributes a typical Samuelson-type information shape, with a volatility term structure that increases as time-to-maturity decreases. This is again not surprising, owing to the mean-reverting nature of the heat rate. Thinking in terms of temperature, a heat wave far from maturity will have little effect on our expectations of prices, however a heat wave close to maturity will have a much bigger impact. The incremental contribution of information dies off as the time horizon in question grows (asymptotically flattening variance scaling law).

The effect of jumps/outages is similar to heat rate/demand, but on the supply side. Knowing that a unit has gone done when we are far from maturity will not affect our expectations nearly as much as when the unit goes down near expiry. This behavior can be seen in the jump amplitude in (2.43). Whatever the intensity of jumps is, far from the maturity the exponential term is close to 1. Of course, the rate at which capacity is restored to the stack (effectively, the mean-reversion rate $\kappa_l$) determines the horizon over which we can say this or that piece of information is unimportant (for forming expectations). A small unit that goes down but comes back quickly will have fewer consequences than a large unit that takes a while to come back online. Note that the presence of jumps close to maturity can make the estimation of volatility for short-dated contracts rather tricky, if not prone to instability. This is a good example of how knowledge of exogenous drivers can be used to temper our analysis of certain effects of interest, if only as a warning system (so to speak).

These are not the only conclusions that we can draw here. As we just saw, each of the constituent stochastic drivers in the dynamics in (2.43) makes its own particular contribution to the overall variability of the power forward over any given time horizon. This fact lays open the possibility of identifying and quantifying the impact (on the dynamics of expected power) on the propagation of information arising from *changes* in these drivers. For example, in markets with liquidly traded gas options, option prices (as in any reasonable efficient market) are typically good projections of realized (gas) volatility. Knowledge of changing demand conditions (*e.g.*, for stationary load patterns, large deviations from normal levels do not die out immediately, but dissipate with some characteristic "half-life," so to speak) and/or supply conditions (knowing the state of outages in a particular generation region) can also be exploited in crafting an understanding of forward power dynamics (if only in qualitative manner).

The point of this discussion is of course not forward power price modeling as such, but rather the point that in general energy markets entail certain (more-or-less) well-known drivers and relationships, and the manner in which information accumulates due to these drivers is often reasonably well understood. This understanding can (and should) be used to delimit the scope of a given econometric problem as much as possible. With this we should begin a more detailed discussion of (variance) scaling laws as a measure of information accumulation.

### 2.2.2 Time scales and robust estimation

We discussed in Subsection 2.1.2 the dangers in relying on asymptotic results in econometric analysis. The behavior of an estimator in an actual, finite size sample can be very different from its large sample behavior. The reality of our situation is that we *always* deal with samples that are small in relation to the amount of structure that we would like to impose on the data. By this we simply mean that we do not doubt actual prices dynamics are driven by stochastic volatility, jumps, *etc.*, however, the amount of data at our disposal simply does not permit robust estimation of such models. In fact, as we will discuss in Section 6.5, even for samples that appear large (say, more than 500 points), many popular estimators can perform quite poorly, *even when we know the true DGP* (and of course we never do).

It is thus generally preferable to consider the distribution of some statistic that is both reflective of a population characteristic of interest, *and* robust in relation to the actual sample size. The statistic should also be amenable to analysis in terms of a random outcome of a *given* size from some population. In other words, we are interested in the *statistic's sampling distribution.*

We will have more to say about sampling distributions when we discuss Wishart processes in Section 6.3. For now, we wish to focus on the question of robust statistics that, if only indirectly, reveal information about the population that otherwise could only be estimated with great unreliability. We have already examined (if only

somewhat briefly) the issue of the comparative performance of MLE for mean-reversion rates as opposed to disturbance variances in Section 2.1.2. We have also noted the behavior of (population) variance over particular time scales as it pertains to the distinction between stationary and non-stationary processes (*e.g.*, the introductory discussion centered about (2.41)). In light of the more robust behavior of variance estimators compared to (say) mean-reversion estimators, we will see that the analysis of variance scaling laws are precisely the alternative we seek to problematic estimations of mean reversion, whose existence in commodity markets (and associated impact on the temporal accumulation of information) is not in doubt.

Before illustrating with actual data, however, we must first indicate some subtleties arising from the inherently conditional nature of the data we are confronted with (that is to say, the kinds of effects we are interested in by nature preclude the ability to speak of unconditional distributions, which are at the heart of much econometric analysis).

### 2.2.3 Jumps and estimation issues

#### *2.2.3.1 Why we should care*

A central concern in any econometric investigation is the size of the available data. As we will continue to stress, size matters, in a relative sense. Depending on the nature of the time scales at work, a given sample may appear large but in reality be too small to extract certain characteristics of interest. We are not simply referring to long-term effects here; even short-term effects that are commonly the focus of many energy market valuation problems can be impacted. This point is worth emphasizing, because it is often thought that the availability of data at very high resolutions (*e.g.*, hourly LMP data for power, or tick data for crude oil) provides samples of sufficient size for robust estimation. After all, a year of daily data comprises a sample of nearly 400 points, and a year of hourly data consists of nearly 10,000 points. However, this sense of security is illusory if the underlying DGP is driven by factors whose characteristic time scales are measured on the order of years. A year of data is still *one year* of data, no matter how finely it is resolved. Even for an entity as nicely behaved as temperature (leaving aside issues of human effects on climate), if certain cycles take place over the order of a century, then a 50-year data set will be inadequate for extracting that kind of information.

Let us proceed to illustrate with a tractable framework that incorporates both multiple time scales and multiple underlying drivers. We will first note some complications arising from jumps, before focusing on more manageable diffusive processes.

*2.2.3.2 The impact of jumps on volatility term structure and scaling*

We have already seen in (2.43) how particular time scales determine how the effects of certain kinds of jump drivers play out. In particular, it was seen how volatility scaling manifests itself in the presence of jumps whose presence is most acutely felt close to maturity. It is worth considering the opposite situation, where jump phenomena have a greater impact *far* from maturity. For simplicity we will focus on heat rates, so *sans* gas consider the following modification of (2.42):

$$\begin{aligned} dh &= \kappa_y(\theta - h)dt + \sigma_h dw_h \\ d\theta &= -\kappa_\theta \theta\, dt + j dq \end{aligned} \tag{2.44}$$

Note the subtle difference: in (2.44), jumps occur *in* the long-term mean, not *off* of it, as in (2.42). This seemingly minor change actually has major ramifications for the underlying scaling law. Proceeding as before, the corresponding forward dynamics[39] can be shown to be

$$\frac{dF^H_{t,T}}{F^H_{t,T}} = -\lambda_\theta E(e^{j\beta} - 1)dt + \sigma_h e^{-\kappa_h(T-t)} dw_h + (e^{j\beta} - 1)dq \tag{2.45}$$

where

$$\beta = \frac{\kappa_h}{\kappa_\theta - \kappa_h}\left(e^{-\kappa_h(T-t)} - e^{-\kappa_\theta(T-t)}\right) \tag{2.46}$$

The first thing to note about the expression in (2.46) is that as time-to-maturity tends zero, so does this modulation factor. Moving away from expiry, this factor initially grows, until eventually dying out. In the extreme case of no mean reversion in jumps (*i.e.*, $\kappa_\theta = 0$), the factor actually asymptotes to 1. The thrust of this behavior is that, close to maturity, jumps in the mean-reversion level have little effect on expectations of future heat rate. However, sufficiently far from maturity, these jumps *do* have a definite effect on expectations. There is thus a very clear contrast between the informational flow pertaining to the dynamics in (2.44) and that in (2.42).

The kind of behavior evinced by (2.44) is typically due to some kind of structural change in the (fundamental) underlying supply conditions driving price formation. For example, a long, unplanned outage (say, for emergency maintenance) of a big, baseload unit (*e.g.*, a nuclear plant) would fit into this category. So would unanticipated weather conditions that remove a significant amount of generation, such as a weak winter drying up hydro facilities. (This situation in fact prevailed in the Western United States in the spring of 2000, during the perfect storm of price spikes that led to outcry [some founded, some unfounded] against trading practices by companies such as Enron.) Such events typically have a greater effect on mid- to long-term expectations, rather than short-term expectations. In contrast to the effects described by the model in (2.43), which manifest themselves as much greater

volatility (if not outright instability) close to maturity, the model in (2.45) exhibits structural breaks in volatility far from expiry.

This discussion, in conjunction with that in Section 2.2.1, seeks to address a familiar question in the context of jumps: when and why do they matter? Broadly, we can answer the first part by saying: it depends on the structural changes taking place in the market environment (by which we mean the endogenous and exogenous factors driving price formation) under consideration. In some situations, jumps are driven by stack outage and have their primary effects close to maturity. In other situations, jumps are driven by structural changes and are chiefly felt far from maturity. Whether these two cases are relevant (and, of course, there is never a clean dividing line) depends on the problem at hand. As to the second part, jumps matter because they exacerbate one of the central challenges confronting econometric analysis of (typically small-sample) energy market data sets: high volatilities. However, in connection to the first query, the enhanced complications depend on the maturities encompassed by the data in question.

The general remedy to these problems is thus to know when jumps might be a problem, and adjust the analysis appropriately. For example, consider the problem of estimating the volatility (really, quadratic variation) that can be captured from delta-hedging some option (on futures). For short-term options, estimation will likely show fairly high volatilities when the jumps are driven by dynamics such as (2.43). These results must be viewed with great care. First, it is well known that delta-hedging options in the presence of jumps is quite challenging. Second, the uncertainty (standard errors, if you will) associated with the estimation of collectible volatility will itself be rather high. These considerations warrant great caution in placing bets on realized volatilities close to maturity. On the other hand, jumps given by processes such as (2.45) generally give rise to overestimates of the volatility that can be collected over longer terms. Data in the sample that happened to pass through the structural change characterized by these kinds of jumps are generally not indicative of the new market structure (especially when the change is relatively non-stationary), and it is probably wise to filter these out (in the plain language sense) of the analysis. (Jumps have somewhat of a corrupting effect in this case.)

In very broad terms, then, we can consider jumps to have rather localized effects, which can be accounted for by increased risk adjustment or proper censoring of data. For the remainder of our discussion here, we will focus on the diffusive case.

#### *2.2.3.3 Multiple scales and econometric impact*

We consider an example introduced by Grzywacz and Wolyniec (2011). Assume we have two (independent) diffusive processes, both mean-reverting with the same reversion level (zero) but different reversion rates:

$$\begin{aligned} dx &= -\kappa_x x dt + \sigma_x dw_x \\ dy &= -\kappa_y y dt + \sigma_y dw_y \end{aligned} \tag{2.47}$$

Imagine, however, that we only observe $z = x + y$ and attempt to estimate a standard AR(1) model from these observations: $z_n = \phi_z z_{n-1} + \varepsilon_n$. The estimator can be written as

$$\begin{aligned}\hat{\phi}_z &= \frac{\langle z_n z_{n-1}\rangle}{\langle z_{n-1}^2\rangle} = \frac{\langle x_n x_{n-1}\rangle + \langle y_n x_{n-1}\rangle + \langle x_n y_{n-1}\rangle + \langle y_n y_{n-1}\rangle}{\langle x_{n-1}^2\rangle + 2\langle x_{n-1} y_{n-1}\rangle + \langle y_{n-1}^2\rangle} \\ &\sim \frac{\hat{\phi}_x \langle x_{n-1}^2\rangle + \hat{\phi}_y \langle y_{n-1}^2\rangle}{\langle x_{n-1}^2\rangle + \langle y_{n-1}^2\rangle}\end{aligned} \tag{2.48}$$

where $\hat{\phi}_{x,y}$ denote the estimators for the discrete-time analogues of the system in (2.47) and we use the (asymptotic) independence of $x$ and $y$ in (2.48) to drop out the cross-term ensemble entities. We thus see from (2.48) that the estimated mean-reversion parameter for the observed variable $z$ is a weighted sum of the individual mean-reversion parameters of the unobserved variables. In particular, the estimator $\hat{\phi}_z$ will be dominated by the component with the *lowest* mean-reversion rate, *i.e.*, the least stationary component. This is due to the presence of the summed squared terms; in a sample of a *given* size, the more non-stationary a variable is, the less applicable is the law of large numbers (as a means of associated sample averages with population properties; we will discuss this issue much more in Section 6.5).

More accurately, the weights reflect the relative contribution of each component to total variance. A sample with a near-unit root process (say) will accumulate variance very slowly (in terms of convergence of the ensemble sums-of-squares), and thus estimation will tend to identify this component. While this is technically the correct diagnosis, it is decidedly unhelpful (indeed, basically incorrect) in ascertaining the short- to medium-term accrual of variance for the process. Ultimately, the issue comes down to the sample size *in relation to the underlying time scales* (which are of course just the counterpart to the operative mean-reversion rates). This example highlights yet again the difficulties of attempting to directly estimate mean reversion.[40]

What about indirect estimation? In fact, it is worth asking which population entities ensemble averages like those in (2.48) correspond to in the presence of mean reversion, for in this case the standard assumption of i.i.d. increments cannot be appealed to. This brings us to another interesting result of Grzywacz and Wolyniec (2011), namely that mean-reversion rates inferred from observed scaling laws are *half* the true rate. To see this fact, consider a sample size $N$ with time $\Delta t$ between observations (focusing only on the component $x$ for simplicity). Using the law of iterated expectations, we have that:[41]

$$
\begin{aligned}
E_0\frac{1}{N}\sum_i(\Delta x_i)^2 &= E_0\frac{1}{N}\sum_i E_{i-1}(x_i - x_{i-1})^2 \\
&= E_0\frac{1}{N}\sum_i [E_{i-1}(x_i - E_{i-1}x_i)^2 + (x_{i-1} - E_{i-1}x_i)^2] \\
&= \frac{1}{N}\sum_i \left[\frac{\sigma_x^2}{2\kappa_x}(1-\exp(-2\kappa_x\Delta t)) + (1-\exp(-\kappa_x\Delta t))^2 E_0 x_{i-1}^2\right] \\
&= \frac{\sigma_x^2}{2\kappa_x}(1-\exp(-2\kappa_x\Delta t)) + \frac{\sigma_x^2}{2\kappa_x}(1-\exp(-\kappa_x\Delta t))^2\frac{1}{N}\sum_i(1-\exp(-2\kappa_x i\Delta t)) \\
&\sim \frac{\sigma_x^2}{2\kappa_x}(1-\exp(-2\kappa_x\Delta t)) + \frac{\sigma_x^2}{2\kappa_x}(1-\exp(-\kappa_x\Delta t))^2 \\
&= \frac{\sigma_x^2}{\kappa_x}(1-\exp(-\kappa_x\Delta t))
\end{aligned}
\tag{2.49}
$$

as claimed.[42] (Note that the stronger the mean-reversion rate is, the larger the sample must be for this asymptotic result to be practically useful.[43]) An example is shown in Figure 2.5.

The result (2.49) can be extended to higher dimensions, and as in the situation considered in (2.48), the problem of less stationary/more non-stationary components in distorting the estimation again arises. (See the Appendix to Chapter 6 for greater elaboration; the result can also be extended to jump processes, *e.g.*, (2.44).) We have thus given some reason to be concerned about the effects that the (relatively) most non-stationary drivers can have on conventional estimation

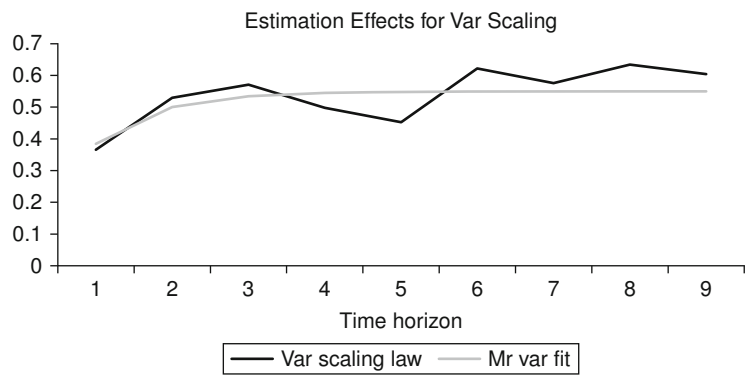



**Figure 2.5** Illustration of non-IDD effects. Variance scaling law for an AR(1) with coefficient 0.3 and disturbance variance 0.25 (per unit time). The equivalent OU process has mean reversion parameter 1.2, but the fitted MR variance scaling law corresponds to a process with reversion rate 0.6

techniques, even those which are generally known to be fairly robust. Nonetheless, variance scaling laws are a very powerful tool for identifying the presence (at least) of the relevant operative time scale for a wide range of processes. We will now turn to specific energy market examples.

### 2.2.4 Spot prices

Spot commodities generally do not trade as such; physical possession must be taken at some point. Spot prices are thus subject to more-or-less fundamental supply-and-demand conditions, reflecting both weather-driven demand factors (to be examined in Section 2.2.6) and generation/supply factors whereby marginal costs of production cannot wander too far without limit. (This latter statement is somewhat inaccurate because generation itself can grow or shrink, but this point is somewhat tangential here.) We would thus expect to see some evidence of variance scaling reminiscent of mean reversion, even though we would also not expect prices as such to be purely stationary. It is useful to conduct the analysis here in terms of average monthly prices. We do this because we wish to abstract as much as possible from localized (say, daily) phenomenon such as jumps, as such effects are not central to the point here. (These effects are best analyzed by conditioning on such entities as monthly averages, because it is only these entities that can have connection to the instruments that actually trade in energy markets, namely futures/forwards.) We start by showing spot time series for two important commodities, Henry Hub natural gas and Brent crude oil. See Figures 2.6 and 2.7.

We note from inspection that neither of these time series could be characterized as stationary, at least in the sense that, say, temperature could be (again, see Section 2.2.6).[44] Nonetheless, the variance scaling laws (for difference of log prices

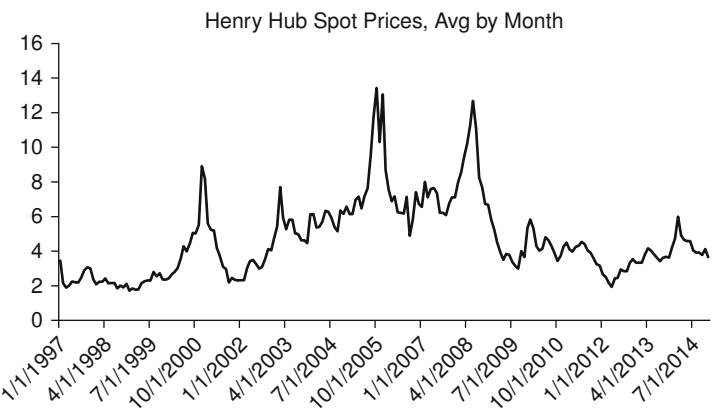



**Figure 2.6** Monthly (average) natural gas spot prices.
*Source*: EIA (www.eia.gov).

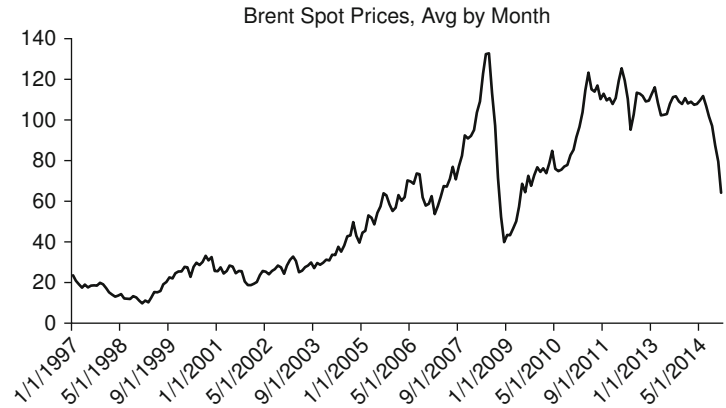



**Figure 2.7** Monthly (average) crude oil spot prices.
*Source*: EIA (www.eia.gov).

[in essence, returns] over increasing time horizons) *do* display some features of stationarity. See Figures 2.8 and 2.9.

Both natural gas[45] and crude oil display tell-tale signs of mean reversion, namely a flattening of the variance over *long enough* time horizons. In truth, the characteristic time scale of the mean reversion is quite long, on the order of two years. There is thus little reason to think that there are short-term opportunities that can be readily exploited, at least on a monthly level. However, even this issue is somewhat beside the point, as spot commodities do not trade as such (again, physical possession must be taken). Forward commodities do trade, and we will consider forwards in the next subsection. What we wish to demonstrate here is that the variance scaling laws exhibited on a spot basis are consistent with volatility collection

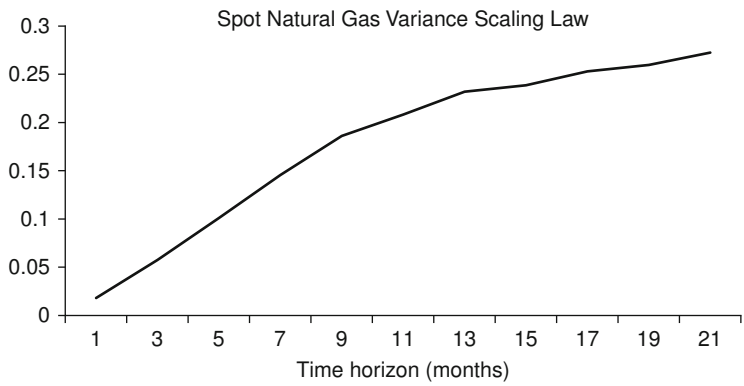



**Figure 2.8** Variance scaling law for spot Henry Hub

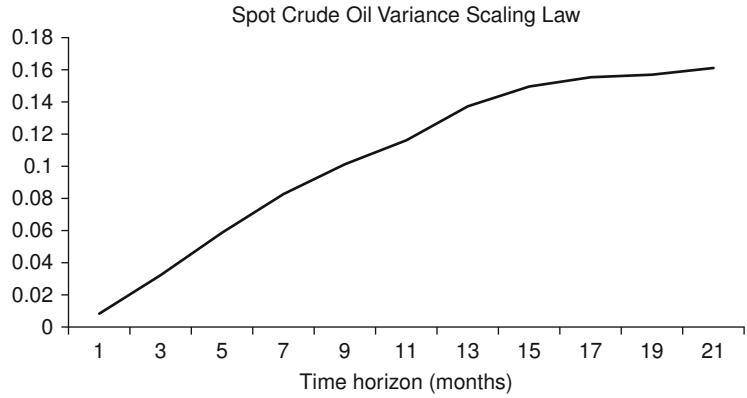



**Figure 2.9** Variance scaling law for spot Brent

on a forward basis. We are referring of course to the term structure of (forward) volatility.

We look at the realized (ATM) option replication volatility (for crude and natural gas) across a set of contracts, for different times-to-maturity. We term this volatility *quadratic variation* (QV). (The methods underlying this analysis will be made clear in Section 3.1.3.) Results are shown in Figures 2.10 and 2.11, for different samples (we omit the period associated with the 2008 financial crisis).

We will see later simple models where a basic mean-reverting log-price spot model gives rise to (implied) forward dynamics exhibiting a volatility term structure (see Section 5.2.7). This result is reflected in Figures 2.10 and 2.11, demonstrating consistency with the scaling law behavior of spot variance shown in Figures 2.8 and 2.9. In truth the picture is somewhat complicated for crude oil. Precrisis, forward crude exhibits typical commodity-like behavior (Samuelson effect/term structure

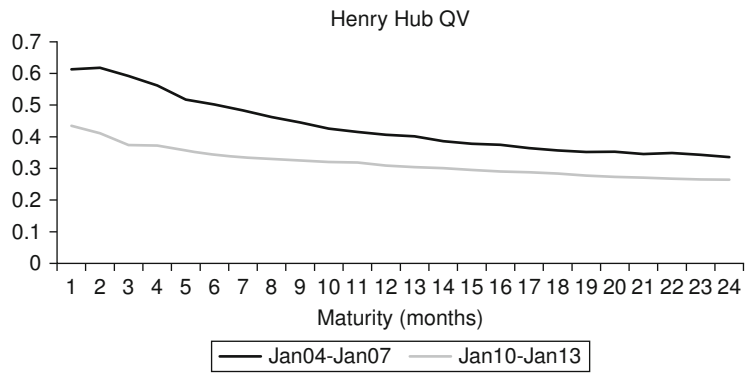



**Figure 2.10** QV/replication volatility term structure, natural gas. Different samples, winter

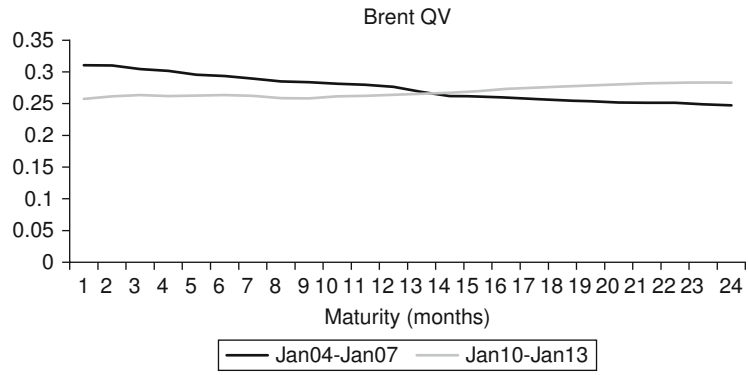



**Figure 2.11** QV/replication volatility term structure, crude oil. Different samples

of volatility). Postcrisis, the volatility term structure is quite flat, almost like equities. (The backend rise reflects the volatility collected through the chaos of the crisis in mid-2008, which would have affected cal-10 contracts.) There appears to be some kind of dissociation between spot and forward crude during this period. We cannot speculate on the reasons for this here, except to note a general pattern of financialization of crude postcrisis.[46]

We now examine the variance scaling laws of forwards as such. With these we begin to anticipate the difference between static and dynamic hedging strategies in commodity markets (essentially, the difference between variance and QV).

### 2.2.5 Forward prices

Unlike spot prices, forward prices[47] *do* arise from transactions/trades (by definition). In liquid futures markets, we expect forward prices to be reasonably informationally efficient, and would not expect that conditioning on current prices will provide exploitable information about futures prices (such as expected values). In other words, we would expect the martingale hypothesis about futures prices to hold. Obviously, we cannot conduct an exhaustive treatment of this subject here.[48] We will illustrate the central ideas as follows. First, we show front month futures time series for Brent crude oil and Henry Hub natural gas, during a rather tumultuous period for commodities (and economic entities in general): January 2008 through December 14. See Figures 2.12 and 2.13.

As with the corresponding spot-price time-series considered in Section 2.2.4, these futures time series show little visual evidence of stationarity. However, it cannot be denied that there are subsamples of these series that have some (superficial, we will see) appearance of mean reversion ("range-bound" in trading parlance). In particular, front-month Brent appears fairly stable in the period April 11–July 14.[49]

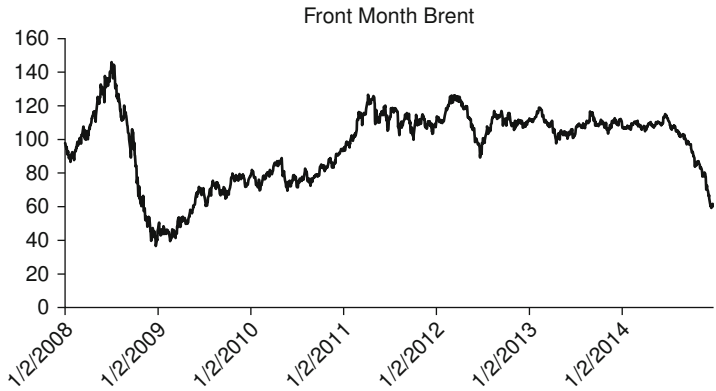



**Figure 2.12** Front month futures prices, crude oil, daily resolution.
*Source*: quandl.com.

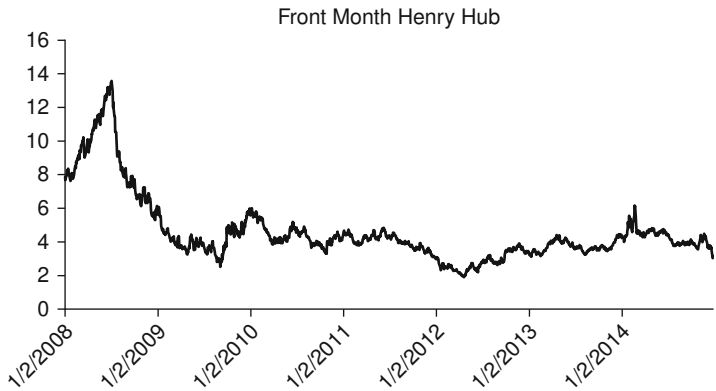



**Figure 2.13** Front month futures prices, natural gas, daily resolution.
*Source*: quandl.com.

One might expect to see some evidence of stationarity in the variance scaling laws, confined to this sample. We examine these in Figures 2.14 and 2.15.

We see, as expected, that forward natural gas displays little evidence of stationary or mean-reverting effects. Forward crude, over the sample in question, *does* show such evidence. However, it is *very* slow mean reversion (if that is in fact an accurate characterization). It is difficult to view forward crude as exploitable on a short-term basis.[50]

### 2.2.6 Demand side: temperature

The prototypical demand driver in energy markets is of course weather, and somewhat more narrowly temperature. Hot summers mean more demand for air conditioning and hence high electricity prices; cold winters mean more demand for

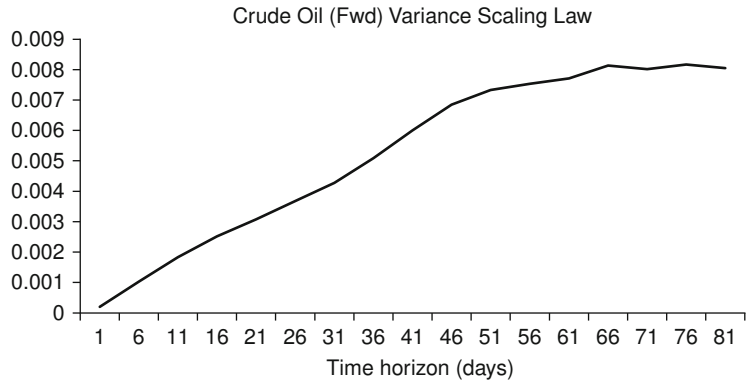



**Figure 2.14** Brent scaling law, April 11–July 14 subsample

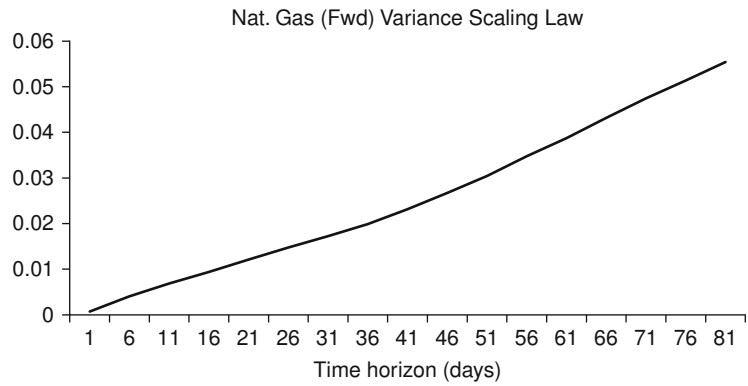



**Figure 2.15** Henry Hub scaling law, April 11–July 14 subsample

heat and hence high natural gas prices.[51] These differing demand patterns translate into well-known seasonal structures of energy price (and volatility) forward curves. (Crude oil markets do not really display seasonal effects.) This predictability reflects the deterministic aspect of temperature, shown in Figure 2.16.

This (apparent) periodicity strongly suggests the presence of stationary effects, and thus that we can meaningfully extract unconditional information by averaging across the sample. In other words, we can deseasonalize the time series. By subtracting the average January, February, *etc.* temperatures from the corresponding raw temperature data, we obtain a time series (of residuals) which displays distinctive patterns associated with mean reversion. We deduce this from the variance scaling law over increasing time horizons, as shown in Figure 2.17.

This result should not be surprising. From everyday experience we know that unusually hot or cold days are typically followed by a day or two of such weather,

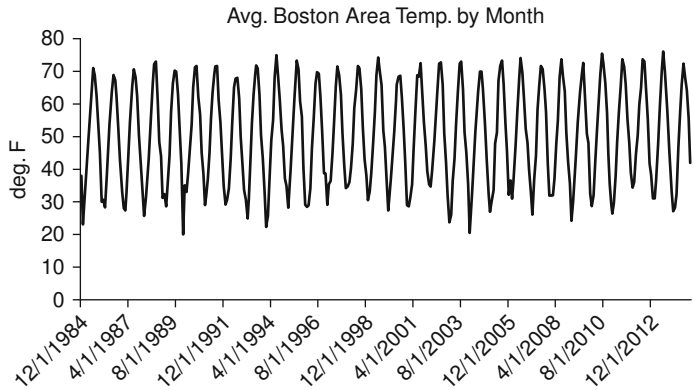



**Figure 2.16** Average Boston area temperatures by month.
*Source*: National Climatic Data Center (www.ncdc.noaa.gov).

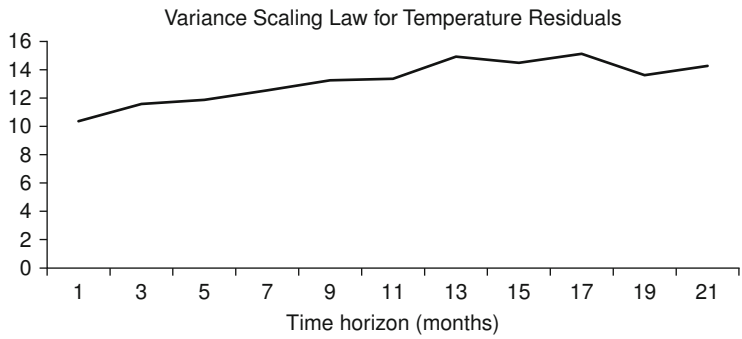



**Figure 2.17** Variance scaling for Boston temperature residuals

meaning that deviations from the norm (and indeed, our experience suggests that we can meaningfully speak of normal weather in the first place) do indeed peter out, but only after some lag period. We generally see this same kind of behavior at different levels of resolution (*e.g.*, monthly). While the characterization of (deseasonalized) temperature as mean-reverting is accurate, there are in fact multiple stationary factors at play here, over differing time scales. By fitting an AR(1) model to these temperature residuals, we typically find values for the auto-correlation parameter in the range 0.2–0.3 and values for the disturbance variance in the range 6–7. Such values give rise to variance scaling laws that asymptote much faster than

the behavior shown in Figure 2.17. Much better fits are given by two-factor models, with (say) independent components and variance scaling laws given by

$$\frac{\sigma_x^2}{2\kappa_x}(1-e^{-2\kappa_x\tau})+\frac{\sigma_y^2}{2\kappa_y}(1-e^{-2\kappa_y\tau}) \tag{2.50}$$

with $\kappa_x \gg \kappa_y$. In other words, the residuals are better described by a quickly reverting component and a more slowly varying transient. We do not address here the issue of the origin of such (relatively more non-stationary) effects, but merely note the clear evidence for their existence.[52]

### 2.2.7 Supply side: heat rates, spreads, and production structure

Like fundamentals on the demand side (*e.g.*, temperature), supply-side entities such as heat rates can also be (broadly) characterized as stationary/mean-reverting. This is not too surprising, owing to the nature of the generation stack (or more accurately, the manner in which *marginal* capacity is brought online/taken offline in response to changes in [stationary] demand[53]). In truth, owing to various complications of marginal cost (such as fuel-switching units in the U.S. Northeast), it is not straightforward to craft a definitive, single measure of *the* market heat rate. However, for our purposes here we can simply show the ratio of a representative (monthly) price and a representative (monthly) gas price; see Figure 2.18.

Visually, (spot) heat rates appear stationary, certainly more stationary than prices themselves. This appearance is borne out by the variance scaling law; see Figure 2.19.[54]

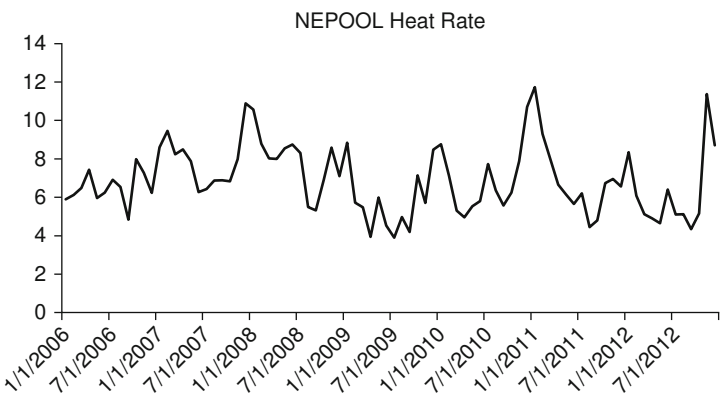



**Figure 2.18** Representative market heat rate (spot). Monthly MA-Hub 5 × 16 price divided by monthly Massachusetts Citygate natural gas price.
*Source*: NE ISO (power) and EIA (gas).

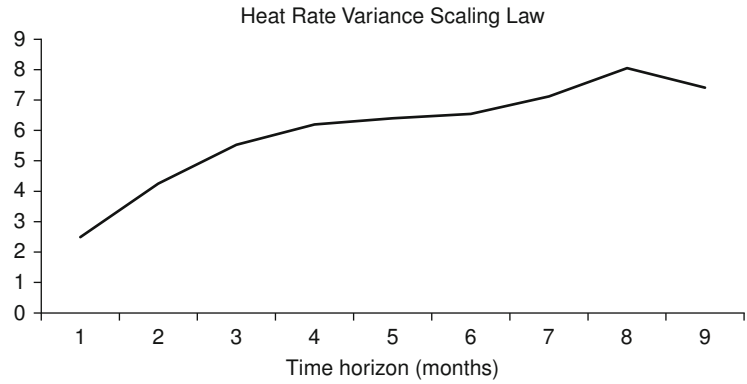



**Figure 2.19** Variance scaling law for spot heat. Same pair as

## 2.3 A recap

Let us review what we have seen thus far, and how it is relevant for the next steps. First, we have argued that the objective of econometrics as it concerns valuation is not formal hypothesis testing as such (however useful that language often may be), but rather the extraction of information about the exposure that results from particular hedging strategies around the product being valued. It is actually good that we do not rely on hypothesis testing, as we saw very simple examples where the results of standard statistical tests would give extremely misleading results regarding inferences concerning certain estimated entities. More than this: flawed conclusions about the nature of the underlying process itself can arise. However, we did see that some entities, namely variances/covariances, often serve as suitable entities of interest in risk assessment as well as possess estimators with reasonably stable characteristics. In addition, we saw the role of variance scaling as revealing the nature of information flows impacting processes of interest, and how they are proxies for entities which are generally *not* robustly estimated (such as mean-reversion rates). We further note here that the role of time horizons in information flows is particularly important in commodity markets.

It is not terribly surprising that standard deviations are so frequently resorted to in statistical analysis. They are generally robustly estimated and often indicative of underlying structures of interest. We speculate that they could potentially be of great use in valuation programs. This is the subject of our next investigation.

# 3 Valuation, Portfolios, and Optimization

## 3.1 Optionality, hedging, and valuation

We now turn to fundamental questions of valuation and hedging, and their close (indeed, inseparable) connection; we will see that they are really two sides of the same coin. We have already introduced a series of (spread) option structures in Chapter 1 in the context of energy market structures, and we will later focus the discussion on such products. However, it should be stressed that the ideas to be presented here are quite general, and can (in principle, at least) be applied to any nonlinear product.

### 3.1.1 Valuation as a portfolio construction problem

#### *3.1.1.1 Payoffs*

With this in mind, let us lay out abstractly the problem we wish to solve. In general we will be presented with a structure with a payoff that can be expressed as

$$\mathfrak{V}(S_{t\leq\tau\leq T};\vartheta_{t\leq\tau\leq T}) \tag{3.1}$$

where $t$ is the current time, $T$ is expiry, $S$ is a spot process (in general vector-valued), and $\vartheta$ denotes some process of controls/actions. The form (3.1) indicates that the payoff depends on the realization of some commodity prices (potentially across some time horizon [not just expiry]), as well as (possibly) some set of actions undertaken across the term of the deal.[1]

Although it may seem obvious, payoffs are central to any valuation problem. If we can relate the payoff to other entities in the market for which (transactable) prices exist, we have gone a long way toward valuing (and equivalently, replicating/hedging) the structure in question. This point is worth stressing because there is an unfortunate tendency to overemphasize explicit modeling of the processes underlying a given structure. This step is often a detour at best and sometimes a barrier to robust valuation at worst. (We will explore these themes in greater detail in Section 3.2 and Chapter 6.) There is a widespread belief that a suitably chosen

price process (capturing certain stylized facts such as jumps or stochastic volatility) that is (somehow) made consistent with ("calibrated to") available market quotes can then be used to value and hedge other products. This common procedure actually has the steps backward: it is value itself that must first be modeled, in such a way as to be robust under as wide a range of price processes as possible. (It goes without saying that we never actually know what processes prevail in reality.) Let us look closer at representations of value.

#### *3.1.1.2 Measures and portfolios*

Generically, we are interested in the expected value of this payoff, conditional on information at the current time:

$$E_t^M \mathfrak{V}(S_{t\le\tau\le T};\vartheta_{t\le\tau\le T}) \tag{3.2}$$

It is important to notice that this conditional expectation is taken with respect to some (as yet unspecified) probability measure $M$, a concept we will elaborate upon in Section 3.2 (and Chapter 5, as well). The point we wish to convey here is that the particular choice of measure is closely related to (indeed, inseparable from) a *specific* hedging/trading strategy put on against the structure in question.[2] In other words, we will be considering specific *portfolios* comprised of a payoff structure (such as $\mathfrak{V}$) and some set of traded instruments. These instruments, which we will denote by $X_k$, have some (market) prices that relate to the underlying spot prices[3] as follows:

$$F_{t,T}^{X_k} = E_t^Q X_k(S(T)) \tag{3.3}$$

For example, for futures we would have $X_k(S) = S_k$,[4] and for options we would have $X_k(S) = (S_k - K)^+$ (for some strike $K$).[5] Note, of course, that $F_{T,T}^{X_k} = X_k(S(T))$.

From (3.3), it can be seen that the prices of these instruments are equal to the expected value of their terminal payoff under a particular probability measure $Q$, which we will term the *pricing measure* (in the customary nomenclature). In other words, they are ($Q-$) martingales.[6] We consequently seek to consider portfolios of the following form:

$$\Pi = \mathfrak{V}(S_{t\le\tau\le T};\vartheta_{t\le\tau\le T}) - V_t + \sum_{\tau,p,k} \Delta_{\tau,p}^k (F_{\tau+\delta_k,T_p}^{X_k} - F_{\tau,T_p}^{X_k}) \tag{3.4}$$

That is to say, the portfolio consists of the structure in question and some specific positions $\Delta_{\tau,p}^k$ in the traded instruments $X_k$, with particular expiries $T_p$ within the time period specified by the structure.[7] The positions are in general dynamic, with the rebalancing period $\delta_k$ being instrument-dependent; some energy futures can be rebalanced relatively frequently (*e.g.*, more than once a day),[8] others less so. (The representation in (3.4) obviously includes static hedges as a special case.) Finally,

we have an initial value or premium $V_t$ that is to be paid out (for purchases) or received (for sales) to acquire exposure to the structure. In truth, this value is properly conceived as a process in its own right, and we shall refer to it as a *value function*. We now raise the following questions:

- Is there a connection between the value function $V_t$ and the (conditional) expected value of the structure payoff in (3.2)?
- If so, what is the relation between the aforementioned measures $M$ and $Q$?
- What is the precise role of the hedge positions $\Delta^k_{\tau,p}$ in (3.4)? How (if at all) do they relate to the value function $V_t$ ?
- How should the actions $\vartheta$ be (optimally) chosen, in light of operational flexibility/constraints?

Ultimately, the portfolio in (3.4) creates a type of *residual* exposure. It is this exposure that we are primarily concerned with. It is important to stress that hedging does not (necessarily) *reduce* risk (as it is commonly claimed); rather, hedging changes the *nature* of risk. Our objective is to understand and account for this (transformed) risk by identifying an appropriate set of entities that encapsulate the exposure through their expectation and manifestation in a *specific* portfolio. We will term such entities *value drivers*, and will have much more to say about them. For now, we simply note that in terms of some residual risk function $\varepsilon$ and a value driver $\upsilon$ and its projection/expectation $\hat{\upsilon}$, the expression (3.4) can be synopsized as

$$\Pi = \varepsilon(\upsilon, \hat{\upsilon}; V_t, \Delta, \vartheta)^9 \tag{3.5}$$

We can thus add another question to our list:

- What is the relationship between the value drivers and the value function, hedges, and controls?

Perhaps an illustration is in order. Consider a special case of the portfolio construction (3.4) in continuous time, with trading in the (observable) underlyings only:

$$\Pi = \mathfrak{V}(S^1_T, S^2_T) - V(S^1, S^2; \hat{\sigma}) + \int_t^T \Delta^1_s dS^1_s + \int_t^T \Delta^2_s dS^2_s \tag{3.6}$$

for some arbitrary payoff function $\mathfrak{V}$ (*e.g.* $(S^2 - S^1)^+$ for a spread option). The parameter $\hat{\sigma}$ is a projection of some (pathwise) property of the joint process $(S^1, S^2)$, *e.g.*, the volatility of the ratio $S^2/S^1$ (as in the case of a heat rate).[10] The value function $V$ (as well as the hedges $\Delta^{1,2}$) depends on this parameter. The precise manner in which the portfolio (3.6) depends on the realization of this (again, pathwise)

property (call it $\sigma$) will determine the nature of the residual risk (3.5). (Consequently, depending on our risk tolerance, we do not project value drivers as such, but rather risk-adjusted projections.) In fact, it is not clear at all that the portfolio creates an exposure to any arbitrary property of the (physical) process, as portfolio construction depends critically as well on the properties of the underlying market, most importantly liquidity. In other words, it is only *specific* properties that can in general be extracted from particular hedging regimens (more than this, from a practical standpoint: only certain properties that can further be robustly estimated).[11] So yet another question is raised:

- What role does the underlying market structure play in portfolio construction (and hence valuation)?

Much of our concern in this work will be with addressing these questions. We will see that these points are all closely intertwined.

#### *3.1.1.3 A unified approach to valuation*

In fact, we can already call attention to two strands of thought that are often viewed as diametrically opposed in the literature but actually share a dual nature: spot vs. forward valuation. Broadly speaking, one typically sees two approaches to valuation. Spot-based valuation, as the name suggests, attempts to model the underlying spot processes against which payoffs, physical operation, and constraints must be applied. Forward-based valuation attempts to represent (or approximate) the structure in question in terms of derivative products that either do trade in the underlying or can be replicated therein. (The two approaches can also be characterized as physically based or financially based, for obvious reasons.)

Both approaches have their strengths and weaknesses. Spot-based methods, while being better tailored to the inherent physical optionality of a given commodity structure, must somehow be linked to (or conditioned upon) existing market information (such as prices and volatilities) and efficiently hedged with market instruments.[12] Forward-based methods obviously have a direct connection to the market, but at the cost of suboptimally representing the underlying physical optionality (if it in fact captures any of it at all).

In truth we regard this dichotomy to be a false one. First of all, in liquid futures markets, it is always more efficient to bet on prices directly (through futures positions) than indirectly via some structured product. That is, regardless of one's view of prices, a structured product hedged with futures creates some non-price exposure representing a new bet, and it is suboptimal to mix this bet with a bet on prices. It is imperative that one be able to clearly identify the sources of profit (or loss) from a particular portfolio. It is in this sense that model "calibration" to futures makes sense, and is indeed mandatory: namely, within the context of a specific portfolio involving these traded instruments. Second, much of the disconnection between

spot-based and forward-based modeling is ultimately an issue of market resolution. In most markets, there is only limited (if any) ability to hedge with instruments settling at the level of individual days (or hours). Thus, flexibility that may appear at such levels from a spot perspective may in fact be largely illusory, if no (operational) means exist for extracting or monetizing this flexibility. We will in fact see examples where, as the level of resolution of traded instruments increases, there is convergence between spot- and forward-based valuations. Finally, we will see that there exists means for bounding forward-based valuations and thus providing an assessment or diagnostic of how well a particular (necessarily lower bound) representation of operational optionality performs. In other words, we can (often) tell how much incremental value we omit with a particular approximation to physical flexibility.

The viewpoints presented here will perhaps become clearer by considering a very well-known example.

### 3.1.2 Black Scholes as a paradigm

Much of what we have just said in the last section can now be considered in the context of standard approaches to the problem of option valuation, a task we now undertake. First we start with some overview. The theory of rational option pricing (by which we mean pricing without reference to subjective preferences) is very well developed, and we will only cover the basics here. The seminal references are of course Black and Scholes (1973) and Merton (1990). The essential idea is conveyed (albeit with insufficient rigor) in Hull (2005). A superb treatment that does not sacrifice clarity for rigor can be found in Baxter and Rennie (1996). We will establish in this section the essential concepts that we will employ throughout the volume. We intend to emphasize the central ideas that are often either obscured by standard and (necessarily) simplified treatments of the subject or sidetracked by excessive technical detail. Specifically, our focus will be on the notion of *replication* of the desired payoff (which in general will be much more complicated than a vanilla option). Replication is impacted both directly by the structure of the underlying market, and indirectly by the features of the actual stochastic processes driving the prices of the assets in question. However, it remains highly instructive to consider option pricing under the usual (again, necessarily) idealized conditions that are standard in the literature, with the caveat that the standard argument has wider applicability beyond these narrow assumptions.

#### *3.1.2.1 Portfolios and hedging*

We start by assuming that the underlying price process follows Geometric Brownian Motion (GBM) under the physical measure:[13,14]

$$\frac{dS}{S} = \mu\, dt + \sigma\, dw \tag{3.7}$$

with the drift $\mu$ and volatility $\sigma$ known constants. Now consider the price $V$ of a call option[15] (with strike price $K$ and [terminal] payoff $(S-K)^+$) on this asset, expiring at time $T$. We want to know what, if any, relationship exists between the price of the option and the properties of the underlying asset. To derive useful results requires that we make certain assumptions about the overall market structure. One assumption almost universally resorted to is that of absence of arbitrage. There exist various ways of formalizing this assumption, but it has a very simple meaning: if the terminal payoffs of two securities are always the same, then the current prices of those securities must be equal ("the law of one price").[16] Another frequently used assumption regards liquidity, the extent to which trading in the underlying asset is possible, and at what cost. The most common such assumption is that the underlying can be dynamically traded costlessly in continuous time. We will discuss in this volume the ramifications of relaxing these assumptions. For now, we will simply review the standard results that can be drawn from them.

We will hold the following portfolio: a (long) static position in an option, and a (short) position $-\Delta_t$ in the underlying. Ignoring discounting effects (for convenience), the portfolio can be written

$$\Pi = (S_T - K)^+ - V - \int_t^T \Delta_s dS_s = \int_t^T (dV_s - \Delta_s dS_s) \tag{3.8}$$

Using basic results from Itô calculus, (3.8) becomes[17]

$$\Pi = \int_t^T \left( \frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} \right)_s ds + \int_t^T (V_S - \Delta_s) dS_s \tag{3.9}$$

Assume we choose the hedge position to satisfy $\Delta_s = V_S$. Then, the standard argument claims that the portfolio is "instantaneously riskless," and hence, by the assumption of no arbitrage, the portfolio (3.9) is equivalent to a risk-free bond. Since we are neglecting interest rates, this means that the option price satisfies the celebrated Black-Scholes (BS) equation:

$$V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} = 0 \tag{3.10}$$

with terminal condition $V(S,T) = (S-K)^+$. We will note here the well-known analytic solution:

$$V(S,t) = S \cdot N\left( \frac{\log(S/K) + \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{T-t}} \right) - K \cdot N\left( \frac{\log(S/K) - \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{T-t}} \right) \tag{3.11}$$

(It is straightforward to show that the coefficient of $S$ in (3.11) is the option delta $V_S$.)

The heuristic argument used to derive (3.10) can be made more rigorous (*e.g.*, Carr [2002]), but we will here provide a more "global" approach, so to speak, that will serve to illustrate approaches that will prove useful when the standard assumptions are relaxed.

*3.1.2.2 Prelude to value drivers*

It is important to understand what was established in the previous subsection. Under the specified assumptions (GBM with known volatility, costless continuous time rehedging, *etc.*), *if* the option trades (and thus has a market price), and *if* the market in question is arbitrage-free, then the option price *must* be the Black-Scholes result. This point can be seen in another way.

Note that, as a convention, the price of the option can be expressed in terms of any other volatility, not necessarily the volatility governing the data-generating process. Indeed, actual market option prices are commonly quoted in terms of the so-called implied volatility, the volatility that recovers the observed price when used in the BS formula. Denoting this volatility by $\sigma_I$, we have (by definition) that

$$V_t + \tfrac{1}{2}\sigma_I^2 S^2 V_{SS} = 0 \tag{3.12}$$

Thus, when we hedge with the BS delta, the portfolio (3.9) can be written as

$$\Pi = \tfrac{1}{2}(\sigma^2 - \sigma_I^2) \int_t^T S_s^2 \frac{\partial^2 V_s}{\partial S^2} ds \tag{3.13}$$

Now note that, by convexity of the payoff function and the fact that the option price is homogenous of degree one,[18] the option "gamma" is strictly positive: $V_{ss} > 0$. Thus, under this hedging strategy the portfolio has a definite sign. If the implied volatility is less than the volatility under the physical measure, then this hedging strategy will yield a profit with certainty. (Similarly, if the implied volatility is greater than the physical volatility, the reverse portfolio [short option, long underlying] will yield a profit with certainty.) Since this contradicts the assumption of non-arbitrage, we cannot have a situation where the market (implied) volatility differs from the volatility of the data-generating process. We have thus established that a *necessary* condition for non-arbitrage in this model is that market option prices satisfy the BS equation. Let us stress this point: we have established here a relation between a market price and a specific property of the data generating process (DGP), as must prevail under a specific market structure.

The situation is rather different when the option in question does *not* trade and hence no market price exists for it; indeed, the very point of valuation is to establish

a price at which one wishes to transact. In the case where the process volatility $\sigma$ is known, the two parties involved in the transaction can readily agree on the price: both sides will take $\sigma_I = \sigma$, as the exposure on either side can be precisely replicated through delta-hedging with this volatility. There is no reason for buyers to bid more than this volatility, and no reason for sellers to ask less. In fact, there appears to be little point in engaging in this transaction. As is well known, in complete markets (such as the BS setting), options are essentially redundant securities. Matters are considerably more interesting in the (incomplete) market case where volatility is not known (and indeed, might not even be deterministic), and this case will be the subject of much of our focus later (in Section 3.2).[19] In this case, some view must be taken regarding the realization of volatility, a view which must be reflected in the pricing *and* hedging, as these two entities (realized vs. projected) manifest themselves in the overall portfolio exposure, as in (3.13).

We must also emphasize here a theme that we will employ throughout this book, namely that valuation of an instrument only makes sense within the context of an overall portfolio that includes a specific hedging strategy. For example, the BS paradigm leading to (3.13) entails the BS delta under a particular volatility. A central part of such strategies (already introduced in Section 3.1.1) is what we will term *value drivers*, or parameters that entail the following:

1. A mapping taking parameters onto values and hedges
2. A projection of their realization over the term of the contract
3. A connection to residual risk around a definite portfolio.

Diagrammatically, we have

$$\begin{aligned}&\text{Portfolio}(t,T) \Rightarrow \\ &\text{Terminal Payoff}(T) - \text{Initial Premium}(t,\hat{\sigma}) - \text{Accrued Hedges}(t,T,\hat{\sigma}) \Rightarrow \\ &\text{Residual Risk}(t,T;\hat{\sigma},\sigma)\end{aligned} \tag{3.14}$$

(Recall the discussion leading up to (3.5).)

In this generic notation, $\sigma$ represents the value driver, which often will be some entity such as (realized) volatility but as will be seen, can be much more general (*e.g.*, a correlation).[20] Its projection is denoted by $\hat{\sigma}$. The functional dependencies make explicit the fact that (over the time horizon of the deal) the (projected) value drivers are a means for producing both the price to pay[21] for entering the deal *and* the hedges to pursue over the course of the deal. Finally, the penultimate cash flow of the portfolio is likewise a function of both this value driver *and* the projection. It is thus stochastic, and hence the task is to not only identify an appropriate value driver, but also to project it in accordance with the residual exposure arising from the (aggregate) portfolio (which is what the residual risk term represents).[22]

In fact, in speaking of projection, we necessarily introduce an econometric aspect to the problem. Namely, a (good) value driver must possess some degree of stability in relation to the (valuation) problem at hand. It must not only be stable across the available information set, but must be (reasonably) expected to retain this stability over the time horizon in question. We have already encountered such issues in Chapter 2 (and we will provide a more rigorous exposition of terms such as "stability" in Chapter 6 when we consider econometrics more abstractly), but essentially we mean that the actual numerical expression (of a value driver) that we extract from a given set of data does not vary wildly when that set changes (*e.g.*, due to the arrival of new information, *etc.*)

With reference to the classic BS model, we saw that, if we priced and hedged an option under the BS functional for which the (market implied) volatility differed from the volatility of the underlying asset, the residual was strictly positive (with certainty!). With the additional constraint that the underlying market was arbitrage-free, we then concluded that the implied volatility could *not* differ from the asset volatility, and thus that the BS price was the "right" one. But again, this result followed from the basic portfolio argument, and not abstract economics. A final point that cannot be emphasized enough: available hedging strategies are driven by the liquidity of the market. We will study in this book various ramifications of departures from perfect liquidity. For now it is sufficient to note that the BS argument depends critically on the existence of such liquidity.

#### *3.1.2.3 Coming back to commodities*

Before tying up a few loose ends, we need to connect the previous discussion to the overall objective of valuation in energy markets. We have already mentioned the role of liquidity in valuation and hedging. We will consider these issues as they pertain to energy markets later in the chapter, but can simply note here the range of traded instruments is rather more truncated in energy markets (indeed, more generally commodity markets) than in financial or equity markets. Thus, the ability to form specific portfolios is correspondingly more limited in energy markets, a fact that has important implications for how we must approach the valuation problem.

There is another important point to be raised here. We note that the drift in (3.7) is completely unspecified (apart from being a constant), and in fact is completely absent in the valuation formula (3.11). (In the standard business school exposition, the drift is "hedged away.") Far from being an irrelevant annoyance, the drift of commodity processes is at the heart of valuation in energy (and commodity) markets. Actually, the constancy of drift is not necessary for the BS argument, and here lies the issue in energy markets. It is precisely because of the physical factors impacting price dynamics in energy markets that the phenomenon of volatility time scales arises. For our purposes right now, we can very broadly characterize this behavior as mean reversion. As we will see, such effects critically affect the kind of value that can be extracted from different hedging strategies. Specifically, a static

strategy will collect less value than a dynamic strategy. We will see in the course of our discussion that in energy markets, it is imperative that we identify the kinds of hedging and trading strategies that are available around a given structure, as this set will determine the kinds of exposure the resulting portfolio creates, and thus what kind of value we can attribute to that structure.

#### *3.1.2.4 A first look at measures*

It is useful to highlight an important concept that we will discuss in greater detail later in Chapter 5. Under the physical measure, prices have the following expectation:

$$E_t^P S_T = Se^{\mu(T-t)} \tag{3.15}$$

This of course means that prices are not *martingales*, which are random variables for which $E_t X_T = X_t$. However, it can be seen from the BS formula that the drift $\mu$ does not appear at all, and in fact from the Feynman-Kac formula, the BS value is an expectation of the terminal payoff with respect to a measure under which prices have no drift.[23] That is, valuation can be formulated in terms of a martingale measure. Specifically, under a measure $Q$[24] for which the dynamics are given by

$$\frac{dS}{S} = \sigma\, dw, \tag{3.16}$$

the option price is given by

$$V = E_t^Q (S_T - K)^+ \tag{3.17}$$

This measure is commonly referred to as the risk-neutral measure, although this is somewhat misleading as the terminology obscures somewhat the notion of relative pricing inherent in this approach to valuation.[25] It is not hard to see that the associated delta (*i.e.*, the dynamic hedging volume) is given by

$$\Delta = E_t^{Q_S} 1(S_T > K) \tag{3.18}$$

where the measure $Q_S$ defines the following dynamics:

$$\frac{dS}{S} = \sigma^2 dt + \sigma\, dw \tag{3.19}$$

(Recall (3.11).) As will be seen later, this measure corresponds to redenominating the underlying market in terms of $S$ (*i.e.*, the asset is now numeraire instead of cash). In other words, under this measure, $S^{-1}$ is a martingale (as is easy to check).

We will have much more to say about changes of measure and representation of value in terms of martingale pricings in Section 3.2 (and further in Chapter 5).

However, what must be stressed now is the fact that these approaches are fundamentally encapsulations of a particular hedging strategy. This can be seen very plainly in the discussion here, where the BS hedging paradigm produces a portfolio that is (pathwise) identically zero. This amounts to saying that the option can be precisely replicated by following this hedging strategy. We will see that these concepts retain great utility even in those cases where such perfect replication is not possible. But it will remain the case that such representations are inextricably connected to some hedging strategy, in particular through the nature of unhedgeable risk. The chief point here is that one cannot first specify a price measure and then compute hedges; rather, the process works the other way around.

We thus see that even within the basic and familiar BS framework, there are a number of subtle and deep points which will form the basis of valuation in much more general settings. Now, we must turn attention to the impact imposed by liquidity constraints, specifically as they concern the kinds of portfolios that can be formed around a particular valuation problem.

### 3.1.3 Static vs. dynamic strategies

#### *3.1.3.1 Quadratic variation vs. variance*

Previously we have assumed that dynamic hedging in the underlying is possible (in continuous time). This is of course unrealistic, especially in energy markets, where there are typically a few backbone or hub markets, and then local, geographically affiliated markets that trade as offsets to this hub, and only (if at all) become independently traded entities over small time horizons (*e.g.*, a year at most). In such cases, it will only be possible to put on a static hedge against an option on such underlyings. (It may only be possible to put on a proxy hedge; this case will be considered later, as well.) The option portfolio would become

$$\Pi = (S_T - K)^+ - V - \int_t^T \Delta_t dS_s = (S_T - K)^+ - V - \Delta(S_T - S) \tag{3.20}$$

Conceptually, the problem posed in (3.20) is really no different than before: assuming we retain a BS pricing functional, we must choose a projected volatility under which to establish the initial premium and initial (static) hedge. As we continually promise, detailed examples will be considered in subsequent sections. What we wish to call attention to here is the fact that the character of the physical price process (or more accurately the scaling of the returns of the physical process) has an enormous impact on the kinds of value that can be extracted via hedging. Heuristically, in the static hedge case (3.20) we are essentially exposed to the variability of the underlying over the global (so to speak) time horizon, as opposed to the variability over local time horizons as in the dynamic hedge case.

Consider the intuition commonly put forth in non-rigorous derivations of the BS equation. Specifically, it is said that the process drift does not matter for valuation because it is canceled out by the delta hedge. This is not really a proper way of phrasing the issue, but it is probably acceptable as a pedagogical device on certain levels. The point to be taken away is that the nature of the hedge cannot be ignored: the drift can impact not only the expected value of an asset, but also the *variance* of that asset over a given time horizon. This is critically important in commodity markets in contrast to equity or financial markets, because of the presence of mean reverting effects. (The reader is again referred to EW for an extensive discussion of the fundamental drivers giving rise to mean reversion in energy markets.) We will have much more to say about this issue, but for now consider the following two models of log-prices:

$$dz = \left(\mu - \tfrac{1}{2}\sigma^2\right) dt + \sigma\, dw \tag{3.21}$$

and

$$dz = \kappa(\theta - z)dt + \sigma\, dw \tag{3.22}$$

Neither process is a martingale, but the instantaneous stochastic dynamics of both processes are the same. However, the variance over finite time horizons (call it $\tau$) differs greatly. For the process in (3.21) (geometric Brownian motion [GBM] with drift), this variance is

$$\sigma^2\tau \tag{3.23}$$

whereas for the process in (3.22) (mean reversion) the variance is given by

$$\sigma^2 \frac{1 - \exp(-2\kappa\tau)}{2\kappa} \tag{3.24}$$

These plainly have a very different character. In particular, for GBM the variance (3.23) grows linearly, while for mean reversion the variance (3.24) becomes asymptotically flat. This is seen in Figure 3.1:

The GBM variance always dominates the mean reverting variance (for the same instantaneous volatility $\sigma$), with the spread increasing with higher mean reversion rate ($\kappa$). The difference between the two expressions illustrates the distinction between quadratic variation and variance. The latter needs little explanation. The former should also be familiar and is essentially a measure of local variation (in an appropriate [probabilistic] limiting sense):

$$QV_\tau = \lim_{h\to 0} \sum_{i=0}^{[\tau/h]-1} (z((i+1)h) - z(ih))^2 \tag{3.25}$$

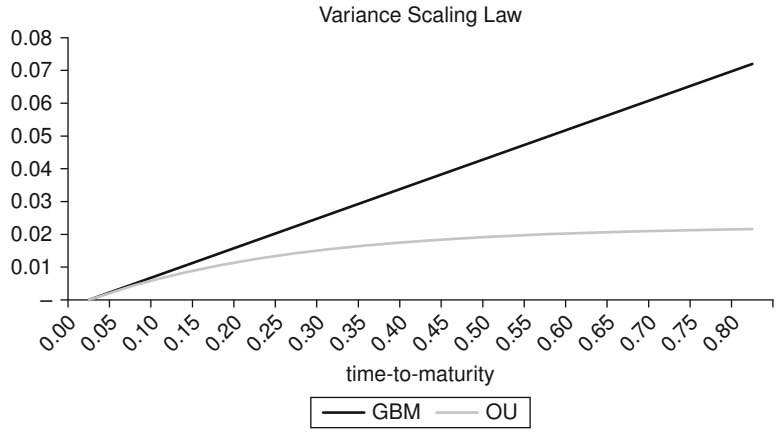



**Figure 3.1** Comparison of variance scaling laws for different processes. Geometric Brownian motion vs. Ornstein-Uhlenbeck (mean reversion). Same instantaneous volatility (30%), mean-reversion rate equal to 2

Note that while the two processes (GBM and mean-reverting) have very different variances, they have the same quadratic variation. This has very important ramifications for the value that can be collected with different hedging strategies under the two processes.

One might expect intuitively (from inspecting the portfolio expression (3.20)) that, when the underlying process follows GBM, local variance (that is to say, quadratic variation) can be captured even from a static hedging strategy, while if the process is mean-reverting, a much smaller variance can be captured (over a given time-to-maturity). It turns out that this intuition is correct. In fact, the situation is a bit richer than this, as will be seen when we consider spread options (options on the difference between two assets, say gas and power). Here, energy prices can exhibit effects of *both* drift and mean reversion, and the variance that can be collected from static hedging strategies depends critically on which asset is statically hedged and which one is dynamically hedged. Basically, if there are any mean reversion effects present, then dynamic hedging is essential for capturing the classic quadratic variation. This point is often overlooked in equity markets, but it is of vital importance in energy markets.

In fact, there is a much more general framework for considering this problem that does not involve a specific mechanism such as mean reversion (although this is a very common mechanism). Alternatively, what really matters here is the presence of *variance scaling laws*, a topic that received much attention in Chapter 2.[26] Obviously, mean reversion manifests itself in a (nonlinear) scaling of variance, while GBM does not. The significance of such scaling lies not only in the impact on

option pricing under different hedging regimes, but crucially in the fact that volatility estimation is typically more robust than direct estimation of mean reversion. Robustness of estimation and identification of value drivers is one of the central themes of this book.

#### *3.1.3.2 Spark spread example*

Let us consider a concrete example from Mahoney and Wolyniec (2012) to show how these effects can interact. A very common model for power prices relates the cost of production (some kind of fuel, call it gas) to the marginal generation unit (a heat rate). Owing to the nature of the supply stack and the principal demand drivers (typically weather),[27] we anticipate that heat rates will be mean-reverting. Fuel prices typically are not.[28] We thus put forth the following model (in terms of log-entities):

$$\begin{aligned} p &= g + h \\ dg &= \mu\, dt + \sigma_g dw_g \\ dh &= \kappa(\theta - h)dt + \sigma_h dw_h \end{aligned} \tag{3.26}$$

with instantaneous correlation represented by $dw_g dw_h = \rho\, dt$. From (3.26), the power dynamics are given by

$$dp = (\mu + \kappa(\theta - p + g))dt + \sigma_p dw_p \tag{3.27}$$

with $\sigma_p^2 = \sigma_g^2 + 2\rho\sigma_g\sigma_h + \sigma_h^2$.[29]

For fuel, the (cumulative) variance over some time horizon $\tau$ is given by the same expression in (3.23): $\sigma_g^2\tau$. The (cumulative) variance of power inherits behavior from (3.23) *and* (3.24). Using techniques that will be thoroughly explained in Chapter 5 (keep in mind the power process is still fundamentally Gaussian), this latter variance is given by

$$\sigma_g^2\tau + 2\rho\sigma_g\sigma_h\frac{1-\exp(-\kappa\tau)}{\kappa} + \sigma_h^2\frac{1-\exp(-2\kappa\tau)}{2\kappa} \tag{3.28}$$

Given the presence of both GBM and mean reversion in this system, we ask a similar question to the one posed in the preceding section: what are the implications for dynamic as opposed to static hedging strategies?

Consider a (spark) spread option between power and gas. (Assume for convenience that units have been chosen so that heat rate has unit expected value.) We

form the following portfolio around this structure (reverting to uppercase variables for non-logs):

$$\Pi = (P_T - G_T)^+ - V(G_t, P_t; \hat{\sigma}) - \int_t^T \Delta_s^P(\hat{\sigma}) dP_s - \int_t^T \Delta_s^G(\hat{\sigma}) dG_s \tag{3.29}$$

for some valuation/hedging volatility $\hat{\sigma}$ (*i.e.*, the value driver). As we will see in Section 7.1, the value function/value driver pair that permits perfect replication of the option is given (as a generalization of the standard BS case) by the so-called Margrabe formula:

$$\begin{aligned} V &= P \cdot N\left(\frac{\log(P/G) + \frac{1}{2}\hat{\sigma}^2\tau}{\hat{\sigma}\sqrt{\tau}}\right) - G \cdot N\left(\frac{\log(P/G) - \frac{1}{2}\hat{\sigma}^2\tau}{\hat{\sigma}\sqrt{\tau}}\right), \\ \hat{\sigma} &= \sigma_h \end{aligned} \tag{3.30}$$

and the deltas $\Delta^{P,G}$ are given by the coefficients of $P$ and $G$ in (3.30) (recall Euler's theorem; note that (3.30) can be expressed in terms of the BS functional as $G \cdot V^{BS}(\frac{P}{G}, 1, \hat{\sigma}, \tau)$).

We now refine our question by asking: does it matter if we only dynamically hedge one of the legs, and if so, which one allows for greater volatility to be collected? For example, consider the following portfolios:

$$\begin{aligned} \Pi^1 &= (P_T - G_T)^+ - V - \Delta_t^P(P_T - P_t) - \int_t^T \Delta_s^G dG_s \\ \Pi^2 &= (P_T - G_T)^+ - V - \int_t^T \Delta_s^P dP_s - \Delta_t^G(G_T - G_t) \\ \Pi^3 &= (P_T - G_T)^+ - V - \Delta_t^P(P_T - P_t) - \Delta_t^G(G_T - G_t) \end{aligned} \tag{3.31}$$

So if (3.29) represents the case of dynamically hedging both legs, the cases in (3.31) represent, respectively, statically hedged power/dynamically hedged gas, dynamically hedged power/statically hedged gas, and both statically hedged. We can now properly phrase our question: which of the (mixed) strategies in (3.31) will collect more value on an expected value basis? We will spare the reader any more suspense and present the results (from quadrature; see Chapter 7) in Figure 3.2:

Figure 3.2 represents the expectations, under the physical measure (*i.e.*, the observed dynamics in (3.26) and (3.27)), of the portfolios in (3.29) and (3.31) when the heat rate value $\sigma_h$ is 30%. (In Chapter 7 we will discuss techniques that can be

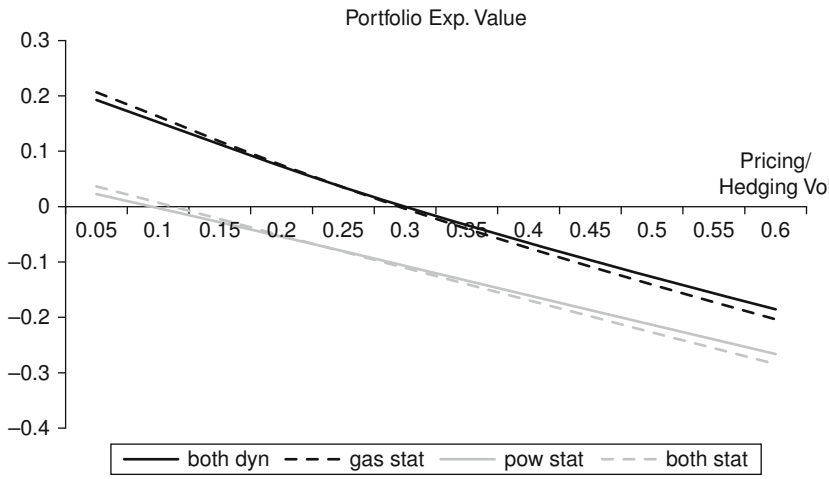



**Figure 3.2** Expected value from different hedging strategies. The true volatility is 30%. Failure to dynamically hedge power (in the model (3.26) and (3.27)) results in lower expected value for the same valuation/hedging volatility

used for facilitating these computations.) It can be seen from this figure that for the full dynamic case, the expected (portfolio) value is positive/negative when valuation and hedging is conducted at valuation volatilities below/above the true value. This result is not too surprising. What is interesting here is that when the power leg is dynamically hedged and the gas leg only statically hedged, the expected value is essentially the same as the full dynamic case. In contrast, when power is only statically hedged (but regardless of how gas is hedged) the expected value is less for a *given* valuation volatility. Alternatively, we can say that we must value (and hedge) at a *lower* volatility to attain a given expected value if we do not dynamically hedge power.

In other words, it is critical that we employ a dynamic hedging strategy regarding power if we hope to collect the volatility we project when bidding on this structure. What distinguishes power from gas here? We will see in Chapter 6 that there is a specific sense (Granger causality) in which gas can be thought of as driving power, and not *vice versa*. More crudely, it is the leg that has (roughly speaking) more mean-reverting effects (even if it is not mean-reverting as such) that has to be actively hedged to extract volatility sufficient to acquire the structure in the first place.[30] The intuition here is that *specific* (in this case, dynamic) hedging strategies are necessary to "suppress" (so to speak) mean reversion (however it is manifested) and extract "optimal" volatility. It is worth noting that this example is not a mere curiosity, but models behavior that we observe in actual energy markets (as we will see in the next subsection). We again contrast this example with the situation encountered in equity markets. Generally speaking, it is far less important in those markets whether dynamic or static hedging strategies are employed for extracting

volatility (indeed, static strategies may be preferable for reducing transaction costs). The situation is very different in energy markets, where the difference between static and dynamic strategies is of overriding importance.

*3.1.3.3 Forward volatility examples*

The example we have considered here is not merely an academic exercise; the behavior in question can be observed in actual markets. Consider the following portfolio:

$$\begin{aligned}\Pi(F_0^1, F_0^2, \sigma_\alpha) = (F_T^2 - \alpha F_T^1)^+ - F_0^1 \cdot V^{BS}\left(\frac{F_0^2}{F_0^1}, \alpha, \sigma_\alpha, T\right) \\ - \sum_{i=0}^{T-1} \Delta_K^{BS}\left(\frac{F^2_{D^1(i)}}{F^1_{D^1(i)}}, \alpha, \sigma_\alpha, \tau_{D^1(i)}\right)(F_{i+1}^1 - F_i^1) \\ - \alpha \sum_{i=0}^{T-1} \Delta_S^{BS}\left(\frac{F^2_{D^2(i)}}{F^1_{D^2(i)}}, \alpha, \sigma_\alpha, \tau_{D^2(i)}\right)(F_{i+1}^2 - F_i^2) \qquad (3.32)\end{aligned}$$

(Recall the valuation formula in (3.30).) In (3.32), $F^{1,2}$ denote futures prices, say gas (1) and power (2), $\alpha$ reflects the (inception) moneyness of the option in question (*e.g.*, $\alpha = \frac{F_0^2}{F_0^1}$ corresponds to ATM), and the hedge volumes $\Delta^{BS}_{S,K}$ denote the BS price and strike deltas, respectively (we will see how these arise in the well-known Margrabe formula for spread option pricing, to be discussed in Section 7.1; we have actually already encountered this result in (3.30)). The time horizon is given by $[0, T]$, and $\tau$ represents a time-to-maturity.

The static or dynamic nature of the hedges in (3.32) is controlled through the index mapping $D$. If $D^{1,2}(i) = i$, then both legs are dynamically hedged; if $D^{1,2}(i) = 0$, then both legs are dynamically hedged. Mixed cases (one leg dynamic, the other static) are handled by, say, taking $D^1(i) = i$ and $D^2(i) = 0$ (corresponding to a static power hedge/dynamic gas hedge).[31] In contrast to the case considered in Subsection 3.1.3.2 (where we looked at expected PnL as a function of hedging volatility), we instead ask, what volatility permits exact replication of the payoff in (3.32)? This amounts to solving the (nonlinear) equation $\Pi(\sigma_\alpha) = 0$ *across paths.*[32,33] A pathwise illustration of this estimator is shown in Figure 3.3.[34]

We carry out these calculations for an ATM PJM-W/Tetco M3 heat rate with up to two years to maturity, using the range of contracts Dec 06–Jul 14. We solve (3.32) for each (contract) time series in the sample, for different months-to-maturity, and average across the summer contracts (Jul–Sep; recall the cautionary tale in endnotes 33 and 34). The results are presented in Figure 3.4. Although the structure is not pristine (the static estimators are typically very noisy, especially for

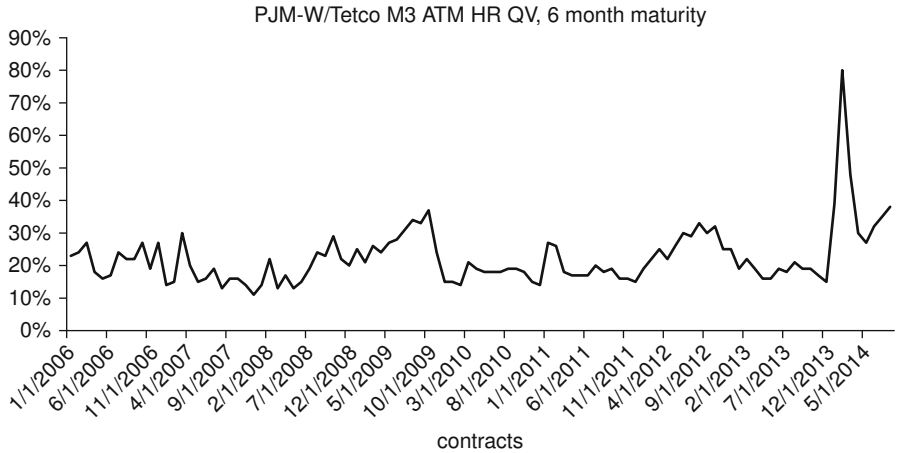



**Figure 3.3** Realized (pathwise) heat rate ATM QV. Using the algorithm in (3.22) (and based on futures data from SNL/Swiss Re). Note the generally stationary appearance, with a spike associated with the extreme weather conditions in the northeast United States in the winter of 2013–14. Output for different levels of moneyness can be similarly obtained; the general pattern will be that OTM/ITM effects are most pronounced closest to maturity. A useful exercise for the reader is to apply the change of measure techniques from Chapter 5 to show that *pathwise*, HR QV is largely invariant to calculations using gas as the unit of account (numeraire).
*Source*: SNL Financial LC. Contains copyrighted and trade secret material distributed under license from SNL. For recipient's internal use only.

shorter times-to-maturity), the main pattern can be seen, most clearly for longer times-to-maturity.

In fact, we can also illustrate the difference between QV (from a dynamic strategy) and variance (from a static strategy). Figure 3.5 shows, for summer contracts in the sample Jan 07–Dec 09, the average volatility collected from dynamic hedging (via (3.32)) vs. the static return volatility for the same (summer) sample. (The latter is essentially the annualized standard deviation of $F_{T,T}/F_{0,T}$ across the sample.) As expected, the dynamic strategy collects a higher volatility than the static strategy, reflecting the presence of mean-reverting effects (not to be confused here with the volatility term structure). We see here a very striking example of the difference between a pricing measure and the physical measure from probability theory.

#### *3.1.3.4 Cash volatility example*

The impact of a static hedge (which is often the only instrument available) can also be seen in an example where the issue is not mean reversion as such, but rather jumps. Consider the following portfolio (recall (3.32)):

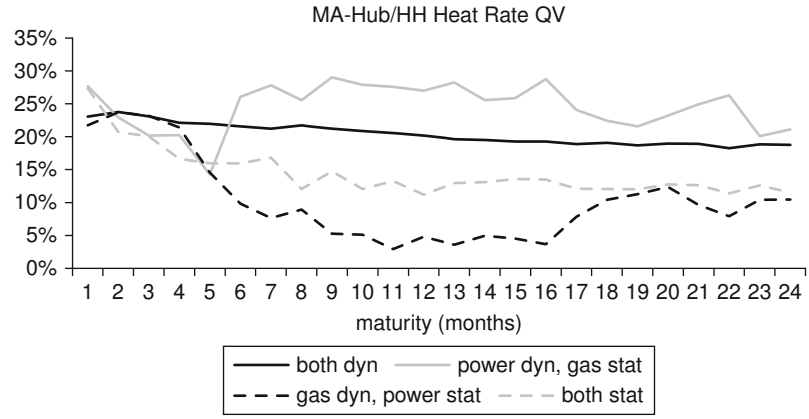



**Figure 3.4** Comparison of volatility collected from different hedging strategies. An analysis of actual (forward) data conforms to the result illustrated in Figure 3.2 for the model in (3.26). Namely, in commodity spread options, the leg that is more fundamentally driven (relative to the time horizon in question) *must* be dynamically hedged in order to optimally capture volatility. (Based on futures data from SNL/Swiss Re; see the disclaimer that applies to Figure 3.3.)

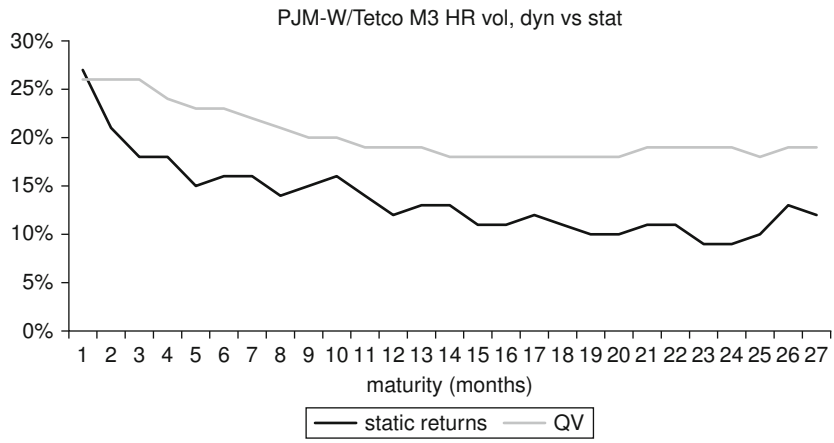



**Figure 3.5** Volatility collected under dynamic vs. static strategies. As expected, mean reversion effects under the physical measure generally imply that a lower vol will be collected under static as opposed to dynamic hedging strategies (associated with a pricing measure). (Based on futures data from SNL/Swiss Re; see the disclaimer that applies to Figure 3.3.)

$$\Pi(F,\sigma_\alpha) = \sum_{i=1}^{N} (S_i - \alpha F)^+ - N \cdot V^{BS}(F, \alpha F, \sigma_\alpha, \tau)$$
$$- \Delta^{BS}(F, \alpha F, \sigma_\alpha, \tau) \cdot \sum_{i=1}^{N} (S_i - F) \tag{3.33}$$

In (3.33), we are considering $N$ daily options (say, within some calendar month) that are struck at the futures price $F$ settled prior to the start of the exercise period; $\tau$ is a representative time-to-maturity, say middle of the month (*e.g.*, 15 days). The parameter $\alpha$ represents the moneyness of the option (or reciprocal moneyness, depending on convention), so ATM corresponds to $\alpha = 1$. Only a static hedge can be put on during this period, which we represent as a BS functional. This functional is parameterized by the replication volatility $\sigma_\alpha$. Obviously, the profit and loss from the portfolio (3.33) depends on the choice of this volatility. The question we ask here is the following: on a pathwise basis, how does the volatility that permits perfect replication (*i.e.*, solves $\Pi(\sigma_\alpha) = 0$) compare with standard measures of (annualized) return volatility? (Recall that we have made no assumptions about the underlying price process.) We present some typical results in Figure 3.6 for the ATM case.

As can be seen from Figure 3.6, the volatility of price returns (by calendar month) is systematically much higher than the volatility that permits perfect replication of the (ATM) option in the portfolio (3.33). In other words, pricing (and hedging) the option in question by projecting return volatilities would greatly overstate the value

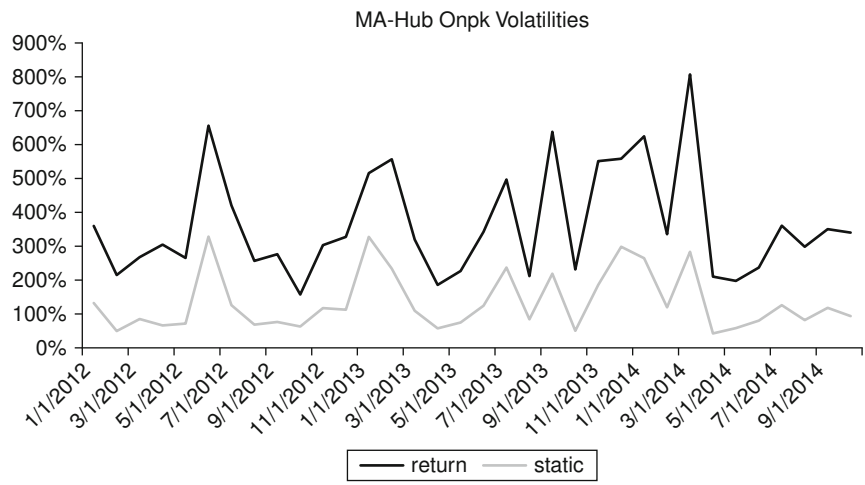



**Figure 3.6** Comparison of volatility collected from static strategy vs. return volatility. (Based on data from NE ISO.) In most energy markets, a daily option cannot be dynamically (*e.g.*, delta) hedged intra-month (apart from not very common cases where balance-of-month ["balmo"] is possible), and the only volatility that can be collected arises from static hedge put on at futures expiry (*e.g.*, a few days before the start of the month in question). This volatility is typically well below the volatility estimated from daily returns. Hence the latter can give a very misleading (*i.e.*, wrong) picture of option valuation. If there are no market instruments that give exposure to daily price changes, then these measures of variability should be employed very carefully in valuation

that can be collected. However, due to the nature of the available (static) hedges there simply is no other way to capture the higher volatility.[35]

Over the time horizons in question, mean reversion does not play a particularly large role; rather, in the case of power, jumps characterize the dynamics, at least in part; recall Figure 1.2. To further illustrate, we simulate paths from a lognormal process (with zero drift), as in (3.16), and compare the replication and return volatilities as was done for actual data in Figure 3.6. The simulated results are shown in Figure 3.7.

As can be seen, both static (cash) volatilities and (annualized) return standard deviations are distributed about the true volatility (50% in this case). The static volatility is extremely noisy and in fact for a fair number of paths is actually above the return volatility, in contrast to the observed data shown in Figure 3.6. Whatever the source of the effects giving rise to the observed pattern (jumps are likely one candidate), the chief point here is that the instruments available for intra-month hedging are quite limited in energy markets, and it is this availability that *directly* drives valuation. But differently, the *indirect* effects of the actual physical price process are of interest, not their actual properties (to the extent that they cannot be exposed through actual hedging strategies).

### 3.1.4 More on dynamic hedging: rolling intrinsic

#### *3.1.4.1 Overview of local time*

Consider the following scenario. Assume the market price of an option with strike $K$ is the (theoretical) BS value. We sell the option and engage in the following trading strategy: if the price of the underlying goes above the strike, hold the intrinsic position, namely long one unit of the underlying and short one unit of a bond with

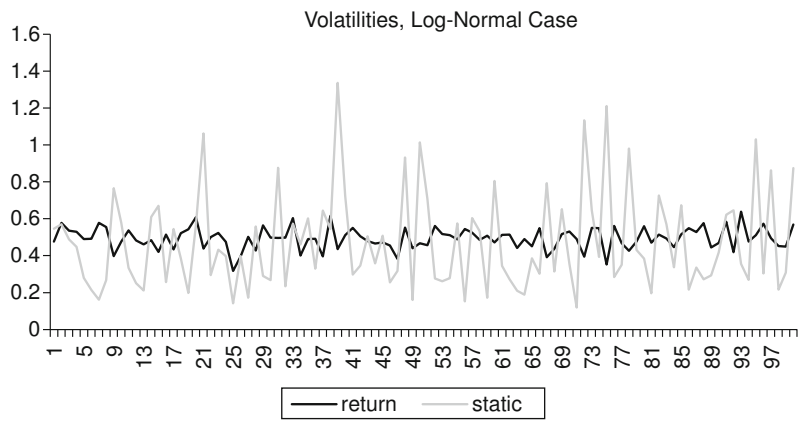



**Figure 3.7** Static vs. return analysis for simulated data. Lognormal process, 50% volatility

face value $K$. If the underlying goes below the strike, exit both positions. This is the so-called stop-loss start-gain strategy. It can also be termed rolling intrinsic. Plainly, this strategy will replicate the terminal payoff, and the initial cost is intrinsic value. The portfolio can be written

$$\Pi = V_{BS} - (S_T - K)^+ + \int_t^T H(S_s - K)dS_s = V_{BS} - (S - K)^+ \tag{3.34}$$

where $H$ is the Heaviside step function: $H(x) = 1$ for $x > 0$ and zero otherwise. Now, there seems to be a contradiction here, for the extrinsic[36] value of the BS option is strictly positive. The above argument suggests that there is a trading strategy that gives certain, positive payoff, in other words, an arbitrage. This is clearly a violation of the BS premises.

However, this paradox is only apparent, as shown by Carr and Jarrow (1990). The above strategy is not in fact self-financing. This follows from the generalization of Itô's formula for convex functions (for which second derivatives do not necessarily exist), known as the Tanaka-Meyer formula. This result can be seen heuristically from a formal application of Itô to the call payoff function:

$$d(S - K)^+ = H(S - K)dS + \tfrac{1}{2}\delta(S - K)dS^2 \tag{3.35}$$

The technical result (for more details we refer the reader to Karatzas and Shreve [1991]) is[37]

$$f(x_t) - f(x_0) = \int_0^t f'(x_s)dx_s + \tfrac{1}{2}\int_{-\infty}^{\infty} L_t(x)f''(x) \tag{3.36}$$

for $x$ a continuous semi-martingale, and where $L$ denotes the *local time* of the process:[38]

$$L_t(x) = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_0^t 1(|x_s - x| < \varepsilon)d\langle x \rangle_s \tag{3.37}$$

With this result, we can see that indeed the stop-loss start-gain strategy is not self-financing, and in fact the contribution due to the local time precisely recovers the missing extrinsic value. The intuition has to do with the fact that the strategy is not measureable with respect to the underlying filtration (that is, say, it is not resolvable with respect to the available information set). At any given time, we only know when the underlying *hits* the strike level, not when it is strictly above or below it. The technical reason has to do with the infinite crossing property of GBM. Once the price hits the strike level from below (say), it is as equally likely to go above as it is to come back down, and in fact will return to this level infinitely often over any

interval of time (however small). Since the trader must make *some* decision when the level is hit, and since he is equally likely to guess wrong as guess right about the next move, some external financing is required, an amount represented by the local time contribution.

Now, we mention this issue because (apart from the inherent theoretical interest), it has relevance for a popular hedging strategy in energy markets, namely that of rolling intrinsic. Specifically, we always hold the intrinsic position against an option position, specifically a long option position (to illustrate). Now, from Tanaka-Meyer, the portfolio (3.34) can be written as

$$\begin{aligned}\Pi &= (S_T - K)^+ - V - \int_t^T H(S_s - K)dS_s \\ &= (S-K)^+ - V + \tfrac{1}{2}\Lambda_{t,T}(S;K) \end{aligned} \tag{3.38}$$

where $\Lambda$ denotes the local time contribution and $V$ is the initial premium. Formally, we can express this term via delta functions:

$$\Pi = (S-K)^+ - V + \tfrac{1}{2}\sigma^2 K^2 \int_t^T \delta(S_s - K)ds \tag{3.39}$$

From (3.39), we can see quite plainly that rolling intrinsic is *not* risk free, as it entails exposure to realized local time (netted against initial premium). We can also understand the interpretation of local time as a measure of the time that the underlying crosses the strike. Carr and Jarrow demonstrate that the expectation of the local time component with respect to the pricing/martingale measure is precisely the BS extrinsic,[39] which of course is just another way of saying that *this* portfolio can be counter-hedged (so to speak) yielding precise replication, which in turn requires that $V$ equal the BS price in an arbitrage-free market.[40]

Thus, we see the ramifications of entering into rolling intrinsic against a structured product: we create exposure to realized local time (akin to the manner in which delta-hedging creates exposure to realized variance). In the context of valuation of structured products, the question then becomes: how do we project the value driver (pertaining to this strategy) such that we can assign an initial price/value to this option (and equally important, conducting the hedging strategy necessary to capture this value)? We will see shortly that in a wide variety of cases rolling intrinsic can recover the same volatility that can be extracted from delta-hedging. Before turning to this topic, we note the following important points. First, the hedge positions will show very big swings, from 100% to 0% (in terms of the deal/contract volume). Thus, when we consider the reality of transaction costs, it can easily be anticipated that such frictions can consume a great deal of extrinsic value. Second,

the distribution of cash flows that will arise from rolling intrinsic will in general have much greater variability than strategies based on incremental changes in position (such as delta-hedging). This will have correspondingly big impacts on risk adjustment. Both of these points argue for the importance of understanding what kinds of value drivers are created (or more accurately, what kinds of exposure are created) from different kinds of hedging strategies, because even though the same magnitude of value (in some broad average sense) may be captured by different strategies, it does not follow that there are no other criteria for identifying one kind of value as superior to another kind.

*3.1.4.2 Example: gas storage*

To illustrate, we consider the example of natural gas storage. A complete discussion of such deals can again be found in EW. For our present purposes we simply note the essential features (again, recall Section 1.2.3). Due to the fundamental fact that, with colder weather in winter as opposed to summer, there is a seasonal structure to demand for natural gas (for heating) and hence differences in seasonal prices. Thus, it makes economic sense to buy summer gas forward and sell winter gas forward. However, to realize this strategy requires that something be done with the gas physically between the two seasons. This is where storage comes in, as a means for taking physical possession of the gas to satisfy the forward transactions: gas must be injected in summer (to satisfy the forward purchase) and withdrawn in winter (to satisfy the forward sale). It should be plain that purchasing[41] a storage facility entails the purchasing of optionality between seasons. More generally, storage is a collection of options between the various months of injection and withdrawal.

We will provide a very simple example here. In doing so, we will have to jump ahead to a topic that we will cover in great detail in Chapter 7, namely the valuation of *spread* options. As the name suggests, these are simply options on the difference between two assets. In the current example, the relevant payoff is

$$(F_{winter} - F_{summer})^+ \tag{3.40}$$

Where $F_s$ denotes the seasonal forward price.[42] In fact, we will jump ahead even further and note that this payoff can be written as

$$F_{summer}\left(\frac{F_{winter}}{F_{summer}} - 1\right)^+ \tag{3.41}$$

which suggests that, if we consider our units of measurement to be denominated in summer gas (really no different economically than switching from dollars to yen), the essential option structure is basically a standard option on the winter/summer ratio. We will see (when we consider change of measure techniques) that this interpretation is correct. For now we will consider the following simple scenario.

Assume the storage facility can be rapidly filled and emptied; in fact in one month's time each way (as is the case with so-called salt-dome facilities). Consider a deal starting in April of the following year. (The typical one-year North American storage deal runs from April to March, with summer/injection months being April–October and winter/withdrawal months being November–March.) If the cheapest month is currently June, and the most expensive month is January, then the optimal intrinsic[43] is clearly: buy June, sell January. We can lock this spread in today by putting on these (forward) positions. However, conditions may change the next day so that July becomes cheapest and February most expensive. We can thus close out the previous positions and put on the new ones. This is the essence of rolling intrinsic. Here we will focus only on hedging strategies up to the start of the injection period.[44]

So, if we are asked to value this basic storage deal, we can consider two strategies. The first is simply the rolling intrinsic we just discussed. We pay some premium to enter the deal, and then follow rolling intrinsic. The second strategy follows from the fact that we can wait until the start of the deal (the start of the injection season) to lock in the optimal spread. Now, consider an option with the following payoff:

$$(\max_i F_{T,T_i} - \min_i F_{T,T_i})^+ \tag{3.42}$$

Here, $T$ denotes the start of the injection period (*e.g.*, April 2013) and $T_i$ represents the various months comprising the deal (*e.g.*, April 2013–October 2014). This expression can be written as

$$\max_{i,j} (F_{T,T_i} - F_{T,T_j})^+ \tag{3.43}$$

That is, a lower bound[45] on the option value inherent in the storage facility is given by the most valuable spread option over all possible pairs of injection/withdrawal ($12.11/2 = 6$ total). Now, *if* this option traded on the market, we could simply sell it, use the premium to enter the storage deal, and be done (again, to the extent that we ignore spot-forward optionality, *etc.*). Typically, however, exotic options such as these do not trade[46] in any energy market, so we have to ask instead: what is the cost of replicating this option? Under generalized BS assumptions of joint lognormality of the underlying processes, this cost is an expected value of the above payoff under a (unique) martingale measure. We will discuss in greater detail various methods for evaluating such expectations in Chapters 5 and 7. For now we simply note that the extension of the BS formula (known as the Margrabe formula) is driven by the ratio volatility given by

$$\sigma^2 = \sigma_i^2 - 2\rho_{ij}\sigma_i\sigma_j + \sigma_j^2 \tag{3.44}$$

where $\rho_{ij}$ denotes the correlation between the two legs. (The intuition here simply follows from the covariance structure of the difference of two Gaussian variables [log-returns]).

So, we compare the following two portfolios (with the same terminal payoff):

1. Rolling intrinsic
2. Delta-hedging the spread options that are optimal at inception.

A comparison of these two portfolios should shed some light on the relation between realized local time (as created through rolling intrinsic) and realized volatility (as created through delta-hedging). We will approach this problem through simulation. (We will discuss simulation methods in some detail in Chapter 7, although we assume familiarity with the basic concepts on the reader's part.) For convenience, we will assume a flat volatility structure across contracts of 30%, and a correlation structure given by

$$\rho_{ij} = \alpha^{|i-j|} \tag{3.45}$$

with $\alpha = 95\%$ (reflecting the characteristic high correlation between neighboring contracts in natural gas markets). A representative forward curve is shown in Figure 3.8.

We simulate the forward curve daily from deal inception to the start of the injection period, computing the relevant hedges for each day along each path. We collect the portfolio results in Figure 3.9.

For 500 simulations (that is, 500 simulated price curve evolutions, through term), the relevant statistics are as follows:[47]

- Rolling Intrinsic: average = 3.55, standard deviation = 0.73
- Delta-hedged Option Portfolio: average = 3.57, standard deviation = 0.27.

As can be seen, both hedging strategies yield the same cash flows (or more accurately the same extrinsic value [intrinsic value in this case is 1.47, basically April

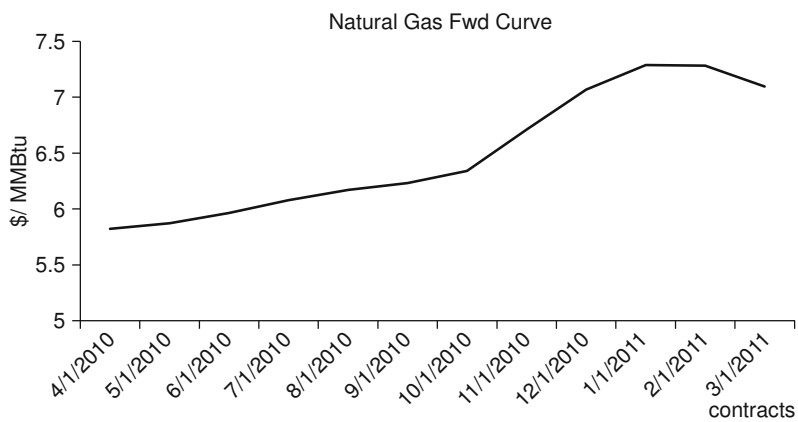



**Figure 3.8** Typical shape of natural gas forward curve

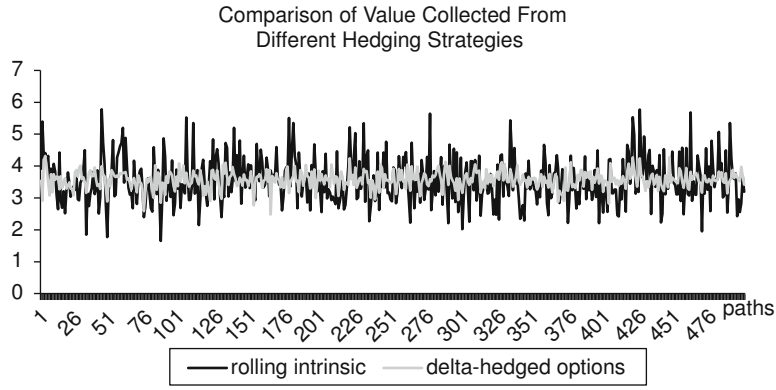



**Figure 3.9** Comparison of cash flows for different storage hedging strategies

injection and January withdrawal]). This reflects the nature of local time as discussed above. But there is a more important point to emphasize here. The *variability* of cash flows is much lower for the delta-hedged option strategy than it is for rolling intrinsic. This is of course due to the comparatively large variability of extrinsic value for any set of realized curve evolution. Zero extrinsic value is certainly a conceivable outcome, for example. One would expect to see large swings in portfolio value under such a strategy. Indeed, note that the underlying positions similarly undergo large swings. At one point in time, one might hold positions of long May and short January, say. The next day, as the curve changes, these may get completely closed out in favor of positions of long July and short December. This behavior has a big implication on the hedge costs that will be faced in reality, as the bid-ask spread (present in all markets) will be repeatedly crossed over large volumes. In contrast, the delta-hedging approach can be expected to incur much smaller (dynamic) costs, as the bid-ask will only be incurred on the *incremental* volumes needed to adjust the deltas.

Still, there is an even larger lesson to be learned here. The comparatively greater variability of cash flows under rolling intrinsic also indicates the correspondingly lower robustness of local time in comparison to another pathwise entity, namely realized volatility/quadratic variation. This is a crucial point, as any structured product valuation problem entails the identification and projection of those entities underlying the hedging strategies that are put in place against the structure in question. We are talking of course about value drivers, the central theme of this chapter to this point. We have here a fairly realistic example (or at least a tractable example of a structure commonly encountered in the industry), in which we see that not only does a particular value driver (in this case, realized variation) recover the same value as an intuitive (and common) hedging strategy, but *also* permits better (in the sense of more robust) projection of value. We will see more examples

throughout the book, but this simple case should suffice to indicate the strength of the approach.

With the proper context now provided, the final (and in a way, the most important) step is to look into the issue of market structure.

### 3.1.5 Market resolution and liquidity

It was necessary to begin the exposition with some abstraction. However, the very first question that must always be asked when approaching a valuation problem is: what is the structure of the underlying market? By this we mean the following:

- What actually trades?
- How far out (maturity-wise) do they trade?
- How frequently can positions be rebalanced, and at what cost?

These questions are actually not unrelated, but it is useful to keep them separate at first. Let us examine them in turn, as they pertain to energy markets.

#### *3.1.5.1 The extent of traded instruments*

The most obvious thing we need to know is what instruments trade in the market in question. It must be stressed that valuation always takes place in a *specific* market, and different markets possess different transactional characteristics. These characteristics are critical to the valuation problem. The valuation of a tolling deal in PJM is different from the valuation of a tolling deal in NEPOOL. Even within a given market, substructures can be identified. The valuation of a toll based on PJM-W is different from one based on PEPCO. By this we do *not* mean that a power plant in two different regions has radically different operational properties. (The basic structure is always the option to convert some fuel to electricity.) Rather, what we mean is the formation of portfolios around tolling arrangements (or any other structured product) depends on the kinds of instruments available for inclusion in the portfolio. This is a simple, but not at all trivial point. As we shall see, the kinds of *residual* exposure that results from an actual portfolio is central for characterizing the measures needed for representing the value of the underlying structure (and by extension, the associated hedges).

By far the most common instruments are futures and options. Owing to their ultimate nature in physical use/consumption, energy markets are typically structured regionally. In the United States, the major power markets of interest reflect both demand patterns and jurisdictional boundaries, *e.g.*, PJM (mid-Atlantic region), NEPOOL (New England), ERCOT (Texas), *etc.* The natural gas pipeline system is likewise based on geographical demarcations, typically originating in the Gulf Coast and delivering to points further north, with various segmentations along the way, *e.g.* TETCO going into the mid-Atlantic and New York, or Algonquin going from

there onto New England. Again, EW is an outstanding source for further information on this kind of market topology.[48] These markets all possess traded futures of varying degrees of liquidity. For example, PJM-W is highly liquid, more geographically localized points such as PEPCO less so. Gas markets display similar gradations, with the primary benchmark (Henry Hub) being extremely liquid and the various basis locations less so (recall from Section 1.2.2 that basis locations such as Tetco M3 trade as adders [positive or negative] to Henry Hub).

In energy markets, option products do trade but they are typically far less liquid and diverse than futures. In U.S. energy markets, for example, as of this writing (2014), only Henry Hub gas, WTI crude,[49] and PJM-W power support *liquid* markets in *monthly* options.[50] PJM-W daily options[51] trade, but they are far less liquid than the corresponding month products (*i.e.*, they have very large bid-ask spreads). Henry Hub daily options do not trade at all.[52] Invariably, the most liquidly traded options are at-the-money (ATM). The ability to include options in portfolio constructions such as (3.4) is fairly limited. When attempting to value a structured product, one must employ great caution, if not outright skepticism, in unthinkingly employing volatility quotes that are not connected to an actual, transactable market (*e.g.*, pulling numbers that are uploaded by traders and marked in an internal risk control system to "calibrate" a model).

It is not our intention here to catalogue the extent of liquidity in energy markets. Rather, our point is that, for any given structured deal, we *must* determine how the (spot) entities underlying the payoff relate to the available (forward) traded instruments that may or may not settle against these entities. For a toll settling against PJM-W (say) we obviously can put on (forward) hedges that correspond to the actual exposure. For a toll settling against some illiquid, physical node within PJM the situation is very different. Then, we *cannot* hedge the exposure in question. We can only put on some kind of proxy ("dirty") hedge, and therefore the value we can attribute to portfolios such as (3.4) *must* take this additional risk source into account. Very crudely (and we will explain in the course of things the precise nature of this statement), the question concerns what price to "plug into" functionals such as Black-Scholes. As with volatility marks, there is a tendency to simply assume that some number that comes out of a formal "system" that has been marked by a trader is adequate for valuation purposes. This very common practice must be guarded against at every step. Of course, good traders can provide reliable insight on this matter. However, good quantitative analysis must also ensure that these marks make sense.

The critical point is that we are forming an actual portfolio about the structure question, and thus we must know what can actually be included in that portfolio. The next question concerns whether traded instruments are available throughout the entire term of the deal.

*3.1.5.2 The (time) maturity of traded instruments*

Any deal takes place over a particular time horizon, some longer, some shorter. For example, a gas storage deal valued in January will typically commence in the upcoming April for a one-year cycle; deals spanning several such seasons are now somewhat rare but certainly not unheard of (especially precrisis). Tolling deals may start a few months from inception and may simply cover the peak summer period or an entire calendar year. Multiyear tolling deals also are rare in the postcrisis era (reflecting the drying up of liquidity and volatility, a topic we will return to), but were once not that uncommon (especially in the western U.S. power market). The point is that the structure in question has some time horizon that may not correspond to the range of traded products over that horizon.

The scenario we envision here is one where at least part of the underlying exposure cannot be (directly) hedged for the entire term of the deal. For example, a long-term toll at, say, PEPCO may only be hedgeable with PEPCO contracts for the first year of the deal. The later years ("back end") of the deal must then be proxy hedged (*e.g.*, with PJM-W). A similar situation can prevail on the gas side, as well, *e.g.*, with a long-term transport deal. A somewhat related scenario is one where the underlying does in fact trade throughout the term, but for longer time horizons may only trade in seasonal or even annual blocks. For example, it may only be possible to put on a July–August power hedge for a toll starting in next year's summer, and only put on individual ("bullet month") hedges closer to the start of the deal. Another example would be buying winter gas contracts a year out and only breaking these out into constituent monthly contracts much closer to the start of the deal term.

We have already seen examples of how dynamically hedged option portfolios can create exposure to realized volatility/QV. It is still possible to capture this kind of exposure (which is generally what is desired from a structured product) even when we cannot cleanly hedge throughout the term. However, one must be very careful to identify *incrementally* the volatilities that are being extracted given the need to stagger (so to speak) the particular hedging instruments employed. Ultimately the issue comes down to aggregation of variance, *e.g.*:

$$\sigma_{agg}^2 \tau = \sigma_{yr}^2 \tau_1 + \sigma_{seas}^2 (\tau_2 - \tau_1) + \sigma_{mnth}^2 (\tau - \tau_1) \tag{3.46}$$

where we can think of total (collected) variance as consisting of the variance collected from trading yearly blocks (over a particular time horizon), plus the variance collected from trading seasonal blocks (again, over some specific time frame), plus finally the variance collected from trading monthly contracts. (There is little point in trying to craft (3.46) in general, abstract terms, as the idea should be clear.)

Obviously, the aggregate volatility in (3.46) need not be equal to the monthly volatility (and will almost certainly be less), and great care must always be exercised

in understanding the resolution of hedging instruments that are available throughout the deal term. Note that we are not simply talking about term structure effects here, which manifest themselves even over the range on which individual contracts trade (the well-known Samuelson effect in commodity markets). Rather, we are talking about decomposing a value driver into components that directly correspond to the market structures that prevail over specific terms encompassing the deal in question. As always, the issue comes down to portfolio formation and the available instruments for such formation.

#### *3.1.5.3 The liquidity of traded instruments*

We have already discussed differences in liquidity as it relates to the extent of traded products. There is another sense in which liquidity is important. This concerns the ability to rebalance positions. Recall the basic portfolio expression in (3.4). In general, dynamic hedging positions can be put on. It may help to recall the famous Black-Scholes valuation framework. In that framework, a standard (European) call option (on a lognormally distributed price process) can be perfectly replicated by a particular dynamic hedging strategy. This strategy entails holding (in continuous time) a position in the underlying equal to the (negative) Black-Scholes delta. (The replication cost is of course the Black-Scholes value.)

The point is that this value can be extracted *only* from a (particular) dynamic hedging/trading strategy. A different (say, static) hedging strategy will not extract this value. This distinction turns out to be critical in energy markets and is far more important than in, say, equity markets. The issue ultimately comes down to the presence of *time scales* in commodity markets and the resulting manner in which volatility accumulates with time horizon. Roughly speaking, it is the presence of mean-reverting effects in commodity markets that give rise to such effects.

Now, there are a few issues to stress. If we are to implement a dynamic hedging strategy, we must first know if such a strategy is feasible. In all markets there are transaction costs. At any instant in time, it is generally more expensive to buy an asset than it is to sell it (the difference being the so-called bid-ask spread that market makers earn). Obviously, the more frequently we change positions (rebalance), the more transaction costs we will incur, and these costs *must* be taken into account. It is thus imperative to understand from the outset whether dynamic hedge strategies can realistically be included in the portfolio construction in (3.4). Of course, as the saying goes, everything has its price, so the issue is not whether rebalancing *can* be done, but rather at what price or cost. In some markets, it may be effectively possible to only put on static hedges around some structured product. For example, it may be reasonably cheap to rebalance PJM-W positions, but prohibitively expensive to do so with PEPCO, say.

In addition, the time horizon in question plays an important role in delineating the extent of (dynamic) liquidity. Some commodities can be rebalanced at tolerable

**Table 3.1** Typical value drivers for selected energy deals

| Hedge | Example deals | Value driver |
|---|---|---|
| Static future/forward | Long-term swap | Heat ratio/ratio |
| Static future/forward | Medium-term tolls, storage, transport | Cumulative variance |
| Dynamic future/forward | Short-term tolls, storage, transport | Quadratic variation |
| Static option, delta-hedged | Short-term tolls, storage, transport | Correlation |
| Static option, delta-hedged | Full requirements/load serving | Convexity |

cost over short time horizons, but not over longer ones. In addition, some commodities are more dynamically liquid than others. One can envision a scenario with a spark spread option where the gas leg is more readily rebalanced than the power leg (or *vice versa*). The situation can have important ramifications for valuation, as we have already seen (recall the examples in Section 3.1.3). Very often, only mixed dynamic-static strategies are viable, so it is critical to make a proper accounting of liquidity in the market in question. .

At this stage, in light of the various instruments we have broadly discussed in the context hedging and valuation (static vs. dynamic forwards, vega/option hedging, *etc.*) and the associated exposures/value drivers, it might be helpful to collect some results here pertaining to energy markets, which we present in Table 3.1. We do not propose to discuss any of them in great detail, although we will do so eventually. Rather, we hope to tie things together at this stage (at least to some extent) by providing some concrete examples that will presumably mesh with the reader's background in the industry (recall we assume some basic familiarity with the standard products).[53]

Note that the simple BS framework serves to illustrate many of these points, even if they are buried in the customary presentation. We will turn now to a few subsidiary (but important) issues before going deeper into the issue of replication and portfolio formation in energy markets.

### 3.1.6 Hedging miscellany: greeks, hedge costs, and discounting

#### *3.1.6.1 Connection of value drivers to greeks: delta and vega hedging*

Keeping with the spirit of jumping ahead to concepts that will be discussed in greater detail later, we note here another example of a value driver. As we have stressed previously, a value driver is an exposure that is created by a particular hedging strategy around a particular structure. We discussed the specific example

of realized and projected volatility in connection with delta hedging of an option. Let us recall the basic result from (3.13):

$$\Pi = \tfrac{1}{2}(\sigma^2 - \hat{\sigma}^2) \int_t^T \hat{\Gamma}_s S_s^2 ds \tag{3.47}$$

Note that the option gamma is not merely an interesting byproduct of the calculation, but plays a central role in the overall portfolio profit and loss.[54] Indeed, all models require back testing to gauge their effectiveness, and gamma would play as critical a role here as the premium and deltas. A more general result can be seen in higher dimensional problems (such as spread options). As we will see when we discuss spread options in the Section 7.1, the basic valuation equation is given by

$$V_t + \tfrac{1}{2}\sigma_1^2 S_1^2 V_{S_1S_1} + \rho\sigma_1\sigma_2 S_1 S_2 V_{S_1S_2} + \tfrac{1}{2}\sigma_2^2 S_2^2 V_{S_2S_2} = 0 \tag{3.48}$$

with terminal condition $V(S_1, S_2, T) = (S_2 - S_1 - K)^+$. Again, we will derive this equation and discuss its properties later. For now we simply note, following steps very similar to the one-dimensional case, that delta hedging both legs of the option under some projected covariance structure (that is, both volatilities and the correlation) creates the following portfolio:

$$\Pi = \tfrac{1}{2}\int_t^T ds((\sigma_1^2 - \hat{\sigma}_1^2)S_1^2\hat{\Gamma}_{11} + 2(\rho\sigma_1\sigma_2 - \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2)S_1S_2\hat{\Gamma}_{12} + (\sigma_2^2 - \hat{\sigma}_2^2)S_2^2\hat{\Gamma}_{22}) \tag{3.49}$$

where the hats denote projected entities and the meaning of the gammas should be clear enough (*e.g.* $\Gamma_{ij} \equiv V_{S_iS_j}$). Now, using Itô's lemma it is not hard to derive the volatility of the ratio $S_2/S_1$:

$$\sigma^2 = \sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2 \tag{3.50}$$

Further, as noted previously, the option premium is expected (by economic intuition) to be homogeneous of degree one in the prices and strike (doubling these, say, should simply double the option price).[55] Thus, by Euler's theorem, we have the following relationships across deltas and gammas:

$$\begin{aligned} V &= S_1\Delta_1 + S_2\Delta_2 + K\Delta_K \\ S_1\Gamma_{11} + S_2\Gamma_{12} + K\Gamma_{1K} &= 0 \\ S_1\Gamma_{12} + S_2\Gamma_{22} + K\Gamma_{2K} &= 0 \\ S_1\Gamma_{1K} + S_2\Gamma_{2K} + K\Gamma_{KK} &= 0 \end{aligned} \tag{3.51}$$

with the subscript $K$ representing a derivative with respect to strike. Using these results, the portfolio expression becomes

$$\Pi = \tfrac{1}{2}\int_t^T ds((\sigma_1^2 - \hat{\sigma}_1^2)S_1 K\hat{\Gamma}_{1K} - (\sigma^2 - \hat{\sigma}^2)S_1 S_2\hat{\Gamma}_{12} + (\sigma_2^2 - \hat{\sigma}_2^2)S_2 K\hat{\Gamma}_{2K}) \quad (3.52)$$

In general, we can see how different kinds of exposures are created from the basic delta-hedging strategy. We can also see how the realization of the value drivers will affect risk adjustment, based on the determinate sign of the various cross-gammas:

$$\begin{aligned} \hat{\Gamma}_{12}, \hat{\Gamma}_{2K} &< 0 \\ \hat{\Gamma}_{1K} &> 0 \end{aligned} \quad (3.53)$$

In particular, for the special case of no fixed strike ($K = 0$) the exposure is due entirely to realized ratio volatility (as we pointed out above): the portfolio makes/loses money when the realized ratio volatility is greater than/less than projected volatility. Observe that, contrary to what one might expect, realized correlation does not necessarily play a role as a value driver. This is a very significant point, for as we shall see in a later chapter, estimation/projection is typically much more robust for volatilities (such as heat rates or ratios) than for correlations.

Note that we do not say that correlation is irrelevant here. In some cases (as we will see), projecting ratio volatility can be conditioned on *traded* market leg volatilities, and in such cases, projected correlation often serves as a usefully robust proxy, an alternative to a regression analysis, say. On the topic of market volatilities, we must mention that quasi-dynamic hedging strategies are not limited to hedges around the legs. Static option positions that are themselves delta hedged can also create exposures to certain value drivers. In other words, *vega* hedging is a viable strategy for valuing structured products.[56] Although we will not derive the results now (consult the Appendix to this chapter), it can be shown that a delta-hedged spread option combined with static vega hedges (for both leg volatilities) will produce a portfolio involving terms explicitly proportional to $\rho - \hat{\rho}$. In other words, this strategy will produce an exposure to realized correlation.

Apart from indicating that value drivers are indeed more general than volatilities, this discussion again demonstrates the central role liquidity plays in our understanding of valuation: the appropriate value drivers are *always* a function of the available hedging instruments. If leg options do not trade, then estimations of correlation can produce very misleading pictures of value, and of course hedges of very dubious quality. This issue becomes even more important when we recognize that there are intermediate levels of liquidity, *e.g.*, markets where monthly options trade but not daily options.

As might be expected, option vegas play a very important role in determining the right hedge volume, and are not free-floating, abstract sensitivities. In fact, in a rather general setting, they are closely related to the gammas, as we will now indicate.

*3.1.6.2 Relations between gamma and vega*

The results that follow will be derived in detail when we consider so-called likelihood ratio methods for computing greeks via Monte Carlo in Chapter 7. For now we simply present the results as a means of illustrating that in fact the gamma and vega coefficients that drive profit and loss are related. Apart from highlighting the role that both entities play in the model reconciliation process, these results show that there can be significant computational savings from exploiting these connections.

First we note, from standard results, the expressions for gamma and vega in the BS setting:[57]

$$\begin{aligned} \Gamma &= \frac{1}{S\sigma\sqrt{\tau}}\frac{1}{\sqrt{2\pi}}e^{-d_1^2/2} \\ \upsilon &= S\sqrt{\tau}\frac{1}{\sqrt{2\pi}}e^{-d_1^2/2} \end{aligned} \tag{3.54}$$

with the usual notation $d_1 = \frac{\log S/K + \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}}$. From (3.54) we get

$$\frac{\upsilon}{\Gamma} = S^2\sigma\tau \tag{3.55}$$

Thus, note that the classic gamma-neutral strategy for hedging two options is also vega-neutral. This result is no accident, and for a general, $n$-dimensional lognormal process we will show later that the follow results hold:

$$\begin{aligned} \upsilon_{kl} &= S_k S_l \sigma_k \sigma_l \tau \Gamma_{kl}, \quad k \neq l \\ \upsilon_{kk} &= S_k \tau \sum_l \rho_{kl}\sigma_l S_l \Gamma_{kl} \end{aligned} \tag{3.56}$$

where the "cross vegas," so to speak, represent sensitivities with respect to a correlation. So, apart from considerable computational savings offered by these results, we see again a connection between the various hedging quantities that arise from the standard greeks. This should not be surprising, as we are operating in a setting where the "natural" choice of pricing functional gives rise to greeks that are operationally meaningful in terms of hedges for the extraction of value. We stress again that the value to be extracted depends on the value driver arising from a particular hedging strategy. We also point out (and will elaborate more when we consider

departures from the standard setting, *e.g.*, incomplete markets) that this pricing functional, as well as the associated greeks/hedge ratios, can still be extremely useful even outside the BS environment.

It is worth saying a few things now about how to account for hedge costs (both static and dynamic), which are an important aspect of any actual deal.

#### *3.1.6.3 Transaction costs*

As is well-known, in any market (be it commodities, equities, *etc.*), trading is not costless and buyers and sellers confront different prices. This difference is the bid-ask spread, and depending on the depth of liquidity, can be either fairly small or quite large. In particular, dynamic hedging in some cases may be prohibitively expensive. Let us see how this plays out. The main ideas can be gleaned from a delta-hedged option portfolio:

$$\Pi = (S_T - K)^+ - V_t - \sum_{i=0}^{T-1} \Delta_i (S_{i+1} - S_i) + c \sum_{i=0}^{T-1} |\Delta_{i+1} - \Delta_i| \tag{3.57}$$

where $c$ is the cost of rebalancing (at each step we must adjust the position in the underlying by $\Delta_{i+1} - \Delta_i$, and this cost is symmetric[58] with respect to (wrt.) buying and selling, hence only the absolute value of the change matters). More generally (although, not realistically), the hedge costs can be taken as state-dependent; for the particular choice of proportional to price (so that $c \to c_i = \gamma S_i$), the well-known formula of Leland, wherein the valuation volatility is adjusted by a term proportional to gamma is obtained (see Wilmott *et al.* [1997]; there is also an excellent discussion in Joshi [2003]). For the more realistic constant transaction costs case, the change of measure techniques to be studied in Chapter 5 can be employed to integrate numerically the expected value of the hedge cost term in (3.57); in fact, we can obtain multidimensional results that generalize the standard one-dimensional Leland result.[59]

The main points we wish to make here about hedge costs are the following:

1. Not all dynamic strategies are created equal: for a given degree of liquidity, it is definitely preferable to follow strategies that require smaller incremental rebalancing. Thus, *e.g.*, (dynamic) strategies such as rolling intrinsic where $\Delta_i$ equals 0 or 1 can potentially be prohibitively costly in high volatility environments (where frequent rebalancing is likely).
2. In light of what we have seen regarding the importance of dynamic strategies for value collection in energy markets, the interplay of time horizons becomes critical for valuation. Increasing the period between rebalancing reduces hedge costs, but at the expense of increasing the horizon over which mean-reversion effects can reduce realized volatility (from the portfolio perspective).

3. Hedge costs must always be accounted for, obviously. However, it is operationally preferable to separate identification of value drivers from transaction costs, as explicit incorporation of the latter typically induces numerical instabilities and thus renders estimation non-robust. Hedge costs should be incorporated after suitable risk adjustment of the value driver, conventionally as a discount to that driver (thus notions of bid-mid-ask valuation are quite separate from the level of risk adjustment underlying the valuation).
4. Somewhat in modification of point number 3 above, for a given value driver there may be cases where it is preferable to hedge at a level of risk adjustment different than the level at which (initial) pricing takes place. Specifically, hedging should take place at a more aggressive level than initial valuation. This effectively creates a term $V(\sigma_H) - V(\sigma_0)$ in (3.57) that can offset (to some degree) dynamic hedge costs. These considerations are most important when we are dealing with value drivers and functionals exhibiting little but (non-zero) extrinsic value, such as natural gas transport deals with low ratio volatilities/high correlations.

*3.1.6.4 The role of discounting*

On the subject of hedging and portfolios, it is important to understand that, for the self-financing strategies under consideration, there *must* be a (dynamic) bond hedge. This point is obscured by the fact that in the presence of zero interest rates, there is of course no pathwise accrual of cash flows from bonds (the dynamics of a bond $B$ would trivially take the form $dB = 0$). This can be seen most clearly in the discrete time case. The cash flows over a path are given by

$$(S_T - K)^+ - V - \sum_{i=0}^{T-1} \Delta_i (S_{i+1} - S_i) \tag{3.58}$$

which can be written as

$$\sum_{i=0}^{T-1} [V_{i+1} - V_i - \Delta_i (S_{i+1} - S_i)] \tag{3.59}$$

Now, by homogeneity and Euler's theorem, we have $V = S\Delta + K\Delta^K$ (where $\Delta^K$ is the strike delta) so the cash flow expression becomes

$$\sum_{i=0}^{T-1} [V_{i+1} - \Delta_i S_{i+1} - \Delta_i^K K] = \sum_{i=0}^{T-1} [(\Delta_{i+1} - \Delta_i) S_{i+1} + (\Delta_{i+1}^K - \Delta_i^K) K] \tag{3.60}$$

Using Euler again, we have $S\Delta_S + K\Delta_S^K = 0$. Consequently, for small enough time steps, the cost associated with the rebalancing of the price hedge is precisely countered by the change in the strike delta, *except at expiration* (that is, for $i = T - 1$

in (3.60)), where the gammas (derivatives of deltas) have a delta-function form[60] and the small time limit only makes sense in an integral (distributional) framework. Thus, the requirement of self-financing, namely that there are no external cash flows but only transfers of money between trading accounts, *implicitly* requires that at any time, we hold a dynamic bond (conventionally paying off $K$ dollars at expiry) position of $\Delta_i^K$. However, this position obviously makes no pathwise contribution (again, except at expiry) due to the fact that $B_{i+1} - B_i = 0$.

Let us stress again that self-financing implies that, after the initial outlay (premium), there are only transfers between trading accounts, with penultimate cash flows depending only on prices (through option payoffs and terminal positions) at expiry. However, as we see from expressions such as (3.13), realized cash flows do have a pathwise aspect, but only through realized vs. projected entities (such as volatility), *not* through cash accruals along any path. Self-financing does *not* mean that the hedging strategy (broadly understood to understand the initial cost and terminal payoffs) pays for itself, only that no additional inflows are necessary after deal inception. (Effectively, the initial premium represents borrowing that is parceled out [so to speak] along a path so that rebalancing requires no [further] external financing.) The upshot of this discussion is that, since it proves notationally quite convenient to neglect discounting and explicit accounting for bond hedges, we will thus do so, with the understanding that we *always* implicitly assume the presence of such hedges in the overall portfolio.

We can now begin to examine the incomplete market case in more detail.

## 3.2 Incomplete markets and the minimal martingale measure[61]

As is well known, option products in the Gaussian, Black-Scholes framework are redundant. *Any* contingent claim can be perfectly replicated by a suitable (dynamic) hedging strategy, from which the value of the option is immediately obtained from absence-of-arbitrage arguments. That is, assuming that rather weak, commonsense conditions hold (*e.g.*, that the law of one price prevails in liquid futures markets [say]), necessary and sufficient results for the price of a contingent claim can be derived. This happy set of events is most definitely *not* the situation with which we are faced in the vast majority of valuation problems, especially in energy and commodity markets. Once we depart from the Gaussian world, structured products *cannot* be perfectly replicated through trading (dynamic or otherwise) in underlying contracts. This challenge applies to even rather simple structures. For example, in most power markets a plain vanilla spark spread option cannot be replicated through trading power and gas contracts, and the structure entails a (risky!) bet on realized heat rate volatility.[62] The situation becomes far more complicated

when we introduce physical constraints that characterize so many energy-based products, such as physical tolls, with their array of start-up costs, ramp rates, and outages.

Happily, it turns out that many of the familiar, standard techniques (which are so often carelessly imported from their use in financial markets) can be brought to bear in the commodity arena, *as long as proper care and understanding of the relevant concepts are employed.* As we continue to emphasize, the correct valuation of a structure entails the construction of suitable portfolios around the structure that enable one to identify an induced exposure that is (presumably) preferable to the exposure that exists in the absence of that portfolio. A central lesson that will be learned is the following: a model that is "wrong" (in the sense of failure to capture known *qualitative* features) can out-perform (in the sense of providing better portfolio profiles) the "right" model. Put differently: *simplicity often trumps structure.*[63]

Let us now provide some examples as well as a theoretical framework for analysis.

### 3.2.1 Valuation and dynamic strategies

It is worth recalling a basic concept that is central to martingale pricing in complete markets (*e.g.*, Black-Scholes). Under suitable technical conditions, the martingale representation theorem states that if $V$ and $S$ are (square-integrable) martingales under some probability measure $Q$ (associated with some filtered probability space), both adapted to the (natural) filtration[64] of a standard Brownian motion (also under $Q$), then there exists a pre-visible[65] process $\Delta_t$ such that

$$V_T = V_t + \int_t^T \Delta_s dS_s = E_t^Q V_T + \int_t^T \Delta_s dS_s \tag{3.61}$$

If $V$ is thought of as the value of some derivative security and $S$ the price of an asset on which that security depends, then it is not hard to see that (3.61) entails a (dynamic) replication strategy for that security, with the $Q$-expectation of the terminal payoff as its value.[66] A very nice discussion of the martingale representation theorem in the context of option pricing can be found in Baxter and Rennie (1996).[67]

There are a few important points to make. First, as noted, (3.61) entails a *dynamic* trading/replication strategy. Second, it is optimal in the (admittedly somewhat trivial) sense of perfectly replicating the derivative security in question. Third, the martingale representation theorem only establishes the *existence* of replication strategy; it says nothing about how to actually construct that strategy. (That is, it does not specify the relationship between the value process and the hedging process.) Fourth, the critical assumption behind (3.61) is that the only source of randomness/uncertainty in the underlying market is Gaussian in nature. (This

is the meaning of requiring that the two $Q$-martingales be adapted to a standard Brownian motion.) Fifth (and most important), the ability to associate the representation in (3.61) with hedging or replication strategies presumes that *all* of the sources of randomness are associated with actual, traded instruments. In other words, market completeness is here intimately connected to a very special assumption about the nature of the market drivers. It is the assumption of market completeness we wish to relax, and to spell out the ramifications thereof. Specifically, we wish to analyze in detail portfolios of the form (3.61), but in *incomplete* markets (that is to say, under the condition that not all of the sources of randomness can be laid off in the market[68]).

Our approach will be to examine the construction of optimal valuation/hedging strategies for linear and nonlinear products in such (incomplete) markets. Our major departure from the existing body of work on the topic is the importance we place on minimal informational requirements in constructing optimal strategies (*i.e.*, pricing measures) in environments with limited information about the nature of the underlying DGP. (The limited information is typically driven by relatively small samples compared to the scale of structural changes in the markets hosting the underlying DGP, but also by the fact that only some of the relevant state processes are observable.) We will develop an abstract characterization of the optimal strategy (pricing measure) and a range of the sufficient statistics required for the sufficient characterization of the optimal strategy. This last point is of crucial importance. The informational requirements of the optimal strategy fall significantly short of the full characterization of the underlying DGP (a point we made explicit in an econometric sense in Chapter 2, and will further pursue in Chapter 6). This result enables us to develop robust optimal pricing/hedging strategies without detailed knowledge of the underlying DGP.

The subsequent discussion will make clear what, exactly, we mean by "optimal." What we are interested in is the behavior of the risk residuals that arise from different (dynamic) hedging strategies, and characterizing a particular strategy as optimal in a suitable sense (*e.g.*, akin to the BS conclusion for (3.61) where the residual is [identically] zero). Valuation then takes place with reference to this optimal strategy (precisely akin to the standard BS case, properly understood). The general thrust of our analysis will be to split the optimization problem into a dynamic and static part; in other words, to separate dynamically hedgeable risks from the unhedgeable ones. Our setting is the analysis of portfolios of the following form:

$$\Pi = \mathfrak{V}(P_T, G_T) - V_t - \int_t^T \Delta_s^G dG_s - \int_t^T \Delta_s^P dP_s \tag{3.62}$$

where $\mathfrak{V}$ is some (typically, but not necessarily, convex) payoff function of two (in our example) tradeables (variables). The integrals in (3.62) represent accumulated

dynamic hedges[69] and $V$ the value at inception (*i.e.*, the initial premium). We are interested in characterizing the "optimal" dynamic strategy and the associated optimal value. It is important to understand that the value and hedges stand in a very definite relation to each other (it is no accident that we explicitly write the time dependence of $V$ in (3.62), both value and hedges are processes). Our objective to find optimal hedging strategies as they relate to the value associated with those strategies. As valuation through measure change is a central concept here, this will be the starting point of our investigation.[70]

### 3.2.2 Residual risk and portfolio analysis

*3.2.2.1 The fundamental theorems of asset pricing (and the question of relevance).*

As already noted in Section 3.1.2, in the BS setting the value of any contingent claim is the expectation of its (terminal) payoff under a particular measure, equivalent to the physical measure under which the actual DGP is observed. To be specific, this measure $Q$ is one for which the underlying tradeables are martingales: $E_t^Q S_T = S_t$[71] This measure is inextricably related to a specific trading/hedging strategy around that claim (indeed, this is precisely its meaning, recall the discussion of the martingale representation theorem pertaining to (3.61)), a point whose essence is actually applicable far outside the original BS premises, as we will see. For now, we will simply state some conventional results that connect measure changes to the issue of valuation:

- First Fundamental Theorem of Asset Pricing: No arbitrage $\Leftrightarrow$ There exists an equivalent measure under which tradeables are martingales
- Second Fundamental Theorem of Asset Pricing: No arbitrage and completeness $\Leftrightarrow$ There exists a unique equivalent measure under which tradeables are martingales.

In the second theorem, "completeness" means that *any* contingent claim can be replicated via appropriate positions in other tradeables. These are very familiar in the usual BS setting: any contingent claim can be replicated by an appropriate (dynamic) hedging strategy (namely, delta-hedging), and the resulting value of the claim is the expectation of the terminal payoff under the only equivalent martingale measure (EMM) for the underlying process (geometric Brownian motion). It must be stressed here that martingale pricing means a *representation* of the claim in terms of a replicating hedging strategy. Pricing and hedging are two sides of the same coin, and valuation here means *relative* valuation.

We are, of course, concerned here with incomplete markets, in which case absence-of-arbitrage arguments are *not* sufficient to establish a rational[72] value for a contingent claim. In general, there is no unique EMM in such cases and martingale

pricing loses its interpretation as an encapsulation of an underlying hedging strategy around the claim in question.[73] As noted in Björk (2009), martingale pricing in incomplete markets amounts to a consistency condition across traded assets.

*3.2.2.2 Minimal martingales and residual risk*

Having said this, martingale pricing in connection with trading strategies remains a compelling framework, and it can be asked if much of the underlying thrust (namely, relative valuation) can be retained. To this end, we briefly discuss here the concept of the so-called minimal martingale measure (MMM), with much greater detail to appear later in this subsection. (See Föllmer and Schweizer [2010] and the references therein for more background.) This is the EMM under which $P$–martingales that are orthogonal to tradeables remain martingales under the new measure. By orthogonal we mean (see the context of continuous semi-martingales in Chapter 5) random variables which have zero quadratic covariation. The intuition here is akin to a regression. Imagine that a derivative security is projected (in some dynamic, but otherwise unspecified, sense) onto the space of tradeables, plus some residual. Abstractly we write:

$$dV = \delta_i dS_i + dM \tag{3.63}$$

with $E_t^P dM = 0$ and $< dS_i, dM >= 0$. Under the MMM $Q$, $E_t^Q dM = 0$ and thus we also have that the value process $V$ is a $Q$–martingale (since the tradeables $S$ are $Q$–martingales). Thus, as in the complete market case, we have an interpretation of martingale pricing as entailing a certain ("optimal", so to speak) hedging strategy. Note that this does *not* imply that there are no other risk factors associated with the residual $M$ for which we need not account; it only means that, as a *formal* representation of value this particular martingale measure is perfectly analogous to the complete market case.[74] Note further that, precisely because of this point (risk adjustment), this expected value is not necessarily the price at which we would want to transact.

MMM clearly provides a very nice framework, not least because it maintains some kind of connection to the original, powerful BS replication argument. (In fact, for a very general set of processes to be considered in Chapter 5, analytical results for deriving the MMM can be obtained.) However, a rather small problem can already be seen: as we noted, the manner in which a derivative security (whose value we seek) is to be projected onto the space of tradeables (and hence deriving the value *and* appropriate hedging strategy) is completely undetermined! What is in fact needed is a methodology that explicitly specifies this projection (*i.e.*, the set of trading strategies around the claim in question) *as the starting point*, and *then* expresses the value of the claim as an appropriate representation of this projection. As we will see, the cost here is that we lose, in some sense, the framework of EMM

pricing. However, this is hardly a catastrophe, since in the case of incomplete markets there is no *necessary* requirement that the value function be an expectation under an EMM. This brings us now to the question of pricing functional, which we approach in an abstract manner before turning to concrete examples.

#### *3.2.2.3 Abstract representation of optimal hedging strategies: pricing functionals and informational efficiency*

Our basic viewpoint that we have emphasized throughout is the notion that valuation is inextricably linked to the question of portfolio optimization, broadly understood. We will provide here a fairly abstract discussion of this theme followed by some specific examples as a means of conveying the general idea. Assume we have a derivative security to which we wish to assign a price/value, in conjunction with a hedging strategy around that security which makes operationally meaningful this assignment. Thus, we are faced with an aggregate portfolio consisting of the security, and some set of tradeables whose hedges/portfolio weights will be denoted by $\Delta_i$. The portfolio dynamics are given by

$$d\Pi = dV - \Delta_i dS_i \tag{3.64}$$

where the value $V$ and hedges $\Delta_i$ (both processes) are to be determined. Let us denote (somewhat heuristically) by $I$ the information set on which valuation and hedging decisions are made. We define a *value function* to be a map from this information set to the set of values and hedges:

$$\Im_\theta : I_t \to (V_t(I_t;\theta), \Delta_s(I_t;\theta)) \tag{3.65}$$

where there may be some kind of parametric dependence represented by $\theta$. (Note that the value function involves *both* value and hedge.) In fact, (3.65) provides a formalization of a concept we have employed throughout, namely a parameterization of *value drivers*. The portfolio dynamics in (3.65) can be written as

$$\Pi = \int_t^T dV_s - \int_t^T \Delta_{i,s} dS_{i,s} = \int_t^T (\partial_t V + \tfrac{1}{2} V_{II} dI_s^2) ds + \int_t^T (V_I dI_s - \Delta_{i,s} dS_{i,s}) \tag{3.66}$$

We leave unspecified for now the dynamics of the information set, as well as the precise dependency of value and hedge on this information. An obvious example would be where the only information used is price information.[75]

We can interpret the terms in (3.66) as follows. The second (ensemble) term represents effectiveness of the hedges. As information changes, the hedges also

change based on whatever appropriate risk criteria are adopted, *e.g.*, local variance minimization:

$$\Delta_s = V_I \frac{\langle dI_s dS_s \rangle}{\langle dS_s^2 \rangle} \tag{3.67}$$

The first term in (3.66) represents the realized (residual) exposure. We will see that we can choose an appropriate form, given the parametric dependence in question, so that a particular exposure is created.[76,77] As an indication, assume that the basic DGP belongs to the class of continuous semimartingales (see Chapter 5) and that the relevant information set is the set of process state variables. Then by Itô's lemma, (3.64) can be written as

$$d\Pi = (V_t + \tfrac{1}{2}\mathfrak{L}V)dt + (V_{S_i} - \Delta_i)dS_i + V_{S'_i}dS'_i \tag{3.68}$$

where the primes denote non-traded entities (such as stochastic volatility) and $\mathfrak{L}$ is a second-order partial differential operator with the following form:

$$\mathfrak{L} = S_i S_j (a_{ij}\partial_{S_i S_j} + 2b_{ij}\partial_{S_i S'_j} + c_{ij}\partial_{S'_i S'_j}) \tag{3.69}$$

#### *3.2.2.4 Value drivers: cause vs. effect*

Now, following the motivation for the MMM in the previous section (*e.g.*, (3.63)), we ask the following question: can we *integrate* the choice of portfolio weights/hedge volumes with the construct of valuation in such a way that the residual portfolio is optimal or efficient in some sense? Note that this question is in some sense in opposition to the manner by which the problem is typically approached, in which some appropriate (in the sense of broadly capturing certain statistical or qualitative characteristics that are deemed relevant) martingale measure is specified, and *then* hedges are determined. It is important to understand that we are not saying anything about absence-of-arbitrage arguments, which are usually invoked to justifying EMM pricing, without first ascertaining the precise sense in which such pricing is meaningful.

In particular, we are free to select value drivers as a means of choosing hedge ratios such that the residual portfolio permits, in an efficient manner (in the sense of exploiting available information [market or otherwise]), a decomposition of the unhedgeable risks that must be accounted for, with the corresponding value drivers as the key components to be determined. This viewpoint is, essentially, the analogue to the MMM framework introduced in (3.63). The meaning of (3.65) is the following: the operational meaning of an appropriate value function is that it facilitates an analysis/understanding of the residual portfolio risk through its corresponding hedging regime (via its dependence on some value drivers). The primary thing to understand is that a value function should be chosen in such a way that we can better understand the resulting exposure (as laid out in (3.66)) than we could with an alternative valuation/hedging program. A very intuitive (and as we will see, often

very suitable) choice for the class of processes under consideration is a projected (not implied) volatility in comparison to realized volatility (or more generally projected covariance structure against realized). As we will also see, a benefit from this approach is robustness of estimation, in comparison to approaches that require the imposition of greater structure.

For example, if we choose the value function and associated (delta) hedges to be functions *only* of the (observable) prices of tradeables, with the value function solving the following Partial Differential Equation (PDE):

$$V_t + \tfrac{1}{2} S_i S_j \hat{a}_{ij} V_{S_i S_j} = 0 \tag{3.70}$$

for some (as yet unspecified) process $\hat{a}$, then the (residual) portfolio process (3.68) becomes

$$d\Pi = \tfrac{1}{2} S_i S_j (a_{ij} - \hat{a}_{ij}) V_{S_i S_j} dt \tag{3.71}$$

From (3.71), it can be seen that the portfolio dynamics are driven by a ("gamma weighted") difference between the realized characteristics ($a$) and the valuation characteristics ($\hat{a}$). In other words, even though our pricing functional in (3.70) is "wrong" (it is essentially based on counterfactual Gaussian dynamics), the resulting portfolio (hedging) stratagem readily provides an understanding of the entailed exposures.

We need to stress here what we are *not* saying. We do not seek an EMM which is "more realistic" than GBM, or which is "better calibrated" to market data. It is a very popular approach in the industry to construct valuations via expectations under some pricing measure, and to derive hedges from these valuations. Typically, appeal is made to the fundamental theorems of asset pricing to justify this endeavor, but this viewpoint is mistaken, frankly. Consider the following simple example. Assume we have a price process following GBM, except that the volatility is unknown. Assume further that the realized volatility (over a given time horizon) is independent of terminal price. Then, as we will see in Section 5.2.2, expectations under *this* measure allow one to separate price from (integrated/cumulative) volatility. Thus, a common approach is to assert that the value of an option on an underlying following this process is given by the following:

$$E_t^Q (S_T - K)^+ = E_t^Q E_\sigma^Q (S_T - K)^+ = E_t^Q V_{BS}(S, K, \sigma) \tag{3.72}$$

where $V_{BS}$ denotes a Black-Scholes functional and $\sigma$ is realized cumulative volatility. The first thing to note is that, even if (stochastic) volatility and price are independent under the physical measure, there is no reason to assume that they would remain so under a pricing measure, in which case it is not even clear how useful (3.72) is on its own terms.[78] But there is a much larger problem with expressions such as (3.72): there is *no* operational meaning assigned to the measure $Q$.

It means nothing to assign value without *also* relating that value to some hedging strategy designed to realize that value.[79]

Failure to understand this basic point can (and has, we can attest) lead to much confusion. In this example, the exposure is clearly to realized volatility (or more accurately variance, but we neglect the distinction here). However, the relevant question is *how* that realization manifests itself. We argue that, since any structured product can only be partially replicated, there will necessarily be exposure arising from the *connection* between valuation and associated hedges (as in (3.71)) and that proper assessment of this exposure (*e.g.*, through optimization of some portfolio characteristic such as variance) *cannot* be addressed by (arbitrarily) choosing some equivalent pricing measure. To understand the distinction, note that from basic convexity considerations that $E_t^Q V_{BS}(S,K,\sigma) \neq V_{BS}(S,K,E_t^Q\sigma)$.[80] This fact can be misleading, because it can erroneously lead to the conclusion that exposure to volatility (the actual risk factor) manifests itself in the initial premium, and not in the context of an overall portfolio in which that premium is but one component. This view (although quite standard) is simply false. These points should become clearer with a concrete example.

*3.2.2.5 Case study: Heston with one traded asset*

Consider the popular Heston stochastic volatility process[81] under the physical measure:[82]

$$\begin{aligned} \frac{dS}{S} &= \mu\, dt + \sqrt{v} dw_1 \\ dv &= \kappa(\theta - v)dt + \sigma\sqrt{v} dw_2 \end{aligned} \tag{3.73}$$

The standard approach to deriving the pricing relation for an option under Heston is as follows (see Wilmott [2000]): assume the underlying ($S$) trades, as well as some option (say, ATM)[83]. Then, to price another option (say, ITM or OTM) with price $V$, we can form the following portfolio:

$$\Pi = V - \Delta S - \Delta^1 V^1 \tag{3.74}$$

where $V^1$ denotes the traded option. Self-financing implies

$$d\Pi = dV - \Delta\, dS - \Delta^1 dV^1 \tag{3.75}$$

All randomness can be eliminated from this portfolio by requiring that (note that we unrealistically assume that variance, although untraded, is nonetheless observable)

$$\begin{aligned} V_S &= \Delta + \Delta^1 V_S^1 \\ V_v &= \Delta^1 V_v^1 \end{aligned} \tag{3.76}$$

Assuming interest rates are zero, by arbitrage conditions we see that

$$V_t + \mathcal{L}V - \frac{V_v}{V_v^1}(V_t^1 + \mathcal{L}V^1) = 0 \tag{3.77}$$

where $\mathcal{L}$ denotes a second-order partial differential operator from the Itô calculus:

$$\mathcal{L}V \equiv \tfrac{1}{2}v(S^2 V_{SS} + 2\rho\sigma S V_{Sv} + \sigma^2 V_{vv}) \tag{3.78}$$

Now, since the second option is a tradeable, it is a martingale under some measure (call it $Q$) so it must satisfy

$$V_t^1 + \mathcal{L}V^1 + \kappa^Q(\theta^Q - v)V_v^1 = 0 \tag{3.79}$$

where for convenience we assume that under the pricing measure the affine form of the underlying process is retained. Now, we can readily see that the primary option price *must* also satisfy this PDE, so in particular is also a $Q$–martingale.[84] As we have already pointed out, martingale pricing amounts to *relative* pricing, a consistency condition across tradable assets (Björk [2009]).

*3.2.2.6 Heston case study: absence of arbitrage to the rescue (not)*

Now, we relax one of these assumptions and consider the case where the second option does not trade. In this case, we *cannot* appeal to arbitrage arguments to deduce the price of the primary option. We can completely eliminate exposure to the underlying but not to the stochastic variance, by taking $\Delta = V_S$. However, as we will see, the *local* variance minimizing hedge includes a vega term, specifically $\frac{\rho\sigma}{S}V_v$, and we will retain this adjustment when we study the global properties of the resulting portfolio. The portfolio evolution then becomes

$$d\Pi = (V_t + \mathcal{L}V)dt + V_v\left(dv - \rho\sigma\frac{dS}{S}\right) \tag{3.80}$$

Now, *purely for convenience*, we may express the value function as an expectation of the terminal payoff under some arbitrary martingale measure (again retaining affine structure), so the portfolio evolution becomes

$$d\Pi = V_v\left(-\kappa'(\theta' - v)dt + dv - \rho\sigma\frac{dS}{S}\right) \tag{3.81}$$

(The primes denote the parameters under this new measure.) The advantage of this choice clearly is to permit a simple connection of the portfolio dynamics to the dynamics of both price and the unhedgeable stochastic variance. However, it

still remains the case that there is residual exposure to that variance that must be accounted for. We can choose (if we like) to represent the value function (and associated hedging strategy) in terms of an EMM expectation, but only *after* we have specified how we want to account for the unhedgeable risk. Without this prior specification, any kind of representation of the value function in terms of martingale pricing is essentially meaningless. Simply put, risk-neutral pricing does not apply in this case.

To highlight this point, suppose that we take the value function to have the following form:

$$V = V_{BS}(S, K, \tau, \hat{\sigma}) \tag{3.82}$$

where $\hat{\sigma}$ is some (projected) volatility at which we price *and* hedge in terms of Black-Scholes (BS) entities (*i.e.*, call values and deltas). By construction, $V_t + \frac{1}{2}\hat{\sigma}^2 S^2 V_{SS} = 0$. In this case, the value function has no "vega" dependence (all partial derivatives wrt. $v$ are zero), so there is no adjustment to the delta and the portfolio evolution becomes

$$d\Pi = \tfrac{1}{2}(v - \hat{\sigma}^2) S^2 V_{ss} dt \tag{3.83}$$

which is very reminiscent of the standard result for hedging with unknown volatility in a Black-Scholes environment (*i.e.*, when volatility is not stochastic). We should stress that this representation of the value function does *not* entail EMM pricing; there is no change of measure that can take the Heston stochastic volatility process into a constant volatility process.

#### *3.2.2.7 Heston case study: portfolio optimization criterion*

This freedom gives us two possible approaches: specify parameters $\kappa'$ and $\theta'$ in an EMM pricing framework, or specify the parameter $\hat{\sigma}$ in a BS pricing framework. (Note that in the EMM framework the covariance structure is the same as under the physical measure.) This specification is made to ensure some property of the portfolio in light of the unhedgeable risk, such as variance minimization. In other words, we consider the following two problems:

$$\min_{\kappa',\theta'} \operatorname{var}(\Pi) = \min_{\kappa',\theta'} \operatorname{var}\left( \int_t^T V_v(-\kappa'(\theta' - v_s)ds + dv_s - \rho\sigma \frac{dS_s}{S_s}) \right) \tag{3.84}$$

and

$$\min_{\hat{\sigma}} \operatorname{var}(\Pi) = \min_{\hat{\sigma}} \operatorname{var}\left( \tfrac{1}{2} \int_t^T (v_s - \hat{\sigma}^2) S_s^2 V_{ss} ds \right) \tag{3.85}$$

Note that both results in some sense express the residual risk in terms of a gamma/vega weighted integral of realized variance over expected/projected variance. In truth, both of these expressions arise from the basic delta-hedged[85] portfolio

$$(S_T - K)^+ - V_t - \int_t^T \Delta_s dS_s \tag{3.86}$$

so it will be convenient to work with this form. To repeat, in each case the (dynamic) hedge is the derivative of the option value function wrt. price. What differs is how we calculate the value (and hedge) at each point at time, and as a consequence, what the residual risk is.

Our basic framework is as follows: we assume we know the data-generating process under the physical measure. That is, we know that the underlying process is Heston, and that we know the underlying parameters. Then, for the two hedging approaches (in terms of EMM and in terms of BS) we look for the valuation/hedging parameters that minimize terminal variance of the resulting portfolio. We will do this via simulation.[86]

#### *3.2.2.8 Heston case study: simulated results*

We use the following parameters:

$$\mu = 0.1, \sigma = 0.3, \rho = -0.6, \kappa = 4, \theta = 0.1 \tag{3.87}$$

and time-to-maturity one year. As we will see in Section 5.2, Heston permits an efficient quadrature-based approach to the calculation of option prices and deltas in terms of an analytical result for the (conditional) characteristic function of the terminal (log) price. To facilitate our objective here, we assume that the stochastic variance is observable, although not traded; this (very heroic) assumption will actually serve to reinforce the point we wish to make. (Of course, in reality the stochastic variance is not observable and the need to filter an estimate of $v$ from observed price data is a major challenge to using continuous-time stochastic volatility models such as Heston, a topic we will address in Chapter 6.)

We will simulate 1,000 paths throughout. For the BS case, we have the following results for portfolio variance, as function of projected (hedging) vol (ATM options) in Figure 3.10:

For the EMM case, we have (as a function of mean reversion rate) in Figure 3.11:

From the simulations it appears that the minimum variance is close in each case. In truth, the BS variance is about 8% higher than the EMM variance (1.45e-4 vs. 1.35e-4). However, the replication cost (the $Q$-expectation) is about 4% higher (0.129 vs. 0.124), giving rise to nearly equal Sharpe ratios for the two portfolios. This is significant, because even in the case where we know everything about the data-generating process under the physical measure, there is no compelling reason to not adopt the model-free BS approach with its "wrong" distributional assumption. It stands to reason, in the usual case where we do not know the process that

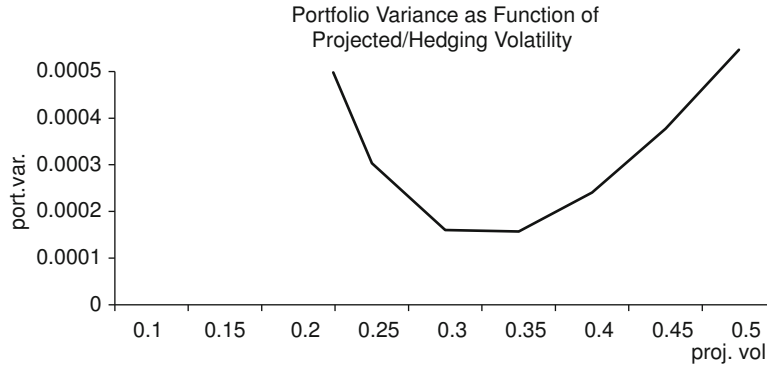



**Figure 3.10** Valuation and hedging with BS functional. Portfolio variance for non-EMM pricing

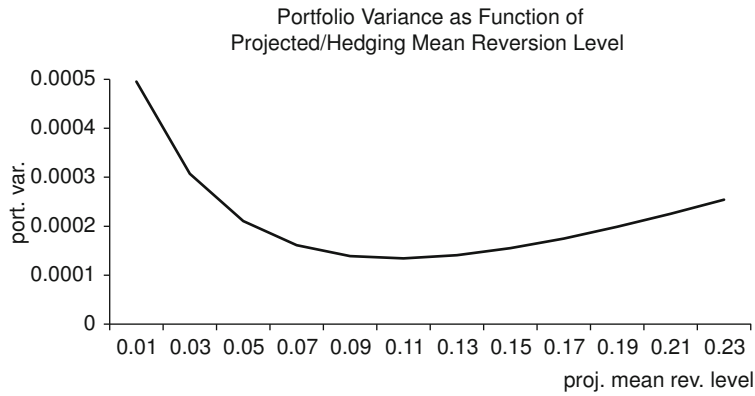



**Figure 3.11** Valuation and hedging with Heston functional. Portfolio variance for EMM pricing

generates the data (much less the values of the underlying parameters), that the BS approach should be superior. The optimal results for the two cases are as follows:

$$\begin{aligned} \hat{\sigma} &= 0.325, \text{ non-EMM case} \\ \hat{\kappa} &= 4, \hat{\theta} = 0.1, \text{ EMM case} \end{aligned} \tag{3.88}$$

(To clarify: the optimal $\hat{\sigma}$ in the non-EMM is the volatility at which we [counterfactually] price and hedge assuming a BS framework. For the EMM case the covariance structure is unchanged relative to the physical measure.) Note that the optimal BS hedging volatility is approximately the square root of the mean reversion level of the stochastic variance. Also note that the optimal hedging parameters for the EMM case appear to correspond to the mean reversion level and rate of the stochastic

variance under the physical measure. As we will see, the locally minimizing EMM parameters are the physical mean reversion rate (4) and level equal to 0.105.

In Figure 3.12 we show the ratio of minimum variances (between non-EMM and EMM cases) as a function of mean reversion rate:

We see again that the reduction in minimal variance (by using EMM valuation/hedging instead of BS) is not large, and becomes less as the mean reversion rate increases. Where the benefit (of EMM-based valuation and hedging) is greatest (in terms of global minimization), is precisely where the problems of estimation are most acute: low mean reversion rates.[87] This is significant, as in practice mean reversion rates are difficult to estimate robustly. These results show that for cases where the stochastic variance exhibits high mean reversion, the incremental benefit of getting a good estimate of the underlying parameters is decreasingly small.[88] Alternatively, we have a measure of the incremental benefit of portfolio variance reduction as the cost of information extraction (*i.e.*, filtering of unobserved stochastic variance) increases. This benefit is of the order of 15–20%, and this reduction prevails *if* filtering provides the exact parameter values (which it generally never does; see Chapter 6). We thus have a practical example arguing against the standard use of EMM valuation, and in terms of non-equivalent (*i.e.*, qualitatively wrong) measures.

#### *3.2.2.9 Heston case study: small sample considerations*

In general we will not know the actual dynamics of the underlying price process, and we will usually have a small sample of data (*e.g.*, forward price histories is some energy markets are no more than four years long). In this section we will investigate some of the implications of this situation. We first note a natural way of looking at

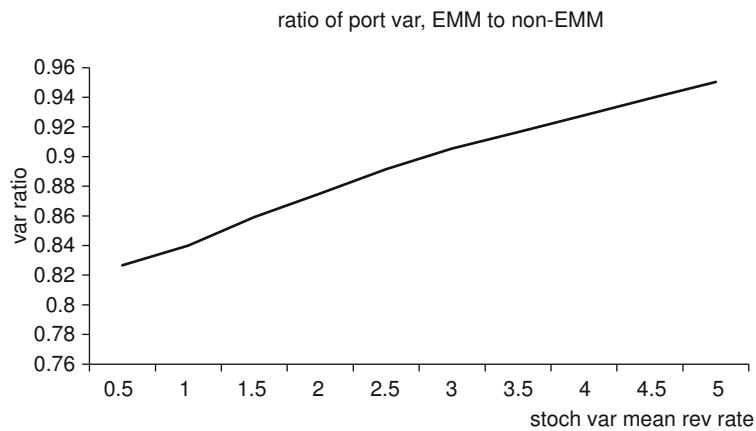



**Figure 3.12** Portfolio variance comparison, EMM vs. non-EMM. As a function of the stochastic variance mean reversion rate (under the physical measure)

the expression driving the minimal variance under the non-EMM case. The P&L, for a given path, is given in this case by

$$P\&L = \frac{1}{2}\int_t^T (v_s - \hat{\sigma}^2) S_s^2 V_{ss} ds \tag{3.89}$$

It seems natural to ask, for a given path, what is the hedging volatility that makes the P&L exactly zero? (Recall that we considered this question in Section 3.1.3 for actual energy market data.) For each price path in the sample, it is straightforward to solve the nonlinear equation that arises from the terminal payoff minus initial premium plus accrued delta hedging proceeds. This yields a sample of replication volatilities, which can be compared against the results of hedging to minimize portfolio variance *across* the price paths in the sample. Assuming we know the underlying price process (as we have done here), both of these entities can be compared to the "known" optimal result (known from a large enough simulation). Typical results (for a set of 50 simulated paths) are shown below in Figure 3.13:

As a function of time to maturity we see the results for:

1. Large sample variance-minimizing volatility (blue)
2. Small sample variance-minimizing volatility (red)
3. Average replication volatility across small sample paths (green).

Generally speaking, pathwise analysis of the replication volatility seems to be a better proxy for the large sample portfolio statistic of interest than the corresponding small sample estimate of that statistic.

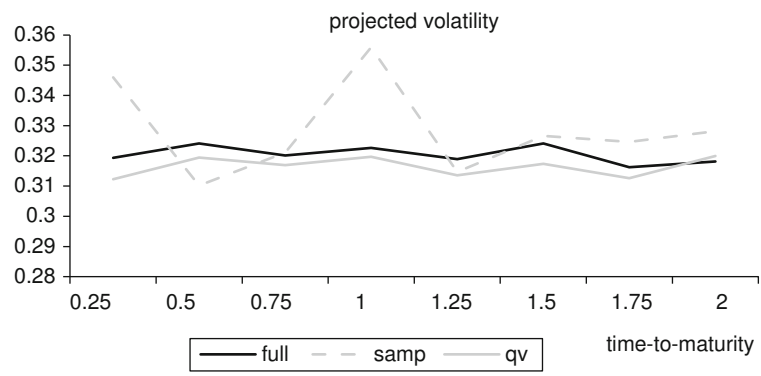



**Figure 3.13** Comparison of volatility projections. Large ("full") and small ("samp") variance minimizing valuation/hedging volatility vs. average of pathwise replication volatility ("QV")

*3.2.2.10 Filtering issues*

We have endeavored to present the best case possible for the Heston valuation function in comparison with a BS functional. This case entailed assuming that the stochastic variance, although not traded (and hence not hedgeable), was nonetheless (perfectly) observable. In reality, of course, we cannot directly observe stochastic variance, and (at best) can only employ some sort of estimate of it in terms of (observable) prices. This problem falls under the general category of (stochastic) filtering, and will be discussed in more detail in Chapter 6 (see also the Appendix to this chapter). Here we wish to only indicate an approach to ascertaining the effects of filtering on portfolio construction.

For a specific example, consider the following (discrete-time) filter:

$$\hat{v}_t = (1-\beta)\hat{v}_{t-1} + \gamma_0 + \gamma_1 \Delta z_t + \gamma_2 (\Delta z_t)^2 \tag{3.90}$$

which has the obvious interpretation of a weighted sum of past projections plus current (observed) log-price deviations. In the continuous-time limit, (3.90) becomes (suitably rescaling the various coefficients)

$$d\hat{v} = (\gamma_0 + \gamma_1(\mu - v/2) + \gamma_2 v - \beta\hat{v})dt + \gamma_1\sqrt{v}dw_1 \tag{3.91}$$

which of course retains affine dynamics. Now consider again the delta-hedged portfolio:

$$P\&L = (S_T - K)^+ - V - \int_t^T \Delta_s dS_s \tag{3.92}$$

Taking expectations under $P$, we get that

$$\begin{aligned} E_t^P P\&L &= E_t^P (S_T - K)^+ - V - \mu \int_t^T E_t^P \Delta_s S_s ds \\ &= E_t^P (S_T - K)^+ - V - \mu \int_t^T E_t^P S_s \cdot E_t^{P_S} \Delta_s ds \end{aligned} \tag{3.93}$$

under a suitable change of numeraire. Recall that we price *and* hedge under some (nonequivalent!) pricing measure $Q$ (*e.g.*, Heston) but with filtered variance "plugged in" to the valuation functional. (See also the Appendix to this chapter.) Thus we have (again employing numeraire change)

$$\Delta_s = E_s^{Q_S} 1(z_T > k) \tag{3.94}$$

where $k \equiv \log K$. Using characteristic function methods (see Chapter 5), (3.94) can be written as

$$\Delta_s = \frac{1}{2\pi} \int_{\Gamma} d\phi \frac{e^{i\phi(z_s - k) + \alpha(\phi)\hat{v}_s + \beta(\phi)}}{i\phi} \tag{3.95}$$

assuming, say, a Heston pricing functional in terms of the projected/filtered variance and using a suitable contour of integration $\Gamma$. (Recall that the actual hedge also consists of a vega correction, which we ignore here for convenience, as it is straightforward to obtain from expressions similar to (3.95).) Thus, the physical expectation in the integrand of (3.93) will entail expectations of the form

$$E_t^{P_s} e^{i\phi z_s + \alpha(\phi)\hat{v}_s} \tag{3.96}$$

which should be tractable owing to the affine dynamics in (3.91). Hence, akin to the static-dynamic problem considered in the lognormal case, the physical expectations should be reducible to quadrature (see Chapter 7), and more readily analyzed (than say through simulation).

Having seen some practical demonstrations of valuation and hedging in incomplete markets, we now consider tools necessary for understanding the operational issues associated with energy structures.

## 3.3 Stochastic optimization

We will discuss in this section various tools broadly associated with so-called optimal (stochastic) control. As usual, our intent is not to supply a thorough account but only to provide the reader with sufficient background to understand applications to energy market valuation problems.[89] For greater details the reader should consult any standard text, *e.g.*, Kall and Stein (1994) for stochastic programming and Nocedal and Wright (2006) for numerical optimization.

### 3.3.1 Stochastic dynamic programming and HJB

#### *3.3.1.1 Control problems and stopping time problems*

Dynamic programming (DP) refers to an optimization problem, where a choice (or action or control) made now affects the decisions that can be made in the future. In stochastic dynamic programming (SDP), these decisions are confronted with uncertainty regarding the relevant drivers for the problem in question. We have already seen several examples, which we recap here:

- American (early exercise) options: the decision to exercise the option now must be compared with the value of holding on to the option (in the hopes that the underlying price will go higher still)

- Natural gas storage: for a given (current) inventory level and prices, a decision can be made to either inject gas (at a cost), withdraw gas (and sell it), or do nothing; in each case, the cost/benefit of the action must be compared with the expected value of storage at the *new* inventory level
- Power plant/tolling: if the unit is currently down (say), a decision must be made to start up the unit (incurring both variable capacity fuel costs and fixed [independent of generation level] start-up costs) or stay down; the comparison is with the expected value of future spark spreads in the same/alternate state (keeping in mind that, once up, no additional start-up costs are required to maintain that state).[90]

These problems can typically be abstractly characterized as follows. For stochastic control problems we look to solve

$$V(S_t, A_t, t) = \sup_a E_t \left( \int_t^T e^{-r(s-t)} g(S_s, A_s; a_s)\, ds + e^{-r(T-t)} f(S_T, A_T) \right),$$
$$A \in \mathfrak{A},\ a \in \mathfrak{a} \tag{3.97}$$

where we consider some finite time horizon $[t, T]$. In (3.97), $S$ denotes some (possibly vectorial) underlying stochastic driver, such as the price of some commodity. $A$ represents the state of the system and $a$ is the policy/control (essentially, the dynamics of the state), assumed adapted to the driver. (In fact, we will assume throughout that we are dealing with Markovian decision rules). In some sense the policy is akin to a stopping time. The two terms in (3.97) represent accumulated gains/losses ($g$) through time plus a terminal payoff ($f$), both state (and driver) dependent (and discounted continuously via the rate $r$, assumed constant). We seek the policy (the manner in which the state changes through time) such that the expectation in (3.97) is maximized. Note that in general there can be constraints on the allowable state, as well as on the dynamics of the state (and hence allowable policies; this is the meaning of the two set memberships following the expectation in (3.97)). For example, in the gas storage problem, there are typically maximum withdrawal and injection rates, maximum and minimum capacity levels (all of which may vary deterministically, *e.g.*, seasonally), and other physical restrictions on operation that must be taken into account; see EW for a greater discussion of such deals.

The perhaps more familiar optimal stopping time problems (*e.g.*, American option valuation) can be framed similarly: Here we wish to solve

$$V(S_t, t) = \sup_{t \le \tau \le T} E_t \left( \int_t^\tau e^{-r(s-t)} g(S_s) ds + e^{-r(\tau - t)} f(S_\tau) \right) \tag{3.98}$$

for some (bounded) stopping time $\tau$ over a finite interval. In the context of the examples considered above, we would have

- American (put) options: $f = (K - S)^+$, $g = 0$
- Gas storage: $f = 0$, $g = -aS$, $a$ is injection/withdraw rate, $A$ continuous, bounded between 0 and facility maximum capacity
- Tolling: $f = 0$, $g = a((P - HR \cdot G - VOM) - (1 - A)X)$, $a = 1$ or 0 (up/down), $A$ discrete, 0 or 1 (on/off)

To be clear, we are omitting many important technical details here; for these we refer to sources like Pham (2010). Our primary purpose is to provide a reasonable amount of intuition and formality to enable the reader to understand some of the applications in energy markets. To further fix matters, we will assume state and driver dynamics of the following (diffusive) form:

$$\begin{aligned} dS &= \mu(S, A)dt + \sigma(S) \cdot dw \\ dA &= adt \end{aligned} \tag{3.99}$$

where we allow the drift of the of driver to be state dependent;[91] of course, for optimal stopping problems, there is no state as such and hence no state dependence in the driver dynamics. (As usual the dot $\cdot$ denotes Hadamard [element-by-element] vector product.)

#### *3.3.1.2 Backward induction as a way forward*

So, we turn attention to the question of how problems such as (3.97) and (3.98) can actually be solved. We will actually have much more to say about this in Chapter 7, but there are some important concepts that must be introduced here. First is the rather important question of how these optimizations are to be carried out. It would seem that these problems are inherently high dimensional and effectively intractable. However, for the class of Markov decision processes that we will confine attention to, this turns out to not be the case. We here appeal to Bellman's (1957) principle, which we quote here: "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."

In other words, the value function equals the maximum over all actions taken now, *plus* the expected value of the value function over later times *given* the state changed induced by that action. This idea permits the use of so-called backward induction to iteratively solve these control problems. To understand the main idea,

consider the following discretization of the control problem (3.97) (for convenience we will drop reference to the set in which the controls must belong):

$$V(S_t, A_t, t) = \sup_a E_t \left( \begin{array}{l} \int_t^{t+\Delta t} e^{-r(s-t)} g(S_s, A_s; a_s)\, ds + \\ \int_{t+\Delta t}^{T} e^{-r(s-t)} g(S_s, A_s; a_s)\, ds + e^{-r(T-t)} f(S_T, A_T) \end{array} \right)$$
$$\approx \max_a (g(S_t, A; a)\,\Delta t + e^{-r\Delta t} E_t V(S_{t+\Delta t}, A + a\Delta t, t + \Delta t)) \quad (3.100)$$

and for the stopping time problem (3.98):

$$V(S_t, t) \approx \max \left( f(S_t), e^{-r\Delta t} \sup_{t+\Delta t \le \tau \le T} E_t \left( \begin{array}{l} \int_t^{\tau} e^{-r(s-t-\Delta t)} g(S_s)\, ds \\ + e^{-r(\tau - t - \Delta t)} f(S_\tau) \end{array} \right) \right)$$
$$= \max \left( f(S_t), e^{-r\Delta t} E_t \sup_{t+\Delta t \le \tau \le T} E_{t+\Delta t} \left( \begin{array}{l} \int_t^{\tau} e^{-r(s-t-\Delta t)} g(S_s)\, ds \\ + e^{-r(\tau - t - \Delta t)} f(S_\tau) \end{array} \right) \right)$$
$$= \max(f(S_t),\ e^{-r\Delta t}(g\Delta t + E_t V(S_{t+\Delta t}, t + \Delta t))) \quad (3.101)$$

Now, the first thing to note is, as discrete-time approximations, (3.100) and (3.101) provide means of valuation via backward induction (we know the terminal values) so long as the underlying expectations and subsequent optimizations can be efficiently (or at least tractably) carried out.[92] We will consider such techniques in Chapter 7. Here, we wish to provide alternative formulations of (3.100) and (3.101) in the continuous-time limit, heuristically considered (leaving the relevant technical details and justifications to Pham [2010]). Using Itô's lemma and the dynamics in (3.99), the stochastic control problem (3.100) becomes[93]

$$V_t + \tfrac{1}{2}\mathfrak{L}V + \max_a (g(S, A; a) + \mu^T V_S + a V_A) - rV = 0 \quad (3.102)$$

where $\mathfrak{L}$ is a second-order partial differential operator associated with the Itô generator of the (driver) process in (3.99). Similarly, for the optimal stopping time problem in (3.101) we find

$$\max \left( f(S) - V, V_t + \mu^T V_S + \tfrac{1}{2}\mathfrak{L}V - rV + g \right) = 0 \quad (3.103)$$

Although of different forms, expressions (3.102) and (3.103) are both referred to as the *Hamilton-Jacobi-Bellman* (HJB) equations for the problem in question. These justly celebrated equations have wide application in mathematical finance in general, and many energy market problems.

A few observations can be made here. First, the optimal stopping time problem solved in (3.103) is the usual variational formulation/linear complementarity condition from American option pricing (*e.g.*, see Wilmott *et al.* [1997]), so the basic idea here should not be completely unfamiliar to most readers. Essentially what (3.103) implies is that there are two regions characterizing the valuation problem: one where early exercise takes place (so that $V = f(S)$) and one where it is better to hold on to the structure in question (so that the usual Itô equation governing martingale price processes holds). Second, the optimal stochastic control extracted from (3.102) involves a relation between the instantaneous cost/benefit from a state change, the state-dependent drift of the driver, and the incremental value of a state change. (The latter is commonly referred to as a shadow price.) For example, in the case of gas storage this term reduces to $\max_a a(V_A - S)$, which has the obvious interpretation that, when the incremental state value is above the current driver price, injection ($a > 0$) is the optimal action, and when the incremental value is below the current driver price, withdrawal ($a < 0$) is the optimal action. Finally, we see a distinction between control and stopping problems (although there are obvious similarities). The problem in (3.102) is linear in the value function but entails a nonlinear optimization (over the control $a$). In contrast, (3.103) is nonlinear in the value function but without explicit reference to the policy.[94]

#### *3.3.1.3 Example: optimal investment*

As a generic example, consider the optimal investment problem from Boguslavsky and Boguslavskaya (2004). The underlying asset follows a standard mean-reverting process with zero mean:

$$dx = -\kappa x dt + \sigma\, dw \tag{3.104}$$

The investor seeks a position $\alpha_t$ in the asset such that his terminal expected (power) utility[95] is maximized. With the mean reversion in (3.104), we can think of this scenario as representing spread/pairs trading in commodity markets. The problem can be expressed as an optimal control problem. Denoting wealth (from self-financing position) by $w$, we have the value of terminal wealth/utility given by

$$\begin{aligned} V(x, w, t) &= \sup_{\alpha} E_t \tfrac{1}{\gamma} w_T^{\gamma} \\ dw = \alpha\, dx &= -\kappa\alpha x dt + \alpha\sigma\, dw \end{aligned} \tag{3.105}$$

with relative risk aversion represented by $1-\gamma$. Invoking HJB, we have the following value function dynamics:

$$V_t - \kappa x V_x + \tfrac{1}{2}\sigma^2 V_{xx} + \sup_{\alpha}\left(-\alpha\kappa x V_w + \tfrac{1}{2}\alpha^2\sigma^2 V_{ww} + \alpha\sigma V_{wx}\right) = 0 \tag{3.106}$$

and of course $V(x, w, T) = \frac{1}{\gamma} w^{\gamma}$. Owing to the quadratic nature of the supremum term in (3.106), the first order optimality condition is easily obtained, and

the resulting PDE (and thus the associated optimal control/position) can be solved analytically; see Boguslavsky and Boguslavskaya (2004) for details.[96] (We will revisit this problem in Section 6.2.2 when we address the issue of filtering out the [typically] unobservable mean-reversion level.)

### 3.3.2 Martingale duality

#### *3.3.2.1 Optimization*

Let us lead into our discussion of duality by recalling the more familiar application in standard optimization theory. Consider the following constrained optimization problem:

$$\begin{aligned} &\max f(x) \\ &\text{st } g_i(x) \leq 0 \end{aligned} \tag{3.107}$$

where $x \in R^n$ and there are $m$ constraints. Now, as is well known from first-year calculus, *equality*-constrained problems can be solved via the standard technique of Lagrange multipliers.[97] However, for the general case of inequality constraints, the issue is a bit subtler, although the more familiar association with Lagrange multipliers is retained to a large degree. We introduce non-negative multipliers $\mu \in R^m_{\geq 0}$ and a so-called Lagrangian given by

$$L(x, \mu) = f(x) - \mu^T g(x) \tag{3.108}$$

Note that, for any feasible point $x$ (that is, a point satisfying the constraints in (3.107)) we must have $L \geq f$. Therefore the maximum of $f$ is bounded from above by $L$, and this prompts us to investigate the so-called *dual* problem[98] given by

$$\begin{aligned} &\min L(x, \mu) \\ &\text{st } \nabla f - \mu^T \nabla g = 0, \\ &\mu \geq 0 \end{aligned} \tag{3.109}$$

where we introduce a stationarity (first-order) condition in (3.109), as any so-called active constraint (one satisfied by equality) must be tangent to the contours of the objective function, as in the usual equality-constrained case. In fact, the famous Karush-Kuhn-Tucker (KKT) conditions provide a necessary criteria for optimality in (3.109), namely that either a constraint is active, *or* the associated Lagrange multiplier is 0; see Nocedal and Wright (2006). (This condition is again reminiscent of the linear complementarity approach for valuing American options.) For a very wide class of problems, namely, convex optimization, the so-called *duality gap* (the difference between optimal solutions of (3.107) and (3.109)) is zero and the

KKT conditions become sufficient, as well. (This class of problems is commonly encountered in mathematical finance, as well.)

*3.3.2.2 Generalizations*

Note that the essence of the duality formulation is to transform an optimization problem over one category (points in $R^n$) into an optimization problem over a *different* category (effectively, the set of active constraints). As this approach often leads to progress in a problem, we speculate as to whether it could be applied to the kinds of stochastic optimization problems we are interested in here. Gratifyingly, the answer turns out to be in the affirmative. We first turn to the important work of Haugh and Kogan (2004, 2008).

Consider the following standard American option pricing problem:

$$V(S_t, t) = \sup_{t \le \tau \le T} E_t(e^{-r(\tau - t)} f(S_\tau)) \tag{3.110}$$

and its discrete-time analogue

$$V(S_t, t) = \max(f(S_t), e^{-r\Delta t} E_t V(S_{t+\Delta t}, t + \Delta t)) \tag{3.111}$$

Now, it should be clear that the discounted value function $V$ is a supermartingale: $V_t \ge e^{-r\Delta t} E_t V_T$ for $t \le T$.[99] This fact leads us to seek approximations to the primal problem (3.110) as follows. For any supermartingale $\pi_t$, we have that

$$\begin{aligned} V(S_t, t) &= \sup_{t \le \tau \le T} E_t(e^{-r(\tau - t)} f(S_\tau) - \pi_\tau + \pi_\tau) \\ &\le \sup_{t \le \tau \le T} E_t(e^{-r(\tau - t)} f(S_\tau) - \pi_\tau) + \pi_t \le E_t \max_{t \le \tau \le T} (e^{-r(\tau - t)} f(S_\tau) - \pi_\tau) + \pi_t \end{aligned} \tag{3.112}$$

The first inequality in (3.112) follows from the optional stopping theorem, and the second should be obvious. Note that in the final expression in (3.112), the supremum over (random) stopping times has been replaced by a (pathwise) maximum over deterministic times. Thus, taking the infimum over supermartingales, we see that

$$V(S_t, t) \le \inf_{\pi} \left\{ E_t \max_{t \le \tau \le T} (e^{-r(\tau - t)} f(S_\tau) - \pi_\tau) + \pi_t \right\} \tag{3.113}$$

Thus, we can already see an analogue with the duality result for standard (deterministic) optimization in (3.109). In fact, since the (discounted) value function is itself a supermartingale (and also dominates the payoff), we also have that

$$V(S_t, t) \le E_t \max_{t \le \tau \le T} e^{-r(\tau - t)} (f(S_\tau) - V_\tau) + V_t \le V_t \tag{3.114}$$

Obviously this point is not particularly helpful in and of itself, as the value function is precisely the thing we are interested in determining. However we do have the very useful result that we can bound from above the American option value, and in particular if we can approximate the value function with supermartingales that are close to the true value, we can attain a tight bound. We will see specific (tolling) applications in Section 4.2 when we consider numerical implementations. The main point to be stressed here is that the original problem in (3.110), an optimization over controls, can be transformed to a new problem in (3.113), which is an optimization over classes of candidate solutions.[100] It thus precisely corresponds to the more familiar duality results from standard optimization theory.

Let us also consider the corresponding results for the control problem in (3.97). We introduce a state-dependent process $h$ (whose properties we will specify shortly) and, using the notation $Ph \equiv h_t + \mu^T h_S + a h_A + \frac{1}{2}\mathfrak{L}h$, we write

$$\begin{aligned}
V(S_t, A_t, t) &= \sup_a E_t \left( \begin{array}{l} \int_t^T (e^{-r(s-t)} g(S_s, A_s; a_s) + P e^{-r(s-t)} h) ds - \\ \int_t^T P e^{-r(s-t)} h ds + e^{-r(T-t)} f(S_T, A_T) \end{array} \right) \\
&= \sup_a E_t \left( \begin{array}{l} \int_t^T e^{-r(s-t)} (g(S_s, A_s; a_s) + (P-r)h) ds - \\ \int_t^T P e^{-r(s-t)} h ds - \int_t^T e^{-r(s-t)} h_S^T \sigma \cdot dw_s + e^{-r(T-t)} f(S_T, A_T) \end{array} \right) \\
&= \sup_a E_t \left( \begin{array}{l} \int_t^T e^{-r(s-t)} (g(S_s, A_s; a_s) + (P-r)h) ds - \int_t^T d e^{-r(s-t)} h ds \\ + e^{-r(T-t)} f(S_T, A_T) \end{array} \right) \\
&= \sup_a E_t \left( \begin{array}{l} \int_t^T e^{-r(s-t)} (g(S_s, A_s; a_s) + (P-r)h) ds + h_t \\ + e^{-r(T-t)} (f(S_T, A_T) - h_T) \end{array} \right) \qquad (3.115)
\end{aligned}$$

Now, let us choose $h$ such that $h_T = f(S_T, A_T)$. Then, proceeding similarly to (3.112), we see that

$$V(S_t, A_t, t) \leq E_t \max_a \left( \int_t^T e^{-r(s-t)} (g(S_s, A_s; a_s) + (P-r)h) ds \right) + h_t \qquad (3.116)$$

from which we get the following upper bound analogous to (3.113):

$$V(S_t, A_t, t) \leq \inf_h \left\{ E_t \max_a \left( \int_t^T e^{-r(s-t)} \begin{pmatrix} h_t + \mu^T h_S + \frac{1}{2}\mathfrak{L}h - rh + \\ g + a_s h_A \end{pmatrix} ds \right) + h_t \right\} \tag{3.117}$$

Continuing the analogy with the early exercise (stopping time) case, we see that a supremum over controls is replaced by an infimum taken over an expectation in terms of pathwise maxima (essentially, perfect foresight execution). It is worth applying this approach to the more general stopping-time problem in (3.98). In this case we find the following upper bound:

$$V(S_t, t) \leq \inf_h \left\{ E_t \max_\tau \begin{pmatrix} \int_t^\tau e^{-r(s-t)} \left( h_t + \frac{1}{2}\mathfrak{L}h - rh + g + \mu^T h_S \right) ds + \\ e^{-r(\tau - t)} \left( f_\tau - h_\tau \right) \end{pmatrix} + h_t \right\} \tag{3.118}$$

Consider now using the value function itself as the argument in the infimum operators in (3.117) and (3.118). Using the equations satisfied by the value function in (3.102) and (3.103), it can readily be seen that the duality gap is zero, again demonstrating that tight upper bounds can be attained by employing good proxies for the actual value function.

#### *3.3.2.3 Example: gas storage*

Again, we will see concrete, numerical illustrations of these concepts in Chapter 4. However, it should already be clear that duality methods provide a very powerful way of crafting optimal control problems. As a short application, consider the following valuation of a natural gas storage facility.[101] Consider again a mean-reverting (log) price with dynamics:

$$\frac{dS}{S} = \kappa(\theta - \log S)dt + \sigma\, dw \tag{3.119}$$

We seek an injection/withdrawal policy (the control; positive for injection, negative for withdrawal) subject to the constraints of the facility. Ignoring discounting effects, the most basic representation of the problem takes the form

$$\begin{aligned} V(S, Q, t) &= \sup_q \left[ -E_t \int_t^T q_s S_s ds \right], \\ dQ &= q dt \\ q_{\min} \leq q \leq q_{\max}, &0 \leq Q \leq Q_{\max} \end{aligned} \tag{3.120}$$

Here $Q$ denotes the cumulative inventory in storage (*i.e.*, current capacity), clearly bounded from above and below.[102] Standard HJB analysis yields the following PDE in terms of controls:

$$V_t + \kappa(\theta - \log S)SV_S + \tfrac{1}{2}\sigma^2 S^2 V_{SS} + \max_q(q(V_Q - S)) = 0 \tag{3.121}$$

Note in (3.121) the familiar interpretation of the so-called shadow price (incremental/marginal value of inventory) as the criterion for injection/withdrawal. We also have the discrete-time version in terms of inventory changes:

$$V(S, Q, t) = \max_{Q'}(-(Q' - Q)S + E_t V(S_{t+\Delta t}, Q', t + \Delta t)) \tag{3.122}$$

The duality result (3.117) becomes

$$V(S_t, Q_t, t) \leq \inf_h \left\{ E_t \max_q \left( \int_t^T \left( \begin{array}{l} h_t + \kappa(\theta - \log S_s)S_s h_S + \tfrac{1}{2}\sigma^2 S_s^2 h_{SS} + \\ q_s(h_Q - S_s) \end{array} \right) ds \right) + h_t \right\} \tag{3.123}$$

Now, the obvious question presents itself: how should we choose the test function (so to speak) $h$ to get a good upper bound? As we have noted, good choices are those that are close to the true value function. A frequently used lower bound for storage valuation is the so-called basket of (spread) options[103] approach. We have already encountered this idea in Section 3.1.4, and more details are provided in EW. The essential idea is that we replicate the storage payoff by buying/selling options such that, regardless of how the options are exercised, no physical constraint of the facility will be violated. The problem thus amounts to a constrained optimization problem.[104]

The particular details of this optimization do not concern us here. The crucial point is that the optimal solution is a linear combination of spread option values, which are of course expectations of terminal value under the appropriate measure (we make no distinction between physical and pricing measure here). As such, their value has the form

$$h_{1,2} = E_t(F_{T_1,T_2} - F_{T_1,T_1})^+ \tag{3.124}$$

where $F_{t,T} = E_t S_T$. As martingales, these value functions satisfy the PDE associated with the Itô generator of the *spot* process.[105] Thus, the corresponding operator terms in the duality expression (3.123) vanish! The duality (upper-bound) result then becomes simply

$$V(S_t, Q_t, t) \leq E_t \max_q \left( \int_t^T q_s (h_Q - S_s) ds \right) + h_t \tag{3.125}$$

The expression in (3.125) has the obvious intuition of adjusting the baseline spread option value by the expected value of (perfect foresight) injection/withdrawal in terms of the shadow price. Note further that the result (3.125) is actually completely general, and does not depend on the particulars of the process (3.119) at all. (We introduced that process simply as a means of providing a familiar context for the analysis.[106]) We will not attempt to quantify now how tight an upper bound (3.125) is. We will simply note that the spread option lower bound is, in practice, typically a good approximation (at least for facilities where the injection/withdrawal rates are not too high, *i.e.*, reservoirs as opposed to salt domes), so we would anticipate the result in (3.125) to be a fairly tight upper bound.[107]

## 3.4 Appendix

### 3.4.1 Vega hedging and value drivers

In this section we explain (somewhat heuristically) how vega hedging a spread option creates exposure to realized correlation. We start (as usual) with the relevant portfolio:

$$\begin{aligned} \Pi = (S_2(T) - S_1(T))^+ - V - \int_t^T \Delta_1(s) dS_1(s) - \int_t^T \Delta_2(s) dS_2(s) \\ - v_1 \left( (S_1(T) - S_1)^+ - C_1 - \int_t^T \Delta'_1(s) dS_1(s) \right) \\ - v_2 \left( (S_2(T) - S_2)^+ - C_2 - \int_t^T \Delta'_2(s) dS_2(s) \right) \end{aligned} \tag{3.126}$$

For simplicity, we assume a zero-strike spread option and that the available leg options are ATM (at inception). The spread option (denoted by $V$) is delta hedged, as well as vega hedged (the leg options are denoted by $C_i$). The vega hedges are taken to be static, but for the particular (fixed) volumes, the leg options are themselves delta hedged. The portfolio (3.126) can be written as

$$\Pi = \tfrac{1}{2}\int_t^T ((\sigma_1^2 - \hat{\sigma}_1^2)S_1^2\Gamma_{11} + 2(\rho\sigma_1\sigma_2 - \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2)S_1S_2\Gamma_{12} + (\sigma_2^2 - \hat{\sigma}_2^2)S_2^2\Gamma_{22})ds$$
$$- \tfrac{1}{2}\upsilon_1\int_t^T (\sigma_1^2 - \hat{\sigma}_1^2)S_1^2\Gamma_1 ds - \tfrac{1}{2}\upsilon_2\int_t^T (\sigma_2^2 - \hat{\sigma}_2^2)S_2^2\Gamma_2 ds \tag{3.127}$$

in terms of realized and projected (denoted by hats) covariances. (Actually we are being a bit sloppy here: the projected [leg] volatilities should actually be taken as the implied volatilities that [by assumption here] exist in the market.) Now, we must say something about the (option) hedge volumes. We take these to be the usual vega-neutral ratios (between spread option and hedge options); using (3.55) and (3.56), these are given by (in terms of inception and projection entities)

$$\begin{aligned}\upsilon_1 &= \frac{S_1(S_1\hat{\sigma}_1\Gamma_{11} + S_2\hat{\rho}\hat{\sigma}_2\Gamma_{12})}{S_1^2\hat{\sigma}_1\Gamma_1} = \frac{\Gamma_{11}}{\Gamma_1} + \hat{\rho}\frac{S_2\hat{\sigma}_2}{S_1\hat{\sigma}_1}\frac{\Gamma_{12}}{\Gamma_1}\\ \upsilon_2 &= \frac{S_2(S_1\hat{\rho}\hat{\sigma}_1\Gamma_{12} + S_2\hat{\sigma}_2\Gamma_{22})}{S_2^2\hat{\sigma}_2\Gamma_2} = \frac{\Gamma_{22}}{\Gamma_2} + \hat{\rho}\frac{S_1\hat{\sigma}_1}{S_2\hat{\sigma}_2}\frac{\Gamma_{12}}{\Gamma_2}\end{aligned} \tag{3.128}$$

With these expressions (as well as the result (3.51)), the portfolio (3.126) becomes

$$\Pi = \tfrac{1}{2}\int_t^T \left(\begin{array}{l}(\sigma_1^2 - \hat{\sigma}_1^2)S_1^2(s)\left(\Gamma_{11}(s) - \frac{\Gamma_{11}}{\Gamma_1}\Gamma_1(s)\right) +\\ 2(\rho\sigma_1\sigma_2 - \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2)S_1(s)S_2(s)\Gamma_{12}(s) -\\ \hat{\rho}\frac{S_2\hat{\sigma}_2}{S_1\hat{\sigma}_1}\frac{\Gamma_{12}}{\Gamma_1}\Gamma_1(s)(\sigma_1^2 - \hat{\sigma}_1^2)S_1^2(s) -\\ \hat{\rho}\frac{S_1\hat{\sigma}_1}{S_2\hat{\sigma}_2}\frac{\Gamma_{12}}{\Gamma_2}\Gamma_2(s)(\sigma_2^2 - \hat{\sigma}_2^2)S_2^2(s) +\\ (\sigma_2^2 - \hat{\sigma}_2^2)S_2^2(s)\left(\Gamma_{22}(s) - \frac{\Gamma_{22}}{\Gamma_2}\Gamma_2(s)\right)\end{array}\right) ds \tag{3.129}$$

or

$$\Pi = \tfrac{1}{2}\int_t^T \left(\begin{array}{l}2(\rho - \hat{\rho})\sigma_1\sigma_2 S_1(s)S_2(s)\Gamma_{12}(s) +\\ (\sigma_1^2 - \hat{\sigma}_1^2)S_1^2(s)\left(\Gamma_{11}(s) - \frac{\Gamma_{11}}{\Gamma_1}\Gamma_1(s)\right) +\\ (\sigma_2^2 - \hat{\sigma}_2^2)S_2^2(s)\left(\Gamma_{22}(s) - \frac{\Gamma_{22}}{\Gamma_2}\Gamma_2(s)\right) +\\ \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2\Gamma_{12}S_1S_2\left(\begin{array}{l}2\left(\frac{\sigma_1}{\hat{\sigma}_1}\frac{\sigma_2}{\hat{\sigma}_2} - 1\right)\frac{S_1(s)}{S_1}\frac{S_2(s)}{S_2}\frac{\Gamma_{12}(s)}{\Gamma_{12}} -\\ \frac{\Gamma_1(s)}{\Gamma_1}\frac{S_1^2(s)}{S_1^2}(\frac{\sigma_1^2}{\hat{\sigma}_1^2} - 1) - \frac{\Gamma_2(s)}{\Gamma_2}\frac{S_2^2(s)}{S_2^2}(\frac{\sigma_2^2}{\hat{\sigma}_2^2} - 1)\end{array}\right)\end{array}\right) ds \tag{3.130}$$

Now, assuming the hedging period is sufficiently small (the typical case in tolling, say, is to rebalance the vega hedges at the start of each month in the deal), we anticipate that the second and third terms in (3.130) will be close to zero. The fourth term is proportional (approximately) to $(\frac{\sigma_1}{\hat{\sigma}_1} - \frac{\sigma_2}{\hat{\sigma}_2})^2$ , and in reasonably efficient volatility markets (where implied volatility is a good projection of realized volatility), we would expect this term to likewise be small. This leaves us with the first term, which is proportional to the difference between realized and projected *correlation*.

### 3.4.2 Value drivers and information conditioning

It is worth introducing some complications to the valuation problem to better draw out the points of interest. These issues will receive greater treatment in Chapter 6 when we discuss filtering; for now the basics will be sufficient.

#### *3.4.2.1 Hedging in the presence of unobservable states*

The quintessential example of a process with an unobservable component is stochastic volatility. There is little doubt that volatilities of prices in actual energy markets (or many other markets) are nonconstant, indeed nondeterministic. (EW should be consulted for further illustration.) A very popular model (which we will make much use of throughout this volume) is due to Heston (1993):

$$\begin{aligned} \frac{dS}{S} &= \mu\, dt + \sqrt{v} dw_1 \\ dv &= \kappa^P(\theta^P - v)dt + \sigma\sqrt{v} dw_2 \end{aligned} \tag{3.131}$$

where the superscript $P$ denotes parameter values under the physical measure (again, the probability measure characterizing the distribution of observed prices). The Brownian terms in (3.131) satisfy the Itô isometry $dw_1 dw_2 = \rho\, dt$. Now, we have already seen in Section 3.1.2 that in the familiar BS case there is a connection between replication and expectation wrt. a measure under which prices are martingales. The condition that the price $S$ (the only tradeable in this example) be a martingale is that it have zero drift; however, in general there is no unique constraint on the drift[108] of the (unobservable) stochastic variance $v$. One possible set of dynamics under some (martingale) measure $Q$ is

$$\begin{aligned} \frac{dS}{S} &= \sqrt{v} dw_1 \\ dv &= \kappa^Q(\theta^Q - v)dt + \sigma\sqrt{v} dw_2 \end{aligned} \tag{3.132}$$

This nonuniqueness is a characteristic feature of so-called incomplete markets, under which replication of derivative securities is not possible. We discussed this situation in detail in Section 3.2.[109] Our interest here is in applying standard martingale pricing to valuing an option under the Heston dynamics in (3.131). This

application will not only reveal the critical role that value drivers play in the valuation problem, but also highlight the general applicability of the idea beyond the standard BS setting. In fact, we will see that the standard approach of appealing to martingale pricing imposes severe informational requirements on the problem, and that we are advised to seek alternatives (while still keeping within the central framework of valuation via portfolio construction).

In general, stochastic variance $v$ and its dynamics are unobservable, so we cannot form valuations or hedges by conditioning directly on variance. We can only extract projections of stochastic variance from observed prices. In other words, we must resort to some (Markovian[110]) filter/estimator, whose dynamics we will represent as

$$\begin{aligned} d\hat{v} &= \mu_v(S,\hat{v})dt + \sigma_v(S,\hat{v})\frac{dS}{S} \\ d\hat{\sigma} &= \mu_\sigma dt + \sigma_\sigma \frac{dS}{S} \\ d\hat{\rho} &= \mu_\rho dt + \sigma_\rho \frac{dS}{S} \\ d\hat{\kappa}^P &= \mu_\kappa dt + \sigma_\kappa \frac{dS}{S} \\ d\hat{\theta}^P &= \mu_\theta dt + \sigma_\theta \frac{dS}{S} \end{aligned} \tag{3.133}$$

and where we omit explicit dependence on the projected parameters (volatility of variance, correlation, and mean reversion rate/level). (We will discuss filtering in much greater detail in Chapter 6.) The main points to take from (3.133) are that the dynamics of the estimators depend on both the prior values of the estimators as well as observed prices, and that the *only* source of stochasticity comes from price, hence the representation in terms of relative price changes. Now, assume we value and hedge a derivative security according to risk-neutral Heston dynamics, using the projected/estimated entities in a martingale formulation. Using (3.132) and invoking Feynman-Kac/Itô, the value function $V$ satisfies

$$V_t + \kappa^Q(\theta^Q - \hat{v})V_{\hat{v}} + \tfrac{1}{2}S^2\hat{v}V_{SS} + \hat{\rho}\hat{\sigma}S\hat{v}V_{S\hat{v}} + \tfrac{1}{2}\hat{\sigma}^2\hat{v}V_{\hat{v}\hat{v}} = 0 \tag{3.134}$$

In other words, we are taking the value function to be the (risk-neutral) martingale expectation; however, we plug the estimates of the (unobserved) stochastic variance into the valuation functional instead of the true value (which we cannot know).[111] Our objective is to choose the (valuation) parameters $(\kappa^Q, \theta^Q)$ appropriately.

An obvious question presents itself: what does this value function mean? As we have stressed, valuation can only take place within the context of a portfolio construction. So, consider the following dynamics of a delta-hedged portfolio:

$$\begin{aligned} d\Pi &= dV - \Delta\, dS \\ &= V_t dt + V_S dS + V_{\hat{v}}\, d\hat{v} + \tfrac{1}{2} V_{SS} dS^2 + V_{S\hat{v}} dS d\hat{v} + \tfrac{1}{2} V_{\hat{v}\hat{v}} d\hat{v}^2 + \cdots - \Delta\, dS \end{aligned} \tag{3.135}$$

where we omit terms involving derivatives wrt. projected process (covariance) parameters, as they will prove somewhat ancillary to our main point. Plainly, we will take the hedge to be the vega-adjusted delta:

$$\Delta = V_S + \frac{\sigma_\nu}{S} V_{\hat{v}} \tag{3.136}$$

Using the valuation PDE in (3.134), the portfolio dynamics (3.135) then becomes

$$d\Pi = \left( \begin{array}{l} V_{\hat{v}}(\mu_\nu - \kappa^Q(\theta^Q - \hat{v})) + \\ \frac{1}{2} V_{SS} S^2 (\nu - \hat{v}) + V_{S\hat{v}} S(\sigma_\nu \nu - \hat{\rho}\hat{\sigma}\hat{v}) + \frac{1}{2} V_{\hat{v}\hat{v}}(\sigma_\nu^2 \nu - \hat{\sigma}^2 \hat{v}) + \\ \cdots \end{array} \right) dt \tag{3.137}$$

The expression (3.137) can be written as

$$d\Pi = \left( \begin{array}{l} V_{\hat{v}}(\hat{\kappa}^P(\hat{\theta}^P - \hat{v}) - \kappa^Q(\theta^Q - \hat{v})) + \\ V_{\hat{v}}(\mu_\nu - \hat{\kappa}^P(\hat{\theta}^P - \hat{v})) + \\ \frac{1}{2} V_{SS} S^2 (\nu - \hat{v}) + \\ V_{S\hat{v}} S((\rho\sigma - \hat{\rho}\hat{\sigma})\hat{v} + \sigma_\nu \nu - \rho\sigma\hat{v}) + \\ \frac{1}{2} V_{\hat{v}\hat{v}}((\sigma^2 - \hat{\sigma}^2)\hat{v} + \sigma_\nu^2 \nu - \sigma^2 \hat{v}) + \\ \cdots \end{array} \right) dt \tag{3.138}$$

*3.4.2.2 Decomposing portfolio (residual) risk*

We can see from each of the constituent terms in (3.138) a combination of effects. In the first vega term, we have the usual (as we will see) difference between the drift of the (unhedgeable) stochastic variance (filtered) under the physical and pricing measures (or more accurately, the estimated/filtered drift). There is also, in the second vega term, the difference between the mean of the filtered variance and the corresponding estimate of the physical drift. In the gamma terms, the term proportional to $V_{SS}$ is the (hopefully by now) familiar difference between realized variance and projected variance. In the next two (gamma) terms, there is the difference between actual parameters (instantaneous covariance between price and stochastic variance and instantaneous variance of stochastic variance) and filtered/estimated parameters, *and* the difference between the variability of the estimators (essentially, estimation error) and actual parameters.

Note that there is no reason to think that the variability of the estimators will correspond to the actual dynamics of the physical process in (3.131). In particular, the cross vega term $V_{S\hat{v}}$ has the contribution $\sigma_\nu \nu - \rho\sigma\hat{v}$. Even if the filtered variance $\hat{v}$

is close to the actual variance $v$, there is no reason to think that the variability $\sigma_v$ will be close to the actual (instantaneous) covariance $\rho\sigma$. Indeed, since the estimator is perfectly correlated with price (by construction), in some sense it *cannot* well capture the joint movements of stochastic variance and price. In other words, there are two sources of risk introduced into the portfolio dynamics (3.138). The first is the usual estimation error between realized and projected entities. The second is the difference between actual dynamics of the unobserved process and the dynamics implied from the estimation dynamics in (3.133). Even if the estimator is good at reducing the former, there is no reason to think it will be good at reducing the latter.

It is reasonable to ask at this stage: what does this mathematics mean? Recall that we are trying to value (and of course hedge) a derivative security through martingale pricing methods that are standard in the literature. Since the underlying process is not (completely) observable, this valuation can only be based on some estimate of the unobservable components (as well as its dynamics). The portfolio process that arises as a result of (dynamically) hedging the security propagates estimation error, both in terms of the direct estimates and the indirect estimates of the (unobserved) process dynamics. The valuation parameters $(\kappa^{\mathbb{Q}}, \theta^{\mathbb{Q}})$ must in some sense do double work: from (3.138) it can be seen that the portfolio entails exposure to the realized (filtered) drift, *plus* the accumulation of filter error that is independent of any valuation. The resulting residual risk must be accounted for through a particular choice of $(\kappa^{\mathbb{Q}}, \theta^{\mathbb{Q}})$ in the valuation function (and associated hedging scheme), and it is not immediately obvious how exactly to do this.

We thus come to our main point here. Akin to the standard BS case (3.13), the portfolio dynamics in (3.138) create an exposure to the realization of some property of price volatility, which must be accounted for through certain parameters in the valuation functional. However, we see that in more general cases of interest (yet still sufficiently simple to prove reasonably tractable), the informational requirements for conducting the standard procedure of martingale pricing are enormous. We must somehow create a filtering regimen (represented abstractly by (3.133)) and then ascertain how the output of this procedure (or more accurately its error) interacts with the value function through a particular choice of parameterization. Neither of these tasks is easy in any way. We are led to wonder if there is an alternative approach.

Fortunately, there is, and we actually have already seen it here. We have actually been a bit misleading here. We are in fact *not* employing standard martingale pricing here. That approach involves pricing in terms of the *actual* unobservables, not some (filtered) estimate. (This fact alone raises questions as to that paradigm's feasibility.) What we have proposed doing here is actually an alternative to martingale pricing, while retaining the formal trappings. Since we have essentially weakened the constraints imposed by strict martingale pricing, we can ask how much further

we can go in relaxing the form of the value function. In fact, we have already answered this question. We saw in Section 3.2 that we can go surprisingly far, and in fact the BS functional performs quite robustly, even in non-BS environments. The reason, we saw, is related to why a Heston functional is problematic, even when the underlying dynamics are *correctly* specified by (3.132): the informational requirements are far less stringent for the BS functional. A good value driver *must* allow for robust conditioning on available information.

Although we have presented this material in fairly general form, we have to stress that the underlying issues that we seek to address are amplified in energy markets, where the nature of price dynamics can cause standard modeling approaches to run headlong into the structure vs. robustness conflict. The fact that, *e.g.*, volatility is stochastic[112] is *not* a reason in and of itself to employ formal stochastic volatility models. Nonetheless, models such as Heston in (3.131) are very useful for a number of reasons, not least their analytical and numerical tractability. We will consider many examples in Chapter 5. Another thing we have seen here is the appearance of well-known greeks (sensitivities to various underlyings) in the valuation problem. This is not an accident, and these entities in fact play an important role in valuation.

# 4 Selected Case Studies

## 4.1 Storage

We have already described the basics of storage deals in Section 1.2.3, and introduced spot-based control problem formulations in Section 3.3.2. (Recall that we also considered storage in Section 3.1.4 in contrasting dynamic hedging strategies [delta-hedging vs. rolling intrinsic].) In fact, in the discussion of control-based approaches, we pointed out that essentially forward-based valuations in terms of baskets-of-spread options provide lower bounds to this generally intractable optimization problem. This connection was not accidental; in Section 3.1.1 we stressed the basic unity of the two valuation paradigms (spot vs. forward). We will now provide a concrete example in support of these claims.

*4.1.0.1 Spot vs. forward modeling*

We will ignore throughout this chapter the distinction between physical and pricing probability measures. Recall the basic (log-) mean-reverting spot model in (3.119), with a (deterministic) seasonality factor $\chi$ introduced:

$$\frac{dS}{S} = (\kappa(\theta - \log S) + \dot{\chi} + \kappa\chi)dt + \sigma\, dw \tag{4.1}$$

The basic relationship between spot and forward prices is given by $F_{t,T} = E_t S_T$. Using methods that will be fully developed in Chapter 5, we will see that the corresponding forward dynamics are given by

$$\frac{dF_{t,T}}{F_{t,T}} = \sigma\, e^{-\kappa(T-t)}\, dw \tag{4.2}$$

with explicit solution

$$F_{t,T} = \exp\left( g e^{-\kappa(T-t)} + \theta(1 - e^{-\kappa(T-t)}) + \chi_T - \chi_t e^{-\kappa(T-t)} + \frac{\sigma^2}{4\kappa}(1 - e^{-2\kappa(T-t)}) \right) \tag{4.3}$$

In Chapter 7 we will see a detailed description of computational approaches to the basic control algorithm in (3.122). Our purpose here is to simply compare this spot-based solution to optimal (static) spread-option allocations based on the (implied) forward representation in (4.3).

#### *4.1.0.2 Relation between daily and monthly valuation*[1]

An alternative means of valuation is to consider, at a monthly level, the value from selling static (monthly) option positions such that total value is maximized while ensuring that, regardless of option exercise, the physical constraints of the system (flow rates, inventory levels, *etc.*) are not violated (see EW for a more detailed discussion). The underlying setup is a linear programming problem in terms of the monthly options across months (*e.g.*, options to inject in a given month and withdraw in some later month); see the Appendix to this chapter for details of the algorithm. These spread options can be computed from the forward curve implied by the underlying spot process (see (4.3)) and using the associated volatility term structure from the dynamics (4.2):

$$\sigma\sqrt{\frac{1-\exp(-2\kappa(T-t))}{2\kappa(T-t)}} \tag{4.4}$$

and using the fact that the movements of the individual forward prices are very nearly perfectly correlated.

Now, we can consider the limit of the monthly option value as we increase the resolution of monthly hedging. In other words, for a given deal term, we can construct a daily forward curve, and then we segment that curve into blocks corresponding to the temporal resolution of the spread options we can sell against the facility. For example, for a 365-day term and a 32-day resolution, we have 12 blocks of "monthly" contracts (11 blocks of length 32 days, 1 block of length 13 days) which can serve as inputs to a monthly storage valuation model. In the limit of daily hedging resolution, we would expect the monthly value to correspond to the daily value.

This is in fact what we see. Consider a specific example with the following parameters (incl. monthly seasonality factor):

$$\begin{aligned}&\sigma = 1.58, \theta = 0.05, \kappa = 0.74, g_0 = 4\\ &\chi = \{1.02, 1.02, 0.56, 0.55, 0.53, 0.49, 0.46, 0.44, 0.48, 0.57, 0.56, 0.73\}\end{aligned} \tag{4.5}$$

Facility max capacity is 1,000,000 MMBtu, max injection rate is 20,000 MMBtu/day, max withdrawal rate is 40,000 MMBtu/day (for convenience we exclude any injection and withdrawal charges and any fuel losses). The implied forward curve at the daily resolution is shown in Figure 4.1. (We assume a single day until the start of the deal.)

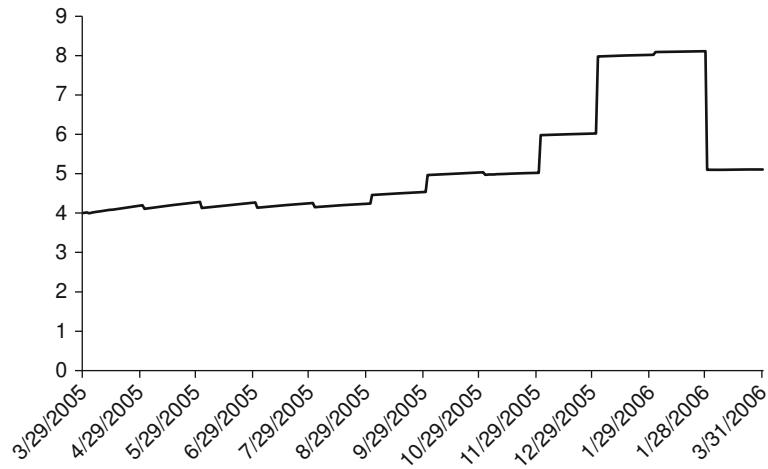



**Figure 4.1** Implied daily curve. Constructed from (4.3) and (4.5). Note that the seasonality factor in this case only has calendar-month dependence

The comparison of daily and monthly value for increasing resolution is shown in Table 4.1 and Figure 4.2.

So we see that, indeed, the monthly value does approach the daily value in the limit of daily hedging. (We were unable to run the monthly valuation for a resolution of one day because of memory constraints; there is an enormous number of spread options to compute in that case. Still, the pattern is clear.) In truth, the daily and monthly valuations do not seem to converge exactly; this is a reflection of the fact that the underlying monthly representation is actually a *lower* bound, but it becomes tighter as the spot mean-reversion rate decreases (or equivalently, as the volatility term structure becomes flatter).

**Table 4.1** Daily and monthly values. Values are in dollars

| | Option value | |
|---|---|---|
| Resolution | Monthly | Daily |
| 32 | 3,754,335 | 4,462,576 |
| 16 | 4,231,060 | 4,462,576 |
| 8 | 4,360,514 | 4,462,576 |
| 4 | 4,394,548 | 4,462,576 |
| 2 | 4,426,909 | 4,462,576 |

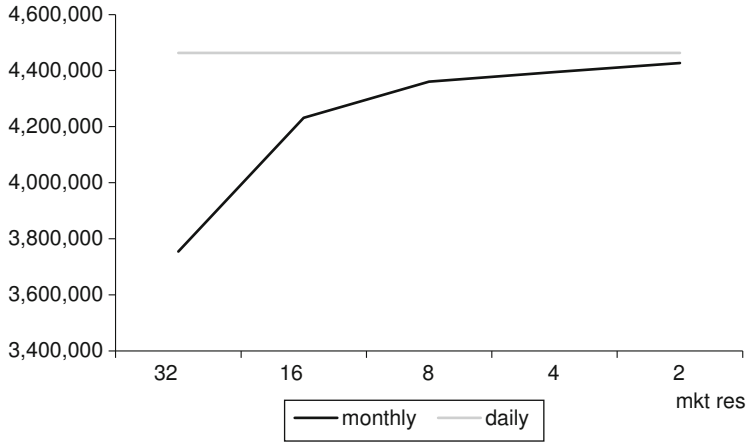



**Figure 4.2** Daily and monthly values. Note that as the resolution of the monthly hedge increases, the monthly value approaches the daily value

*4.1.0.3 Assessment*

Obviously, actually existing spot prices are considerably more complex than the simple model given by (4.1). And of course, as we saw in Chapter 2 (to be further confirmed in Chapter 6), even if prices *were* so well described, it is unlikely that we would be able to robustly estimate the underlying mean-reversion rates (and levels) from available data. However, as the simple example considered here reveals, we can expect to obtain reasonable lower bounds to the storage valuation problem through a suitable spread option representation in terms of forwards. The relevant inputs to these spread option valuation functionals are of course the ratio volatilities between pairs of injection and withdrawal months. These ratio volatilities of course manifest the underlying mean reversion through the term structure of volatility. In practice, the algorithms laid out in Section 3.1.3, for estimating realized QV, can provide a robust alternative to detailed spot- and control-based valuations (in conjunction of course with the [static] optimization algorithms provided in the Appendix to this chapter). Again, the idea is valuation through replication of good lower bounds.[2]

## 4.2 Tolling

Tolling deals were introduced in Section 1.2.1. Although the core embedded optionality in such deals is a spread option between electricity and fuel (typically natural gas, at least in the United States), there is an enormous amount of operational complexity that makes the valuation problem far more interesting (and challenging). In particular, even approximations to the problem are sufficiently computationally

taxing that simulation methods, of which we generally caution against overuse, are a necessary resort (see Section 7.5). These techniques are further hampered by the fact that simulation, by its essentially forward-looking nature, is ill equipped to handle control problems, whose solution generally requires backward-proceeding methods such as dynamic programming.

A common approach to this difficulty is so-called regression-based techniques for optimal control problems with simulation. These will be discussed in greater detail in Section 7.6. For the purposes here, one should simply recall the modern definition of conditional expected value as a projection onto a certain sigma-algebra (roughly speaking, the set of events generated by the entity being conditioned upon). This notion of projection has obvious affinities with regression. A further point that will be elaborated upon in Chapter 7 is that tolling problems generally have *multiple* exercise rights, *e.g.*, the right to start up only for onpeak periods and shut down during offpeak periods. The basic duality results introduced in Section 3.3 must then be extended, and we discuss in Chapter 7 important work due to Meinshausen and Hambly (2004).

After this (admittedly very brief) overview of the main computational issues,[3] we will now look more closely at tolling problems.

#### *4.2.0.1 Control problems: application to tolling*

While the Meinshausen-Hambly results are clearly an important extension of the basic upper-bound duality result, some modifications are required for control problems, as undertaken by Gyurkó *et al.* (2011). It proves more convenient to analyze the problem directly in terms of (state-dependent) value, as opposed to marginal value. We will illustrate with an example from tolling. (We will only focus on the essential features of the problem, and so ignore realistic aspects such as hourly dispatch capability, ramp-up effects, outages, emissions, heat rate curves, *etc.*) Although we have already encountered the basic structure of these deals, we will recap the central aspects here.

The essential structure of a tolling deal is a spread option between power revenue and fuel costs. There are in fact additional costs, both variable (proportional to generation level) and fixed (*e.g.*, start-up costs [including volumes of fuel] necessary to bring the unit to a minimum level of generation). In addition, part of the spread optionality includes the ability to generate incrementally above some minimum level of output. The basic form for the payoff from starting up on a given day[4] can be represented as

$$24\left(\begin{array}{l} C_{\min}(P - HR_{\min}G - VOM)+ \\ (C_{\max} - C_{\min})(P - HR_{\text{inc}}G - VOM)^+ \end{array}\right) - SC - F \cdot G \qquad (4.6)$$

where the notation is as follows:

- $P$, $G$: power and gas prices (resp.)
- $C_{\min}$, $C_{\max}$: generation levels at min and max capacity (resp.)
- $HR_{\min}$, $HR_{\max}$: unit heat rates at min and max capacity
- $HR_{\text{inc}}$: incremental heat rate, equal to $\frac{HR_{\max}C_{\max}-HR_{\min}C_{\min}}{C_{\max}-C_{\min}}$
- $VOM$: variable O&M
- $SC$: start-up cost
- $F$: start-up fuel amount

The presence of volume-independent costs and min-run operational levels means the optimal dispatch problem cannot be reduced to a series of (daily) spread options. For example, it may be optimal to run the plant during periods of negative price spreads so as to avoid start charges for periods where spreads are very high. We can craft the problem as an optimal control, normalizing by $24C_{\max}$ and using the notation $\alpha = C_{\min}/C_{\max}$, $X = SC + F \cdot G$,[5] and $Z = P - HR \cdot G - VOM$ (for notational convenience we ignore the distinction between heat rates at different operational levels). The problem is then stated as[6]

$$V_t(G_t, P_t, u_{t-1}) = \sup_{u_t,\ldots,u_T} E_t^Q \sum_{s=t}^{T} u_s(\alpha Z_s + (1-\alpha)Z_s^+ - (1-u_{s-1})X_s) \tag{4.7}$$

where dispatch occurs over days indexed by $t, \ldots, T$ and we have chosen an appropriate pricing measure $Q$. The control $u$ indicates the state of the unit: 0 means the unit is off, 1 means the unit is off. (We implicitly assume here that there are no limitations on start-ups/shut-downs.) The optimal control at a given time will depend on the power and gas prices at that time, as well as the state of the unit at the previous time. We can write this expression in the usual form amenable for backward induction:

$$\begin{aligned} &V_t(G_t, P_t, u_{t-1}) \\ &\quad = \max_{u_t(G_t,P_t,u_{t-1})} \{u_t(\alpha Z_t + (1-\alpha)Z_t^+ - (1-u_{t-1})X_t) + E_t^Q V_{t+1}(G, P, u_t)\} \\ &\quad = \max \left\{ \begin{array}{c} u_{t-1}(\alpha Z_t + (1-\alpha)Z_t^+) + E_t^Q V_{t+1}(G, P, u_{t-1}), \\ (1-u_{t-1})(\alpha Z_t + (1-\alpha)Z_t^+ - X_t) + E_t^Q V_{t+1}(G, P, 1-u_{t-1}) \end{array} \right\} \end{aligned} \tag{4.8}$$

#### 4.2.0.2 *State-dependent value proxies*

In the second equation in (4.8), we see that the optimal decision is between staying in the current state (*e.g.*, remaining on) and switching (*e.g.*, starting up). We now adapt here some of the duality results from Gyurkó *et al.* (2011). Introduce a family of state-dependent $Q$–martingales $M_t^{u_t}$. Then we have that

$$
\begin{aligned}
&V_t(G_t, P_t, u_{t-1}) \\
&= \sup_{u_t,\ldots,u_T} E_t^Q \left[ \sum_{s=t}^{T-1} (H(G_s, P_s; u_{s-1}, u_s) - M_{s+1}^{u_s} + M_s^{u_s}) + H(G_T, P_T; u_{T-1}, u_T) \right] \\
&\leq E_t^Q \max_{u_t,\ldots,u_T} \left[ \sum_{s=t}^{T-1} (H(G_s, P_s; u_{s-1}, u_s) - M_{s+1}^{u_s} + M_s^{u_s}) + H(G_T, P_T; u_{T-1}, u_T) \right]
\end{aligned} \tag{4.9}
$$

where

$$
H(G_T, P_T; u_{t-1}, u_t) = u_t(\alpha Z_t + (1-\alpha) Z_t^+ - (1-u_{t-1}) X_t) \tag{4.10}
$$

Thus we have

$$
V_t(G_t, P_t, u_{t-1}) \leq \inf_M E_t^Q \max_{u_t,\ldots,u_T} \left[ \begin{array}{l} \displaystyle\sum_{s=t}^{T-1} (H(G_s, P_s; u_{s-1}, u_s) - M_{s+1}^{u_s} + M_s^{u_s}) + \\ H(G_T, P_T; u_{T-1}, u_T) \end{array} \right] \tag{4.11}
$$

To see that the duality gap is again zero, consider the Doob-Meyer decomposition of the value function:

$$
\tilde{M}_{t+1}^{u_t} = \tilde{M}_t^{u_t} + V_{t+1}(G_{t+1}, P_{t+1}, u_t) - E_t^Q V_{t+1}(G, P, u_t) \tag{4.12}
$$

Using *this* martingale, we see that the infimum in (4.11) is bounded above by

$$
E_t^Q \max_{u_t,\ldots,u_T} \left[ \begin{array}{l} \displaystyle\sum_{s=t}^{T-1} (H(G_s, P_s; u_{s-1}, u_s) - V_{s+1}(G_{s+1}, P_{s+1}, u_s) + \\ E_s^Q V_{s+1}(G, P, u_s)) + H(G_T, P_T; u_{T-1}, u_T) \end{array} \right] \tag{4.13}
$$

Now, since

$$
V_t(G_t, P_t; u_{t-1}) \geq u_t(\alpha Z_t + (1-\alpha) Z_t^+ - (1-u_{t-1}) X_t) + E_t^Q V_{t+1}(G, P, u_t) \tag{4.14}
$$

we see that the infimum is further bounded by

$$
\begin{aligned}
&E_t^Q \max_{u_t,\ldots,u_T} \left[ \begin{array}{l} \displaystyle\sum_{s=t}^{T-1} (V_s(G_s, P_s, u_{s-1}) - V_{s+1}(G_{s+1}, P_{s+1}, u_s)) + \\ H(G_T, P_T; u_{T-1}, u_T) \end{array} \right] \\
&\quad = E_t^Q \max_{u_t,\ldots,u_T} [V_t(G_t, P_t, u_{t-1}) - V_T(G_T, P_T, u_{T-1}) \\
&\qquad + H(G_T, P_T; u_{T-1}, u_T)] \leq V_t(G_t, P_t, u_{t-1})
\end{aligned} \tag{4.15}
$$

where the last inequality in (4.15) follows from the fact that the value function always dominates any constituent payoff term. Thus we see that the duality gap is zero for this special choice of martingale, just as in the conventional American option case. So we again see a natural approach to getting a good upper bound: with a suitable lower-bound approximation, we can find the martingale component (via Doob-Meyer) then perform the above pathwise optimization (see also Andersen and Broadie [2004][7]).

#### *4.2.0.3 More tolling: lower and upper bounds*[8,9]

Having outlined how good upper bounds on the value function can be obtained, we now turn attention to some approximate (lower-bound) valuations that not only offer computational feasibility and provide a means for obtaining the desired upper bounds, but also allow decomposition of the tolling value in components that can be related to instruments that trade in reasonably liquid markets. These would certainly include forward prices and futures in most developed power and gas markets, but in many cases options (typically at-the-money [ATM], for both monthly and daily exercise).

To provide some context, we note that, as an option, the value from the physical tolling contract comes from the ability to reverse operational decisions and to commit to specific operational strategies based on current information. For example, we may decide to start the plant on a given day, and leave it on for the remainder of some period (say, a week). Or, we may decide to shut the plant down on a Friday night and leave it off for the weekend, starting back up on Monday. In general, as described above, this kind of Markov control problem will depend on the current state of the unit (including the number of state changes up to the current time) as well as the current value of prices. We can anticipate that this exercise surface will be quite complicated, but some intuitive, suboptimal control choices readily suggest themselves.

Assume we are at the start of a month. Then, we can elect to run the unit in one of the following three ways:

1. If economical, run at min capacity for the whole month, with the option to ramp up to max capacity each day.
2. If economical, run at min capacity each day, and shut down until the start of the next day; we have the option to ramp up to max capacity in the relevant temporal block (*e.g.* on-peak and off-peak) within each day.
3. If economical, run at max capacity for each temporal block of each day, and shut down at the end of that block.

Note that in modes 2 and 3, start-up costs are incurred each day. That is to say, price spreads must be sufficiently high to overcome variable and start-up costs. (Thus mode 3 is simply a collection of block [on-peak and off-peak] spread options, with strike price being VOM plus baked-in start charges.) In mode 1, there is flexibility

to remain on if the unit is already up (thus avoiding a single start charge that would effectively be prorated over the entire month). Based on price and volatility information at the start of the month in question, we can choose to operate in the mode yielding the highest value. In other words, the value is given by

$$\max(E_0^Q V_1, E_0^Q V_2, E_0^Q V_3) \tag{4.16}$$

where the constituent payoffs $V_{1,2,3}$ are given in the Appendix to this chapter.[10]

The expression in (4.16) obviously represents a lower-bound valuation, as it encompasses only a subset of possible operational flexibility. The central question then is: how good of a (lower) bound is it? Before addressing this issue, we note that we can always bind the value from above by considering perfect foresight operation across *paths*. This amounts to taking the (trivial) zero martingale in the duality expression (4.11). Our question can then be rephrased as determining where the true value is situated between these two bounds. For this we turn to the martingale duality techniques we have just discussed.

As an example, we perform an actual valuation of a specific power plant. Unit characteristics are shown in Table 4.2:

We do not strive for any particular realism for the prices and volatilities/correlations, but we take not unreasonable values for all months (so no seasonality, but forward vols are allowed to decline by 5% per month to mimic the Samuelson effect), as displayed in Table 4.3:

As with storage, in practice actual inputs would arise from the estimators outlined in Section 3.1. Depending on the particular market structure, correlations will either be implied from projected heat rate volatilities or directly estimated; the critical caveat is whether or not vega hedges can be put on (*i.e.*, whether or

**Table 4.2** Representative operational characteristics for tolling

| HR_min (MMBtu/MWh) | HR_max (MMBtu/MWh) | VOM ($/MWh) | SC ($) | F (MMBtu) | C_min (MW) | C_max (MW) |
|---|---|---|---|---|---|---|
| 10 | 8 | 2.5 | 10,000 | 1,000 | 50 | 100 |

**Table 4.3** Representative price and covariance data for tolling. Constant across contract months (Jul12–Jun15); no seasonal structure, but Samuelson effect for volatility term structure

| Fwd prices | | | Monthly vols | | | Monthly corrs | | | Cash vols | | | Cash corrs | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Onpk | Offpk | Gas | Onpk | Offpk | Gas | Onpk-offpk | Onpk-gas | Offpk-gas | Onpk | Offpk | Gas | Onpk-offpk | Onpk-gas | Offpk-gas |
| ($/MWh) | ($/MWh) | ($/MMBtu) | | | | | | | | | | | | |
| 50 | 35 | 7 | 0.3 | 0.3 | 0.4 | 0.6 | 0.8 | 0.8 | 1.8 | 1.7 | 0.5 | 0.5 | 0.6 | 0.6 |

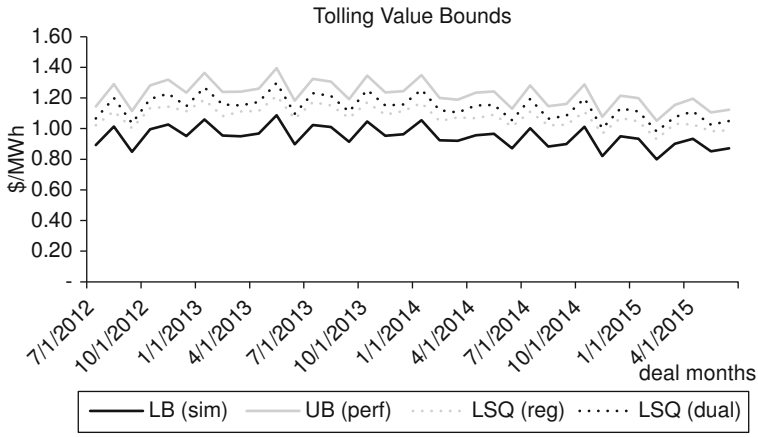



**Figure 4.3** Bounded tolling valuations. Comparison of three-strategy dispatch option (4.16) (lower bound), perfect foresight dispatch (upper bound), least-squares Monte Carlo, and martingale duality

not leg volatilities trade; see also the Appendix to Chapter 3). The three-year deal term begins one month from inception (we assume that the setting is the summer of 2012). The results from the upper- and lower-bound calculations, least-squares (LSQ) Monte Carlo (using basis functions of the form (7.199) and (7.200)), and the duality formulation (using the LSQ valuation as martingale proxy) are presented in Figure 4.3.[11]

We observe the following points. First, the tightness of the regression-based result and the corresponding duality result indicate that this particular set of basis functions is in fact a good choice: the duality gap is small. The obvious drawback of least-squares Monte Carlo is that the choice of regressors is typically arbitrary and we seldom know how good (or bad) they are; in conjunction with duality techniques we see we have a useful diagnostic. Second, we have a good idea of where the true value lies relative to the lower bound (three-strategy dispatch) and upper bound (perfect foresight/lookback dispatch). This is important because these two valuations are both simple and intuitive, and fairly straightforward to implement (as is the important calculation of the associated hedges).[12] However, in practice we rarely know how far off these bounds are from the true value, and it would be good to have a useful assessment (beyond the natural choice of taking the midpoint valuation).

Finally, we come to perhaps the most important point: we now have some idea of how much physical, operational detail is necessary for a given tolling problem. Put differently, we can assess the incremental cost of excluding certain operational characteristics. Roughly speaking, from Figure 4.3 we see that the lower bound captures about 85–90% of the true value. That is to say, these three basic operational strategies account for a fairly high percentage of value. It is probably possible to conceive of other operational modes that can be incorporated in a tractable

manner in payoffs such as (4.16). The question, as always, is whether it is worth doing so. We see here a manner of gauging the incremental value. We thus have here a means of bridging market-based, financial valuations with physical, spot-based valuations.[13]

## 4.3 Appendix

### 4.3.1 (Monthly) Spread option representation of storage

We introduce the following notation:

- $F$: forward prices
- $O$: option prices
- $N$: # months
- $V$: positions in spread options
- $U$: positions in max options
- $X$: positions in min options

Represent the portfolio payoff as

$$\sum_{i=1}^{N}\sum_{j=1}^{i}[V_{ij}(F_i - F_j)^+ + U_{ij}\max(F_i, F_j) - X_{ij}\min(F_i, F_j)] \tag{4.17}$$

In other words, sell spread options and max options and buy min options against the facility (for simplicity fuel and commodity costs have been omitted but it should be clear how to include them); note that diagonal elements in this ensemble ($i = j$) correspond to forward positions. We require that, regardless of how these options are exercised, no physical constraint will be violated. Introducing this (presumably clear) notation for facility constraints, this framework gives rise to the following optimization problem (a linear program):

$$\max \sum_{i=1}^{N}\sum_{j=1}^{i}[V_{ij}O_{ij}^{spread} + U_{ij}O_{ij}^{\max} - X_{ij}O_{ij}^{\min}]$$

s.t.

$$\sum_{j=i+1}^{N} V_{ji} + \sum_{j=i+1}^{N} X_{ji} + \sum_{j=1}^{i} X_{ij} \le MIQ_i$$

$$\sum_{j=1}^{i-1} V_{ij} + \sum_{j=i+1}^{N} U_{ji} + \sum_{j=1}^{i} U_{ij} \le MWQ_i$$

$$
\begin{aligned}
InvMin_i &\le C_0 + \sum_{k=1}^{i}\left[\sum_{j=k}^{N}(V_{jk}+X_{jk}) - \sum_{j=1}^{k}(V_{kj}+U_{kj})\right] \le InvMax_i \\
InvMin_i &\le C_0 + \sum_{k=1}^{i}\left[\sum_{j=1}^{k}X_{kj} - \sum_{j=k}^{N}U_{jk}\right] \le InvMax_i
\end{aligned}
\tag{4.18}
$$

That is, the optimal allocation of options are constrained such that, under the worst-case situation where they are all exercised, we have that

1. Injections from spread option sales and min option purchases do not exceed maximum monthly injection quantities.
2. Withdrawals from spread-option sales and max-option sales do not exceed maximum monthly withdrawal quantities.
3. Cumulative constraints (inventory requirements less initial capacity) are not violated, in two senses:
   a) Total inflows against subsequent months minus total outflows from preceding months
   b) Total inflows from preceding months minus total outflows against subsequent months (spread options cannot enter into this balance).

### 4.3.2 Lower-bound tolling payoffs

Refer back to (4.16) and the associated discussion. The payoffs for the three basic operational strategies (min-run/ramp to max, daily start-up/shut-down, and block start-up/shut) are given (resp.) by

$$
\begin{aligned}
V_1 &= \beta(\alpha(P_{on} - HR_{\min}G - VOM) + (1-\alpha)(P_{on} - HR_{inc}G - VOM)^+) \\
&\quad + (1-\beta)(\alpha(P_{off} - HR_{\min}G - VOM) + (1-\alpha)(P_{off} - HR_{inc}G - VOM)^+) \\
V_2 &= \begin{pmatrix} \alpha(P_{day} - HR_{\min}G - VOM) - SC - F.G \\ (1-\alpha)\begin{pmatrix} \frac{2}{3}(P_{on} - HR_{inc}G - VOM)^+ + \\ \frac{1}{3}(P_{off} - HR_{inc}G - VOM)^+ \end{pmatrix} \end{pmatrix} \\
&\quad + (1-\beta)(P_{off} - HR_{\max}G - VOM - SC - F.G)^+ \\
V_3 &= \beta\begin{pmatrix} \left(\frac{2}{3}(P_{on} - HR_{\max}G - VOM) - SC - F.G\right)^+ \\ \left(\frac{1}{3}(P_{off} - HR_{\max}G - VOM) - SC - F.G\right)^+ \end{pmatrix} \\
&\quad + (1-\beta)(P_{off} - HR_{\max}G - VOM - SC - F.G)^+
\end{aligned}
\tag{4.19}
$$

where $\beta$ is the ratio of on-peak to off-peak days, and we assume the usual breakout in U.S. markets of 16 on-peak hours and 8 off-peak hours on a peak day. Note

that the daily start and block start options have a common component for off-peak days. (For convenience we assume the unit was already on at the start of the period in question, so there is no need to include one-day start charges in the min-run strategy.) It should be clear how fuel switching can be incorporated in this framework.

# 5 Analytical Techniques

## 5.1 Change of measure techniques

We will now discuss in detail a powerful set of techniques based on changes of probability measure. These techniques are extremely useful for reducing the computational complexity of pricing a wide variety of structured products, as well as identifying and extracting the essential features of many valuation problems. By this latter point we mean not simply numerical aspects (important as they are), but reducing a problem so that those statistics sufficient for establishing robust valuations (*i.e.*, value drivers) can be obtained.

We have already seen two examples of these techniques. The first is the standard Black-Scholes framework, where the replication cost of an option is given by the conditional expectation of the terminal payoff under a probability measure under which the price of the underlying asset is a martingale. The second is the Margrabe formula for pricing a spread option, where the option price, denominated in terms of one of the underlying leg assets, is also given by a conditional expectation of the redenominated payoff. But this redenominated asset (the ratio of the two legs) depends *only* on a specific combination of the constituent volatilities and correlation (namely, the ratio volatility). Let us provide a bit of overview.

### 5.1.1 Review/main ideas

Although our purpose here is not to provide a complete treatment of the relevant, technical concepts from measure theoretic probability, it is appropriate to provide a quick outline and definitions.[1] Two probability measures $P$ and $Q$ are said to be equivalent if they share a common sample space $\Omega$ and a common sigma algebra $\mathfrak{F}$ of events,[2] and the following holds for any event $A$:

$$P(A) > 0 \Leftrightarrow Q(A) > 0$$

In other words, events that are possible (impossible) under $P$ are possible (impossible) $Q$. (That is, they define the same null sets.) The so-called Radon-Nikodym

theorem states that there is a measureable function $\xi$ (that is, measureable with respect to the common sigma algebra of the two probability measures) such that

$$Q(A) = E^p \xi 1_A \tag{5.1}$$

where $1_A$ denotes the indicator function with respect to the event $A$. The function $\xi$ is commonly referred to as the Radon-Nikodym (RN) derivative and is usually written as

$$\frac{dQ}{dP} = \xi \tag{5.2}$$

By taking the measure of the universal event we immediately see that any candidate RN derivative must satisfy $E^p \xi = 1$, which in the context of stochastic processes amounts to the requirement that the RN derivative, conceived of as a process, must be a $P-$martingale.

#### *5.1.1.1 A concrete example*

To dispense with some of the abstraction, it is worth stressing that many classes of processes that we are interested in will retain their basic structural form under the kinds of measure changes that are important for valuation purposes. To fix matters, we will here state a collection of basic foundational results for Gaussian processes, which will be derived in the course of this section. The central message is the manner in which means and covariances change under specific measure changes, which can broadly be thought of in terms of re-denominating the basic units in which some market are reckoned.

Assume we have $n$ assets which have the following dynamics under a measure $Q$:

$$dz_i = -\tfrac{1}{2}\sigma_i^2 dt + \sigma_i dw_i$$

where $dw_i dw_j = \rho_{ij} dt$. Assume further we have a contingent claim with payoff of the form

$$Y_T = e^{z_k(T)} X_k(e^{z_i(T)})$$

where $X_k$ is a multilinear form. Then for the following changes of measure

$$\frac{dQ_p}{dQ} = e^{z_p(T) - z_p}$$

we have the following results:

(i). $z_k(T) \sim N(z_k + (\rho_{pk}\sigma_p\sigma_k - \frac{1}{2}\sigma_k^2)\tau, \sigma_k)$ under $Q_p$ where $\tau = T - t$, or in Stochastic Differential Equation (SDE) form, $dz_k = (\rho_{pk}\sigma_p\sigma_k - \frac{1}{2}\sigma_k^2)dt + \sigma_k dw_k$

(ii). $z_k(T) - z_1(T) \sim N_{n-1}(z_k - z_1 - \frac{1}{2}(\sigma_1^2 - 2\rho_{1k}\sigma_1\sigma_k + \sigma_k^2)\tau, \Sigma_{ij}\tau)$ under $Q_1$, where $N_{n-1}$ denotes the multidimensional normal cumulative distribution function (CDF) and the covariance matrix is given by $\Sigma_{ij} = \sigma_1^2 - \rho_{1i}\sigma_1\sigma_i - \rho_{1j}\sigma_1\sigma_j + \rho_{ij}\sigma_i\sigma_j$; in SDE form $d\tilde{z}_k = -\frac{1}{2}\Sigma_{kk}dt + \sqrt{\Sigma_{kk}}d\tilde{w}_k$ where $\tilde{z}_k \equiv z_k - z_1$ and $d\tilde{w}_k$ is a standard Brownian motion with $d\tilde{w}_i d\tilde{w}_j = \Sigma_{ij}dt/\sqrt{\Sigma_{ii}\Sigma_{jj}}$

(iii). $E_t^Q Y_T = e^{z_k} E_t^{Q_k} X_k(e^{z_i(T)-z_k(T)})$

(iv). For the special case where $X_k$ can be expressed in the form $\mathbb{I}(A_{ij}^k z_j \leq b_i^k)$, the expectations in part (iii) can be evaluated in terms of multidimensional normal CDFs using the fact that normality is retained under linear transformations and the known results for the means and covariances from the change of measure results in part (ii).

#### *5.1.1.2 Girsanov and beyond*

Technicalities aside, computations involving change of measure are greatly facilitated (in a very wide range of applications/models) via characteristic functions. (We will see the central role characteristic functions play in the construction of the canonical class of Lévy process models in Section 5.2.1) In fact, the utility of the characteristic function formulation goes well beyond this theoretical elegance. This point can be seen in the derivation of the well-known Girsanov transformation, which plays a central role in standard BS option pricing theory. Suppose that, under the physical (*i.e.*, observed) measure $P$, log prices are given by GBM with drift:

$$dz = (\mu - \sigma^2/2)dt + \sigma\, dw \tag{5.3}$$

Now, if we seek a pricing measure $Q$ under which prices are martingales (from which meaningful [in whatever sense] option prices, say, can be derived), we must have[3]

$$E_t^Q e^{z_T} = e^z \tag{5.4}$$

Now, the RN process must be a $P-$martingale, so must take the form $\frac{dQ}{dP} = \exp(-\frac{1}{2}\alpha^2\tau + \alpha w_\tau)$ for some $\alpha$, where $\tau \equiv T - t$. Thus we have

$$E_t^Q e^{z_T} = e^{z-\frac{1}{2}\alpha^2\tau+(\mu-\sigma^2/2)\tau} E_t^P e^{(\alpha+\sigma)(w_T-w)} = e^{z+(\mu+\alpha\sigma)\tau} \tag{5.5}$$

using standard results for the characteristic function of a normal random variable.[4] Thus, the condition that $z$ be a $Q$–martingale becomes

$$\alpha = -\frac{\mu}{\sigma} \tag{5.6}$$

which is of course the familiar Girsanov result for change of measure for GBM (as well as illustrating the economic intuition of the Sharpe ratio for weighting preferences in the appropriate portfolio).

This standard textbook derivation of Girsanov can be augmented by the following example, which is quite simple but also serves to introduce some underlying subtleties. A basic intuition that must be grasped is that a measure change essentially changes the (probabilistic) weightings given to the paths of a process; a measure change *cannot* change the fundamental nature of the process itself.[5] We saw in the context of Girsanov that a measure change served to change the drift of the underlying process, but not its diffusive structure. This point holds true in general. So, for example, a jump process will remain some kind of jump process under a measure change, and if the process is suitably simple, should retain some basic features (such as remaining a Poisson process, in particular). To see this, consider the logarithm $z$ of a price process (under the physical measure $P$) with only linear drift and jumps (*i.e.*, no diffusion term):

$$dz = (\mu - \lambda k)dt + jdq \tag{5.7}$$

where the jump intensity is denoted (as usual) by $\lambda$ and where $k = E(e^j - 1)$. To (helpfully) fix matters we suppose that the jump amplitude (given by $j$) is binomially distributed: the log process shifts up by $u$ with probability $p$ and shifts down by $d$ with probability $1 - p$ (so that the price process goes up to $e^u$ or down to $e^d$); in between jumps the process simply trends linearly. Again, we seek a $P$-martingale process induces a measure change $Q$ under which the price process is a $Q$-martingale. A suitable candidate $\xi$ must have the following $P$-dynamics:

$$d\xi = -\lambda k' dt + j' dq \tag{5.8}$$

for some new jump amplitude $j'$ and where $k' \equiv E(e^{j'} - 1) = P(e^{u'} - 1) + (1 - p)(e^{d'} - 1)$ for some new up/down moves $u'$ and $d'$. Using basic results of Itô calculus for jumps and (5.2), we have that

$$E_t^Q e^{z_T} = e^{-\xi} E_t^P e^{z_T + \xi_T} = \exp(z + (\mu - \lambda k - \lambda k' + \lambda E(e^{j+j'} - 1))(T - t)) \tag{5.9}$$

The terms in the exponent proportional to the time horizon in (5.9) can be written as

$$\begin{aligned} &\mu - \lambda k - \lambda k' + \lambda E(e^j e^{j'} - e^{j'} + e^{j'} - 1) \\ &\quad = \mu - \lambda k + \lambda E e^{j'}(e^j - 1) = \mu - \lambda k + \lambda' E'(e^j - 1) \end{aligned} \tag{5.10}$$

where $\lambda' = \lambda E e^{j'}$ and the prime on the expectation operator in (5.10) denotes *another* measure change induced by the RN derivative $e^{j'}/Ee^{j'}$. Now, the first thing to note is that the requirement that the price process becomes a martingale under the measure change *cannot* be uniquely determined; from (5.10) we only require that the jump amplitude of the RN process satisfy $\mu - \lambda k + \lambda' E'(e^j - 1) = 0$. This

reflects the fact that, outside of the realm of complete, Gaussian markets, the pricing measure exists but in general is nonunique (the so-called second fundamental theorem of asset pricing). Also, although we have established the necessary martingale conditions, we have not yet said anything about the dynamics of the price process under the new measure, as in the Girsanov case. For this, we appeal to the intuition that the process will retain its (pure) jump form and, using basic results for the arrival times of a Poisson process, we see that

$$\begin{aligned} E_t^Q 1(\tau_1 > T) &= e^{-\xi} E_t^P e^{\xi_T} 1(\tau_1 > T) \\ &= e^{-\xi+\xi-\lambda k'(T-t)-\lambda(T-t)} = \exp(-\lambda E e^{j'}(T-t)) = e^{-\lambda'(T-t)} \end{aligned} \tag{5.11}$$

where $\tau_1$ denotes the arrival time of the first jump. (5.11) demonstrates that under the new measure, the jumps retain their Poisson structure with new intensity $\lambda'$.[6] For further development of these points, see the excellent articles by Benninga *et al.* (2002) and Schroder (1999).[7]

With the relevant context established, we can now turn to some actual examples.

### 5.1.2 Dimension reduction/computation facilitation/estimation robustness

#### *5.1.2.1 Basic spread structures*

First we consider a simple (but very common) structure in energy markets, a spread option. The terminal payoff is given by

$$(S_2(T) - \alpha S_1(T))^+ \tag{5.12}$$

for two assets $S_1$ and $S_2$ and some weighting factor $\alpha$.[8] We are concerned with valuing expectations of the form

$$E_t^Q (S_2(T) - \alpha S_1(T))^+ \tag{5.13}$$

under an appropriate (pricing) measure $Q$. Now, (5.13) can be written as

$$E_t^Q (S_2(T) - \alpha S_1(T))^+ = E_t^Q (S_1(T) \left( \frac{S_2(T)}{S_1(T)} - \alpha \right)^+ = E_t^Q S_1(T) (\tilde{S}(T) - \alpha)^+ \tag{5.14}$$

with $\tilde{S}(T) \equiv \frac{S_2(T)}{S_1(T)}$. Under the following change of measure (recall that prices of tradeables are martingales under the pricing measure):[9]

$$\frac{dQ_1}{dQ} \equiv \frac{S_1(T)}{S_1} \tag{5.15}$$

the expectation in question can be written as

$$S_1 E_t^{Q_1}(\tilde{S}(T)-\alpha)^+ \tag{5.16}$$

What has been done here essentially is to change the basic underlying units in which prices are reckoned; that is, we are taking asset 1 to be the numeraire. Denominated in these units, the underlying option price becomes simply a $Q_1-$expectation. It is important to stress that this change of measure/change of numeraire does not change the basic interpretation of (redenominated) prices as martingales, *viz.*

$$E_t^{Q_1}\tilde{S}(T) = E_t^{Q}\frac{S_1(T)}{S_1}\frac{S_2(T)}{S_1(T)} = \frac{1}{S_1}E_t^{Q}S_2(T) = \tilde{S} \tag{5.17}$$

We immediately see that the dimensionality of the problem has been reduced from two to one. The computational benefits should be obvious, but more importantly the essential features driving the value of the option are more readily extracted and delineated. However, all of this depends on actually being able to evaluate the expectation under this new measure, so we will consider this issue with some specific assumptions about the underlying process. It turns out the measure change is feasible under a rather rich class of processes, but for now we will focus on the usual lognormal case.[10]

*5.1.2.2 Martingale dynamics*

First, it is worth starting with a generalization of a more basic issue already discussed, namely the change from the observable, physical measure (denoted by $P$) to the pricing measure ($Q$). Assume log prices $z$ follow joint Brownian motion with drift:

$$\begin{aligned} dz_1 &= \left(\mu_1 - \tfrac{1}{2}\sigma_1^2\right)dt + \sigma_1 dw_1 \\ dz_2 &= \left(\mu_2 - \tfrac{1}{2}\sigma_2^2\right)dt + \sigma_2 dw_2 \end{aligned} \tag{5.18}$$

with (in terms of the usual Itô isometry) $dw_1 dw_2 = \rho dt$. We seek an RN process such that, under the associated change of measure, the log prices are exponential martingales (which amounts to the condition that the linear drifts $\mu_i$ are absent). We assume this process $\xi$ has the following $P-$dynamics:

$$d\xi = -\tfrac{1}{2}(\alpha_1^2 + 2\rho\alpha_1\alpha_2 + \alpha_2^2)dt + \alpha_1 dw_1 + \alpha_2 dw_2 \tag{5.19}$$

Let us again highlight a useful application of (conditional) characteristic functions, by more generally determining the dynamics of the log prices under this new measure. Let

$$f = E_t^Q e^{i\phi_1 z_1(T) + i\phi_2 z_2(T)} = e^{-\xi}E_t^P e^{\xi(T) + i\phi_1 z_1(T) + i\phi_2 z_2(T)} \tag{5.20}$$

Thus, the entity $\tilde{f} \equiv e^{\xi} f$ is a $P-$martingale, and thus from Itô's lemma[11] must satisfy the following PDE:

$$\begin{aligned}
&\tilde{f}_t + (\mu_1 - \tfrac{1}{2}\sigma_1^2)\tilde{f}_{z_1} + (\mu_2 - \tfrac{1}{2}\sigma_2^2)\tilde{f}_{z_2} - \tfrac{1}{2}(\alpha_1^2 + 2\rho\alpha_1\alpha_2 + \alpha_2^2)\tilde{f}_{\xi} \\
&\quad + \tfrac{1}{2}\sigma_1^2\tilde{f}_{z_1 z_1} + \rho\sigma_1\sigma_2\tilde{f}_{z_1 z_2} + \tfrac{1}{2}\sigma_2^2\tilde{f}_{z_2 z_2} \\
&\quad + \tfrac{1}{2}(\alpha_1^2 + 2\rho\alpha_1\alpha_2 + \alpha_2^2)\tilde{f}_{\xi\xi} + \sigma_1(\alpha_1 + \rho\alpha_2)\tilde{f}_{z_1\xi} + \sigma_2(\alpha_2 + \rho\alpha_1)\tilde{f}_{z_2\xi} = 0
\end{aligned} \tag{5.21}$$

Now, substituting for $\tilde{f}$, we find that[12]

$$\begin{aligned}
&f_t + \left(\mu_1 + \sigma_1(\alpha_1 + \rho\alpha_2) - \tfrac{1}{2}\sigma_1^2\right) f_{z_1} + \left(\mu_2 + \sigma_2(\alpha_2 + \rho\alpha_1) - \tfrac{1}{2}\sigma_2^2\right) f_{z_2} \\
&\quad + \tfrac{1}{2}\sigma_1^2 f_{z_1 z_1} + \rho\sigma_1\sigma_2 f_{z_1 z_2} + \tfrac{1}{2}\sigma_2^2 f_{z_2 z_2} = 0
\end{aligned} \tag{5.22}$$

Consequently, the $Q-$dynamics of the log prices are given by

$$\begin{aligned}
dz_1 &= \left(\mu_1 + \sigma_1(\alpha_1 + \rho\alpha_2) - \tfrac{1}{2}\sigma_1^2\right) dt + \sigma_1 dw_1 \\
dz_2 &= \left(\mu_2 + \sigma_2(\alpha_2 + \rho\alpha_1) - \tfrac{1}{2}\sigma_2^2\right) dt + \sigma_2 dw_2
\end{aligned} \tag{5.23}$$

Note the confirmation of the general intuition that under a measure change for standard BM, the covariance structure will be unchanged. We also see the conditions that must hold for the measure change to take prices into martingales:

$$\begin{aligned}
\mu_1 + \sigma_1(\alpha_1 + \rho\alpha_2) &= 0 \\
\mu_2 + \sigma_2(\alpha_2 + \rho\alpha_1) &= 0
\end{aligned} \tag{5.24}$$

We see that for uncorrelated assets the appropriate transformation is a pair of separate one-dimensional Girsanov transformations. This linear system can be easily solved, and we will see a number of additional applications later (*e.g.*, importance sampling for simulations, discussed in Section 7.5). Note in particular that in this case the martingale measure is unique.

#### *5.1.2.3 Numeraires and greeks*

Having established the relation between the physical and pricing measures, we can now turn attention to the measure change associated with the change of numeraire in (5.15). Characteristic functions again prove a very fruitful approach to some of the computational issues involved in the problem. Ultimately (from (5.16)) we are concerned with the distribution of the ratio $\tilde{S}$ under the measure $Q_1$. By the Fourier inversion theorem, the characteristic function will provide us with the necessary information. So consider (in terms of log-differences instead of price ratios)

$$f = E_t^{Q_1} e^{i\phi(z_2(T) - z_1(T))} = e^{-z_1} E_t^{Q} e^{(1-i\phi)z_1(T) + i\phi z_2(T)} \tag{5.25}$$

which in terms of standard results becomes

$$f = e^{-z_1+(1-i\phi)(z_1-\frac{1}{2}\sigma_1^2\tau)+\frac{1}{2}(1-2i\phi-\phi^2)\sigma\phi_1^2\tau+i\phi(z_2-\frac{1}{2}\sigma_2^2\tau)-\frac{1}{2}\phi^2\sigma_2^2\tau+\rho\sigma_1\sigma_2\tau\phi(i+\phi^2)}$$
$$= e^{i\phi((z_2-z_1)-\frac{1}{2}(\sigma_1^2-2\rho\sigma_1\sigma_2+\sigma_2^2)\tau)-\frac{1}{2}\phi^2(\sigma_1^2-2\rho\sigma_1\sigma_2+\sigma_2^2)\tau} \tag{5.26}$$

Thus, under the measure with leg 1 as numeraire, the price ratio (exponential of the log-price difference) is a martingale, and lognormally distributed with volatility equal to the ratio volatility $\sigma = \sqrt{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}$ (recall the discussion of storage spread options in Section 3.1.4). Consequently, the expectation in (5.16) gives an option price of

$$S_1 BS(\tilde{S}, \alpha, \sigma) \tag{5.27}$$

which is precisely the Margrabe formula previously encountered in Chapter 3 (and derived alternatively in Chapter 7).

Although this is a simple example, it should suffice to show how change of measure techniques can greatly facilitate the derivation and computation of many option pricing problems. A very general situation is to evaluate expectations of payoffs of the following form: $F(S_1, S_2, \ldots, S_N) = S_1 F(1, \frac{S_2}{S_1}, \ldots, \frac{S_N}{S_1})$. In many cases of interest the distribution of the reduced set of assets (redenominated using asset 1 as numeraire) can be determined under the measure change in (5.15), reducing the computational complexity by one dimension. More than this: the sufficient statistics needed for robust valuation (*i.e.*, the identification of value drivers) are often made structurally clear from this reduced form. The ratio volatility is a case in point. In many situations (and the structure of the underlying market, *e.g.*, the breadth of liquidity, is the crucial factor) it is *not* necessary to estimate and project separately leg volatilities and correlations. The ratio volatility is all that is needed (a fortuitous situation, as volatilities typically allow for more robust estimation than correlations).

Measure changes also offer benefits in the determination of option greeks, which are often produced naturally in the computation of option value. Note that the standard pricing formula can be rewritten as follows:

$$\begin{aligned}
&E_t^Q(S_2(T) - S_1(T))^+ \\
&= E_t^Q S_2(T) 1(S_2(T) > S_1(T)) - E_t^Q S_1(T) 1(S_2(T) S_1(T)) \\
&= S_2 E_t^{Q_2} 1(S_2(T) > S_1(T)) - S_1 E_t^{Q_1} 1(S_2(T) > S_1(T)) \\
&= S_2 E_t^{Q_2} 1(\tilde{S}(T) > 1) - S_1 E_t^{Q_1} 1(\tilde{S}(T) > 1)
\end{aligned} \tag{5.28}$$

so that greeks are simply read off from the coefficients of the initial asset prices (via Euler's theorem for homogenous functions of degree one). We also see here that the

problem amounts to computations of binary options under the appropriate measure, and that these binaries are of one less dimension than the original problem. This trick can be generalized to a wide range of problems, as we will now see.

### 5.1.3 Max/min options

As an application, consider an option on the maximum of two assets. The terminal payoff is given by

$$(\max(S_1, S_2) - K)^+ \tag{5.29}$$

Similar to the decomposition used in (5.28), the pricing problem can be written as

$$\begin{aligned}
&E_t^Q(\max(S_1, S_2) - K)^+ \\
&\quad = E_t^Q(S_1(T) - K)1(S_1(T) > S_2(T), S_1(T) > K) \\
&\qquad + E_t^Q(S_2(T) - K)1(S_2(T) > S_1(T), S_2(T) > K) \\
&\quad = S_1 E_t^{Q_1} 1(S_1(T) > S_2(T), S_1(T) > K) + S_2 E_t^{Q_2} 1(S_2(T) > S_1(T), S_2(T) > K) \\
&\qquad - K(E_t^Q 1(S_1(T) > S_2(T), S_1(T) > K) + E_t^Q 1(S_2(T) > S_1(T), S_2(T) > K))
\end{aligned} \tag{5.30}$$

Thus, in terms of log prices, the option valuation problem becomes one of calculating expectations of the form

$$E_t^{Q_i} 1(z_i(T) - z_j(T) < a, z_i(T) < b) \tag{5.31}$$

We discuss various methods for computing such expectations later. First, we note a few more general options with payoffs of similar form that commonly appear in energy markets.

- Fuel switching in tolling (option to choose cheaper of two fuels with which to generate electricity): $(S_3 - \min(S_1, S_2))^+$
- Pipeline segmentation in transport (option to flow gas to the higher priced of two delivery points): $(\max(S_2, S_3) - S_1)^+$
- Multiple injection/withdrawal points in storage: $(\max(S_3, S_4) - \min(S_1, S_2))^+$.

For example, the fuel-switching case will have constituent terms such as

$$E_t^Q S_3(T) 1(S_3(T) > S_2(T), S_2(T) < S_1(T)) \tag{5.32}$$

which, in terms of the numeraire $S_3$, becomes

$$S_3 E_t^{Q_3} 1(\tilde{S}(T) < 1, \tilde{S}_2(T) < \tilde{S}_1(T)) \tag{5.33}$$

which is precisely of the form (5.31). In general, it proves useful to work in terms of log prices, and the category of max/min options can be reduced to calculating expectations of the form

$$E_t^Q e^{\alpha_i z_i(T)} 1(A_{kj} z_j(T) < b_k) \tag{5.34}$$

where the summation convention of repeated indices is adopted. We now discuss in detail how to carry out these kinds of calculations.

### 5.1.4 Quintessential option pricing formula

*5.1.4.1 The basic setting*

Valuation of options with payoff structures such as those in (5.34) was studied in Skipper and Buchen (2003). The idea is to use change of measure techniques as follows. We have that

$$\begin{aligned} E_t^Q e^{\alpha_i z_i(T)} 1(A_{kj} z_j(T) < b_k) &= E_t^Q e^{\alpha_i z_i(T)} \cdot E_t^{Q_\alpha} 1(A_{kj} z_j(T) < b_k) \\ &= E_t^Q e^{\alpha_i z_i(T)} \cdot E_t^{Q_\alpha} 1(\tilde{z}_k(T) < b_k) \end{aligned} \tag{5.35}$$

where we define $\tilde{z}_k \equiv A_{kj} z_j$ and the new measure $Q_\alpha$ by

$$\frac{dQ_\alpha}{dQ} = \frac{e^{\alpha_i z_i(T)}}{E_t^Q e^{\alpha_i z_i(T)}} \tag{5.36}$$

So the first issue is to determine the distribution of $\tilde{z}_k$ under this measure. We again turn to the characteristic function as a means of carrying out the calculations that are needed to flesh out the essential results. We have that

$$f = E_t^{Q_\alpha} e^{i\phi_k \tilde{z}_k(T)} = \frac{E_t^Q e^{(\alpha_i + i\phi_k A_{ki}) z_i(T)}}{E_t^Q e^{\alpha_i z_i(T)}} \tag{5.37}$$

Note that, to this point, we have said nothing about the distribution of $z$ under the pricing measure (or any measure for that matter). The result in (5.37) is thus quite general, and can be applied whenever we know the characteristic function of the primary random variables in question. As we have mentioned (and will pursue in great detail later), there is a very wide and rich class of processes for which the characteristic function is known. We are referring of course to Lévy processes, and, armed with the connection of stochastic volatility models to time-changed Brownian motions, we will devote much attention to the canonical class of affine jump diffusions (which are not themselves of Lévy type). The common thread in all of this is the facilitation of analysis through the (conditional) characteristic function.

*5.1.4.2 The standard Gaussian case*

An obvious choice to consider is the case of joint normality. In this case the log prices satisfy

$$dz_i = -\tfrac{1}{2}\sigma_i^2 dt + \sigma_i dw_i \tag{5.38}$$

with $dw_i dw_j = \rho_{ij} dt$. Now we apply the same line of reasoning used in the derivation of (5.24). From (5.37) we have that $\tilde{f} \equiv f \cdot E_t^Q e^{\alpha_i z_i(T)}$ is a $Q-$martingale and so satisfies the following PDE:

$$\tilde{f}_t - \tfrac{1}{2}\sigma_1^2 \tilde{f}_{z_i} + \tfrac{1}{2} X_{ij} \tilde{f}_{z_i z_j} = 0 \tag{5.39}$$

where $X_{ij} \equiv \rho_{ij}\sigma_i\sigma_j$. Note also that

$$E_t^Q e^{\alpha_i z_i(T)} = e^{\alpha_i (z_i - \frac{1}{2}\sigma_i^2 \tau) + \frac{1}{2} X_{ij}\alpha_i\alpha_j \tau} \tag{5.40}$$

Substituting the expression for $\tilde{f}$ yields

$$f_t + (-\tfrac{1}{2}\sigma_i^2 + X_{ij}\alpha_j) f_{z_i} + \tfrac{1}{2} X_{ij} f_{z_i z_j} = 0 \tag{5.41}$$

Thus, the effect of the measure change on the asset dynamics is to the leave the covariance structure unchanged (not surprising) and to shift the (vector) mean by the vector $X\alpha$. This is of course just a generalized Girsanov transformation, with (5.24) being a special case. So we further find that

$$\begin{aligned} &E_t^{Q_\alpha} e^{i\phi_k \tilde{z}_k(T)} \\ &\quad = E_t^{Q_\alpha} e^{i\phi_k A_{kj} z_j(T)} = e^{i\phi_k A_{kj}(z_j - \frac{1}{2}\sigma_j^2 \tau + X_{jm}\alpha_m \tau) - \frac{1}{2} A_{kj} X_{jn} A_{mn} \phi_k \phi_m \tau} \end{aligned} \tag{5.42}$$

and thus obtain the $Q_\alpha-$mean and covariance of $\tilde{z}$.[13] As an example, consider the case of a numeraire change, say in terms of asset 1. Then $\alpha_i = \delta_{i1}$ and $A$ is an $N-1$ by $N$ matrix of the form

$$A_{ij} = \begin{cases} -1, & j = 1 \\ 1, & j = i+1 \\ 0 & \text{otherwise} \end{cases} \tag{5.43}$$

Thus, the $Q_\alpha-$drift and covariance (resp.) are given by

$$\begin{aligned} \tilde{\mu} &= z_k - z_1 - \tfrac{1}{2}(\sigma_1^2 - 2\rho_{k1}\sigma_1\sigma_k + \sigma_k^2)\tau \\ \tilde{\Sigma}_{mn} &= (\sigma_1^2 - \rho_{k1}\sigma_1\sigma_k - \rho_{m1}\sigma_1\sigma_m + \rho_{km}\sigma_m\sigma_k)\tau \end{aligned} \tag{5.44}$$

(Compare with the results in Carmona and Durrleman [2005].) Of course, (5.44) illustrates once again that a change of measure that takes the form of a change of numeraire retains the martingale property of (redenominated) tradeables. Note again the reappearance of the ratio vol in this valuation (of course there is an associated correlation structure for more than two assets).

Consequently, in the joint normal case, we see that payoffs that can be decomposed into constituent terms such as (5.35) are reduced to the evaluation of expectations of the form

$$E1(\tilde{z} < b) = N_n(b; \tilde{\mu}, \tilde{\Sigma}) \tag{5.45}$$

where $N_n$ is the $n-$dimensional cumulative normal distribution function:

$$N_n(b; \tilde{\mu}, \tilde{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det \tilde{\Sigma}}} \int_{-\infty}^{b} dz \exp\left(-\tfrac{1}{2}(z-\tilde{\mu})^T \tilde{\Sigma}^{-1}(z-\tilde{\mu})\right) \tag{5.46}$$

Now, in truth this may not seem like any great triumph, as the evaluation of multidimensional integrals is in general quite challenging numerically. We will consider some techniques for addressing this problem in Chapter 7, but in practice five dimensions is about as high as one can get without having to resort to simulation methods. Still, there is little doubt that the formal approach stemming from (5.35) is quite nice and in fact quite useful in many applications. Finally, we stress yet again that the applicability extends beyond the usual case of joint normality.

Change of measure techniques are also very useful for obtaining certain symmetry results pertaining to a popular structure in energy markets, namely options struck against some average price (*i.e.*, Asian options), as we will now see.

### 5.1.5 Symmetry results: Asian options

Because of the large volatilities commonly encountered in energy markets (see EW for a useful accounting of the issue), it is often desirable to structure products around the average of an asset price during some period (be it a month, a year, or longer), in order to "smooth out" price spikes and other regime shifts that characterize such markets. The instrument of choice here is the so-called Asian option, whose payoff is the average price over some period against some strike. Such a payoff can either be of fixed strike variety, where the payoff is given by

$$\left(\frac{1}{T-t}\int_t^T S_S ds - K\right)^+ \tag{5.47}$$

or floating strike variety, with payoff

$$\left(k \cdot S_T - \frac{1}{T-t}\int_t^T S_s ds\right)^+ \tag{5.48}$$

These options are not terribly easy to value, even in the standard case of GBM. The seminal paper on the topic is Geman and Yor (1993), who, using rather obscure results for Bessel processes, are able to compute the Laplace transform of the option price for the fixed strike case, which can then be inverted.[14] Note that the above payoffs refer to continuously averaged prices; in general the payoff will be on the discrete (*e.g.*, daily) average. For the discrete case, approximation techniques that take advantage of the particular properties of GBM (in particular, analytical results for conditional normal variables) are available and reasonably effective; see Nielsen (2001) for an excellent overview.[15] The salient results can also be found in EW.

The concern here is with the relationship between the fixed and floating strike options with payoffs given in (5.47) and (5.48). This is an interesting question to resolve, as most of the available analytical techniques are for the fixed strike case. This problem was first solved by Henderson and Wojakowski (2002) for GBM via an application of Girsanov's theorem. We will take a related but slightly different approach that will prove useful for extending the results to more general Lévy processes. We write the general floating strike valuation problem as

$$V = e^{-r(T-t)}E_t^Q\left(k \cdot S_T - \frac{1}{T-t}\int_t^T S_s ds\right)^+ = e^{-r(T-t)}E_t^Q S_T\left(k - \frac{1}{T-t}\int_t^T \frac{S_s}{S_T} ds\right)^+ \tag{5.49}$$

where we explicitly introduce discounting effects. In fact, it is important here that we specify the form of the $Q$–dynamics, noting the nonzero drift:

$$\frac{dS}{S} = (r-q)dt + \sigma\, dw \tag{5.50}$$

(We will see that a particularly interesting aspect of the symmetry result to be derived only arises when interest rates and dividends are present in the risk-neutral process.) Now, we apply the (by now) familiar numeraire change of measure given by $\frac{dQ_s}{dQ} = \frac{e^{-(r-q)(T-t)}S_T}{S}$[16] to get

$$V = Se^{-q(T-t)}E_t^{Q_s}\left(k - \frac{1}{T-t}\int_t^T \frac{S_S}{S_T} ds\right)^+ \tag{5.51}$$

Plainly, (5.51) is highly reminiscent of the fixed strike (put) problem, although the dividend rate has taken the place of the discount rate. However, the entity being averaged is not a price as such, and in fact is not even a price redenominated in terms of the asset $S$ (as it is expressed in terms of units of the *terminal* price). Thus, it would seem that the numeraire change/measure change has not achieved much benefit. Fortunately this is not the case. Let us write the $Q$–dynamics of the log price as

$$dz = \left(r - q - \tfrac{1}{2}\sigma^2\right) dt + \sigma\, dw \tag{5.52}$$

Now, the problem amounts to determining what the dynamics of $e^{z_s - z_T}$ under the measure change given by $\frac{dQ_S}{dQ} = e^{z_T - z - (r-q)(T-t)}$. To answer this question, we again appeal to characteristic functions as a convenient means of elucidating the transformation of the process. The characteristic function of the ratio (in (5.51)) under this measure is given by

$$E_t^{Q_S} e^{i\phi(z_s - z_T)} = e^{-z-(r-q)(T-t)} E_t^Q e^{i\phi z_s + (1-i\phi) z_T} \tag{5.53}$$

Using the law of iterated expectations and the results for the characteristic function of Gaussian processes, we find that

$$\begin{aligned}
E_t^{Q_s} e^{i\phi(z_s - z_T)} &= e^{-z-(r-q)(T-t)} E_t^Q e^{i\phi z_s} E_s^Q e^{(1-i\phi) z_T} \\
&= e^{-z-(r-q)(T-t)} E_t^Q e^{i\phi z_s} e^{(1-i\phi)(z_s + (r-q-\frac{1}{2}\sigma^2)(T-s)) + \frac{1}{2}(1-i\phi)^2\sigma^2(T-s)} \\
&= e^{-z-(r-q)(T-t) + (1-i\phi)(r-q-\frac{1}{2}\sigma^2)(T-s) + \frac{1}{2}(1-i\phi)^2\sigma^2(T-s)} E_t^Q e^{z_s} \\
&= e^{-(r-q)(T-t) + (1-i\phi)(r-q-\frac{1}{2}\sigma^2)(T-s) + \frac{1}{2}(1-i\phi)^2\sigma^2(T-s) + (r-q)(s-t)} \\
&= e^{i\phi(q-r-\frac{1}{2}\sigma^2)(T-s) + \frac{1}{2}\phi^2\sigma^2(T-s)}
\end{aligned} \tag{5.54}$$

So, what has been established in (5.54) is the following. Under the measure change induced by the terminal price "numeraire" (properly discounted), the price ratio $\frac{S_s}{S_T}$ appearing in the averaging term of the payoff in (5.51) is GBM, with expected value $e^{(q-r)(T-s)}$,[17] volatility $\sigma$, and the roles of the discount rate and dividend rate interchanged. Indeed, (5.54) establishes that the $Q$–dynamics of $\tilde{z}_{t,T} \equiv z_t - z_T$ depend only on the time difference $T - t$, so that the average can be written as

$$\int_t^T e^{\tilde{z}_{s,T}} ds = \int_t^{T-t} e^{\tilde{z}_{0,s}} ds \tag{5.55}$$

This is of course a continuously summed ensemble of lognormal variables over the same effective time horizon as the original fixed strike problem. Thus, valuation of

the floating strike Asian call option represented by (5.51) is equivalent to valuing a corresponding *fixed* strike Asian *put* option, with the correspondence given by

$$V^{\text{call}}_{\text{float}}(S,k,r,q) = SV^{\text{put}}_{\text{fixed}}(1,k,q,r) \tag{5.56}$$

Of course, by homogeneity, the result in (5.56) can be expressed in terms of the current price $S$ and strike $kS$. The volatility of course is unchanged. Hence, we have established the desired association between fixed and floating strike Asian options, and furthermore have shown the great utility of change of measure techniques – specifically in conjunction with characteristic functions as a computational framework – in making the result clear.[18]

These results are in fact valid for much more general processes (*e.g.*, those with jumps); see the Appendix to this chapter. The central feature, as we have noted, is the ability to form tractable results via the (conditional) characteristic function of the underlying process. We thus now turn to a discussion of a very broad class of processes (two, actually) for which such an approach is feasible.

## 5.2 Affine jump diffusions/characteristic function methods

Let us re-emphasize the central theme of this work: valuation through the identification of appropriate exposures via the construction of specific portfolios. Achieving this objective obviously requires the proper tools, and one such powerful tool is the language of stochastic calculus,[19] specifically in the analysis of portfolio dynamics. There is of course a very well-developed theory here, with so-called semi-martingales[20] as the centerpiece. We wish to employ this machinery as widely as possible, while simultaneously encompassing a class of stochastic models whose breadth is no wider than necessary. It is a basic (and frankly, overstated) truism that actually existing price processes, especially in energy markets, do not follow Gaussian laws. Although we shall demonstrate that it is often neither necessary nor desirable for a model to explicitly account for *all* characteristics of an underlying price process, it is important nonetheless to have a framework that allows one to answer (to an acceptable degree) the question of which characteristics *do* matter. To this end, we retain the basic foundation of analysis based on semi-martingales.[21]

### 5.2.1 Lévy processes

#### *5.2.1.1 The starting point*

We recall a very elementary property of Brownian motion: stationary, independent increments. It is natural to inquire as to how far this basic aspect can be extended. The answer is given by the class of stochastic processes known as Lévy processes,

which we have already discussed previously. The essence of such processes is that they include not just familiar Brownian motion with drift, but also a very rich class of (discontinuous) jump processes. Furthermore, the discontinuous class encompasses not only well-known Poisson processes, but also a less well-known category that is reminiscent, in some ways, of (continuous) diffusive motion. We repeat the definition here for ease of exposition.

*Definition*: A Lévy process is a stochastic process $X_t$[22] satisfying:

(i). $X_0 = 0$ a.s.
(ii). The increments $X_{t_i} - X_{t_j}$ and $X_{t_{i'}} - X_{t_j}$ for $t_{j'} > t_{j'} \geq t_i > t_j$ are independent
(iii). The increments $X_{t_i - X_{t_j}}$ are stationary; *i.e.* $X_{t_i} - X_{t_j}$ is distributed as $X_{t_i - t_j}$
(iv). $X$ is continuous in probability.[23]

A deeply important and useful result is the famed Lévy-Khintchine representation, which provides an explicit expression for the characteristic function of a Lévy process: $Ee^{i\phi X_t} = e^{t\psi(\phi)}$, where the so-called characteristic exponent $\psi$ is given by[24]

$$\psi(\phi) = i\phi\alpha - \tfrac{1}{2}A\phi^2 + \int_{\mathbb{R}/\{0\}} (e^{i\phi x} - 1 - i\phi x h(x))\nu(dx) \tag{5.57}$$

where $h$ is some function concentrated around the origin, *e.g.*, $h(x) = 1(|x| < 1)$ is a very common choice. Before explaining what the entity $\nu$ in the integrand of (5.57) represents, we point out that the first two terms clearly correspond to Brownian motion with drift (per unit of time), as claimed. To understand what the integral in (5.57) represents, first assume that the expression $h(x)$ is absent and that $\nu(x)$ is integrable across the *entire* real line (*i.e.*, including the origin). Then, it is not hard to see that the integral term corresponds to a Poisson jump, with intensity (arrival rate) of jumps given by $\int_{\mathbb{R}} \nu(dx)$ (again, per unit of time) and amplitude (distribution) of jumps given by $\nu(dx)/\int_{\mathbb{R}} \nu(dx)$. So, as claimed we see that (Poisson) jumps are included in the class of Lévy models.

#### 5.2.1.2 *A more general category of jumps*

However, great subtlety is introduced when we consider the case where $\nu(x)$ is *not* integrable. Here, we must refer to $\nu(x)$ by its proper name: the Lévy measure of the process. It has the interpretation that the expected number of jumps within the time interval $[0, t)$ with amplitudes inside some (Borel) set $B$ is given by

$$E\{\#\text{jumps} \in B \text{ up to time } t\} = t\int_B \nu(dx) \tag{5.58}$$

Particularly nice intuition is provided by Lewis (2001) in terms of random measures. We define a counting process which represents the number of jumps within

some time interval $[0, t]$ which have size within an interval $B$ (excluding the origin); this process is a Poisson process with mean given by (5.58). It is precisely the lifting of the restriction of integrability of the Lévy measure across the entire real line that gives rise to a very rich class of jump processes.

To understand this class better, consider the following decomposition of (5.57) for the usual case of indicator function truncation:

$$\begin{aligned}\int_{\mathbb{R}/\{0\}} (e^{i\phi x} - 1 - i\phi x 1_{|x|<1}) \nu(dx) &= \int_{|x|\geq 1} (e^{i\phi x} - 1)\nu(dx) \\ &+ \int_{0<|x|\leq 1} (e^{i\phi x} - 1 - i\phi x)\nu(dx) \\ &= \int_{|x|\geq 1} (e^{i\phi x} - 1)\nu(dx) + \sum_n \int_{\frac{1}{2^{n+1}} \leq |x| \leq \frac{1}{2^n}} (e^{i\phi x} - 1 - i\phi x)\nu(dx) \end{aligned} \tag{5.59}$$

The ensemble in (5.59) is clearly associated with a superposition of (standard) Poisson jumps, specifically a single term representing "large" amplitude jumps (*i.e.*, those greater than 1 in absolute value), and an infinite (but countable) collection of "smaller and smaller" amplitude jumps of "faster and faster" intensities (arrival rates). More accurately, the $n$th term in the summation in (5.59) is the characteristic function[25] of a pure martingale jump process (note the compensator term $-i\phi x$ in the integrand) whose amplitude is confined to the interval $\varepsilon_n = \{x : \frac{1}{2^{n+1}} \leq |x| \leq \frac{1}{2^n}\}$, with intensity given by (recall the trick used in the ordinary Poisson case) $\lambda_n = \nu(\varepsilon_n)$. In the case of non-integrability of the Lévy measure, divergence[26] implies that these intensities tend to infinity as the interval of support for the amplitude shrinks.

#### *5.2.1.3 A recap*

We have thus outlined a foundational result, namely the Lévy-Itô decomposition, which states that any Levy process is decomposable into three independent processes:

1. A Brownian motion with drift
2. A pure jump (point) process, specifically a compound Poisson process
3. A pure jump (martingale) process with a countable number of jumps within any finite interval.

In general, then, Lévy processes can be thought of as Brownian motion with drift, *plus* a summation (broadly conceived) of jumps within the time interval in question and across all possible jump sizes. Roughly speaking, there is a non-separable, "two dimensional" structure to jumps in general. If the Lévy measure is integrable across the entire real interval, then we can think of jumps as being two (separable) independent "one dimensional" entities: first, as a jump as such (*i.e.*, without reference to jump size), and second (conditionally on a jump occurring) as a jump of particular (stochastic) size. In other words, in this case (and this case only) the jump process is the familiar compound Poisson process, with associated jump intensities and amplitudes.

In general though, jumps can have a much richer structure, and this is the meaning of the third class of jumps in the Lévy-Itô decomposition. This class is commonly referred to as *infinite activity* jumps, because (over any time horizon) they correspond to a countable number of arbitrarily small-sized jumps.[27] Examples of such pure jump processes as they apply to financial research are the variance Gamma process[28] see (Madan *et al.* [1998]), and the normal inverse Gaussian process (see Barndoff-Nielsen and Shephard [2012]). Now, there is little doubt that this is an appealing theoretical notion, both in terms of greater generality and the fact that this representation does permit for a fair amount of analytical results (*e.g.*, in terms of option pricing). Also, there is the obvious intuition that a process that can exhibit jumps of arbitrary size over any time interval is strongly reminiscent of Brownian motion itself. In other words, in the interest of both theoretical sparsity and completeness, there might be good reason to model prices as *purely* jump processes, as this should include as special cases the usual compound Poisson process (directly) *and* Brownian motion (indirectly). This approach has been advocated by Carr *et al.* (2002) and Geman (2002).[29]

While Lévy processes undoubtedly provide a rich and indeed beautiful generalization of familiar Brownian motion, they unfortunately do not capture all of the aspects that we would like from a model. In particular, they do not directly incorporate stochastic volatility. Let us be clear about our objection: we do not propose building elaborate models that "realistically" capture many known (stylized) facts about observed price series. (*E.g.*, we strongly advise against estimating jump diffusions, much less stochastic volatility models, from available data, for reasons the reader will discover in Chapter 6.) We have emphasized this viewpoint throughout. Rather, our objective is to have a modeling framework that is both flexible and rich enough to enable us to understand the interplay between valuation and hedging that forms the basis of any trading strategy underlying portfolio construction. (Indeed, in a real sense valuation *is* portfolio construction.) The fact remains that stochastic volatility *is* a feature we would like to include in our modeling building blocks, and so we must turn attention to how, if at all, it can be integrated with Lévy modeling.

### 5.2.2 Stochastic volatility

*5.2.2.1 Some background*

As the name suggests, stochastic volatility models are models for which the variance of a stochastic process is itself random. We will see quite general examples in the next subsection on affine jump diffusions (*e.g.*, the well-known Heston model), but we can note in passing other models that are popular in the literature such as Constant Elasticity of Variance (CEV) (probably better characterized as a local volatility model) and Stochastic Alpha Beta Rho (SABR).[30] There is a good discussion of stochastic volatility in EW, where certain effects of interest in energy markets are discussed. In the interest of completeness we outline some of the reasons why jumps alone might be insufficient in modeling efforts.

- The so-called inverse leverage effect in commodity (chiefly energy) prices, where volatility (both realized and implied) tends to increase with the level of prices (typically represented by treating movements of stochastic volatility and price as negatively correlated)
- Different patterns in excess kurtosis (fatness of tails relative to normality) distinguishing jumps from stochastic volatility, as a function of time horizon; jumps are typically characterized by decreasing excess kurtosis as the time horizon increases, while stochastic volatility gives rise to the opposite pattern
- Similar differences in the implied volatility surface for processes driven by jumps and by stochastic volatility; the smile/skew in the former case tends to be steep for prompt contracts and a flat for longer-dated contracts, while the latter tends to have a flat prompt structure and steepness further out (and in truth both effects can be present simultaneously).

Our focus here will be more narrow, primarily to show how such models can be tied in with Lévy processes, and to lead into the canonical class of affine jump diffusions where these effects can be more readily modeled (at, of course, the cost of losing some generality).

*5.2.2.2 Time change results*

The first thing to note is that a common feature of stochastic volatility models is that the associated log-price processes no longer possess independent increments. (Typically the *joint* process dynamics remain Markovian, but this is a different issue.) Thus, we can immediately anticipate that such log-price models cannot be of Lévy type. However, we hold out some hope that stochastic volatility models can be related to Lévy processes, given the appeal of such models laid out in the previous section. Progress can be made from the following intuition. We have characterized volatility as a measure of information accumulation over specific time horizons. (Recall the discussion in Section 2.2) Note that nothing in this conception rules out

the possibility of this time horizon being random itself. One thinks of the common situation where unanticipated news events suddenly impact prices. In energy markets, weather would be a typical case, although other examples include unplanned outages of a large (baseline) generation unit or an announcement of a major gas pipeline development.[31] It should be recalled that pure jump processes such as the aforementioned variance Gamma can also be crafted in terms of a time-changed Brownian motion. In other words, in this model deterministic time is replaced by a "stochastic clock," or ordinary time becomes "business time." Thus, we anticipate that our conception of volatility can be generalized: volatility is still a measure of information flow, but we allow the relevant time horizon over which this information accumulates to be stochastic, as well.

In fact, the above intuition has some basis in fact. It turns out that *every* semimartingale can be written as a time-changed Brownian motion (Monroe [1978]; see also Karatzas and Shreve [1991]). These considerations lead us to examine the use of time-changed Lévy processes as a means for modeling stochastic volatility. We start with some Lévy process $X_t$ and a random time[32] $T_t$. We then introduce a new process $Y$ obtained by evaluating $X$ at $T$: $Y_t \stackrel{\Delta}{=} X_{T_t}$. To understand the nature of the process $Y$, it is natural to consider its characteristic function. Under the assumption that the time change $T$ is independent of the underlying process $X$, we have that

$$\begin{aligned} Ee^{i\phi Y_t} &= Ee^{i\phi X_{T_t}} \\ &= E(E(e^{i\phi X_u}|T_t = u)) = Ee^{T_t \psi_X(\phi)} = \mathcal{L}_T(-\psi_X(\phi)) \end{aligned} \tag{5.60}$$

using (5.57), and where $\mathcal{L}$ is the Laplace transform of the random time. Note that in general the Lévy status of $X$ is not retained. We will consider a special class of random times in Section 8.1.4 in the context of dependent Lévy processes known as subordinators, which are pure jump processes with only positive jumps. In this case, a subordinated Lévy process remains a Lévy process. However, (5.60) provides a general framework of ascertaining the distribution of stochastic volatility models when the requisite Fourier and Laplace transforms are known.[33]

#### *5.2.2.3 An illustration of conditioning without independence*

Let us consider an example. First we will establish a useful but (seemingly) not widely used result. Consider the problem of finding a conditional expectation. We have that

$$\Pr(x_2|x_1) = \frac{\Pr(x_1, x_2)}{\Pr(x_1)} = \frac{\frac{1}{(2\pi)^{n_1+n_2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\phi_1 d\phi_2 f(\phi_1, \phi_2) e^{-i\phi_1^T x_1 - i\phi_2^T x_2}}{\frac{1}{(2\pi)^{n_1}} \int_{-\infty}^{\infty} d\phi_1 f(\phi_1, 0) e^{-i\phi_1^T x_1}} \tag{5.61}$$

where $f$ denotes the characteristic function (Fourier transform) of the process: $f(\phi) = Ee^{-i\phi^T x}$. Here $x = (x_1 \;\; x_2)^T$ is some (random) vector and the indices 1 and

2 denote some partition of this vector. Using the result $\delta(x = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} d\phi e^{i\phi^T x})$, we find that (assuming justification of interchanging orders of integration)

$$E(e^{-i\phi^T x_2}|x_1)\frac{\int_{-\infty}^{\infty} d\phi_1 f(\phi_1,\phi)e^{-i\phi_1^T x_1}}{\int_{-\infty}^{\infty} d\phi_1 f(\phi_1,0)e^{-i\phi_1^T x_1}} \tag{5.62}$$

Thus, as long as the characteristic function is known, the conditional characteristic function can be evaluated, at least numerically. Note that the dimension of problem is reduced to the dimension of the nonconditioned variable.[34]

However, our chief concern here is with stochastic volatility models. As we have done before, we will introduce some concepts that will be properly developed in subsequent sections. In that spirit, consider the popular stochastic volatility model of Heston (1993):

$$\begin{aligned} dz &= -\tfrac{1}{2}\sigma^2 v dt + \sigma\sqrt{v}dw \\ dv &= \kappa(1-v)dt + \sqrt{v}dw' \end{aligned} \tag{5.63}$$

with $dwdw' = \rho dt$ and in contrast to the conventional representation, the stochastic variance is normalized with its long-term value set to one. Now, with $\nu$ unitless we can define the following random time:

$$\mathcal{T}_T = t + \int_t^T v_s ds \tag{5.64}$$

We introduce the following auxiliary processes:[35]

$$\begin{aligned} dV &= vdt \\ d\Omega &= \sqrt{v}dw \\ d\Omega' &= \sqrt{v}dw' \end{aligned} \tag{5.65}$$

(Note the association between cumulative [integrated] variance and [accumulated] stochastic time.) We then have that[36]

$$z_T = z - \tfrac{1}{2}\sigma^2 \int_t^T v_s ds \sigma \int_t^T \sqrt{v_s}dw_s = z - \tfrac{1}{2}\sigma^2(V_T - V) + \sigma(\Omega_T - \Omega) \tag{5.66}$$

Now, if the (log) price and stochastic variance are uncorrelated ($\rho = 0$), it should be intuitively clear that the $\Omega$ term is a time-changed standard Brownian motion,

evaluated at $\mathcal{T}_T - t$. In this case, the (log) price is indeed a time-changed Lévy process (Brownian motion with drift).[37] In the general case (nonzero correlation) we cannot make this conclusion, so this particular stochastic time change cannot be the one implied by the aforementioned result of Monroe (1978). However, we can still ask if we can somehow relate the Heston model to Lévy processes.

To answer this question, we apply the results for conditional expectations derived above. Specifically, we ask about the distribution of $\mathbf{\Omega}$ given $V$. To use our conditional characteristic function results, have to know the underlying unconditional characteristic functions. As will be seen when we discuss affine processes, the joint characteristic function can be written

$$E_t e^{i\phi_0 V_T + i\phi_1 \Omega_T + i\phi_2 \Omega'_T} = e^{i\phi_1 V + i\phi_1 \Omega + i\phi_2 \Omega' + \beta(t;\phi_0,\phi_1,\phi_2)v + \gamma(t;\phi_0,\phi_1,\phi_2)} \tag{5.67}$$

where un-subscripted variables represent time–$t$ (*i.e.*, current) values, and the coefficients $\beta$ and $\gamma$ satisfy the following ordinary differential equations (ODEs):

$$\begin{aligned} \dot{\beta} + (-\kappa + i\rho\phi_1 + i\phi_2)\beta + \tfrac{1}{2}\beta^2 + i\phi_0 - \tfrac{1}{2}\phi_1^2 - \tfrac{1}{2}\phi_2^2 - \rho\phi_1\phi_2 = 0 \\ \dot{\gamma} + \kappa\beta = 0 \end{aligned} \tag{5.68}$$

with $\beta(T) = \gamma(T) = 0$. Thus,[38]

$$\begin{aligned} &E_t(e^{i\phi_1(\Omega_T - \Omega)} | V_T - V, \Omega'_T - \Omega') \\ &= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\phi_0 d\phi_1 e^{-i\phi_0(V_T - V) - i\phi_2(\Omega'_T - \Omega') + \beta(\phi_0,\phi_1,\phi_2)v + \gamma(\phi_0,\phi_1,\phi_2)}}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\phi_0 d\phi_1 e^{-i\phi_0(V_T - V) - i\phi_2(\Omega'_T - \Omega') + \beta(\phi_0,0,\phi_2)v + \gamma(\phi_0,0,\phi_2)}} \end{aligned} \tag{5.69}$$

While this may look intractable, note from the governing ODEs that only certain combinations of the Fourier variables occur. In particular, the following change of variables is suggested:

$$\begin{aligned} \phi_2 &= \tilde{\phi}_2 - \rho\phi_1 \\ \phi_0 &= \tilde{\phi}_0 - i\tfrac{1}{2}\rho_s^2\phi_1^2 \end{aligned} \tag{5.70}$$

where $\rho_s^2 = 1 - \rho^2$. Thus, a given $\beta$ with nonzero $\phi_1$–dependence can be transformed into the solution of an equivalent ODE with $\phi_1 = 0$. Using this substitution in the numerator of the above conditional expectation, and adjusting

integration contours appropriately (we will discuss applications of contour integration to option pricing in greater detail in the next subsection), we find that

$$e^{-i\phi_0(V_T-V)-i\phi_2(\Omega'_T-\Omega')+\beta(\phi_0,\phi_1,\phi_2)v+\gamma(\phi_0,\phi_1,\phi_2)} \rightarrow$$
$$e^{-i\phi_0(V_T-V)-i\phi_2(\Omega'_T-\Omega')+\beta(\phi_0,0,\phi_2)v+\gamma(\phi_0,0,\phi_2)} e^{i\phi_1\rho(\Omega'_T-\Omega')-\frac{1}{2}\rho_s^2\phi_2^2} \tag{5.71}$$

from which we can conclude

$$E_t(e^{i\phi_1(\Omega_T-\Omega)}|V_T-V,\Omega'_T-\Omega') = e^{i\phi_1\rho(\Omega'_T-\Omega')-\frac{1}{2}\rho_s^2\phi_2^2} \tag{5.72}$$

In other words, conditional on the stochastic time (change) $V_T - V$ and the auxiliary variable $\Omega'_T - \Omega'$, the auxiliary variable $\Omega_T - \Omega$ is normal. Note that when there is no correlation, this latter auxiliary variable is precisely a time-changed Brownian motion, a fact that we have now established rigorously rather than intuitively.[39] This sort of idea can serve as the basis of exact simulations of such processes (*i.e.*, without recourse to discretizing the underlying SDE in (5.63); see Broadie and Kaya [2006]).

#### *5.2.2.4 Ornstein-Uhlenbeck (OU) based modeling*

We thus see that there is a connection between the Heston model and a suitably interpreted Lévy process. We will briefly discuss another approach to modeling stochastic volatility with Lévy processes. Consider the following non-Gaussian Ornstein-Uhlenbeck (OU) process:

$$dy = -\kappa y dt + dL \tag{5.73}$$

where $L$ is a Lévy process (typically with no drift and positive increments, *i.e.*, a subordinator). Letting $\tilde{y} = e^{\kappa t}y$, we see that $d\tilde{y} = e^{\kappa t}dL$ and thus

$$Y_T = e^{-\kappa(T-t)}y + \int_t^T e^{-\kappa(T-s)}dL_s \tag{5.74}$$

and consequently

$$E_t e^{i\phi y_T} = e^{i\phi e^{-\kappa(T-t)}y} E_t e^{i\phi\int_t^T e^{-\kappa(T-s)}dL_s} = e^{i\phi e^{-\kappa(T-t)}y+\int_t^T \psi(\phi e^{-\kappa(T-s)})ds} \tag{5.75}$$

where $\psi$ is the characteristic exponent of $L$.[40] This approach is used by Barndorff-Nielsen and Shephard (2001) to construct a wide class of stochastic volatility models.

#### 5.2.2.5 *Modeling: generality and sufficiency*

We have described to this point how the commonly used model of Brownian motion can be extended in its essentials (primarily, stationary, independent increments) to incorporate features such as jumps. We have further indicated how this generalization can be itself extended (via time changes) to include additional features such as stochastic volatility. It remains the case, however, that we would like a modeling framework that more transparently integrates both aspects (jumps and stochastic volatility) while retaining (indeed, potentially enlarging) tractability. There is in fact a category of stochastic processes that allow great flexibility in modeling jumps and stochastic volatility, and in fact other behaviors such as mean reversion. We are referring to the canonical class of affine jump diffusions. These models also provide a very useful means for ascertaining the relationship between spot and forward dynamics, and between dynamics under different measures. It is in fact quite straightforward to craft and analyze portfolio dynamics (the foundation of our approach to valuation) using affine jump diffusions. In addition, they share with Lévy processes the fact that many computations and analyses can be carried out conveniently via Fourier methods (*i.e.*, through characteristic functions). The principal drawback of such models is that they take a very specific form, and thus lose much of the generality offered by Lévy models (and their time-changed extensions).

However, this is not a problem as such. Our ultimate objective is to understand how much certain features of actually existing energy markets matter, or more accurately, *do not* matter. The question that must always be asked is whether a particular modeling representation is sufficient to answer various questions put in front of it. For example, we would generally not want to value an option using a jump diffusion model that we fit to data. However, we *would* very much like to understand the effects of pricing an option driven by a jump process, when we ignore jumps and concentrate on effects that are more robustly appraised, such as the diffusive structure. Generality is nice, but is often not necessary for understanding what we need to know for a particular problem (and indeed, can often be a barrier to such understanding). We have taken the time to introduce and discuss Lévy processes (and their extensions through time changes) because (apart from the fact that they are an interesting and important topic in their own right) we can better see how the effects they are intended to capture through a very general framework can be *sufficiently* provided through a more restricted, canonical form. With these points in mind, we now turn to a fuller discussion and applications of affine jump diffusions.

### 5.2.3 Pseudo-unification: affine jump diffusions

*5.2.3.1 Basic structure*

We will consider now the general class of affine jump diffusions. These processes have dynamics that can be generically written as

$$dz_i = (A_{ij}z_j + b_i)dt + \sigma_i^k \sqrt{z_k} dw_i^k + \sigma_1^0 dw_i^0 + j_{ik} dq_k \tag{5.76}$$

where the usual convention of summation over a repeated index not on the left-hand side is adopted.[41] Here $w$ denotes a standard Brownian motion (diffusive), and $q$ is a Poisson jump process (discontinuous). We further assume that the covariance structure of the diffusive components is given by

$$dw_i^k dw_j^l = \delta_{kl}\rho_{ij}^k \tag{5.77}$$

and that the $i$-th jump intensity takes the form

$$\lambda_i = \lambda_{i0} + \lambda_{ij} z_j \tag{5.78}$$

The jumps are independent, but note that the number of term processes need not be equal to the number of state variables. Consequently, the specification in (5.76) allows for a sort of joint dependence across jumps. Note further that the "aggregate" covariance structure, so to speak, becomes (as an Itô isometry)

$$E(\sigma_i^k \sqrt{z_k} dw_i^k + \sigma_i^0 dw_i^0)(\sigma_j^k \sqrt{z_k} dw_j^k + \sigma_j^0 dw_j^0) = (z_k X_{ij}^k + X_{ij}^0)dt \tag{5.79}$$

where

$$X_{ij}^k \equiv \rho_{ij}^k \sigma_i^k \sigma_j^k \tag{5.80}$$

*5.2.3.2 Governing equations*

In light of (5.79), (5.78), and the drift term in (5.76), it should be clear why these processes are referred to as affine: they are fully described by entities that are linear in the underlying state variables plus a constant. This structure permits the characteristic function to be determined by solving a system of (nonlinear) ordinary differential equations (ODEs). (We have already seen the great computational utility of employing characteristic functions in the context of change of measure; we will now see yet more applications.) This is very significant, as the problem (for a general process) would ordinarily entail having to solve a partial integro-differential equation (PIDE) in several dimensions, a very daunting task to say the least. To appreciate the utility, let us explicitly indicate that the dynamics in (5.76) take place under a specific measure, which we denote by $Q$. Then we are interested in the following:

$$f = E_t^Q e^{i\varepsilon_n \phi_n z_n(T)} \tag{5.81}$$

where $\varepsilon_n$ is sort of an indicator/flag only taking on values 0 or 1. Thus, (5.81) could represent the characteristic function of a single state variable or of the entire joint ensemble or of particular combinations of variables. Since $f$ is a $Q$–martingale, by using the Itô calculus extended for jumps (see Tankov and Cont [2003] for a thorough discussion), we find that it satisfies the following PIDE:

$$f_t + (A_{ij}z_j + b_i)f_{z_i} + \tfrac{1}{2}(X^k_{ij}z_k + X^0_{ij})f_{z_iz_j} + (\lambda_{k0} + \lambda_{kl}z_l)E^Q(f(z_i + j_{ik}) - f) = 0 \quad (5.82)$$

and of course $f(z,T) = e^{i\varepsilon_n\phi_n z_n}$. Now, due to the affine structure, we can seek a solution to (5.82) of the following form:

$$f = e^{\alpha_0(t;\phi)+\alpha_k(t;\phi)z_k} \quad (5.83)$$

Substituting (5.83) into (5.82) and collecting coefficients of $z_k$, we get the following system of ODEs:

$$\begin{aligned} \dot{\alpha}_k + A_{ik}\alpha_i + \tfrac{1}{2}X^k_{ij}\alpha_i\alpha_j + \lambda_{ik}K_i(\alpha) = 0 \\ \dot{\alpha}_0 + b_i\alpha_i + \tfrac{1}{2}X^0_{ij}\alpha_i\alpha_j + \lambda_{i0}K_i(\alpha) = 0 \end{aligned} \quad (5.84)$$

where

$$K_i(\alpha) \equiv E^Q(e^{\alpha_m j_{mi}} - 1) \quad (5.85)$$

and the terminal conditions are

$$\begin{aligned} \alpha_k(T) = i\varepsilon_k\phi_k \\ \alpha_0(T) = 0 \end{aligned} \quad (5.86)$$

Thus, we see that calculation of the characteristic function in (5.81) has been reduced to solving the system of ODEs in (5.84)–(5.86). These will in general be nonlinear, but in the absence of jumps, the quadratic structure gives the system a Riccati form. We will see shortly a number of nontrivial cases where this system can be solved analytically. In general, though, this system can be readily solved numerically, although we will see examples where care must be exercised (*e.g.*, volatility scaling laws on time horizons of different orders of magnitude[42]). Therefore, the kinds of models represented in (5.76) are understandably popular. Of course, knowing the characteristic function is tantamount (in theory at least) to knowing the distribution of the underlying process (via Fourier inversion). However, it turns out that many applications of interest – specifically option valuation – can be carried out directly in terms of the characteristic function, without recourse to any inversion, as we will now see.

### 5.2.4 General results/contour integration

*5.2.4.1 The primary calculation*

Consider the following valuation problem:

$$E_t^Q(e^{z_T} - K)^+ \tag{5.87}$$

In terms of indicator functions, (5.87) can be written as

$$E_t^Q(e^{z_T} - K)^+ = E_t^Q e^{z_T} 1(z_T > \log K) - K \cdot E_t^Q 1(z_T > \log K) \tag{5.88}$$

Consider the second expectation in (5.88):

$$E_t^Q 1(z_T > \log K) = \int_{-\infty}^{\infty} dz_T \Pr(z_T|z) 1(z_T > \log K) \tag{5.89}$$

Now, let us assume the (conditional) characteristic function of the process $z$ is known. Denoting the characteristic function by $f$ and using the Fourier inversion formula, (5.89) becomes

$$E_t^Q 1(z_T > \log K) = \int_{-\infty}^{\infty} dz_T 1(z_T > \log K) \frac{1}{2\pi} \int_{-\infty}^{\infty} d\phi f(\phi) e^{-i\phi z_T} \tag{5.90}$$

Let us emphasize that the integral wrt. $\phi$ in (5.90) is one-dimensional, even though the characteristic function $f$ will in general be multidimensional in terms of the underlying state variables[43] (see (5.83)). Now, we would like to interchange the order of integration in (5.90) (for reasons that will soon become apparent). However, as the limits of integration stand, this is not possible as $e^{-i\phi z_T}$ is not integrable along the entire real $\phi-$axis. This problem can be remedied by choosing a contour of integration in the Fourier inversion that lies below the real $\phi-$axis, say with negative imaginary part. See Figure 5.1.

We thus write eq. (5.90) as

$$E_t^Q 1(z_T > \log K) = \int_{-\infty}^{\infty} dz_T 1(z_T > \log K) \frac{1}{2\pi} \int_{\Gamma} d\phi f(\phi) e^{-i\phi z_T} \tag{5.91}$$

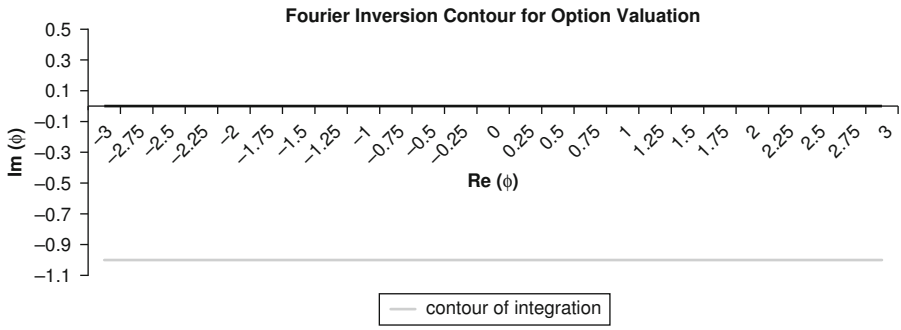



**Figure 5.1** Contour for Fourier inversion. The contour must lie below the real axis to ensure convergence of integrals such as (5.91) (as analogues of (5.90))

where $\Gamma$ is a contour like the one in Figure 5.1. Interchanging the order of integration (which is now justified) in (5.91), we get

$$E_t^Q 1(z_T > \log K) = \frac{1}{2\pi} \int_\Gamma d\phi f(\phi) \int_{-\infty}^{\infty} dz_T e^{-i\phi z_T} 1(z_T > \log K)$$

$$= \frac{1}{2\pi} \int_\Gamma d\phi f(\phi) \int_k^{\infty} dz_T e^{-i\phi z_T} = \frac{1}{2\pi} \int_\Gamma d\phi f(\phi) \frac{e^{-i\phi k}}{i\phi} \tag{5.92}$$

where $k \equiv \log K$. Now, it is customary in the literature to shift the contour $\Gamma$ back to the real $\phi-$axis, and when taking into account the pole at $\phi = 0$, we get

$$E_t^Q 1(z_T > \log K) = \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} d\phi \mathrm{Re}\left[ f(\phi) \frac{e^{-i\phi k}}{i\phi} \right] \tag{5.93}$$

We would like to point out that, although it is common to approach the computation in (5.93) as a real integral, it is actually not necessary, and we will see examples (in Chapter 7) where it is actually preferable numerically to directly perform the integration in the complex plane in (5.92).

*5.2.4.2 Primary calculation (continued): change of measure*

Turning now to the first expectation in (5.88), change of measure techniques prove very useful. We have that

$$E_t^Q e^{z_T} 1(z_T > \log K) = E_t^Q e^{z_T} \cdot E_t^{Q_z} 1(z_T > \log K) \tag{5.94}$$

where $Q_z$ is a measure defined by $\frac{dQ_z}{dQ} = e^{z_T}/E_t^Q e^{z_T}$. Now, if we can determine the characteristic function of $z$ under this measure, then the same argument used to derive (5.93) can be applied here as well. Of course, this is a very easy task:

$$f^z(\phi) = E_t^{Q_z} e^{i\phi z_T} = \frac{E_t^Q e^{i(\phi-i)z_T}}{E_t^Q e^{z_T}} = \frac{f(\phi - i)}{f(-i)} \tag{5.95}$$

Now, in most cases of interest $Q$ will be a valuation measure under which $z$ is an exponential martingale, so the option value can be written in standard form as

$$E_t^Q(e^{z_T} - K)^+$$
$$= e^z \left( \tfrac{1}{2} + \frac{1}{\pi} \int_0^\infty d\phi \mathrm{Re}\left[ \frac{f(\phi - i)}{f(-i)} \frac{e^{-i\phi k}}{i\phi} \right] \right) - K \left( \tfrac{1}{2} + \frac{1}{\pi} \int_0^\infty d\phi \mathrm{Re}\left[ f(\phi) \frac{e^{-i\phi k}}{i\phi} \right] \right) \tag{5.96}$$

So, the option valuation problem amounts to a pair of quadratures, as long as the characteristic function of the underlying process is known. We will have a bit more to say on quadrature in Chapter 7, but it suffices for now to note that in practice the problem in (5.96) proves amenable to standard techniques such as Gauss-Laguerre.

#### *5.2.4.3 Additional applications*

In fact, although it does not appear to be widely known, this technique can be applied to a wider class of structures often encountered in energy markets, specifically cases where it is more natural to consider the underlying DGP directly, and not transformed via logarithms. Obvious examples include basis options (as in natural gas markets) or options on realized variance. The basic valuation problem in this case is

$$E_t^Q(z_T - k)^+ \tag{5.97}$$

which of course can be written as

$$E_t^Q z_T 1(z_T > k) - k \cdot E_t^Q 1(z_T > k) \tag{5.98}$$

We will continue to restrict attention to cases where the (conditional) characteristic function of the underlying process $z$ is known. Then, the second term in (5.98) presents no problems, as we have just seen. However, the first term does not appear directly amenable to the change of measure techniques that proved so fruitful in arriving at (5.96), precisely because $z$ itself may not represent a valid numeraire. However, these methods can be employed indirectly. Consider the following expectation:

$$E_t^Q e^{\alpha z_T} 1(z_T > k) \tag{5.99}$$

Proceeding as in (5.94), we find that

$$E_t^Q e^{\alpha z_T} 1(z_T > k) = E_t^Q e^{\alpha z_T} \cdot E_t^{Q\alpha} 1(z_T > k) \tag{5.100}$$

where the new measure $Q_\alpha$ is defined via $\frac{dQ_\alpha}{dQ} = \frac{e^{\alpha z_T}}{E_t^Q e^{\alpha z_T}}$. So, to evaluate the expression in (5.99) we need to know the characteristic function of $z$ under this measure, which is obtained as follows:

$$f^\alpha(\phi) = E_t^{Q_\alpha} e^{i\phi z_T} = \frac{E_t^Q e^{(\alpha + i\phi) z_T}}{E_t^Q e^{\alpha z_T}} = \frac{f(\phi - i\alpha)}{f(-i\alpha)} \tag{5.101}$$

an obvious generalization of (5.95). Thus, following (5.92) we have that

$$E_t^Q e^{\alpha z_T} 1(z_T > k) = \frac{1}{2\pi} \int_\Gamma d\phi f(\phi - i\alpha) \frac{e^{-i\phi k}}{i\phi} \tag{5.102}$$

for an appropriately chosen contour $\Gamma$. Now, assuming sufficient regularity of the integrand, upon taking a derivative wrt. $\alpha$ under the integral in (5.102) and setting $\alpha$ to 0 yields

$$E_t^Q z_T 1(z_T > k) = \frac{-i}{2\pi} \int_\Gamma d\phi f'(\phi) \frac{e^{-i\phi k}}{i\phi} \tag{5.103}$$

Consequently the ensemble in (5.98) can be evaluated via characteristic functions, as long as the derivatives entailed in (5.103) can be readily evaluated. In fact, for the class of affine jump diffusions considered above, the characteristic function has the form $f = \exp(\alpha_i(\phi) z_i + \alpha_0(\phi))$ so that derivatives wrt. $\phi$ are given by $(a_i z_i + a_0) f$ where $a \equiv \frac{\partial \alpha}{\partial \phi}$ and, using (5.84), satisfy the following (linear) system of ODEs:

$$\begin{aligned} \dot{a}_k + A_{ik} a_i + a_i X_{ij}^k \alpha_j + a_n \lambda_{ik} E^Q j_{ni} e^{\alpha_m j_{mi}} = 0 \\ \dot{a}_0 + b_i a_i + a_i X_{ij}^0 \alpha_j + a_n \lambda_{i0} E^Q j_{ni} e^{\alpha_m j_{mi}} = 0 \end{aligned} \tag{5.104}$$

with terminal conditions suitably derived from (5.86).[44] We thus see that this framework for option valuation encompasses a fairly rich class of both processes and structures.

It is important to stress that the result in (5.96) is quite general (as is (5.103)). It applies for *any* process for which the characteristic function is obtainable. It thus applies to the class of Lévy processes, as well as the class of affine jump diffusions presently under discussion. Let us consider some specific examples now. (We will work in terms of log prices throughout, unless explicitly noted.)

### 5.2.5 Specific examples

*5.2.5.1 Black-Scholes*

The standard Black-Scholes model needs very little introduction, so we simply write the dynamics of the log-price here:

$$dz = -\tfrac{1}{2}\sigma^2 dt + \sigma\, dw \tag{5.105}$$

The characteristic function satisfies

$$f_t - \tfrac{1}{2}\sigma^2 f_z + \tfrac{1}{2}\sigma^2 f_{zz} = 0 \tag{5.106}$$

with $f(z, T) = \exp(i\phi z)$. The solution is easily seen to be

$$f = e^{i\phi(z - \frac{1}{2}\sigma^2\tau) - \frac{1}{2}\sigma^2\phi^2\tau} \tag{5.107}$$

(We will take $\tau = T - t$) throughout. Now, using the standard result,

$$\tfrac{1}{2} + \frac{1}{2\pi}\int_{-\infty}^{\infty} d\phi \frac{e^{i\alpha\phi - \beta^2\phi^2/2}}{i\phi} = N\left(\frac{\alpha}{\beta}\right) \tag{5.108}$$

It is readily seen that the option pricing formula (5.96) becomes

$$e^z N\left(\frac{z - k + \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}}\right) K \cdot N\left(\frac{z - k - \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}}\right) \tag{5.109}$$

which is of course the Black-Scholes formula. For a related application, consider an option on a Gaussian process (*i.e.*, the so-called Bachelier model):

$$dz = \sigma\, dw \tag{5.110}$$

with characteristic function $f = e^{i\phi z - \frac{1}{2}\sigma^2\phi^2\tau}$. For an option with payoff $(z-k)^+$, we have (from (5.103) with any contour $\Gamma$ below the real $\phi$ –axis) that

$$\begin{aligned} E_t^Q(z_T-k)^+ &= \frac{1}{2\pi}\int_\Gamma d\phi(z+i\phi\sigma^2\tau)e^{i\phi z-\frac{1}{2}\sigma^2\phi^2\tau}\frac{e^{-i\phi k}}{i\phi} - kN\left(\frac{z-k}{\sigma\sqrt{\tau}}\right) \\ &= (z-k)N\left(\frac{z-k}{\sigma\sqrt{\tau}}\right) + \frac{1}{2\pi}\sigma^2\tau\int_\Gamma d\phi e^{i\phi(z-k)\frac{1}{2}\sigma^2\phi^2\tau} \\ &= (z-k)N\left(\frac{z-k}{\sigma\sqrt{\tau}}\right) + \frac{1}{2\pi}\sigma\sqrt{\tau}e^{-\frac{1}{2}(\frac{z-k}{\sigma\sqrt{\tau}})^2}\int_\Gamma d\phi e^{-\frac{1}{2}(\phi-i\frac{z-k}{\sigma\sqrt{\tau}})^2} \\ &= (z-k)N\left(\frac{z-k}{\sigma\sqrt{\tau}}\right) + \sigma\sqrt{\tau}\varphi\left(\frac{z-k}{\sigma\sqrt{\tau}}\right) \end{aligned} \tag{5.111}$$

which of course can be verified by direct calculation.

*5.2.5.2 Merton*

Merton's well-known jump diffusion model (see Merton [1990]) can be written as

$$dz = \left(-\tfrac{1}{2}\sigma^2 - \lambda k\right)dt + \sigma\, dw + j dq \tag{5.112}$$

where $\lambda$ is the jump intensity (so that the probability of a jump in an infinitesimal time step $dt$ is $\lambda dt$ ) and $k \equiv E^Q(e^j - 1) = e^{\mu_j + \frac{1}{2}\sigma_j^2} - 1$ for normally distributed jump amplitudes. The characteristic function $f$ now satisfies a PIDE:

$$f_t - \left(\tfrac{1}{2}\sigma^2 + \lambda k\right)f_z + \tfrac{1}{2}\sigma^2 f_{zz} + \lambda E^Q(f(z+j) - f) = 0 \tag{5.113}$$

again with terminal condition $f(z,T) = \exp(i\phi z)$. The solution is easily found to be

$$f = \exp\left(i\phi z + \tau(-i\phi(\lambda k + \tfrac{1}{2}\sigma^2) - \tfrac{1}{2}\sigma^2\phi^2 + \lambda(e^{i\phi\mu_j - \sigma_j^2\phi^2/2} - 1))\right) \tag{5.114}$$

Now, substituting (5.114) into the basic option pricing formula (5.96), expanding the exponential in terms of $\lambda e^{i\phi\mu_j - \sigma_j^2\phi^2/2}$, and using the result (5.108), we recover the well-known infinite series result

$$e^{-\tilde{\lambda}\tau}\sum_{n=0}^{\infty}\frac{(\tilde{\lambda}\tau)^n}{n!}\left(e^z N\left(\frac{z-k+r_n+\frac{1}{2}v_n^2}{v_v}\right) - Ke^{-r_n}N\left(\frac{z-k+r_n-\frac{1}{2}v_n^2}{v_v}\right)\right) \tag{5.115}$$

where

$$\begin{aligned} r_n &= -\lambda k\tau + n(\mu_j + \sigma_j^2/2) \\ v_n^2 &= \sigma^2\tau + n\sigma_j^2 \\ \tilde{\lambda} &= \lambda(k+1) \end{aligned} \tag{5.116}$$

*5.2.5.3 Heston*

A very well-known (and popular) stochastic volatility model is due to Heston (1993), and is probably the first published research to make use of characteristic function methods for option pricing. (We have already appealed to Heston several times to illustrate earlier points.) The model is given by

$$\begin{aligned} dz &= -\tfrac{1}{2}v dt + \sqrt{v}dw_z \\ dv &= \kappa(\theta - v)dt + \sigma\sqrt{v}dw_v \end{aligned} \tag{5.117}$$

where $dw_z dw_v = \rho\, dt$ (as an Itô isometry). The PDE satisfied by the characteristic function is

$$f_t - \tfrac{1}{2}vf_z + \kappa(\theta - v)f_v + \tfrac{1}{2}vf_{zz} + \rho\sigma\, vf_{zv} + \tfrac{1}{2}\sigma^2 vf_{vv} = 0 \tag{5.118}$$

with the usual terminal condition. We look for a solution of the form

$$f = \exp(i\phi z + C(t;\phi)v + D(t;\phi)) \tag{5.119}$$

with $C(T;\phi) = D(T;\phi) = 0$. These coefficients can be seen to satisfy the following system of ODEs:

$$\begin{aligned} \dot{C} + (i\rho\sigma\phi - \kappa)C + \tfrac{1}{2}\sigma^2 c^2 - \tfrac{1}{2}\phi(\phi + i) &= 0 \\ \dot{D} + \kappa\theta C &= 0 \end{aligned} \tag{5.120}$$

These can be solved analytically by the standard transformation for Riccati equations,[45] yielding

$$\begin{aligned} C &= -\frac{2m_1}{\sigma^2}\frac{1 - e^{(m_2 - m_1)\tau}}{1 - \frac{m_1}{m_2}e^{(m_2 - m_1)\tau}} \\ D &= -\frac{2\kappa\theta}{\sigma^2}\left(m_1\tau + \log\left(\frac{1 - \frac{m_1}{m_2}e^{(m_2 - m_1)\tau}}{1 - \frac{m_1}{m_2}}\right)\right) \end{aligned} \tag{5.121}$$

where $m_{1,2}$ are the roots of the quadratic equation

$$m^2 + (\kappa - i\phi\rho\sigma)m - \frac{\sigma^2}{4}\phi(\phi + i) = 0 \tag{5.122}$$

Computationally, a particularly convenient form of (5.121) is the following:

$$\begin{aligned} C &= -\frac{\phi(\phi+i)}{g+\gamma\,\frac{1+E}{1-E}} \\ e^D &= \frac{2^p \exp\left(\frac{1}{2}p(g-\gamma)\tau\right)}{\left(1+E+\frac{g}{\gamma}(1-E)\right)^p} \end{aligned} \tag{5.123}$$

where $g \equiv \kappa - i\rho\sigma\phi, \gamma \equiv \sqrt{g^2+\sigma^2\phi(\phi+i)}$, $p = \frac{2\kappa\theta}{\sigma^2}$, and $E = \exp(-\gamma\tau)$. Note that for $\phi = -i$ the results in either (5.121) or (5.123) give no contribution to the characteristic function, confirming that $z$ is a $Q$–exponential martingale, as we would expect. With the results in (5.123), the option value can be obtained via Gauss-Laguerre quadrature. It should be pointed out here that potential difficulties arise due to the (implicit) presence of branch-cut singularities (arising from complex logarithms and/or complex powers). We will say a bit more about this issue in Chapter 7 on numerics, but here we simply note the work of Lord and Kahl (2008).[46]

#### 5.2.5.4 *Schwartz*

Apart from option pricing, the methods under discussion greatly facilitate the extraction of certain structures from a wide class of problems. As an example, we consider a general class of processes originally popularized by Schwartz (see, *e.g.*, Schwartz [1997] or Gibson and Schwartz [1990]). A simple representation will suffice here. We take the underlying process dynamics to be

$$\begin{aligned} dz &= \kappa(\theta - z)dt + \sigma_z dw_z \\ d\theta &= \mu\, dt + \sigma_\theta\, dw_\theta \end{aligned} \tag{5.124}$$

with $dw_z dw_\phi = \rho\, dt$. The log-price $z$ is a mean-reverting process with a nonstationary stochastic mean. These features are fairly common in energy markets. An example is the generation stack in electricity markets. The marginal generation unit (that sets the market heat rate) is mean reverting: demand tends to be driven by fundamental, mean-reverting entities such as temperature (so that during heat waves, say, the market-clearing heat rate is set by a unit higher up on the stack, and when the heat wave subsides the marginal unit comes back down the stack).[47] However, there are also long-term, non-stationary capital effects governing the overall growth of the stack (that is, the level to which the marginal unit reverts). Due to capital substitution effects across the economy as a whole, we anticipate that the marginal heat and long-term mean are negatively correlated: $\rho < 0$. (See EW for a full discussion of the generation stack in power modeling.)

Although (5.124) is a linear, Gaussian process, it is still nonstandard and it is not immediately clear what type of probabilistic behavior it gives rise to. However,

characteristic function methods prove quite helpful in deriving this structure. The characteristic function $f = E_t e^{i\phi z_T}$ satisfies

$$f_t + \kappa(\theta - z)f_z + \mu f_\theta + \tfrac{1}{2}\sigma_z^2 f_{zz} + \rho\sigma_z\sigma_\theta f_{z\theta} + \tfrac{1}{2}\sigma_z^2 f_{\theta\theta} = 0 \tag{5.125}$$

Looking for a solution of the form $f = \exp(\alpha z + \beta\theta + \gamma)$ we get the following ensemble:

$$\begin{aligned} \dot{\alpha} - \kappa\alpha &= 0 \\ \dot{\beta} - \kappa\alpha &= 0 \\ \dot{\gamma} + \mu\beta + \tfrac{1}{2}\sigma_z^2\alpha^2 + \rho\sigma_z\alpha\beta + \tfrac{1}{2}\sigma_\theta^2\beta^2 &= 0 \end{aligned} \tag{5.126}$$

with $\alpha(T) = i\phi$, $\beta(T) = \gamma(T) = 0$. These ODEs are readily solved to get

$$\begin{aligned} \alpha &= i\phi e^{-\kappa\tau} \\ \beta &= i\phi(1 - e^{-\kappa\tau}) \\ \gamma &= i\phi\mu\left(\tau - \frac{1 - e^{-\kappa\tau}}{\kappa}\right) - \frac{1}{2}\phi^2\left(\tilde{\sigma}^2\frac{1 - e^{-2\kappa t}}{2\kappa} + 2(\rho\sigma_z - \sigma_\theta)\sigma_\theta\frac{1 - e^{-\kappa t}}{\kappa} + \sigma_\theta^2\tau\right) \end{aligned} \tag{5.127}$$

where $\tilde{\sigma}^2 = \sigma_z^2 - 2\rho\sigma_z\sigma_\theta + \sigma_\phi^2$. (Note that this entity is reminiscent of the ratio volatility that plays a central role in spread option pricing.) From the quadratic structure (in $\phi$) of (5.127), we see that indeed the process $z$ is Gaussian. Of primary interest here is the volatility structure, identifiable as the quadratic term in the expression for $\gamma$ in (5.127). This can be seen to possess features of both a mean-reverting process (asymptoting variance) and a nonstationary process (linearly growing variance). Typical behavior of the volatility can be seen in Figure 5.2.

Figure 5.2 is somewhat mislabeled as showing a "term structure," which is not really meaningful as (5.124) is better thought of as a spot process, in which case this figure is really showing a variance scaling law (annualized in terms of volatility). However, we will see shortly how the implied forward dynamics are readily obtained from such models, and that the term "structure" is indeed captured here. The main point to be taken here is that the arresting of the volatility drop-off (the so-called Samuelson effect) and eventual increase and leveling off is not an illusory effect and is commonly seen in commodity markets, an exemplification being power-gas (forward) heat rates. (It can be discerned in the examples from Section 3.1.3) Again, this effect is due to the capital structure of the particular market, and manifests itself over different time horizons in different markets. For example, the effect typically appears around maturities of two to three years in electricity markets, and four to five years in gas transport markets. We again call attention to a

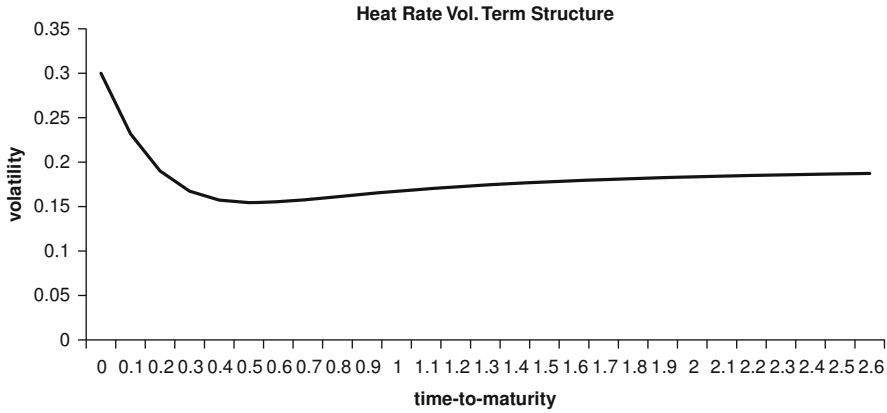



**Figure 5.2** Volatility term structure for mixed stationary/non-stationary effects

theme that will be prevalent throughout this work, namely the idea of volatility as representing information flows over different time scales. This example here, although quite simple, nonetheless captures important features of actually existing markets.

### 5.2.6 Application to change of measure

*5.2.6.1 The quintessential formula, revisited*

We have already seen in the derivation of (5.96) how change-of-measure techniques facilitated the underlying argument by identifying the two constituent expectations as being special cases of a more general computation. This categorization was possible because of the ability to exploit the particular form of the characteristic function of the underlying process. Affine jump diffusions permit a rather general use of change of measure, which, as we will see later, can prove quite useful in various numerical applications. In fact, let us consider a generalization of the quintessential option pricing formula of Section 5.1.

The basic problem we are interested in is the following:

$$E_t^Q e^{a_k z_k(T)} 1(K_{ij} z_j(T) \leq L_i) \tag{5.128}$$

This can be written as

$$E_t^Q e^{a_k z_k(T)} E_t^{Q_a} 1(K_{ij} z_j(T) \leq L_i) \tag{5.129}$$

where the new measure $Q_a$ is defined through the RN derivative

$$\frac{dQ_a}{dQ} = \frac{e^{a_k z_k(T)}}{E_t^Q e^{a_k z_k(T)}} \tag{5.130}$$

Consequently, the problem in (5.129) requires knowledge of the distribution of the matrix random vector $\tilde{z}(T) = Kz(T)$ under the measure $Q_a$. To determine this, we consider

$$f = E_t^{Q_a} e^{i\phi_k \widehat{z}_k(T)} = \frac{E_t^Q e^{(a_j + i\phi_k K_{kj}) z_j(T)}}{E_t^Q e^{a_i z_i(T)}} \tag{5.131}$$

Now, for the class of affine jump diffusions, both of these expectations can be obtained. We have

$$\begin{aligned} E_t^Q e^{(a_j + i\phi_k K_{kj}) z_j(T)} &= e^{\alpha_k' z_k + \alpha_0'} \\ E_t^Q e^{a_i z_i(T)} &= e^{\alpha_k z_k + \alpha_0} \end{aligned} \tag{5.132}$$

where $\alpha'$ and $\alpha$ both satisfy the system (5.84) with the following terminal conditions:

$$\begin{aligned} &\alpha_k'(T) = a_k + i\phi_j K_{jk}, \alpha_0'(T) = 0 \\ &\alpha_k(T) = a_k, \alpha_0(T) = 0 \end{aligned} \tag{5.133}$$

Now, in some sense, we are done. We have the characteristic function of $\tilde{z}$ under $Q_a$ via

$$E_t^{Q_a} e^{i\phi_k \hat{z}_k(T)} = e^{(\alpha_k' - \alpha_k) z_k + \alpha_0' - \alpha_0} \tag{5.134}$$

and thus we can calculate (in principle) the expectation $E_t^{Q_a} 1(\tilde{z}_j(T) \leq L_i)$ by a fairly straightforward extension of the approach taken in the one-dimensional case. We will have much more to say about some of the issues involved in the Chapter 7, when we will see how appropriate use of contour integration can be of great utility. Here we briefly give an indication of the essence of the calculation in two dimensions.

#### *5.2.6.2 Higher dimensions*

Consider the following expectation:

$$E1(z_1 < \gamma_1, z_2 < \gamma_2) \tag{5.135}$$

Assuming we know the characteristic function $f$ of the joint process, (5.135) can be written

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dz_1 dz_2 1(z_1 < \gamma_1, z_2 < \gamma_2) \frac{1}{(z\pi)^2} \int_{\Gamma_1} \int_{\Gamma_2} d\phi_1 d\phi_2 f(\phi_1, \phi_2) e^{-i\phi_1 z_1 - i\phi_2 z_2} \tag{5.136}$$

where the contours $\Gamma_{1,2}$ are chosen to permit interchanging the orders of integration:

$$E1(z_1 < \gamma_1, z_2 < \gamma_2) = \frac{1}{(z\pi)^2} \int_{\Gamma_1} \int_{\Gamma_2} d\phi_1 d\phi_2 f(\phi_1, \phi_2) \frac{e^{-i\phi_1\gamma_1}}{-i\phi_1} \frac{e^{-i\phi_2\gamma_2}}{-i\phi_2} \tag{5.137}$$

which is entirely analogous to (5.92). Of course, by moving the contours back to the real $\phi$ -axes and taking into account the poles, a form similar to (5.93) can be obtained. Again, though, we will see that this step is neither necessary nor desirable.

*5.2.6.3 Restricted forms*

Although this general result is definitely useful, it would still be of interest to say something about specific structures induced by measure changes in (5.130). Here we must consider some special cases owing to the fundamentally nonlinear nature of the dynamics in (5.76), in contrast to the linear, Gaussian case. A natural choice would be to take $K = I$ (the identity matrix), so that (5.131) amounts to determining the dynamics of the original process under the change of measure. We look for a solution to (5.132) of the form

$$\alpha'_k = \alpha_k + \beta_k, \alpha'_0 = \alpha_0 + \beta_0 \tag{5.138}$$

with $\beta_k(T) = i\phi_k$, $\beta_0(T) = 0$. Now using (5.84), the system of equations for $\alpha'$ become

$$\begin{aligned} \dot{\alpha}_k + \dot{\beta}_k + A_{ik}(\alpha_i + \beta_i) + \tfrac{1}{2} X^k_{ij}(\alpha_i + \beta_i)(\alpha_j + \beta_j) + \lambda_{ik} K_i(\alpha + \beta) = 0 \\ \dot{\alpha}_0 + \dot{\beta}_k + b_{ik}(\alpha_i + \beta_i) + \tfrac{1}{2} X^0_{ij}(\alpha_i + \beta_i)(\alpha_j + \beta_j) + \lambda_{i0} K_i(\alpha + \beta) = 0 \end{aligned} \tag{5.139}$$

We also have that

$$\begin{aligned} K_i(\alpha + \beta) &= E^Q(e^{(\alpha_m + \beta_m)j_{mi}} - 1) \\ &= E^Q e^{\alpha_m j_{mi}}(e^{\beta_m j_{mi}} - 1) + E^Q(e^{\alpha_m j_{mi}} - 1) \\ &= E^Q e^{\alpha_m j_{mi}} E^{Q^i_\alpha}(e^{\beta_m j_{mi}} - 1) + E^Q(e^{\alpha_m j_{mi}} - 1) \end{aligned} \tag{5.140}$$

with the new measures $Q^i_\alpha$ defined through the RN derivatives

$$\frac{dQ^i_\alpha}{dQ} = \frac{e^{\alpha_m j_{mi}}}{E^Q e^{\alpha_m j_{mi}}} \tag{5.141}$$

Using the fact that $\alpha$ satisfies the system (5.84), the ensemble (5.139) becomes

$$\begin{aligned}\dot{\beta}_k + \beta_i(A_{ik} + X^k_{ij}\alpha_j) + \tfrac{1}{2}X^k_{ij}\beta_i\beta_j + \tilde{\lambda}_{ik}\tilde{K}_i(\beta) = 0\\ \dot{\beta}_0 + \beta_i(b_i + X^0_{ij}\alpha_j) + \tfrac{1}{2}X^0_{ij}\beta_i\beta_j + \tilde{\lambda}_{i0}\tilde{K}_i(\beta) = 0\end{aligned} \tag{5.142}$$

where under the induced measure change we have

$$\begin{aligned}\tilde{\lambda}_{ik} &= \lambda_{ik}E^Q e^{\alpha_m j_{mi}}\\ \tilde{K}_i(\beta) &= E^{Q^i_\alpha}(e^{\beta_m j_{mi}} - 1)\end{aligned} \tag{5.143}$$

Consequently, the characteristic function (5.134) is of the form $\exp(\beta_k z_k + \beta_0)$ where $\beta$ satisfies the same kind of Riccati system as in (5.84). Thus, the process $z$ remains an affine jump diffusion under the new measure $Q_a$. However, the drift and jump dynamics are changed. The drift changes as

$$\begin{aligned}A_{ik} &\to A_{ik} + X^k_{ij}\alpha_j\\ b_i &\to b_i + X^0_{ij}\alpha_j\end{aligned} \tag{5.144}$$

and the jump intensities and distribution of jump amplitudes changes are laid out in (5.141) and (5.143). (Note, not surprisingly, that the covariance structure is unchanged.) We thus see great commonality with, and much generalization of, the results for the pure diffusive case (*e.g.*, in (5.42)) as well as the Lévy case (*e.g.*, in (5.208)). There is a complication, however: the new dynamics will now have explicit time dependence, as the entities $\alpha$ are of course nonconstant in the general case. Similarly, the induced jump measure change in (5.141) will also be impacted. However, we have arrived at our desired result: the measure change (5.130) retains the affine structure of the underlying system. In Chapter 7 we will address the question of how, numerically, the kinds of binary option valuations entailed in (5.128) can be obtained. The point to be taken now is that, armed with the ability to carry out such calculations, more general expectations can be similarly obtained via change of measure.

### 5.2.7 Spot and implied forward models

It should be clear that a great many features of actually existing energy markets can be incorporated in the affine class of processes in (5.76), chief among them mean reversion and jumps (but also, obviously, stochastic volatility). These features arise due to various physical and operational constraints that impact energy markets. Examples include mean-reverting temperature driving demand for electricity, as well as equilibrating supply/demand constraints along the generation stack. In addition,

the actual form of this stack gives rise to jumps during conditions of stressed supply. As usual, we direct the reader to EW for a complete discussion. The main point we wish to emphasize here is that these physical constraints refer to effects at the *spot* level of price dynamics. However, any valuation/hedging problem can only take place in terms of (financially traded) forwards or futures. We do not propose here to investigate in any thorough sense the relationship between spot and futures prices. We only note here that there *is* such a relationship, and we explain how the class of affine jump diffusions offers a convenient framework for illustrating this relationship, as well as the critical question of what kinds of value drivers arise from different kinds of hedging strategies around futures.

#### 5.2.7.1 *Affine dynamics*

So we start by specifying a spot process with dynamics under two measures, $P$ and $Q$, where, in line with standard notation, $P$ represents the measure governing the actual, observed data-generating process (*i.e.*, the physical measure) and $Q$ represents a measure under which expectations relevant for valuation and hedging takes place (*i.e.*, the pricing measure). We further assume that the spot process remains affine under both of these measures. We thus write the dynamics in (5.76) as

$$dz_i = (A_{ij}^P z_j + b_i^P)dt + \sigma_i^k \sqrt{z_k} dw_i^k + \sigma_i^0 dw_i^0 + j_{ik} dq_k \tag{5.145}$$

under the physical measure and

$$dz_i = (A_{ij}^Q z_j + b_i^Q)dt + \sigma_i^k \sqrt{z_k} dw_i^k + \sigma_i^0 dw_i^0 + j_{ik} dq_k \tag{5.146}$$

under the pricing measure. The covariance structure is of course unchanged, while the jump intensities change so that $\lambda_i^P = \lambda_{i0}^P + \lambda_{ij}^P z_j$ under $P$ and $\lambda_i^Q = \lambda_{i0}^Q + \lambda_{ij}^Q z_j$ under $Q$. In addition, the jump amplitudes differ under the two measures, and we indicate this through the (pseudo) moment generating function $K_i^{P,Q}(\alpha) = E^{P,Q}(e^{\alpha_{mj} mi} - 1)$. Now, we write the futures price as the expectation of spot under the pricing measure:[48]

$$F_{t,T} = E_t^Q S_T \tag{5.147}$$

Working in terms of log assets, for those entities which represent log prices we have

$$F_{t,T}^i = E_t^Q e^{z_i(T)} \tag{5.148}$$

Since $z$ is an affine process, the log-forward price $f$ (we drop the superscript $i$ for convenience) satisfies

$$f = \alpha_i^Q z_i + \alpha_0^Q \tag{5.149}$$

where $\alpha^Q$ satisfies a system such as (5.84) and the superscript $Q$ highlights the fact that the parameters of the $Q$–process are being used. Generically, we have that the

dynamics of the log forward are

$$df = (\dot{\alpha}_k^Q z_k + \dot{\alpha}_0^Q)dt + \alpha_i^Q dz_i \tag{5.150}$$

Since there are different spot dynamics under the two measures, we clearly see there are likewise different forward dynamics, as well. Using (5.84), we see that (5.150) can be written as

$$df = -\begin{pmatrix} z_k(A_{ik}^Q \alpha_i^Q + \frac{1}{2}X_{ij}^k \alpha_i^Q \alpha_j^Q + \lambda_{ik}^Q K_i^Q(\alpha^Q)) + \\ b_i^Q \alpha_i^Q + \frac{1}{2}X_{ij}^0 \alpha_i^Q \alpha_j^Q + \lambda_{i0}^Q K_i^Q(\alpha^Q) \end{pmatrix} dt + \alpha_i^Q dz_i \tag{5.151}$$

Now, using the different dynamics under $P$ and $Q$, we see that (5.151) becomes

$$\begin{aligned} df = &-\tfrac{1}{2}\alpha_i^Q(X_{ij}^Q z_k + X_{ij}^0)\alpha_j^Q dt \\ &- K_i^Q(\alpha^Q)(\lambda_{ik}^Q z_k + \lambda_{i0}^Q)dt \\ &+ \alpha_i^Q((A_{ik}^M - A_{ik}^Q)z_k + b_i^M - b_i^Q)dt \\ &+ \alpha_i^Q(\sigma_i^k \sqrt{z_k} dw_i^k + \sigma_i^0 dw_i^0 + j_{ik} dq_k) \end{aligned} \tag{5.152}$$

where $M$ can denote either $P$ or $Q$. Some specific examples will be most useful in illustrating how particular behavior in spot is manifested in the forwards. However, even at an abstract level several important points can be discerned.

*5.2.7.2 Ramifications for pricing and valuation*

First, note that the log forward retains the affine structure under both measures. Of course, the forward can be obtained directly via (5.149), but there are often applications where the dynamics are of interest and these are handily provided by (5.152). In particular, the usual affine machinery for option valuation is completely applicable here (with the minor complication that there is now explicit time dependence involved). An example would be monthly storage valuation, where spread options involving monthly legs of differing times to maturity are used as hedging instruments. Specifically, some injection (say) decisions will be made in conjunction with some decisions to withdraw in the future, and the latter kind of decision will involve entering into (short) forward positions that will have nonzero times to maturity. The basic valuation structure is thus of the form

$$E_t^Q(F_{T_1,T_2} - F_{T_1,T_1})^+ \tag{5.153}$$

with $T_2 > T_1 > t$.[49] Thus we see the need for understanding the "incremental" (so to speak) dynamics of $F$, which can be obtained from (5.152). We further note here that the typical dichotomy one often sees between forward-based and spot-based

valuations is in truth largely a false one. The fact that forward and spot prices *must* be related (albeit usually in a more complicated manner than (5.148)) means that there must be some kind of relationship between respective valuations in terms of them. One reason (there are others) that the different approaches give rise to different values is the nature of the resolution of the two entities: forwards typically apply over lower resolution time blocks than spot. One advantage of the affine framework in (5.146) and (5.152) is that examples can be constructed in which the temporal resolution is the same for both forward and spot representations, and the convergence of the two valuations can be demonstrated. (We considered such examples in Chapter 4.)

Second, we can see how the dynamics in (5.152) depend critically on which measure is being considered. Under $Q$, the linear drift terms vanish and the jump compensator terms are precisely such that the log forward is an exponential $Q$–martingale, as expected from (5.148). In fact, we explicitly have

$$\begin{aligned}\frac{dF}{F} &= (K_i^M(\alpha^Q)(\lambda_{ik}^M z_k + \lambda_{i0}^M) - K_i^Q(\alpha^Q)(\lambda_{ik}^Q z_k + \lambda_{i0}^Q))dt \\ &\quad + \alpha_i^Q((A_{ik}^M - A_{ik}^Q)z_k + b_i^M - b_i^Q)dt \\ &\quad + \alpha_i^Q(\sigma_i^k\sqrt{z_k}dw_i^k + \sigma_i^0 dw_i^0) + (e^{\alpha_i^Q j_{ik}} - 1)dq_k \end{aligned} \tag{5.154}$$

Now consider the forward dynamics under the physical ($P$) measure. In this case, the drift terms do *not* vanish.[50] Both the linear drift and the jump compensator terms are present in (5.154). This means that the evolution of the forwards under the physical measure will inherit (at least in time-modulated form) many of the characteristics of the physical spot process. In particular, there can be something like mean-reverting effects in financially traded forwards, even though there is *no* explicit mechanism for effecting such mean reversion! It should go without saying that this situation can have enormous effects on valuation, depending on the kind of hedging strategy that is adopted (and hence, what kind of exposures are created). Put differently, the kinds of value drivers that are relevant can be very different depending on the particular hedging strategy employed.

#### *5.2.7.3 Example: variance scaling laws*

Some specific examples will serve to illustrate these points. Consider a mean-reverting log price, with respective $P$– and $Q$–dynamics:

$$\begin{aligned} dz &= \kappa^P(\theta^P - z)dt + \sigma\, dw \\ dz &= \kappa^Q(\theta^Q - z)dt + \sigma\, dw \end{aligned} \tag{5.155}$$

with $\kappa^Q \neq \kappa^P$. The log forward is easily seen to be

$$f_{t,T} = ze^{-k^Q(T-t)} + \theta^Q(1 - e^{-k^Q(T-t)}) + \frac{\sigma^2}{4k^Q}(1 - e^{-2k^Q(T-t)}) \tag{5.156}$$

The $Q$–dynamics of the forward are

$$\frac{dF_{t,T}}{F_{t,T}} = \sigma\, e^{-k^Q(T-t)}\, dw \tag{5.157}$$

while the $P$–dynamics are

$$\frac{dF_{t,T}}{F_{t,T}} = e^{-\kappa^Q(T-t)}(\kappa^P\theta^P - \kappa^Q\theta^Q - (\kappa^P - \kappa^Q)z)dt + \sigma\, e^{-\kappa^Q(T-t)}\, dw \tag{5.158}$$

There are a few things of note here. First, the evolution of the forward in (5.157) gives rise to a term structure for volatility:

$$\tilde{\sigma}_{t,T} = \sigma\sqrt{\frac{1 - \exp(-2k^Q(T-t))}{2\kappa^Q(T-t)}} \tag{5.159}$$

This model thus illustrates the well-known Samuelson effect that is often seen in commodity markets, namely the run-up in volatility as time to maturity decreases. This behavior is a reflection of the general nature of volatility as a measure of accumulation of information over particular intervals of time. The intuition in this case is that, due to mean reversion, significant movements in spot are of diminishing importance for the determination of future spot (and hence expectation of future spot) as the time horizon under consideration increases. Alternatively, significant movements are of increasing importance for expectations of future spot over smaller time horizons (as there is correspondingly less time for those movements to equilibrate). Indeed, in terms of the rate of change of variance with respect to time we have

$$\frac{\partial v_{t,T}}{\partial t} = \sigma^2 e^{-2k^Q(T-t)} \tag{5.160}$$

indicating how the flow of information increases incrementally the closer one is to expiration. Note of course that a critical factor here is the strength of mean reversion; as $\kappa^Q \downarrow 0$ the term structure of volatility flattens and the accumulation of information is uniform over all time horizons.

The next thing to see is that, under the physical measure, the dynamics of the forward inherit many of the properties of the spot. In this simple model, since the relationship between log forward and log spot is affine (via (5.156)), by simple substitution it can be seen that the $P$–dynamics of the forward in (5.158) are mean

reverting, albeit with some time modulation. This fact has obvious significance as far as the value that can be collected from different hedging strategies with the forward. In general, a static hedging strategy will collect the cumulative variance under the physical measure, while dynamic hedging is necessary to extract the quadratic variation given by (5.159). Although the affine structure makes it straightforward to determine the cumulative variance of the forward over any time interval, it is actually trivial to calculate this variance to expiry, since $F_{T,T} = E_T^Q e^{z_T} = e^{z_T}$. Thus, the cumulative variance over the total term is just the cumulative variance of spot under the physical measure, which for convenience we annualize as a volatility:

$$v_{t,T} = \sigma \sqrt{\frac{1 - \exp(-2\kappa^P (T-t))}{2\kappa^P (T-t)}} \tag{5.161}$$

Since in general $\kappa^Q < \kappa^P$, the term structures in (5.159) and (5.161) will be quite different.[51] An example is shown in Figure 5.3.

The effect shown in Figure 5.3 is not simply a theoretical curiosity, but is in fact observed in actual commodity markets. We saw real examples and discussed these kinds of phenomena in greater detail in Section 3.1.3. Having laid out a specific example (additional ones can be found in the Appendix to this chapter), we now turn to a more general formalization of these ideas.

### 5.2.8 Fundamental drivers and exogeneity

As we have stressed in Chapter 3, the particular characteristics of commodity markets make the distinction between static and dynamic hedging strategies critically important. Specifically, we pointed out that the particular nature of fundamental relationships between traded assets determines which of those assets must be dynamically hedged to extract optimal value of a structured product depending on all

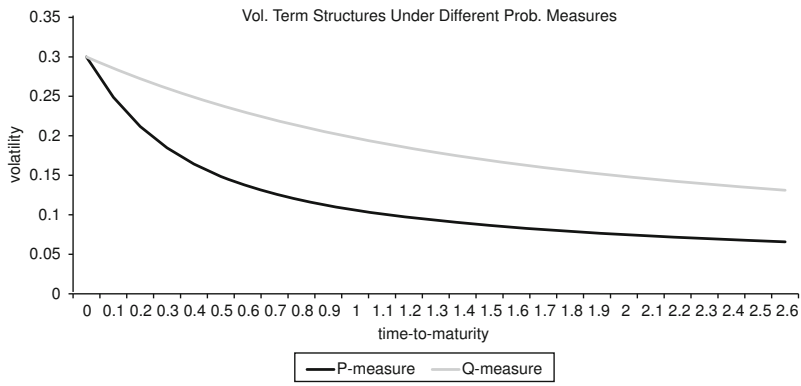



**Figure 5.3** Volatility term structure for static vs. dynamic hedging strategies

of the assets. In this section we will describe some of the features of energy price models that give rise to such effects. We should note here that we will revisit (and in fact expand upon) these points in the next chapter when we discuss econometric issues.

#### 5.2.8.1 *Background examples*

First, let us review (from Section 3.1.3) a basic model of the joint (spot) dynamics of (log) gas and power prices:

$$\begin{aligned} dg &= \mu_g dt + \sigma_g dw_g \\ dh &= k_h(\theta - h)dt + \sigma_h dw_h \\ p &= h + g \end{aligned} \tag{5.162}$$

Again, the basic intuition is clear: gas prices are essentially diffusive input costs, with the ultimate product determined by supply and demand factors (namely, the generation stack and weather, respectively) that are mean reverting.[52] (In general there is some [typically negative] instantaneous correlation between gas and [market-clearing] heat rate.) We have previously noted that the equilibrating relationship between power and gas prices impacts the value that can be extracted from dynamically as opposed to statically trading those assets. However, there is another point to be raised regarding the structure of the simple model in (5.162): specifically, it can be seen that the equilibrating relationship[53] between power and gas is driven by the heat rate, which is not itself dependent on that relationship. In the parlance of economics, the heat rate is an *exogenous* variable (in common language terms, it is a variable that is in some sense external to a model, and may be thought of as given) while power and gas are *endogenous* variables (again in common language, the variables that the model is trying to explain).

For further elucidation, consider the Schwartz-type model in (5.124), where the mean-reversion level is itself stochastic. The basic, fundamental relationship between power and gas in (5.162) still holds, even though the overall dynamics of the heat rate are very different. Indeed, in this case the heat rate is itself subject to an equilibrating relationship reflecting the capital structure of the *entire* economy (*i.e.*, not just the commodity markets). Consider also the important role temperature and weather effects play in the formation of prices. Clearly such effects help determine the equilibrium level of prices (over the appropriate time scale) but are not themselves determined by this equilibrium state.

#### 5.2.8.2 *Formalization*

To formalize some of these points, consider the following model (motivated by Pesaran *et al.* [2000]):[54]

$$d\begin{pmatrix} x \\ y \end{pmatrix} = \left[\begin{pmatrix} b_x \\ b_y \end{pmatrix} + \begin{pmatrix} A_{xx} & 0 \\ A_{xy} & A_{yy} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\right] dt + \begin{pmatrix} \sigma_x \cdot dw_x \\ \sigma_y \cdot dw_y \end{pmatrix} \tag{5.163}$$

where $\cdot$ denotes Hadamard (element-by-element) matrix multiplication. In (5.163) $x$ is an $m$-dimensional vector of exogenous variables and $y$ is an $n$-dimensional vector of endogenous variables. Note that the exogenous variables can be analyzed independently of the endogenous variables, but not *vice versa* (of course). Clearly, models such as (5.162) and (5.220) can be fit into the framework of (5.163). We can construct additional log-price processes via affine relationships of the form

$$z = c_x^T x + c_y^T y \tag{5.164}$$

We thus have a generalization of the basic power-gas-heat rate model of (5.162). Typically, (5.163) and (5.164) refer to spot relationships, and of course we are primarily concerned with forward relationships since much of the portfolio formation (both static and dynamic) that is the backbone of valuation will be conducted in terms of forward-traded entities (such as futures).

*5.2.8.3 Underlying structure: more variance scaling laws*

Thus, we consider expectations of the form[55]

$$F_{t,T} = E_t e^{z_T} \tag{5.165}$$

Plainly, the characteristic function methods we have developed in this chapter are ideally suited for evaluating expressions such as (5.165). Let us stress that the system (5.163) is, owing to its essential linearity, jointly Gaussian and so (again invoking linearity) expectations such as (5.165) can be computed in a number of ways. We appeal to characteristic functions here because they greatly facilitate these computations, and because they have application across a wide range of (non-Gaussian) processes. So, looking for a solution of the form $F_{t,T} = \exp(\gamma_x^T x + \gamma_y^T y + \gamma_0)$, we obtain the following system of ODEs:[56]

$$\begin{aligned} &\dot{\gamma}_x + A_{xx}^T \gamma_x + A_{yx}^T \gamma_y = 0, \quad \gamma_x(T) = c_x \\ &\dot{\gamma}_y + A_{yy}^T \gamma_y = 0, \quad \gamma_y(T) = c_y \end{aligned} \tag{5.166}$$

Following the approach laid out in Section 5.2.7, we see that the forward dynamics (under the pricing measure) are given by

$$\frac{dF_{t,F}}{F_{t,T}} = \gamma_x^T \sigma_x \cdot dw_x + \gamma_y^T \sigma_y \cdot dw_y \tag{5.167}$$

(Compare with (5.215).) Although it will not be possible to provide an analytic solution to (5.167), some important general points can be made.

As is well known, a linear system of ODEs can in principle be solved by the following transformation (although in practice it is generally not a good idea to do so; see the entertaining account in Moler and van Loan [2003]):

$$\dot{x} = Ax = VJV^{-1}x \Rightarrow \dot{y} = Jy \tag{5.168}$$

where $y = V^{-1}x$ and $V$ is the matrix of generalized eigenvectors of $A$, with $J$ the corresponding Jordan block form of eigenvalues.[57] It can thus be seen from (5.166) that the modulation factors acting on the volatility terms in (5.167) will depend on the eigenvalue/eigenvector structure of the matrices $A$ in the (spot) dynamics (5.163). In fact, the forward dynamics will in general depend on the interplay of the eigenstructures, so to speak, owing to the (inhomogeneous) term $A_{yx}^T \gamma_y$ in the first set of equations in (5.166). A specific example has already been encountered in (5.215). We have already discussed the reality of the volatility term structure in commodity markets, and we have here a general model that gives rise to this (Samuelson) effect, by relating that structure to the underlying (joint) behavior of both fundamental (that is, non-price) drivers and other price components. But of course (as we have emphasized), volatility is a measure of information accumulation over specific intervals of time, and thus we also have here an illustration of how information accumulation over different time scales that is specific to various physical drivers (as encapsulated by their underlying eigenvalue/eigenvector structure) manifests itself in the dynamics of financial contracts that settle against functions of those drivers.

The importance of these dynamics (specifically, the time scales over which different effects combine) can be seen when we consider the evolution under the physical measure. Again following Section 5.2.7, we find that the physical dynamics are given by

$$\begin{aligned}\frac{dF_{t,F}}{F_{t,T}} = &\begin{pmatrix} \gamma_x^T(b_x^P - b_x^Q + (A_{xx}^P - A_{xx}^Q)x) + \\ \gamma_y^T(b_y^P - b_y^Q + (A_{yx}^P - A_{yx}^Q)x + (A_{yy}^P - A_{yy}^Q)y) \end{pmatrix} \\ &+ \gamma_x^T \sigma_x \cdot dw_x + \gamma_y^T \sigma_y \cdot dw_y \end{aligned} \tag{5.169}$$

where superscripts $P$ and $Q$ denote the (in general different) drift terms under the physical and pricing measures, respectively. Compare with (5.158).[58] We have stressed repeatedly the different value that accrues from static as opposed to dynamic hedging strategies, and (5.169) illustrates the manner in which the relevant time scales manifest themselves in the drift of forward prices. In general, the precise nature of the dynamics will be complicated and depend on the eigenstructure of the

underlying drivers, but at the heart of the process are the operative time scales of these drivers.

We have already said much about the critical importance of time scales in Chapter 2, in particular on the interplay of prices and their fundamental drivers operating on differing time scales.[59] The discussion here has served to present these concepts (which we have encountered several times already) from the angle of techniques ordinarily applied to change-of-measure techniques (namely, via characteristic functions). We have to this point emphasized the *difference* in value that accrues from different hedging strategies (*e.g.*, static vs. dynamic) in the formation of specific portfolios around structured products. We now revisit the subject of Chapter 3, namely the *representation* of value and how, again in terms of trading strategies, valuation depends on the construction of portfolios from which particular kinds of exposure are created (or more accurately, how one kind of exposure is replaced by another, presumably preferable, exposure[60]).

### 5.2.9 Minimal martingale applications

In this section we provide a more detailed discussion of issues previously raised in Section 3.2 on pricing in incomplete markets via the minimal martingale measure (MMM).

#### *5.2.9.1 Continuous time/affine case*

We start here with a process following our familiar affine jump diffusion (*sans* jumps), under the physical ($P$) measure (recall (5.76) from Section 5.2.3):

$$dz_i = (A_{ij}z_j + b_i)dt + \sigma_i^k\sqrt{z_k}dw_i^k + \sigma_i^0 dw_i^0 \tag{5.170}$$

where as usual the summation convention over repeated indices not on the LHS is adopted and we again take $dw_i^k dw_j^l = \delta_{kl}\rho_{ij}^k$ and adopt the notation $X_{ij}^k \equiv \rho_{ij}^k\sigma_i^k\sigma_j^k$. We assume there are $N$ state variables, of which the first $N' \leq N$ represent log prices of traded assets. We now look for a Radon-Nikodym (RN) process[61] of the following form:

$$d\xi = -\tfrac{1}{2}Mdt + \alpha_i^k\sigma_i^k\sqrt{z_k}dw_i^k + \alpha_i^0\sigma_i^0 dw_i^0 \tag{5.171}$$

where $M = z_k\alpha_i^k X_{ij}^k\alpha_j^k + \alpha_i^0 X_{ij}^0\alpha_j^0$, so $\xi$ is an exponential $P$–martingale This form is chosen for convenience; in general the factors $\alpha$ will be state dependent (*i.e.*, the RN process will not itself be affine). Now, we are interested in the state dynamics under a new measure $Q$ defined by

$$\frac{dQ}{dP} = e^{\xi_T - \xi} \tag{5.172}$$

To derive these dynamics, consider the characteristic function of $z(T)$ under $Q$:

$$f = E_t^Q e^{i\phi_k z_k(T)} = e^{-\xi} E_t^P e^{\xi_T + i\phi_k z_k(T)} \tag{5.173}$$

Thus, $\tilde{f} \equiv f e^{\xi}$ is a $P$–martingale and so satisfies[62]

$$\begin{aligned} &\tilde{f}_t + (A_{ij} z_j + b_i)\tilde{f}_{z_i} + \tfrac{1}{2}(z_k X_{ij}^k + X_{ij}^0)\tilde{f}_{z_i z_j} \\ &\quad - \tfrac{1}{2} M\tilde{f}_{\xi} + \tfrac{1}{2} M\tilde{f}_{\xi\xi} + (z_k X_{ij}^k \alpha_j^k + X_{ij}^0 \alpha_j^0)\tilde{f}_{z_i \xi} = 0 \end{aligned} \tag{5.174}$$

By substitution, we find that the coefficient of $f_{z_i}$, and hence the $Q$–drift, is given by

$$A_{ik} z_k + b_i + z_k X_{ij}^k \alpha_j^k + X_{ij}^0 \alpha_j^0 \tag{5.175}$$

Thus, the condition that tradeables are $Q$–martingales becomes

$$A_{ik} z_k + b_i + z_k X_{ij}^k \alpha_j^k + X_{ij}^0 \alpha_j^0 + \tfrac{1}{2}(z_k X_{ii}^k + X_{ii}^0) = 0, 1, \ldots, N' \tag{5.176}$$

The first thing to note is that in general (*i.e.*, for $N' < N$), there is no unique solution to the equations in (5.176), indicating again the general non-uniqueness of equivalent martingale measures (EMM) in incomplete markets. To specify a choice, we consider the MMM. This is defined by the change of measure under which martingales that are orthogonal to tradeables under $P$ remain martingales under $Q$. Orthogonality here means that the product of two martingales (or more generally the martingale components [*e.g.*, arising from Doob-Meyer] of two random variables) is itself a martingale, or equivalently, that they have zero quadratic covariation. As will be seen, this amounts to requiring that martingales that are orthogonal to tradeables are also orthogonal to the RN process.

*5.2.9.2 MMM conditions*

So, consider the following exponential $P$–martingale:

$$dx = -\tfrac{1}{2} M' dt + \beta_i^k \sigma_i^k \sqrt{z_k} dw_i^k + \beta_i^0 \sigma_i^0 dw_i^0 \tag{5.177}$$

Orthogonality with tradeables implies that

$$z_k X_{ij}^k \beta_j^k + X_{ij}^0 \beta_j^0 = 0, \; i = 1, \ldots, N' \tag{5.178}$$

Now, we have to require that $x$ is an exponential $Q$–martingale. We can proceed as we did in (5.175) for the change of measure. The adjustment to the $P$–drift of $x$ is given by the coefficient of the $x\xi$ cross term of a similar PDE to the one in

(5.174). But this adjustment must be zero to ensure that $x$ remains an exponential $Q$–martingale, so we have that

$$z_k \alpha_i^k X_{ij}^k \beta_j^k + \alpha_i^0 X_{ij}^0 \beta_j^0 = 0 \tag{5.179}$$

These requirements mean that we must have, $\forall k$, that

$$\begin{aligned} \alpha_i^k &= \alpha_i^0, i = 1, \ldots, N' \\ \alpha_i^k &= 0, i = N' + 1, \ldots, N \end{aligned} \tag{5.180}$$

It is worth noting the form the dynamics of the RN process in (5.171) now take:

$$d\xi = -\tfrac{1}{2} M dt + \alpha_i^0 (\sigma_i^k \sqrt{z_k} dw_i^k + \sigma_i^0 dw_i^0) \tag{5.181}$$

In other words, the randomness driving this process is a linear combination of the various sources of randomness driving each traded state variable. The condition (5.176) that tradeables be $Q$–martingales becomes

$$(z_k X_{ij}^k + X_{ij}^0)\alpha_j^0 + A_{ik} z_k + b_i + \tfrac{1}{2}(z_k X_{ii}^k + X_{ii}^0) = 0, \; i = 1, \ldots, N' \tag{5.182}$$

and we have a unique solution (as many equations as unknowns). After these equations are solved, the $Q$–dynamics of the non-traded state variables (*e.g.*, stochastic variance) can be obtained.

*5.2.9.3 Heston-type examples*

As an example, consider (yet again) Heston:

$$\begin{aligned} dz &= (\mu - v/2)dt + \sqrt{v} dw_1 \\ dv &= \kappa(\theta - v)dt + \sigma \sqrt{v} dw_2 \end{aligned} \tag{5.183}$$

There is only one tradeable, and the martingale condition becomes

$$v\alpha_1^0 + \mu = 0 \tag{5.184}$$

from which the $Q$–drift of stochastic variance is given by

$$\kappa(\theta - v) + v\rho\sigma\alpha_1^0 = \kappa\left(\theta - \frac{\rho\sigma\mu}{\kappa} - v\right) \tag{5.185}$$

so that the physical mean reversion rate is unchanged, but the level is adjusted. Note that if the tradeable is already a $Q$–martingale ($\mu = 0$), or if stochastic variance

is orthogonal to price ($\rho = 0$), then the $Q$–dynamics of variance are unchanged under the MMM. In fact, affinity is not necessarily retained under the MMM, as the following "augmented Heston" example shows. Consider this process:[63]

$$\begin{aligned} dz &= (\mu - (\nu X_{11}^2 + X_{11}^0)/2)dt + \sigma_1^2\sqrt{\nu}dw_1^2 + \sigma_1^0 dw_1^0 \\ dz &= \kappa(\theta - \nu)dt + \sigma_2^2\sqrt{\nu}dw_2^2 + \sigma_2^0 dw_2^0 \end{aligned} \tag{5.186}$$

In this case, the stochastic variance $Q$–drift is given by

$$\kappa(\theta - \nu) - \mu\frac{\nu X_{21}^2 + X_{21}^0}{\nu X_{11}^2 + X_{11}^0} \tag{5.187}$$

and only for rather uninteresting cases (*e.g.*, Heston-only or deterministic volatility) will the structure remain affine.

#### *5.2.9.4 Optimal hedge ratios*

An important question, obviously, concerns the nature of the hedges that the MMM corresponds to. We have already seen (in (3.63)) the notion of the MMM as a measure under which orthogonal projections of the value function onto the space of tradeables produce residuals that are (zero mean) martingales under *both* physical and pricing measure. This property permitted (formal) application of EMM pricing techniques. There is in fact another interpretation that is more closely tied to a portfolio optimization problem, specifically variance minimization.

We will consider rather general (diffusive) dynamics, with tradeables denoted by $S$ and non-traded state variables denoted by $x$. The portfolio dynamics around some structured product (with value function $V$) and positions $-\Delta$ in the tradeables can be written as

$$d\Pi = (V_t + \tfrac{1}{2}\mathfrak{L}V)dt + (V_{S_i} - \Delta_i)dS_i + V_{x_i}dx_i \tag{5.188}$$

(Recall (3.68).) Now, taking the value function $V$ to be an expectation under some (as yet unspecified) pricing measure $Q$, we take hedges of the form $\Delta_i = V_{s_i} + \Gamma_{ij}V_{xj}$ for some (also as yet unspecified) matrix $\Gamma$ (of less than full rank). Using the PDE solved by $V$ (recall the example in (3.79)), (5.188) can be written as

$$\begin{aligned} d\Pi &= -\mu_{x_i}^Q V_{x_i}dt - \Gamma_{ij}V_{x_j}dS_i + V_{x_i}dx_i \\ &= V_{x_i}(\mu_{x_i}^P - \mu_{x_i}^Q - \Gamma_{ji}\mu_{S_j}^P)dt + V_{x_i}(\sigma_{x_i}dw_{x_i} - \Gamma_{ji}\sigma_{S_j}dw_{S_j}) \end{aligned} \tag{5.189}$$

The notation in (5.189) should be clear: both entities (tradeables and non-tradeables) have some drift terms under both measures,[64] and some (correlated) diffusive structure. The requirement that the hedge residual in (5.189) have zero expectation/drift under the physical measure can be written in matrix form as

$$\Gamma^T\mu_S^P = \mu_x^P - \mu_x^Q \tag{5.190}$$

Obviously, this condition is not sufficient to determine the hedge component associated with sensitivity to non-tradeables given the pricing measure, nor is it very useful in selecting or even characterizing the pricing measure given some hedge component. We must turn to some other criteria to seek a connection between a pricing measure and a hedging regime.

A natural choice is minimum residual variance. We can only hope to attain tractable results with a local measure, since the cumulative variance involves an integration with "vega" weighting $V_x$. From (5.189), this requires that

$$\Gamma = X_{SS}^{-1} X_{Sx} \tag{5.191}$$

in terms of the process covariances. (This is essentially a regression analysis.) While this argument fixes the (optimal) hedge volume, it in general leaves (5.190) violated, and more importantly still says nothing about the nature of the pricing measure. However, it should be clear from generalizations of (5.175) and (5.182) that the MMM is precisely that measure for which the solution (5.191) implies (5.190).[65] In other words, MMM is the local variance optimizing measure for which the residual error has expectation zero under the physical measure.

As an example, the MMM-optimal hedge for Heston is given by

$$\Delta_{\text{opt}} = V_S + \frac{\rho\sigma}{S} V_v \tag{5.192}$$

using (5.184) and (5.185). This expression has obvious parallels with the familiar vega-adjusted delta hedge in the presence of volatility skew.

#### 5.2.9.5 *Connection with entropy*

An interesting result is that for these (affine) processes, we can show that the MMM corresponds to the Minimal Entropy Measure (MEM) (in general the MMM corresponds to the minimal *reverse* entropy measure). That is, we consider measure changes that minimize the entity

$$E^P\left(\frac{dQ}{dP}\log\frac{dQ}{dP}\right) = E^Q\left(\log\frac{dQ}{dP}\right) \tag{5.193}$$

(The entropy measure will be seen to have econometric applications in Section 6.5.) Now, the drift of $\xi$ under $Q$ simply changes sign, and we are faced with minimizing

$$E_t^Q(\xi_T - \xi) = E_t^Q \int_t^T \tfrac{1}{2}(z_k \alpha_i^k X_{ij}^k \alpha_j^k + \alpha_i^0 X_{ij}^0 \alpha_j^0) ds \tag{5.194}$$

such that

$$A_{ik}z_k + b_i + z_k X^k_{ij}\alpha^k_j + X^0_{ij}\alpha^0_j + \tfrac{1}{2}(z_k X^k_{ii} + X^0_{ii}) = 0,\ i = 1,\ldots,N' \tag{5.195}$$

Note that since the RN process is not in general affine, we cannot resort to the usual machinery associated with such processes. However, we note that the integrand in (5.194) is a quadratic form, so we can require that it be minimized subject to the martingale constraint for tradeables. In particular, note that the integrand can be written in matrix form as

$$((\alpha^1)^T \cdots (\alpha^0)^T) \begin{pmatrix} z_1 X^1 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & X^0 \end{pmatrix} \begin{pmatrix} \alpha^1 \\ \vdots \\ \alpha^0 \end{pmatrix} \tag{5.196}$$

Now, using Lagrange multipliers, it is easy to show that the solution to the following minimization problem:

$$\begin{aligned} &\min \tfrac{1}{2} x^T A x \\ &\text{st } Cx = b \end{aligned} \tag{5.197}$$

with $A$ symmetric, positive definite, is given by

$$x_* = A^{-1}C^T(CA^{-1}C^T)^{-1}b \tag{5.198}$$

In our case, $A$ is block-diagonal with inverse

$$\begin{pmatrix} (z_1 X^1)^{-1} & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & (X^0)^{-1} \end{pmatrix} \tag{5.199}$$

and $C$ has the form

$$\begin{pmatrix} (z_1 X^1)_1 & \cdots & (X^0)_1 \\ (z_1 X^1)_2 & \cdots & (X^0)_2 \\ \vdots & \vdots & \vdots \end{pmatrix} \tag{5.200}$$

where the subscripts denotes rows of the corresponding submatrices. Thus, $A^{-1}C^T$ has the form

$$\begin{pmatrix} e_1 & e_2 & \cdots \\ e_2 & e_2 & \cdots \\ \vdots & \vdots & \end{pmatrix} \tag{5.201}$$

where $e_i$ is a unit vector of dimension $N$ with $e_{ij} = \delta_{ij}$. Thus, the optimal allocation will have the form

$$\begin{pmatrix} \hat{\alpha}^1 \\ \vdots \\ \hat{\alpha}^0 \end{pmatrix} \tag{5.202}$$

with $\hat{\alpha}_i^k = \hat{\alpha}_i^0$ for each $k$ and fixed $i = 1, \ldots, N'$ and $\hat{\alpha}_i^k = 0$ for fixed $k$ and $i > N'$. But this is exactly the same form as the condition for the MMM.[66]

Having established a general framework and methodology for valuation in terms of portfolio formation (implicitly entailing specific trading strategies), we now turn attention to an important facet of structured products that is pervasive in energy markets, namely the fact that many such structures are associated with the state of some system (*e.g.*, a power plant being on or off), and that valuation must optimize (in the appropriate sense) with regards to decisions that affect that state.

## 5.3 Appendix

### 5.3.1 More Asian option results

#### *5.3.1.1 Extension to Lévy processes*

Recall the discussion from Section 5.1.5, on Asian option symmetries for Brownian processes. Note that the crux of the argument developed there is two-fold: namely the ability to employ characteristic functions (to elucidate the change of measure employed in (5.51)), and the independence of increments of the underlying process (that enabled us to conclude that the average was retained in the payoff via (5.55)). It is reasonable to speculate whether the fixed-floating symmetry holds for more general processes for which each of those two features is present. We are of course referring to Lévy processes. In terms of log prices, we recall the basic argument laid out in (5.49) and (5.51):

$$\begin{aligned} V_{\text{float}}^{\text{call}} &= e^{-r(T-t)} E_t^Q e^{z_T} \left( k - \frac{1}{T-t} \int_t^T e^{z_s - z_T} ds \right)^+ \\ &= e^{-r(T-t)} E_t^Q e^{z_T} E_t^{Q_z} \left( k - \frac{1}{T-t} \int_t^T e^{z_s - z_T} ds \right)^+ \end{aligned} \tag{5.203}$$

with the measure change given by $\frac{dQ_z}{dQ} = \frac{e^{z_T}}{E_t^Q e^{z_T}}$. We are again confronted with the question: what is the distribution of the entity $\tilde{z}_{t,T} \equiv z_t - z_T$ under this measure change? To answer this question, recall the basic result for the characteristic

function of a Lévy process from Section 5.2.1:

$$E_t^Q e^{i\phi z_T} = \exp\left(z + (T-t)\left(i\phi a - \frac{1}{2}\phi^2\sigma^2 + \int_{\mathbb{R}\setminus\{0\}} (e^{i\phi x} - 1 - i\phi x h(x))\mu(dx)\right)\right) \tag{5.204}$$

Note that the requirement that the log-asset $z$ be an exponential $Q$–martingale[67] becomes

$$a + \frac{1}{2}\sigma^2 + \int_{\mathbb{R}\setminus\{0\}} (e^x - 1 - xh(x))\mu(dx) = r - q \tag{5.205}$$

We introduce the notation $J(\phi) \equiv \int_{\mathbb{R}\setminus\{0\}} (e^{i\phi x} - 1 - i\phi x h(x))\mu(dx)$. Proceeding as before, we seek the characteristic function of $\tilde{z}$ under the new measure:

$$\begin{aligned}
E_t^{Q_z} e^{i\phi\tilde{z}_{s,T}} &= e^{-z-(r-q)(T-t)} E_t^Q e^{(1-i\phi)z_T + i\phi z_s} = e^{-z-(r-q)(T-t)} E_t^Q e^{i\phi z_s} E_s^Q e^{(1-i\phi)z_T} \\
&= e^{-z-(r-q)(T-t)} E_t^Q e^{z_s(T-s)((1-i\phi)a + \frac{1}{2}(1-i\phi)^2\sigma^2 + J(-\phi-i))} \\
&= e^{-(r-q)(T-t)+(T-s)((1-i\phi)a+\frac{1}{2}(1-i\phi)^2\sigma^2+J(-\phi-i))+(s-t)(a+\frac{1}{2}\sigma^2+J(-i))} \\
&= e^{(T-s)(-(r-q)+(1-i\phi)a+\frac{1}{2}(1-i\phi)^2\sigma^2+J(-\phi-i))} \\
&= e^{(T-s)(-(r-q)+a+\frac{1}{2}\sigma^2-i\phi(a+\sigma^2)-\frac{1}{2}\phi^2\sigma^2+J(-\phi-i))}
\end{aligned} \tag{5.206}$$

Now, the term involving $J$ in the exponent in final entity in (5.206) appears to dash our hopes here, but this proves to not be the case. For we have

$$\begin{aligned}
J(-\phi - i) &= \int_{\mathbb{R}\setminus\{0\}} (e^{(1-i\phi)x} - 1 + (-1 + i\phi)xh(x))\mu(dx) \\
&= \int_{\mathbb{R}\setminus\{0\}} (e^x e^{-i\phi x} - e^x + e^x - 1 + (-1 + i\phi)xh(x))\mu(dx) \\
&= \int_{\mathbb{R}\setminus\{0\}} (e^x(e^{-i\phi x} - 1) + i\phi x h(x))\mu(dx) + \int_{\mathbb{R}\setminus\{0\}} (e^x - 1 - xh(x))\mu(dx)
\end{aligned} \tag{5.207}$$

Thus (5.206) becomes

$$E_t^{Q_z} e^{i\phi\tilde{z}_{s,T}} = \exp((T-s)(i\phi\tilde{a} - \frac{1}{2}\phi^2\sigma^2 + \int_{\mathbb{R}\setminus\{0\}} (e^{i\phi x} - 1 - i\phi x\tilde{h}(x))\mu(dx))) \tag{5.208}$$

where $\tilde{h}(x) = e^x h(-x)$, $\tilde{\mu}(dx) = e^{-x}\mu(-dx)$[68], and $\tilde{a} = -a - \sigma^2$. Therefore, the log-difference process involved in the option payoff in (5.203) remains a Lévy process under the measure $Q_z$. The main impact of the measure change is to adjust the linear drift (as in the usual Brownian case) *and* to modify the Lévy measure as specified above.[69] Note that the expected value is

$$E_t^{Q_z} e^{\tilde{z}_{s,T}} = \exp((T-s)(-a - \frac{1}{2}\sigma^2 + \int_{\mathbb{R}\setminus\{0\}} (1 - e^{-x} - xh(-x))\mu(dx))) = e^{(q-r)(T-s)} \tag{5.209}$$

So, as happened with the standard GBM case, the role of interest rate and dividend rate are switched. We thus see that the fixed-floating symmetry that existed in the Black-Scholes environment *also* prevails in the much more general Lévy case.[70] Of course, these symmetry results, while very interesting, do not provide any guidance on the actual pricing of Asian options under Lévy processes.[71] (Obviously, they do reduce the scope of the problem to the valuation of *either* fixed or floating strike options.) On this issue, there appear to be few viable results, apart from rather dissatisfactory approximations such as moment matching/Edgeworth expansions. We refer the reader to recent work by Cai and Kou (2011); see also Albrecher (2004).

#### *5.3.1.2 Further extensions*

It is worth exploring briefly the possibility of dropping at least the independent increments assumption in these arguments. Ultimately, we are concerned with evaluating expectations of the following form:

$$E_t^{Q_z} e^{i\phi(z_s - z_T)} = \frac{E_t^Q e^{(1-i\phi)z_T + i\phi z_s}}{E_t^Q e^{z_T}} = \frac{E_t^Q e^{i\phi z_s} E_s^Q e^{(1-i\phi)z_T}}{E_t^Q e^{z_T}} \tag{5.210}$$

Now, if we assume that the characteristic function of $z$ takes the form $E_t^Q e^{i\phi z_T} = e^{\alpha(T-t;\phi)z + \beta(T-t;\phi)}$, then (5.210) can be written as

$$\frac{e^{\beta(T-s;\phi)} E_t^Q e^{(i\phi + \alpha(T-s;\phi))z_s}}{e^{\alpha(T-t;-i)z + \beta(T-t;-i)}} \tag{5.211}$$

The expectation in the numerator in (5.211) has the form of a characteristic function, although it clearly is not. Still, it raises hope that there may be classes of processes rich enough that certain structures are preserved under measure changes to make such computations in (5.211) feasible.[72] We now begin investigating such questions.

### 5.3.2 Further change-of-measure applications

*5.3.2.1 Additional energy market applications*

It is worth considering the case introduced in (5.124), of a mean-reverting spot with a stochastic (non-stationary) mean. In fact, to illustrate our continuing theme of volatility over different time scales, we make the following modification (to the $Q$–dynamics):

$$\begin{aligned} dz &= \kappa_z(\theta - z)dt + \sigma_z dw_z \\ d\theta &= \kappa_\theta(\bar{\theta} - \theta)dt + \sigma_\theta dw_\theta \end{aligned} \tag{5.212}$$

so that the stochastic mean is now also mean reverting, but we may suppose it mean reverts at a (much) slower rate than the log-asset itself (*i.e*,. $\kappa_\theta \ll \kappa_z$). The forward $F_{t,T} = E_t^Q e^{z_T}$ has the form $\exp(\alpha(t)z + \beta(t)\theta + \gamma(t))$ where the relevant ODEs are

$$\begin{aligned} \dot{\alpha} - \kappa_z \alpha &= 0 \\ \dot{\beta} - \kappa_\theta \beta + \kappa_\theta \alpha &= 0 \end{aligned} \tag{5.213}$$

with $\alpha(T) = 1, \beta(T) = 0$. (Compare with (5.126).) The solution is readily found to be

$$\begin{aligned} \alpha &= e^{-\kappa_z(T-t)} \\ \beta &= -\frac{\kappa_z}{\kappa_z - \kappa_\theta}(e^{-\kappa_z(T-t)} - e^{-\kappa_\theta(T-t)}) \end{aligned} \tag{5.214}$$

(Of course for $\kappa_\theta = 0$ the structure in (5.127) is recovered.) It is also not hard to see that the forward dynamics (again, under $Q$) are given by

$$\frac{dF_{t,T}}{F_{t,T}} = \sigma_z e^{-\kappa_z(T-t)} dw_z - \sigma_\theta \frac{\kappa_z}{\kappa_z - \kappa_\theta}(e^{-\kappa_z(T-t)} - e^{-\kappa_\theta(T-t)})dw_\theta \tag{5.215}$$

In this case, the forward dynamics has a term with (local) volatility modulated by a monotonically decreasing (as a function of time-to-maturity) exponential function as in (5.157), due to mean reversion in (log) spot. Now, however, there is a second term due to the stochasticity of the mean reversion level. For nonzero mean reversion rate of the stochastic level (*i.e.*, $\kappa_\theta \neq 0$), the modulation of volatility ($\beta$ in (5.214)) is *not* monotone, but has a general shape like that shown in Figure 5.4.

For nonzero $\kappa_\theta$ the modulation factor increases initially and then peaks before asymptoting to zero. (Note that for the non-stationary mean case, where $\kappa_\theta = 0$, there is no peak and the modulation factor is monotonically increasing, aymptoting to one.) The location of the peak (as a function of time to maturity) depends on the relative sizes of the mean reversion rates. The interpretation is as follows. If the

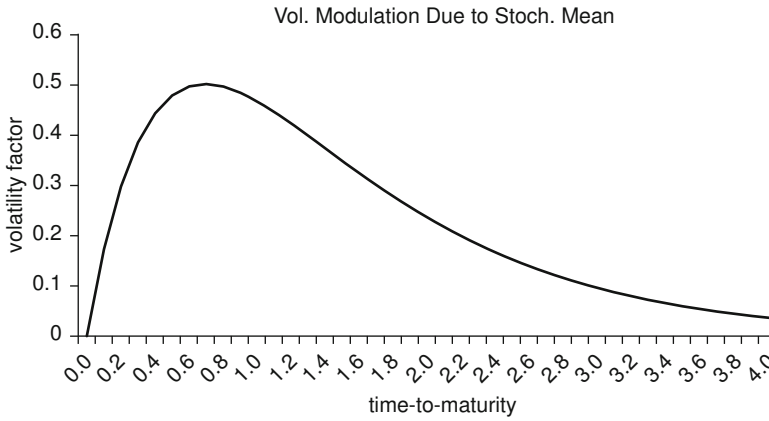



**Figure 5.4** Volatility modulation factor for mean-reverting stochastic mean

mean reversion levels moves, then the expectation of the level to which spot reverts will in general also move in the same direction. However, if this movement in the mean happens sufficiently far from maturity, then (due to mean reversion in the stochastic mean), the stochastic mean will tend to revert to its equilibrium level $\bar{\theta}$ over the term, in which case the expectation of spot will experience little movement (since the expectation is that spot will still tend to come near the equilibrium level). This is to be contrasted with the non-stationary mean case, where movements of the stochastic mean impart *permanent* changes in expectations.

Another example is the forward process for the Heston model in (5.117). It proves more interesting to modify the original Heston process to include mean reversion (under the $Q$ measure):

$$\begin{aligned} dz &= \left(\chi - \kappa_z z - \tfrac{1}{2}v\right) dt + \sqrt{v} dw_z \\ dv &= \kappa(\theta - v)dt + \sigma\sqrt{v} dw_v \end{aligned} \tag{5.216}$$

The forward price $F_{t,T} = E_t^Q e^{z_T}$ satisfies

$$F_t + (\chi - \kappa_z z - \tfrac{1}{2}v)F_z + \kappa(\theta - v)F_v + \tfrac{1}{2}vF_{zz} + \rho\sigma vF_{zv} + \tfrac{1}{2}v\sigma^2 f_{vv} = 0 \tag{5.217}$$

Using a solution of the form $F = \exp(\alpha(t)z + \beta(t)v + \gamma(t))$ we find (not surprisingly) that $\alpha = e^{-\kappa_z(T-t)}$ and that the $Q$–dynamics are given by

$$\frac{dF_{t,T}}{F_{t,T}} = e^{-\kappa_z(T-t)}\sqrt{v} dw_z + \beta\sigma\sqrt{v} dw_z \tag{5.218}$$

where $\beta$ satisfies (making the substitution $\tau = T - t$)

$$\dot{\beta} = (\rho\sigma e^{-\kappa_z \tau} - \kappa)\beta + \tfrac{1}{2} e^{-k_z \tau}(e^{-\kappa_z \tau} - 1) \tag{5.219}$$

with $\beta(0) = 0$. We now see why the introduction of mean reversion makes the problem interesting. If $\kappa_z = 0$ the inhomogeneous term in (5.219) vanishes, and given the initial condition it is easy to see that $\beta = 0$. Thus, the dynamics in (5.218) are just the Heston spot dynamics, which is not surprising since under the $Q$ measure $z$ is an exponential martingale. Thus (again, not surprisingly), there is no volatility term structure unless there is mean reversion in spot. Unfortunately, for $\kappa_z \neq 0$ the equation (5.219) does not appear to be solvable analytically (although it can of course be transformed to a linear second-order ODE like any other one-dimensional Riccati equation), as in the regular Heston case. Still, it presents no problems numerically and we can discern the relevant features of the forward dynamics. Typical behavior is shown in Figure 5.5.

Note that there some similarities with the case of a mean-reverting stochastic mean considered above, and in fact there is a similar interpretation of the effect. Sufficiently close to maturity (where "sufficiently" is dependent on the relative mean reversion rates), a big move in stochastic variance will take a while to revert to its long-term mean. Thus, over that term the (log) spot will be correspondingly more volatile, which for a mean-reverting process means a greater probability of overshooting/undershooting *its* long-term mean. The aggregate effect is greater volatility around the expectation of terminal spot. (This effect is of course absent in the case of non-stationary spot, hence as expected $\beta = 0$ and there is no contribution to variability of expected spot from variability of stochastic variance as such – *i.e.*, an effect dependent on $\sigma$.)

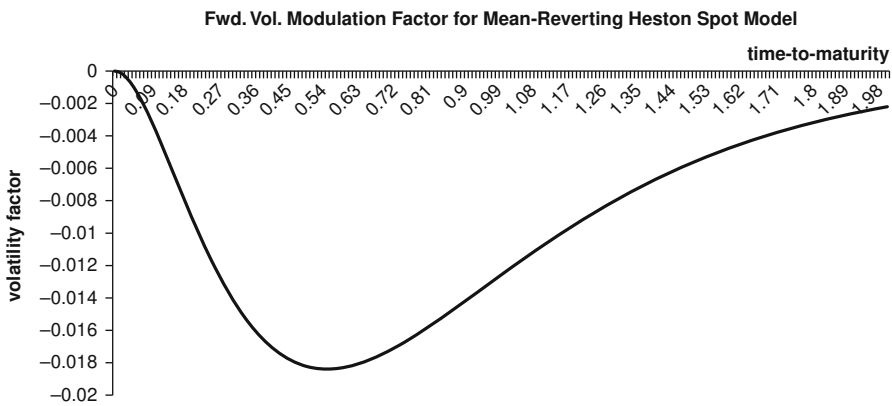



**Figure 5.5** Forward volatility modulation factor for stochastic variance in a mean-reverting spot model

As a final example, consider the following model of spikes motivated by Hikspoors and Jaimungal (2007). The log-(spot) price $z$ is decomposed into two components, one a mean-reverting diffusive process, the other a pure jump process with level-dependent drift (under measure $Q$):

$$\begin{aligned} z &= y + 1 \\ dy &= \kappa_y(\theta - y)dt + \sigma\, dw \\ dl &= -\kappa_l l dt + j dq \end{aligned} \tag{5.220}$$

with the jump term being a standard, Merton-type Poisson process with normally distributed amplitudes.[73] The intuition here is clear: the log asset (say, a log-heat rate) is "usually" mean reverting and continuously valued, but during "extreme" conditions (such as during outages) the heat rate can (briefly) spike up and (quickly) come back down. The forward price is given by $F_{t,T} = E_t^Q e^{y_T + l_T}$ and satisfies

$$F_t + \kappa_y(\theta - y)F_y - \kappa_l l F_l + \tfrac{1}{2}\sigma^2 F_{yy} + \lambda E^Q(F(l+j) - F) = 0 \tag{5.221}$$

It is easy to show that $F_{t,T} = \exp(ye^{-\kappa_y(T-t)} + le^{-\kappa_l(T-t)} + \alpha(T-t))$[74] and that the $Q$–dynamics are given by

$$\frac{dF_{t,F}}{F_{t,F}} = -\lambda(E^Q e^{je^{-\kappa_l(T-t)}} - 1)dt + \sigma\, e^{-\kappa_y(T-t)} dw + (e^{je^{-\kappa_l(T-t)}} - 1)dq \tag{5.222}$$

# 6 Econometric Concepts

## 6.1 Cointegration and mean reversion

As we endeavored to emphasize throughout Chapter 2, the reality of non-stationarity (in light of relevant time scales) demands that great care be exercised in analyzing energy-market data (challenging in its own right in light of the hammer and anvil of data sparsity/high volatility). In fact, one may reasonably wonder if there is *any* extent to which the well-established tools developed for stationary time series have relevance to real-world data. However, it turns out that there is an important concept, known as cointegration, which allows for the investigation of *stationary* relationships between entities that are individually non-stationary. In some sense it is probably not surprising that such a concept has viability, especially in energy markets. After all, there are a number of fundamental drivers (*e.g.*, weather-driven demand, stack structure, *etc.*) that place certain physical constraints on (joint) price formation. Obviously, there are different time scales over which these effects occur, but the relevant effect is to constrain the extent to which certain price *relationships* (*e.g.*, price-gas ratios) can vary. This is the essence of a cointegrating relationship, and we will now discuss some of the specific techniques that can be brought to bear on such problems. In particular, we point out how they relate to certain, more robust, methods we have already presented.

### 6.1.1 Basic ideas

First, we present the very basic concepts. A non-stationary variable $y$ is said to be integrated of order 1 (denoted by $I(1)$) if its first differences ($\Delta y_t \equiv y_t - y_{t-1}$) are stationary. A vector of non-stationary processes is said to be cointegrated if there is some linear combination of them that *is* stationary. Since a linear combination of stationary variables is of course stationary, in general a cointegrating relationship (if it exists) will not be unique; at a minimum some sort of normalization or identification in terms of one of the entities can be imposed. In addition, there can in general be more than one cointegrating relationship.[1] Let us now consider some of these issues in a bit more operational detail.

*6.1.1.1 Common stochastic drivers*

Consider an $N$–dimensional process $y$ which is non-stationary and $I(1)$. The process is said to be cointegrated if there exists an $N \times h$ (nonzero) matrix $A$ (with $h \leq N$) such that each component of $z \equiv A^T y$ is stationary. We will assume that the columns of $A$ are linearly independent (that is, the rank of $A$ is $h$). There are said to be $h$ cointegrating relationships, corresponding to the columns of $A$. (An obvious special case is $h = N$ , in which case the process $y$ is in fact stationary.) The actual relationships are not unique, as stationarity is persevered under the transformation $A \to AC^T$ for any $h \times h$ (nonzero) matrix $C$. This fact can be used to write the cointegrating relationships in a useful form. Write the matrix $A$ as[2]

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \tag{6.1}$$

where $A_1$ is $h \times h$ and $A_2$ is $(n-h) \times h$. Now, choose $C$ st. $CA_1^T = I$ (the $h \times h$ identity matrix) and let $\Gamma \equiv -A_1^{-T} A_2^T$. Then, partitioning $y$ in the obvious way, the cointegrating relationship can be expressed as

$$z_t = \tilde{z}_t + Ez_t = (I \quad -\Gamma) \begin{pmatrix} y_{1t} \\ {\scriptstyle h \times h} \\ y_{2t} \\ {\scriptstyle (n-h) \times h} \end{pmatrix} \tag{6.2}$$

Since $z$ is stationary, it is meaningful to write the unconditional expectation in (6.2) (so that $\tilde{z}$ is a zero mean stationary process).

We now write

$$y_{1t} = \Gamma y_{2t} + Ez_t + \tilde{z}_t \tag{6.3}$$

Since the first difference of $y$ is stationary, we anticipate that we can write the following dynamics:

$$\Delta y_{2t} = \mu_{2t} + \varepsilon_{2t} \tag{6.4}$$

where the drift $\mu_{2t}$ can be thought of as the *conditional* expectation of $\Delta y_{2t}$ and $\varepsilon_{2t}$ can be thought of as some white-noise process. Thus, in this representation the process $y$ consists of two different types of entities: $n - h$ purely non-stationary components (*e.g.*, diffusions with drift, as in (6.4)) and $h$ (non-stationary) components that stand in a linear relationship with the purely non-stationary drivers (as in (6.3)), such that the resulting residual is stationary. The entities $y_2$ can be characterized as common stochastic trends. These are fundamentally different from the probably more familiar case of deterministic trending (*e.g.*, trend-stationary linear growth), although in any finite sample the two effects can appear very similar.

It is important to keep these two categories distinct in any econometric analysis. For example, running standard regressions on non-stationary time series can lead

to very misleading results. Evidence can be found for relationships that do not exist. Let us consider an example.

#### *6.1.1.2 Spurious regressions I: when zero is not zero*

Following the old adage that a picture is worth a thousand words, we will illustrate two seemingly similar scenarios with very different econometric ramifications. Assume we have two processes (call them $x$ and $y$) that are unrelated. Process $x$ is non-stationary, say, a random walk/Brownian motion. Now, in the first scenario, $y$ is simply (stationary) white noise. In the second scenario, $y$ is also a (non-stationary) Brownian motion. In both cases $y$ is independent of $x$. In plain language terms, there is no relation between the two processes, and we would, perhaps naively, expect familiar econometric analysis to bear out this impression. It turns out our naiveté would be exposed.

To demonstrate, we simulate two pairs of time series: (1) a random walk and an independent white noise, and (2) two independent random walks. We further have two different sample sizes for each pair, 500 and 2,000. For each of these 4 combinations, we generate 200 scenarios, and compute the standard OLS regression coefficient between the constituent series. Our uninformed intuition would anticipate that (a) the estimator would be distributed about zero (since there is no underlying relation) and (b) the estimator would become more narrowly dispersed with increasing sample size. It turns out that we would be (partly) wrong: the intuition holds true only in the former case (of white noise regressed against a random walk). Figures 6.1 and 6.2 make clear the distinction between the two cases.[3]

Figure 6.1 indicates that the estimator, although noisy, is indeed distributed about zero, *and* that it becomes less noisy as the underlying sample size increases. These two points are actually not unrelated. It is precisely because of the fact that the estimator variance decreases (as the sample grows) that we can speak of it as converging

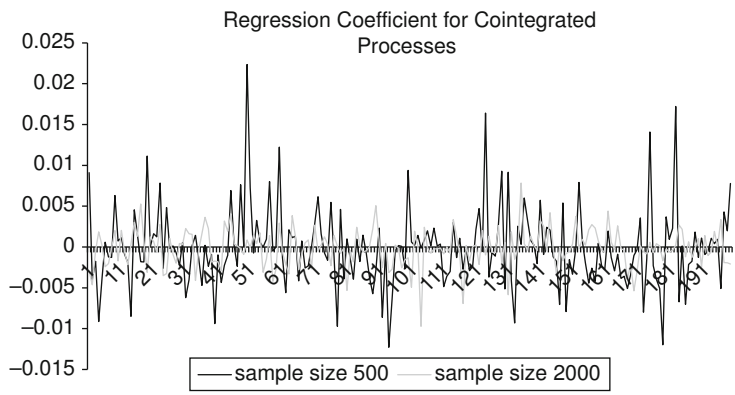



**Figure 6.1** OLS estimator, "cointegrated" assets

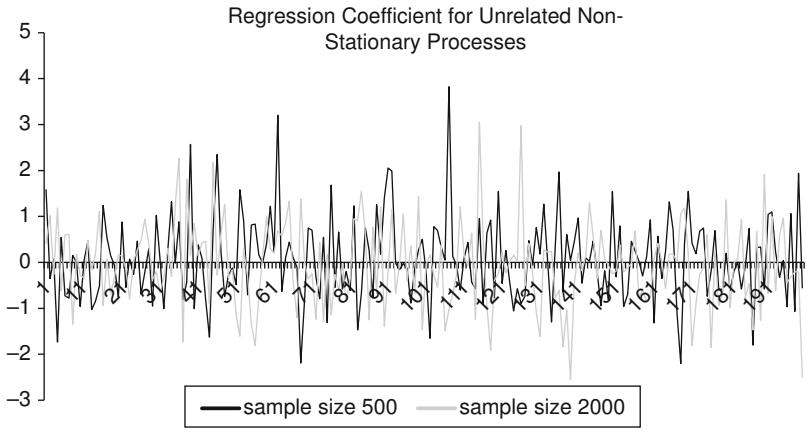



**Figure 6.2** OLS estimator, non-cointegrated assets

(in either the common language sense or the appropriate strict probabilistic sense) to something (presumably, an actual population entity). In fact, the variance of the estimator decreases proportionately with sample size, as conventional diagnostics would predict (*i.e.*, the standard deviation for the 2,000-size sample is half of that for the 500 sample size). Put differently, the probability of the estimator producing a numerical value very far from zero (and thus giving the [false] impression of a real relationship) decreases as more data points are added.

Figure 6.2 provides a good illustration of these points by contrast. Although the estimator appears distributed about zero, this is misleading. Visually, the distribution of the estimator does not seem very different between the two sample sizes. There does not appear to be any sense of convergence of the estimator. This would be problematic in its own right, but given the (again, visually) much greater variability of the estimator in this case than the previous one, concerns can (and should) actually be raised that almost *any* numerical value could be produced by the estimator for a given sample, and potentially falling within the range for which conventional diagnostics would (completely falsely) identify the result as indicative of a real population property. Unlike the previous case, there is in fact no reason to think that adding information (by increasing the sample size) will decrease the probability that we are misled into thinking we have identified a real relationship.

These concerns prove to be well founded, and we will now turn to the theoretical explanation.

#### *6.1.1.3 Spurious regressions II: a bit of theory*

We will elucidate the underlying theory with a very simple example. Consider (again) the following correlated Brownian motion:

$$\begin{aligned} \Delta x_t &= \sigma_x \varepsilon_t^x \\ \Delta y_t &= \sigma_y \varepsilon_t^y \end{aligned} \tag{6.5}$$

with the disturbances $\varepsilon$ having unit variance and correlation $E\varepsilon_t^x \varepsilon_t^y = \rho$. Clearly, the dependency structure in (6.5) is described by the correlation between the differences. However suppose that, this fact unbeknownst to us, we decide to apply OLS to some realization of the time series $x$ and $y$ (*i.e.*, we look for a relationship between levels instead of returns). That is, we look to estimate a model of the following form:[4]

$$y_t = \gamma\, x_t + u_t \tag{6.6}$$

with $u$ some assumed (normal) disturbance. The OLS estimator is given by

$$\hat{\gamma} = \frac{\langle x_t y_t \rangle}{\langle x_t^2 \rangle} \tag{6.7}$$

Now, paralleling the argument used in the analysis of unit root tests (*e.g.*, (2.22)), the large sample limit of (6.7) can be written as

$$\hat{\gamma} \sim \frac{\sigma_y}{\sigma_x} \frac{\int_0^1 w_s^x w_s^y ds}{\int_0^1 (w_s^x)^2 ds} \tag{6.8}$$

where $w_t^{x,y}$ are correlated Brownian motions. By decomposing in terms of conditional relationships between $x$ and $y$, (6.8) can be written as

$$\hat{\gamma} \sim \frac{\sigma_y}{\sigma_x} (\rho + \sqrt{1-\rho^2} \frac{\int_0^1 w_s^x v_s ds}{\int_0^1 (w_s^x)^2 ds}) \tag{6.9}$$

where $v$ is a Brownian motion independent of $w^x$.

We see an important point here: (6.9) does *not* involve explicit dependence on the sample size, as in, say, the (stationary) AR(1) case in (2.21). This implies that, for an arbitrary realization of (6.5), *no matter how large the sample*, the OLS estimator will have a nonvanishing probability of lying in *any* arbitrary interval. In other words, the estimator does not converge to anything. This situation should be contrasted with the stationary case, where the probability of the estimator lying arbitrarily far from the true model values is vanishingly small (as the sample size increases).

For example, for the uncorrelated case ($\rho = 0$), there is obviously no relation between the two time series. Yet, the OLS estimator can lie arbitrarily far from zero, no matter how large the sample size. This is the essence of spurious regressions: statistically significant results for an inconsistent estimator. Naively running regressions on non-stationary time series can yield evidence of relationships that simply do not exist. (See Proposition 18.2 in Hamilton [1994] for the multidimensional generalization of (6.9).)

#### *6.1.1.4 A sneak peek at error-correction models*

Note that there is a rather trivial case where OLS *does* yield consistent results here: the case of perfect correlation or anti-correlation (*i.e.*, $\rho = \pm 1$). Here, the underlying relationship is simply $y_t = \frac{\sigma_y}{\sigma_x} x_t$. In fact, as we move away from the correlated Brownian case and assume there *is* a linear relationship such as (6.6), then OLS will be a consistent estimator:

$$\hat{\gamma} = \gamma + \frac{\langle x_t u_t \rangle}{\langle x_t^2 \rangle} \tag{6.10}$$

although of course the (asymptotic) diagnostics will be different from the case of standard OLS (the estimator in (6.10) is said to be super-consistent, converging like $T^{-1}$ instead of the usual $T^{-1/2}$; see Hamilton [1994]). In other words, if $\begin{pmatrix} -\gamma & 1 \end{pmatrix}^T$ is the sole cointegrating vector for the process $\begin{pmatrix} x_t & y_t \end{pmatrix}^T$, OLS can validly be used to estimate it.[5]

In general, some specific mechanism must be introduced in the drift of the process dynamics to exhibit a nontrivial stationary relationship in conjunction with individual non-stationary behavior. For example, we could extend (6.5) like

$$\begin{aligned} \Delta x_t &= b_1(y_{t-1} - \gamma\, x_{t-1}) + \sigma_x \varepsilon_t^x \\ \Delta y_t &= b_2(y_{t-1} - \gamma\, x_{t-1}) + \sigma_y \varepsilon_t^y \end{aligned} \tag{6.11}$$

from which we see that

$$\Delta(y_t - \gamma\, x_t) = (b_2 - \gamma\, b_1)(y_{t-1} - \gamma\, x_{t-1}) + \sqrt{\sigma_y^2 - 2\gamma\,\rho\sigma_x\sigma_y + \gamma^2\sigma_x^2}\,\varepsilon_t \tag{6.12}$$

with $\varepsilon \sim N(0,1)$. Thus if $|1 + b_2 - \gamma\, b_1| < 1$, the entity $y_t - \gamma\, x_t$ is stationary while $x$ and $y$ are individually non-stationary.[6] The kind of model in (6.14), where non-stationary dynamics are driven by both Brownian terms and deviations from some stationary (equilibrium, so to speak) relationship, will be considered in greater detail in Section 6.1.3. Before doing so, however, we turn attention to the idea of causality in econometrics, and its connection to structure within cointegrated systems.

### 6.1.2 Granger causality

A natural question to ask here is the following: how does cointegration relate (if at all) to the more familiar dependency category of correlation, and indeed, more deeply to causality? We certainly have no intention of delving into any philosophical issues here. Our interest here is simply in the informational structure (if any) that can be exploited when forming conditional expectations of some stochastic process. Very broadly, the issue can be framed as follows: does knowing something about one variable tell us something about another variable (and in what sense)? In other words, we are seeking to characterize *incremental* information.

#### *6.1.2.1 A standard heat-rate example*

To show the essential ideas, consider the following special case of the model (3.26):

$$\begin{aligned} \Delta g_t &= \sigma_g \varepsilon_t^g \\ \Delta p_t &= -\kappa(p_{t-1} - g_{t-1}) + \sigma_p \varepsilon_t^p \end{aligned} \tag{6.13}$$

which can be thought of as a simple heat-rate model, with $g$ and $p$ log gas and log-power prices, respectively, and with a cointegrating relationship $h \equiv p - g$ representing the stationary log-heat rate. The long-term variance of the log-heat rate is finite and given by[7]

$$\frac{\sigma_h^2}{2\kappa - \kappa^2} \tag{6.14}$$

where $\sigma_h^2 = \sigma_g^2 - 2\rho\sigma_g\sigma_p + \sigma_p^2$. The cumulative correlation between gas and power is given by

$$\begin{aligned} \rho_T &= \frac{Cov(g_T, p_T)}{\sqrt{Var(g_T)\,Var(p_T)}} \\ &= \frac{Var(g_T) + Cov(g_T, h_T)}{\sqrt{Var(g_T)(Var(g_T) + 2Cov(g_T, h_T) + Var(h_T))}} \end{aligned} \tag{6.15}$$

Owing to the fact that the long-term heat rate variance is finite while the long-term gas variance grows linearly with time, we immediately see that $\rho_T \to 1$ as $T \to \infty$.

We see in this simple example a general principle: a cointegrating relationship is the counterpart of very strong *long-term* correlation.[8] (Of course, over any finite time horizon, power and gas can effectively de-correlate, which is another way of saying that the heat rate can blow out; the ramifications of this kind of behavior depends on the strength of mean reversion $\kappa$ [as well as the magnitude of the heat rate vol $\sigma_h$], which again reflects the importance of time scales in the problem.)

However, there is another aspect to the dynamics of the dependency. Rewrite (6.13) in AR form:

$$\begin{aligned} g_t &= g_{t-1} + \sigma_g \varepsilon_t^g \\ p_t &= p_{t-1} - \kappa(p_{t-1} - g_{t-1}) + \sigma_p \varepsilon_t^p \end{aligned} \tag{6.16}$$

We see from (6.16) that (log) power plays no role in projecting future values of (log) gas, while (log) gas *does* play a role in projecting future values of (log) power. In other words, the relationship between power and gas is not symmetrical. The equilibrium condition that characterizes the joint system (namely, the [stationary] heat rate) affects the dynamics of power, but not gas.

This notion can be formalized by the concept of *Granger causality*. Granger causality has nothing to do with causality in any philosophical sense, but is rather a (testable) econometric property. So, we say that a time series $x_t$ Granger-causes a time series $y_t$ if the conditional expectation of $y_t$ given the history of $x$ *and* $y$, has lower variance than the conditional expectation given the history of $y$ only:

$$Var(E(y_t|y_{t-1}, \ldots, x_{t-1}, \ldots)) < Var(E(y_t|y_{t-1}, \ldots)) \tag{6.17}$$

(Alternatively, if the variance relation in (6.17) holds as an equality, $y$ is said to be exogenous [in a time-series sense] wrt. $x$.) Technically, we are sneaking in some assumptions about stationarity when we invoke unconditional variances in (6.17) (the same point applies to the formulation in terms of mean-squared forecast error, as in Hamilton [1994]), however this is tangential to the main point. The issue concerns the benefit from conditioning on one data set as opposed to another.

#### *6.1.2.2 Stochastic trends and more hints of error correction*

Operationally (and again implicitly invoking some degree of stationarity), Granger causality is commonly tested by running joint hypothesis tests (*e.g.*, standard $F$–tests from OLS) on regressions of lagged $x$ and lagged $y$ against contemporaneous $y$. These are not of great concern to us. We simply note that dynamics such as (6.16) capture this aspect. In fact, we consider the following generalization of the simple model in (6.14), in conjunction with the stochastic trend representation of (6.3) and (6.4):

$$\Delta \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}_t = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} (y_1 - \Gamma y_2)_{t-1} + \varepsilon_t \tag{6.18}$$

for $h$ cointegrating relationships and with an appropriate block structure in the various vectors and matrices. We thus see a means for identifying which of the stochastic trends ($y_2$), if any, can be regarded as Granger causative: those components corresponding to the rows of $B_2$ with all zeros have dynamics that are independent of the cointegrating relationship.

The issue of Granger causality is not merely a theoretical curiosity, but actually has very real implications for valuation and hedging. Mahoney and Wolyniec (2012) demonstrate that, in valuing spread options on cointegrated pairs, it is imperative that the Granger-caused leg be dynamically hedged. (This example was in fact considered here in Section 3.1.3.) The volatility that is collected with a statically hedged Granger-caused leg will be lower in general than that collected with the dynamic strategy. (The Granger-causative leg can be statically hedged.) As a final example of an application to trading, consider the following portfolio:

$$\Pi_{t,T} = y_{1T} - y_{1t} - \Gamma(y_{2T} - y_{2t}) = \xi_T - \xi_t \tag{6.19}$$

where $\xi \equiv y_1 - \Gamma y_2$ is a stationary process with (unconditional) mean zero.[9] Consequently, (6.19) provides a means of detecting trading opportunities. Because $\xi$ effectively has a directional bias (due to its implicit mean reversion), we can assess the probability that the current spread is high (or low, in which case the sign of the positions in (6.19) would need to be reversed) relative the holding period. This simple example again displays the ubiquitous nature of time scales.[10]

We can now turn to a more systematic discussion of the dynamic interplay of equilibrium and diffusive effects in cointegrated systems.

### 6.1.3 Vector Error Correction Model (VECM)

*6.1.3.1 Basic framework*

We start with a basic vector autoregressive (VAR)[11] process:

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots \Phi_p y_{t-p} + \varepsilon_t \tag{6.20}$$

where $y$ is an $N \times 1$ vector process (so $\Phi_i$ are $N \times N$ matrices) and the random error term is serially uncorrelated but has contemporaneous covariance structure given by

$$E\varepsilon_t \varepsilon_t^T = \mathbf{\Omega} \tag{6.21}$$

In operator notation, we can write (6.20) as

$$(I - \Phi_1 L - \Phi_2 L^2 - \cdots \Phi_p L^p) y_t = \varepsilon_t \tag{6.22}$$

where of course $L$ is the lag operator: $Ly_t = y_{t-1}$. Now, one might suppose that the operator representation in (6.22) is invertible (say, formally in terms of an infinite series representation as in the one-dimensional case) if the eigenvalues of the constituent matrices $\Phi_i$ are sufficiently bounded. This is in fact the case, although the properly derivable result (see Hamilton [1994]) is that the roots of

$$\det(I - \Phi_1 z - \Phi_2 z^2 - \cdots \Phi_p z^p) = 0 \tag{6.23}$$

lie outside the unit circle.[12] (We shall consider in greater detail the roles eigenvalues play in connecting certain important features of these kinds of processes later.) If this condition is satisfied, then (6.22) is invertible, yielding a Wold-type (convergent infinite series) moving average representation. In other words, the (vector) process $y$ is stationary. We shall be interested in situations where some roots of (6.23) lie *on* the unit circle, in which case the process is non-stationary.[13]

*6.1.3.2 An appropriate representation and factorization*

It turns out that it is convenient to rewrite (6.20) in a form more reminiscent of continuous-time stochastic processes. To this end, in matrix form we write (6.22) as

$$\begin{pmatrix} I & -\Phi_1 & -\Phi_2 & \cdots & -\Phi_p \end{pmatrix} \begin{pmatrix} I & I & 0 & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ 0 & I & -I & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & I & -I & \cdots & -I \end{pmatrix}$$
$$\begin{pmatrix} I & -I & 0 & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ 0 & I & -I & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & -I \end{pmatrix} \begin{pmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{pmatrix} = \varepsilon_t \tag{6.24}$$

(Note that the product of the $(p+1)N \times (p+1)N$ matrices[14] is the identity.) Consequently (6.24) becomes

$$\Delta y_t = \Gamma_0 y_{t-1} + \Gamma_1 \Delta y_{t-1} + \cdots \Gamma_{p-1} \Delta y_{t-p+1} + \varepsilon_t \cdot \tag{6.25}$$

where $\Delta y_t = y_t - y_{t-1}$ is the differencing operator and

$$\begin{aligned} \Gamma_{p-1} &= -\Phi p \\ \Gamma_{p-2} &= -(\Phi_{p-1} + \Phi_p) \\ &\vdots \\ \Gamma_1 &= -(\Phi_2 + \cdots + \Phi_p) \\ \Gamma_0 &= -I + (\Phi_1 + \cdots + \Phi p) \end{aligned} \tag{6.26}$$

This is the so-called VECM representation of the VAR. In this form, the dynamics of the differences are driven not just by the previous (lagged) differences, but also by the previous value of the level, through the term $\Gamma_0 y_{t-1}$ . This latter entity can be

thought of as a sort of equilibrium level driving the dynamics. Of particular interest is the case where there exists some matrix $A$ such that $A^T y$ is stationary. In general $A$ will be $N \times h$, where $h \leq N$ is the number of cointegrating relationships. In such a case the equilibrium coefficient can be written as[15]

$$\Gamma_0 = -BA^T \tag{6.27}$$

for some $N \times h$ matrix $B$. (The sign choice is simply conventional.) Thus, in general, $\Gamma_0$ will not have full rank. The task of estimating the model in (6.25) entails satisfying the restriction in (6.27) and conducting the appropriate diagnostics.

*6.1.3.3 ML estimation*

We will outline Johansen's maximum likelihood approach to estimating VECMs because, apart from the intellectual value of the inherent ingenuity entailed, it is useful to understand some of the underlying concepts and structures as they will be relevant when we relate cointegrating relationships to variance scaling laws. In addition, it can be seen how the underlying approach is quite flexible and readily extended to situations not always presented in detail in standard treatments. For simplicity, we will consider the single-lag case, so that $p = 1$ in (6.20) and (6.25); it should be clear how the idea is extended to higher lags. It is useful to first review some properties of the ML estimator for an unrestricted VAR, which we have already considered in (2.25), where we noted the estimator of the matrix coefficient in (2.27). The log-likelihood function is given by

$$\mathcal{L} = -\frac{1}{2}\sum_t (y_t - \Phi_1 y_{t-1})^T \boldsymbol{\Omega}^{-1} (y_t - \Phi_1 y_{t-1}) - \frac{TN}{2}\log 2\pi - \frac{T}{2}\log\det\boldsymbol{\Omega} \tag{6.28}$$

where $T$ is the number of data points. Now, it is not hard to show that the ML estimator of the covariance matrix is simply related to the realized residuals $\varepsilon_t \equiv y_t - \hat{\Phi}_1 y_{t-1}$ via

$$\hat{\boldsymbol{\Omega}} = \frac{1}{T}\sum_t \varepsilon_t \varepsilon_t^T \tag{6.29}$$

(A useful result here, which will be used again in the discussion of the likelihood ratio method in Monte Carlo in the next chapter, is that $\frac{\partial}{\partial A}\log\det A = A^{-T}$ for a general matrix $A$.) We further note, from the fact that (adopting the summation convention) $\varepsilon_t^i \boldsymbol{\Omega}_{ij}^{-1} \varepsilon_t^j = \varepsilon_t^i \varepsilon_t^j \boldsymbol{\Omega}_{ji}^{-1} = \mathrm{Tr}(\varepsilon_t \varepsilon_t^T \boldsymbol{\Omega}^{-1})$ and the linearity of the trace operator, that the optimal value of the likelihood function can be written as

$$\mathcal{L} = -\frac{TN}{2} - \frac{TN}{2}\log 2\pi - \frac{T}{2}\log\det\boldsymbol{\Omega} \tag{6.30}$$

This form will prove efficacious for deriving results in the restricted case of cointegrating relationships represented by (6.27).

*6.1.3.4 Concentrating the likelihood*

To this end, consider ML estimation of the model in (6.25) (again for $p=1$), subject to the constraint in (6.27). The likelihood function is given by

$$\mathcal{L} = -\frac{1}{2}\sum_t (\Delta y_t + BA^T y_{t-1})^T \Omega^{-1} (\Delta y_t + BA^T y_{t-1}) - \frac{TN}{2}\log 2\pi - \frac{T}{2}\log\det\Omega \tag{6.31}$$

The idea now is to "concentrate" the likelihood function, by optimizing over different levels, employing subsequently finer resolutions of information. To this end, first assume we know the optimal estimate of $A$. Then the maximization posed by (6.31) is a standard VAR, from which we get

$$\hat{B}(A) = -\langle \Delta y_t y_{t-1}^T \rangle A (A^T \langle y_{t-1} y_{t-1}^T \rangle A)^{-1} \tag{6.32}$$

For the covariance structure, from (6.30) we see that the problem becomes minimization of

$$\begin{aligned} \det\hat{\Omega} &= \det\left(\frac{1}{T}\langle (\Delta y_t + BA^T y_{t-1})(\Delta y_t + BA^T y_{t-1})^T \rangle\right) \\ &= \det(\Sigma_{uu} + BA^T \Sigma_{vu} + \Sigma_{uv} AB^T + BA^T \Sigma_{vv} AB^T) \end{aligned} \tag{6.33}$$

where $\Sigma_{uu}$ is the sample variance of $\Delta y_t$, $\Sigma_{vv}$ is the sample variance of $y_{t-1}$, and $\Sigma_{uv} = \Sigma_{vu}^T$ is the sample covariance between $\Delta y_t$ and $y_{t-1}$. This structure can be further simplified by introducing the following factorization:

$$\begin{aligned} \Sigma_{uu} &= FF^T \\ \Sigma_{vv} &= HH^T \\ \Sigma_{uv} &= FRH^T \end{aligned} \tag{6.34}$$

where $R$ is a diagonal matrix. The representation in (6.34) arises from the so-called canonical correlation problem, which seeks a linear combination between two random variables such that the correlation is maximized subject to unit variance; see Hamilton (1994) for more details. The relevant result here is that the solution $F$ of (6.34) arise as (columns of) eigenvectors of the generalized eigenvalue problem

$$\Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{uv}^T x = \lambda \Sigma_{uu} x \tag{6.35}$$

with the eigenvectors normalized by $x^T \Sigma_{uu} x$, and the diagonal elements of $R$ being the square roots of the eigenvalues. (A similar result prevails for $H$ with $u$ and $v$

reversed; the eigenvalues are unchanged.) Consequently, the objective now becomes minimization of

$$\det(I + \tilde{B}\tilde{A}^T R + R\tilde{A}\tilde{B}^T + \tilde{B}\tilde{A}^T\tilde{A}\tilde{B}^T) \tag{6.36}$$

with $\tilde{A} \equiv H^T A$ and $\tilde{B} \equiv F^{-1}B$. With this notation, (6.32) can be written as

$$\tilde{B} = -R\tilde{A}(\tilde{A}^T\tilde{A})^{-1} \tag{6.37}$$

The final step is to minimize

$$\det(I - R\tilde{A}(\tilde{A}^T\tilde{A})^{-1}\tilde{A}^T R) = \det(I - \tilde{A}^T R^2 \tilde{A}(\tilde{A}^T\tilde{A})^{-1}) = \frac{\det(\tilde{A}(I - R^2)\tilde{A})}{\det(\tilde{A}^T A)} \tag{6.38}$$

wrt. $\tilde{A}$. (In the second equation in (6.38) we have employed Sylvester's determinant identity.) As is well known, cointegrating relationships are nonunique up to linear transformations, so we are free to impose some sort of normalization, which for convenience we take to be $\tilde{A}^T\tilde{A}$. Then (6.38) becomes

$$\det(I - A^T R^2 A) \tag{6.39}$$

(Recall that $A$ is $N \times h$, so although it is column orthogonal, it is not necessarily row orthogonal.) Although we will not show it here (see Johansen [1988]), the optimal solution is to take the columns of $A$ to be standard unit vectors such that the $h$ largest diagonal elements[16] of $R$ are selected, so to speak. Assuming $R$ is ordered so that the diagonal elements are in decreasing order (upper left to lower right), this can be achieved by taking the columns of $U$ to be the first $h$ unit vectors $e_i$. For example, the $3 \times 2$ case would be

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{6.40}$$

(The intuition here is that, in the unrestricted case, the determinant in (6.39) is $\prod_i (1 - r_{ii}^2) \equiv \prod_i (1 - \lambda_i)$ so in the restricted case we would expect only the $h$ largest eigenvalues to be present.) Finally, the ML estimator of the (instantaneous) covariance matrix is

$$\hat{\Omega} = \Sigma_{uu} - \Sigma_{uv} A A^T \Sigma_{vu} \tag{6.41}$$

#### *6.1.3.5 Some more asymptotics*

This outline of Johansen's (1988, 1991) ML estimation of VECM should serve to convey the central ideas. It is not hard to include additional lags or deterministic terms (such as drifts or seasonal dummies); see Hamilton (1994) for more details.

It is probably also worth saying something about the (asymptotic) diagnostics. Note that the entities of interest are the eigenvalues in (6.35). The matrix in question is given by

$$\Sigma_{uu}^{-1}\Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{uv}^{T} = \langle \Delta y_t \Delta y_t^T \rangle^{-1} \langle \Delta y_t y_{t-1}^T \rangle \langle y_{t-1} y_{t-1}^T \rangle^{-1} \langle \Delta y_t y_{t-1}^T \rangle^T \tag{6.42}$$

Now, following the heuristics employed when discussing the Dickey-Fuller diagnostics for unit root processes (*e.g.*, in (2.38)) and invoking the implicit decompositions in (6.34), we have the following:

$$\begin{aligned}
\langle \Delta y_t \Delta y_t^T \rangle^{-1} &\sim \left( TF \int_0^1 dw_s dw_s^T F^T \right)^{-1} = T^{-1} F^{-T} \int_0^1 dw_s dw_s^T F^{-1} \\
\langle \Delta y_t y_{t-1}^T \rangle &\sim T \cdot F \int_0^1 dw_s w_s^T H^T \\
\langle y_{t-1} y_{t-1}^T \rangle^{-1} &\sim \left( TH \int_0^1 ds \cdot w_s w_s^T H^T \right)^{-1} = T^{-1} H^{-T} \left( \int_0^1 ds \cdot w_s w_s^T \right)^{-1} H^{-1}
\end{aligned} \tag{6.43}$$

where (it turns out) $w$ denotes a standard, $(n-h)$–dimensional Brownian motion; see Johansen (1988, 1991) for much greater detail. Thus (due to the similarity transformation involving $F$), the eigenvalues of the estimation are asymptotically distributed as the eigenvalues of

$$\int_0^1 dw_s w_s^T \left( \int_0^1 ds \cdot w_s w_s^T \right)^{-1} \int_0^1 w_s dw_s^T \tag{6.44}$$

Obviously, the random variable in (6.44) is nonstandard, and the critical values must be obtained via simulation; see MacKinnon *et al.* (1999). (It is worth noting that for the one-dimensional case, (6.44) reduces to the square of the DF statistic in (2.22), so Johansen's procedure also provides an alternative means of testing for the presence of unit roots.) Recall that any cointegrating relationship is nonunique up to a linear transformation, so hypothesis testing of cointegration typically entails testing whether there is *a* relationship (as opposed to none), or one *more* relationship (given a particular number of relationships). It should go without saying that great care must be used in employing asymptotic results in testing for cointegration; see Mallory and Lence (2012).

In light of this last point, we ask whether there are alternative, more robust means for detecting multiple cointegrating relationships, given the clear power and relevance of the concept. It turns out that there is indeed such an approach, and it ties in with the variance scaling laws we have previously emphasized.

### 6.1.4 Connection to scaling laws

Consider the following $n$–dimensional process:

$$dz_i = (A_{ij}z_j + b_i)dt + \sigma_i dw_i \tag{6.45}$$

With $dw_i dw_j = \rho_{ij}dt$ and we adopt the notation $X_{ij} \equiv \rho_{ij}\sigma_i\sigma_j$. This is a Gaussian process, and the structure can be discerned from the characteristic function. In particular the covariance structure (over some interval $\tau$) is extracted from the coefficients of the Fourier variables in

$$\int_0^\tau ds \cdot \alpha_i X_{ij} \alpha_j \tag{6.46}$$

where $\alpha$ satisfies

$$\dot{\alpha} = A_{ki}\alpha_\kappa,\ \alpha_i(0) = i\phi_i \tag{6.47}$$

Now, use the Jordan decomposition[17] to write in matrix form $A = PJP^{-1}$ where $P$ is a nonsingular matrix and $J$ is a diagonal matrix of the eigenvalues of $A$. Making the substitution $\alpha = P^{-T}\beta$ we get

$$\dot{\beta} = J\beta, \beta(0) = iP^T\phi \tag{6.48}$$

Thus, $\beta = iKP^T\phi$ where $K = e^J = \text{diag}(e^{-\kappa_i\tau})$ and $-\kappa$ are the eigenvalues of $A$ (also the diagonal elements of $J$)[18]. So, $\alpha = iP^{-T}KP^T\phi$ and the covariance matrix is given by

$$\int_0^\tau ds \cdot PK\tilde{X}KP^T \tag{6.49}$$

where $\tilde{X} = P^{-1}XP^{-T}$. Now, from the form of $K$ we introduce the matrix $E$ with elements given by

$$E_{jn} = \frac{1 - e^{-(\kappa_j + \kappa_n)\tau}}{\kappa_j + \kappa_n} \tag{6.50}$$

so that the covariance matrix is given by

$$P(\tilde{X} \circ E)P^T \tag{6.51}$$

where $\circ$ denotes the Hadamard product (element-by-element multiplication).

Now, what does all of this have to do with cointegration? Well, following the Johansen framework, assume that $z$ is non-stationary and that $A$ can be written as $A = RH^T$ where $R$ and $H$ are $n \times h$ matrices such that the product $H^T z$ is stationary. Letting $\xi = H^T z$, (6.45) implies that

$$d\xi_k = (H_{ik} R_{il} \xi_l + H_{ik} b_i) dt + H_{ik} \sigma_i dw_i \tag{6.52}$$

From the preceding analysis, stationarity of $\xi$ implies that the eigenvalues of $H^T R$ must all be negative. Furthermore, non-stationarity of $z$ implies that at least one eigenvalue of $A(= RH^T)$ must be zero. For example, a possible form for $J$ is

$$\begin{pmatrix} -\kappa_1 & 0 & 0 & 0 \\ 0 & -\kappa_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{6.53}$$

with $\kappa_{1,2} > 0$. (Note that the nonzero eigenvalues of $RH^T$ must be the same as those of $H^T R$.) Thus the presence of a zero eigenvalue is a necessary, but not sufficient condition for non-stationarity. The matrix $P$ must also have a particular form, which is not terribly restrictive. It simply has to allow linearly growing terms in *all* of the diagonal terms of the covariance structure in (6.51) *E.g.*, if $P$ was the identity matrix, then clearly $z_{1,2}$ would simply be mean reverting and thus stationary. For the example in (6.53), we have that

$$E = \begin{pmatrix} \frac{1-e^{-2\kappa_1 \tau}}{2\kappa_1} & \frac{1-e^{-(\kappa_1+\kappa_2)\tau}}{\kappa_1+\kappa_2} & \frac{1-e^{-\kappa_1 \tau}}{\kappa_1} & \frac{1-e^{-\kappa_1 \tau}}{\kappa_1} \\ & \frac{1-e^{-2\kappa_2 \tau}}{2\kappa_2} & \frac{1-e^{-\kappa_2 \tau}}{\kappa_2} & \frac{1-e^{-\kappa_2 \tau}}{\kappa_2} \\ & & \tau & \tau \\ & & & \tau \end{pmatrix} \tag{6.54}$$

Now, the matrix in (6.53) can be factored as

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} -\kappa_1 & 0 & 0 & 0 \\ 0 & -\kappa_2 & 0 & 0 \end{pmatrix} \tag{6.55}$$

Thus, we see that the choice

$$H^T = \begin{pmatrix} -\kappa_1 & 0 & 0 & 0 \\ 0 & -\kappa_2 & 0 & 0 \end{pmatrix} P^{-1}. \tag{6.56}$$

gives the desired linear combination that yields stationary variables. For in this case the linearly growing terms in (6.54) are killed off, giving the resulting ensemble a mean-reverting covariance structure. (Recall the preservation of Gaussianity under linear transformations.)

Up to this stage, we have assumed a particular structure for $A$ in the dynamics of $z$ and shown how this is essentially equivalent to the standard VECM cointegration formalism, and that this formalism is consistent with the underlying covariance scaling laws. We would like to proceed in reverse now, and see if we can in general infer any information about the underlying cointegrating relationships from the covariance scaling laws. To this end, consider the scaling law from (6.51). First, note that for small time horizons, the matrix $E$ in (6.50) is approximately a matrix of all ones, in which case the scaling law is simply the local dynamics of the underlying diffusion in (6.45), namely $X$. Thus there is little information to be gained at this time scale (not surprisingly). However, for very large time horizons, the matrix $E$ becomes (again continuing with the example at hand)

$$E \sim \tau \begin{pmatrix} 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 \\ & & 1 & 1 \\ & & & 1 \end{pmatrix} \tag{6.57}$$

Consequently, the covariance matrix in (6.51) will have zero eigenvalues (two in this example). Note that the expression in (6.51) does not entail a similarity transform because in general $P$ is not orthogonal. However, it is easy to show in general that two matrices $W$ and $V$ related via $W = PVP^T$ will share any eigenvalues that are *zero*. Note further that any zero eigenvalues must be inherited from a structure of the form (6.57), as the local covariance structure $X$ will in general be positive definite.

We thus have a useful (and presumably robust) criterion for the *number* of cointegrating relationships: the number of zero eigenvalues of the long-term covariance scaling law matrix. (See Mahoney [2015b] for examples consistent with Johansen's test.)

## 6.2 Stochastic filtering

### 6.2.1 Basic concepts

A general problem encountered with many estimation problems is the fact that certain stochastic drivers of interest are not directly observable, and hence cannot be explicitly included in any kind of estimation procedure (such as ML). An obvious example is stochastic volatility. Consequently, any viable econometric technique

must operate *only* on observable data, while at the same time taking into account the dynamics implied by the model being tested against that data. This is the task of stochastic filtering, which, as the name suggests, is a means of extracting information about unobservable state processes. An example in the context of ML estimation should convey the essential ideas.

Recall that in ML estimation we optimize, with respect to a set of a model parameters, the log-likelihood function over the data set: $\max_{\theta} \mathcal{L}(z;\theta)$ where $\theta$ are the model parameters, $z$ is the data set, and the log-likelihood function is given by

$$\mathcal{L}(z;\theta) = \sum_{t} \log \Pr_{\theta}(z_t | I_{t-1}) \tag{6.58}$$

The notation here refers to the fact that the conditional density is parameter-dependent, and the probability of an observation at a given time is dependent on the information $I$ over previous times. For example, this information set might the previous history of prices, *e.g.*, $I_t = \{z_{0:t}\} \equiv \{z_0, \ldots, z_t\}$[19]. In general, this data is generated by some stochastic process $x$, which is only partially observable, so we must relate the distribution of observables to the distribution of the state process. It is assumed that there is some (non-invertible!) transformation between state variables and observables (often termed the measurement equation).

Bayesian methods prove quite useful here. We have the following steps (essentially an application of the Chapman-Kolmogorov equation):

$$\begin{aligned}
\Pr(x_{t+1|I_t}) &= \int_{\Omega} dx_t \Pr(x_t, x_{t+1} | I_t) = \int_{\Omega} dx_t \Pr(x_{t+1} | x_t) \Pr(x_t | I_t) \\
\Pr(z_{t+1} | I_t) &= \int_{\Omega} dx_{t+1} \Pr(x_{t+1}, z_{t+1} | I_t) = \int_{\Omega} dx_{t+1} \Pr(z_{t+1} | x_{t+1}) \Pr(x_{t+1} | I_t) \\
\Pr(x_{t+1} | I_{t+1}) &= \frac{\Pr(x_{t+1} | I_t) \Pr(z_{t+1} | x_{t+1})}{\Pr(z_{t+1} | I_t)}
\end{aligned} \tag{6.59}$$

where $\Omega$ is the sample space of the state process $x$. (In the perhaps more familiar context of Bayesian reasoning, the ensemble in (6.59) can be characterized as prediction [the first equation] and updating [the third equation], along with an appropriate normalization [the second equation]) Alternatively, we can even more succinctly write (6.59) in terms of joint (rather than marginal) densities as

$$\Pr(x_{0:n+1} | z_{0:n+1}) = \frac{\Pr(z_{n+1} | x_{n+1}) \Pr(x_{n+1} | x_n)}{\int dx_{n:n+1} \Pr(z_{n+1} | x_{n+1}) \Pr(x_{n+1} | x_n) \Pr(x_n | z_{0:n})} \Pr(x_{0:n} | z_{0:n}) \tag{6.60}$$

There are a few assumptions implicit in (6.59) that we should make explicit here. First, the dynamics of the underlying state process are Markovian (this is used in the second equation of (6.59)). Second, conditioned on the current state of the process, the observation is independent of previous observations (this is used in the first equation of (6.59)). It is further assumed that both the transition density $\Pr(x_{t+1}|x_t)$ for the dynamics of the state and the conditional observation density $\Pr(z_t|x_t)$ are known (from the model in question). For example, recall the popular Heston stochastic volatility model (or any other such model). As we saw in Chapter 5, the transition density can be retrieved from the conditional characteristic function (which has a known analytical form). The measurement equation is trivial, since the price *is* the observation and the observation density is simply a delta function. In this context we could write (6.59) as

$$\begin{aligned}
\Pr(S_{t+1}, v_{t+1}|S_{0:t}) &= \int_0^\infty dv_t \Pr(S_{t+1}, v_{t+1}|S_t, v_t) \Pr(v_t|S_{0:t}) \\
\Pr(S_{t+1}|S_{0:t}) &= \int_0^\infty dv_{t+1} \Pr(S_{t+1}, v_{t+1}|S_{0:t}) \\
\Pr(v_{t+1}|S_{0:t+1}) &= \frac{\Pr(S_{t+1}, v_{t+1}|S_{0:t})}{\Pr(S_{t+1}|S_{0:t})}
\end{aligned} \tag{6.61}$$

Note that (6.61) makes clear that, although the joint dynamics (price and stochastic variance) are Markovian, the dynamics of price *alone* are not (the dynamics depend on the entire history, not just the most recent observation).

The recursive nature of the filter should be clear, and for given initial densities of the state and observation, the ensemble in (6.59) can be iteratively carried out to build up the terms needed in the log-likelihood function in (6.58). It should be equally clear, however, that this is computationally formidable task, as there are multidimensional integrations involved. The compact form of (6.59) or (6.61) is deceptively simple. There are a number of techniques commonly used for tackling this problem, such as particle filters and Markov-Chain Monte Carlo (MCMC). We will not discuss these (essentially sampling/simulation based methods) here, but merely direct the reader to the relevant literature.[20] Instead we will consider a case (linear, Gaussian) where the recursion can be performed analytically, namely the very well-known Kalman filter.

### 6.2.2 The Kalman filter and its extensions

#### *6.2.2.1 The classic filter*

A very popular technique originally due to Kalman (1960) and bearing his name is, at heart, a (deceptively[21]) simple application of conditional expectation for normal

variables. We will demonstrate the essential features here. We start with a Gaussian vector process $z$, partitioned into $x$ and $y$:

$$\begin{aligned} z &\sim N(\mu, \Sigma) \\ z &= \begin{pmatrix} x \\ y \end{pmatrix} \\ y|x &\sim N(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{yx}^{T}) \end{aligned} \tag{6.62}$$

Now, we suppose that there is some state process governing the overall system dynamics:

$$x_{k+1} = Fx_k + w_{k+1}, w \sim N(0, Q) \tag{6.63}$$

with $w$ independent of $x$. We further suppose that these state variables are not directly observable (*e.g.*, only some of them are, or only particular functions of them are, *etc.*). Obvious examples would include price dynamics driven by stochastic volatility, with the former but not the latter being observable. We express this assumption by introducing a so-called measurement/observation equation:[22]

$$z_{k+1} = Hx_{k+1} \tag{6.64}$$

Note that time-inhomogeneity can easily be introduced into all of these relationships, as well as sources of uncertainty in the observation equation ("measurement error"), but we omit them here for convenience. Now, we look for the dynamics of the filtered/estimated value of the state variable. Denoting this entity by a hat, we have

$$x_k = \hat{x}_k + v_k, v_k \sim N(0, P_k) \tag{6.65}$$

(Since the underlying problem is linear, the Gaussian structure is retained.) The optimal (in terms of variance-minimizing) value of the filtered/projected state variable is simply its expectation conditional on the current information set:

$$\hat{x}_k = E_k x_k \equiv E(x_k|z_k, I_{k-1}) \tag{6.66}$$

with the associated conditional variance/uncertainty $P_k$. Using the standard results above for conditional normals (and assuming state noise and filter noise are

uncorrelated), we get the following relations:

$$\begin{aligned}
z_{k+1} &= Hx_{k+1} = HFx_k + Hw_{k+1} \\
E_k x_{k+1} &= F\hat{x}_k \equiv \tilde{x} \\
E_k z_{k+1} &= HF\hat{x}_k = H\tilde{x} \\
x_{k+1} - E_k x_{k+1} &= F(x_k - \hat{x}_k) + w_{k+1} \\
z_{k+1} - E_k z_{k+1} &= HF(x_k - \hat{x}_k) + Hw_{k+1} \\
\Sigma^{xx} &= FP_k F^T + Q \equiv \tilde{P} \\
\Sigma^{zz} &= HFP_k F^T H^T + HQH^T = H\tilde{P}H^T \\
\Sigma^{xz} &= FP_k F^T H^T + QH^T = \tilde{P}H^T
\end{aligned} \tag{6.67}$$

Assembling these, we find that

$$\hat{x}_{k+1} = E(x_{k+1}|z_{k+1}, I_k) = \tilde{x} + \tilde{P}H^T(H\tilde{P}H^T)^{-1}(z_{k+1} - H\tilde{x}) \equiv \tilde{x} + K(z_{k+1} - H\tilde{x}) \tag{6.68}$$

where $K$ is known as the Kalman gain matrix. The $(k+1)$–conditional variance is given by

$$P_{k+1} = \tilde{P} - \tilde{P}H^T(H\tilde{P}H^T)^{-1}H\tilde{P} = (I - KH)\tilde{P} \tag{6.69}$$

It is interesting to note that in the case under consideration (namely, no measurement error), the following relation holds: $HP_k = 0 \cdot \forall k$. (Sensibly, we also have that $H\hat{x}_k = z_k$.) This is not too surprising: there are fewer observables than state variables, so we would expect there to be some combinations ("preferred directions" if you will) of the state variables that have zero variance (owing to the lack of measurement error and the underlying linearity of the stochastic structure). In other words, the (square) covariance matrix of the projected state variable is singular (degenerate, in fact; it has less than full rank). We would further anticipate this state of affair manifesting itself in non-robustness of estimation, as the uncertainty associated with *individual* components of the state variable become quite noisy in the face of having to balance out each other (in terms of some linear combinations).[23]

So, starting from an initial (unconditional) estimate of the state variable's mean and variance, the recursion and filter can proceed. It is a useful exercise to verify that the general recursive filter in (6.59) recovers the Kalman filter results for the

special case where all the conditional densities are Gaussian. The primary tool is the fact that under convolutions Gaussian structures are retained:

$$\begin{aligned}
&\int_{-\infty}^{\infty} dx \frac{1}{\sqrt{(2\pi)^N \det V}} e^{-\frac{1}{2}(z-Ax-a)^T V^{-1}(z-Ax-a)} \\
&\qquad \frac{1}{\sqrt{(2\pi)^M \det W}} e^{-\frac{1}{2}(x-By-b)^T W^{-1(x-By-b)}} \\
&\quad = \frac{1}{\sqrt{(2\pi)^N \det \Omega}} e^{-\frac{1}{2}(z-A(By+b)-a)^T \Omega^{-1}(z-A(By+b)-a)}
\end{aligned} \tag{6.70}$$

where $z$ is $N \times 1$, $x$ and $y$ are $M \times 1$, and the various matrices have appropriate dimensions. The result in (6.70) is obtained from suitably completing the square in $x$ and the covariance matrix $\Omega$ is given by

$$\Omega = (I - ACD^{-1}C^T A^T V^{-1})^{-1} V \tag{6.71}$$

where $CC^T = W$ and $D = I + C^T A^T V^{-1} AC$. Using the Woodbury matrix identity, (6.71) can be written as

$$\Omega = (I + AC(D - C^T A^T V^{-1} AC)^{-1} C^T A^T V^{-1}) V = V + AWA^T \tag{6.72}$$

Using these results in (6.59), we see that if $x_{t+1}|x_t \sim N(F, Q)$ and $x_t|I_t \sim N(\hat{x}_t, P_t)$, then $x_{t+1}|I_t \sim N(\tilde{x}_t, \tilde{P}_t)$ (prediction) $z_{t+1}|I_t \sim N(H\tilde{x}_t, H\tilde{P}_t H^T)$ and (normalization), where $\tilde{x}_t = F\hat{x}_t$ and $\tilde{P}_t = Q + FP_t F^T$. Finally, we get $x_{t+1}|I_{t+1} \sim N(\tilde{x}_t + K_t(z_{t+1} - H\tilde{x}_t), (I - K_t H)\tilde{P}_t)$ (update), where $K_t = \tilde{P}_t H^T (H\tilde{P}_t H^T)^{-1}$. These are of course the Kalman filter results in (6.68) and (6.69).

#### *6.2.2.2 ML estimator*

As another example, consider an application to calculation of the likelihood function for maximum likelihood estimation. The entity of interest is (note the dependence on the [assumed known] initial state)

$$\Pr(x_0, Z_1, \ldots, Z_T) = \prod_{i=1}^{T} \Pr(z_i | z_{i-1}, \ldots, x_0) \tag{6.73}$$

The non-Markovian structure of the problem (or more accurately, of the observables) can clearly be seen in (6.73), where a conditional density of the observable does not depend only on the observable at the prior time, but on *all* previous observables. It proves interesting to analyze the path in (6.73) *en toto*, as the (by now) familiar Kalman structure will again fall out.

From the state dynamics (6.63) and observable relation (6.64) we get

$$\begin{aligned} z_1 &= Hw_1 + HFx_0 \\ z_2 &= Hw_2 + HFw_1 + HF^2x_0 \\ &\vdots \\ z_n &= HF^n x_0 + H\sum_{i=1}^{n} F^{n-i}w_i \\ &\vdots \end{aligned} \tag{6.74}$$

Owing to the special form of (6.74), the covariance matrix of the joint density in (6.73) has a block structure, with upper diagonal elements given by

$$\Sigma_{mn} = H\sum_{i=1}^{m} F^{m-i}QF^{T(n-i)H^T} = HQ_mF^{T(n-m)}H^T \tag{6.75}$$

for $n \geq m$ and where $Q_0 = 0$, $Q_i = FQ_{i-1}F^T + Q$. This covariance structure has the following factorization: $\Sigma = XX^T$, where

$$X = \begin{pmatrix} HC & 0 & 0 & \cdots \\ HFC & HC & 0 & \cdots \\ HF^2C & HFC & HC & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{6.76}$$

and $CC^T = Q$. Note that in general $X$ is not a square matrix, since $H$ is not square (fewer observables than states). Thus, this factorization cannot be used to facilitate inversion of the covariance matrix in the density function (this is precisely how the filtering problem manifests itself).

Fortunately we do not require the actual Cholesky factorization, as the special structure can be suitably exploited. In fact, we do not even need the actual inverse of the Cholesky factors. Denoting the block Cholesky structure by $(C_{ij})$ and its (similarly lower triangular) inverse by $(D_{ij})$, the log-likelihood function in (6.73) entails calculation of the entity $Z^TZ$, where the vector $Z$ has elements given by $Z_n = \sum_{i-1}^{n} D_{ni}(z_i - \mu_i)$, where the (unconditional) mean $\mu$ can be read off from (6.74). From this it follows that $z_n - \mu_n = \sum_{i=1}^{n} C_{ni}Z_i$ and that $Z_n = C_{nn}^{-1}(z_n - \mu_n - \sum_{i=1}^{n-1} C_{ni}Z_i)$. Thus, the factorization of the aggregate density is broken out into conditional terms,

with in particular $C_{nn}C_{nn}^T$ representing the conditional covariance at $n$th step. These covariances can be calculated iteratively[24] from the Cholesky algorithm[25] as follows:

$$\begin{aligned}
C_{11}C_{11}^T &= HQ_1H^T \equiv H\tilde{Q}_1H^T \\
C_{22}C_{22}^T &= H(Q_2 - FQ_1H^T(C_{11}C_{11}^T)^{-1}HQ_1F^T)H^T \\
&= H(Q + F(Q_1 - Q_1H^T(C_{11}C_{11}^T)^{-1}HQ_1)F^T)H^T \equiv H\tilde{Q}_2H^T \\
C_{33}C_{33}^T &= H\begin{pmatrix} Q_3 - F^2Q_1H^T(C_{11}C_{11}^T)^{-1}HQ_1F^{2T} - \\ F(Q_2 - FQ_1H^T(C_{11}C_{11}^T)^{-1}HQ_1F^T(C_{22}C_{22}^T)^{-1}) \\ H(Q_2 - FQ_1H^T(C_{11}C_{11}^T)^{-1}HQ_1F^T)^TF^T \end{pmatrix}H^T \\
&= H(Q + F(\tilde{Q}_2 - \tilde{Q}_2H^T(C_{22}C_{22}^T)^{-1}H\tilde{Q}_2)F^T)H^T \equiv H\tilde{Q}_3H^T
\end{aligned} \tag{6.77}$$

So proceeding inductively, we see that we recover the basic Kalman algorithm.

*6.2.2.3 Nonlinear extensions I: the basics*

Although the Kalman filter is unquestionably elegant, its underlying assumptions of linearity (both in terms of state dynamics and relation between state and observable) are obviously rather limiting, and a generalization to nonlinear problems is desirable. Let us craft the relationships in (6.63) and (6.64) as follows. The state dynamics and measurement relationship can be written generally as

$$\begin{aligned}
x_{n+1} &= f(x_n, w_{n+1}) \\
z_{n+1} &= h(x_{n+1})
\end{aligned} \tag{6.78}$$

with $f$ and $g$ nonlinear functions in their arguments and where $w$ is (for convenience) taken to be Gaussian noise.[26] An obvious first approach to filtering the system in (6.78) is to simply linearize (via Taylor expansion) the nonlinear entities about the current value of the projected state variable:

$$\begin{aligned}
x_{n+1} &\approx f(\hat{x}_n, 0) + f_x(\hat{x}_n, 0)^T(x_n - \hat{x}_n) + f_w(\hat{x}_n, 0)^T w_{n+1} \\
z_{n+1} &\approx h(f(\hat{x}_n, 0)) + h_x(f(\hat{x}_n, 0))^T(x_{n+1} - f(\hat{x}_n, 0))
\end{aligned} \tag{6.79}$$

Standard (linear) Kalman filtering can then be applied to the (approximate) system in (6.79). This approach is known as the Extended Kalman Filter (EKF). (In essence the EKF approximates the densities in the generic filter (6.59) with Gaussian forms.)

*6.2.2.4 Nonlinear extensions II: the sweet (non-) smell of success*

Although simple and intuitive, the EKF has a serious flaw, namely that it is only as good as the underlying linear (first order) approximation. For highly nonlinear problems the EKF can perform rather poorly.[27] In addition, calculation of the

gradients in (6.79) is potentially expensive. An alternative approach, the oddly-named Unscented Kalman Filter (UKF) directly uses the nonlinear dynamics and observable relationship (so no linearizations are involved[28]), and in fact shares certain affinities with quasi-maximum likelihood estimation (to be studied in Section 6.5.2) and quasi-Monte Carlo integration (see Section 7.5.3). To understand the idea, we first introduce the unscented transform (unsurprisingly).

Consider a (vector) random variable $x$ with mean $\mu_x$ and covariance matrix $\Sigma_x$. We are interested in finding approximations to the mean and variance of the variable $y = f(x)$, related to $x$ through some nonlinear transformation. (Linearization would yield $\mu_y \approx f(\mu_x)$ and $\Sigma_y \approx f_x(\mu_x)^T \Sigma_x f_x(\mu_x)$.) Introduce weights $w_i$ and so-called sigma points $x_i$ satisfying[29]

$$\begin{aligned} \sum_i w_i &= 1 \\ \sum_i w_i x_i &= \mu_x \\ \sum_i w_i (x_i - \mu_x)(x_i - \mu_x)^T &= \Sigma_x \end{aligned} \tag{6.80}$$

We then take the desired approximations to be

$$\begin{aligned} \mu_y &\approx \sum_i w_i f(x_i) \\ \Sigma_y &\approx \sum_i w_i (f(x_i) - \mu_y)(f(x_i) - \mu_y)^T \end{aligned} \tag{6.81}$$

In other words, the full nonlinearity is used in the approximation of the mean and covariance, as opposed to deriving these entities from an approximation of the nonlinearity. Note that for affine transformations, these approximations are exact, *regardless* of the particular choice of weights and sigma points. For $2n+1$ sigma points in $n$–dimensions, a convenient choice is

$$\begin{aligned} x_0 &= \mu_x, \quad w_0 = \frac{\kappa}{n+\kappa} \\ x_i &= \mu_x + \sqrt{n+\kappa} L_i, \quad w_i = \frac{1}{2}\frac{1}{n+\kappa} \\ x_{n+i} &= \mu_x - \sqrt{n+\kappa} L_i, \quad w_{n+i} = \frac{1}{2}\frac{1}{n+\kappa} \end{aligned} \tag{6.82}$$

for $i = 1, \ldots, n$, where $\kappa$ is a suitable (positive) constant and $L_i$ is the $i^{\text{th}}$ column of the Cholesky factor of $\Sigma_x$: $LL^T = \Sigma_x$.

The basic idea and motivation behind the unscented transform is that it is often better to approximate the probabilistic impact (*e.g.*, via approximate Gaussianity) of a nonlinear relationship than to try to approximate the relationship itself. Put differently, it may be advisable to get good approximations for, say, the first two moments induced by a nonlinear transformation, than to try to ascertain the impact of the nonlinearity on the distribution as a whole. This is the essential feature of UKT: to use an unscented transformation (*e.g.*, (6.82)) to extract the means and covariances of the distributions underlying the filtering algorithm in (6.59). (These distributions can of course be extracted exactly in the linear case, owing to the *true* Gaussian nature in that case.)

To illustrate, we consider a special case of the nonlinear system (6.78), which is linear in the disturbance:

$$\begin{aligned} x_{n+1} &= f(x_n) + w_{n+1} \\ z_{n+1} &= h(x_{n+1}) \end{aligned} \tag{6.83}$$

with $w$ Gaussian noise. The next-step predicted mean and variance are given by $\tilde{x}_n = E_n x_{n+1} = E_n f(x_n)$ and $\tilde{P}_n = E_n(x_{n+1} - \tilde{x}_n)(x_{n+1} - \tilde{x}_n)^T$. These can be computed via unscented transform, with the prior step's projected means and variances ($\hat{x}_{n-1}$ and $P_{n-1}$, resp.) used as the basis of an ensemble such as (6.82) to form the necessary sigma points. Updating then takes place, with the measurement mean ($E_n z_{n+1} = E_n h(x_{n+1})$) and measurement variance and covariance ($K_{xz} = E_n(x_{n+1} - E_z x_{n+1})(z_{n+1} - E_z z_{n+1})^T$ and $K_{zz} = E_n(z_{n+1} - E_z z_{n+1})(z_{n+1} - E_z z_{n+1})^T$) similarly being computed via unscented transform. From these we get the Kalman gain $K_{xz}K_{zz}^{-1}$ and the familiar recursion in (6.68) and (6.69), albeit calculated very differently from the standard linear case.[30] To reiterate: the idea is that we approximate with Gaussians (via mean and covariance calculations based on the actual nonlinearity) the non-Gaussian densities induced by the nonlinear structure, as opposed to using exact Gaussian entities based on linear approximations to the nonlinear system.

#### 6.2.2.5 *Continuous-time heuristics*

It is of interest to consider, if only non-rigorously, the continuous-time limit of the Kalman filter. In the dynamics (6.63), let $F = I + A$, where $A$ and also $Q$ are $O(\Delta t)$. Then the process dynamics become

$$\Delta x_{k+1} = A x_k + w_{k+1} \tag{6.84}$$

and the measurement equation can be written as

$$\Delta z_{k+1} = H \Delta x_{k+1} = HA x_k + H w_{k+1} \tag{6.85}$$

So, with respect to the filtration (information set) generated by $(z_k, \hat{x}_{k-1})$[31] we have

$$\begin{aligned} E_k \Delta z_{k+1} &= HA\hat{x}_k \\ E_k \Delta z_{k+1} \Delta z_{k+1}^T &= HA(P_k + \hat{x}_k \hat{x}_k^T) A^T H^T + HQH^T \end{aligned} \tag{6.86}$$

Call this (observation) filtration $\mathfrak{I}_z$. From the Kalman filter/recursion,

$$\hat{x}_{k+1} = F\hat{x}_k + K(z_{k+1} - HF\hat{x}_k) \tag{6.87}$$

which in turn implies

$$\Delta \hat{x}_{k+1} = A\hat{x}_k + K(z_{k+1} - H\hat{x}_k - HA\hat{x}_k) = A\hat{x}_k + K(\Delta z_{k+1} - HA\hat{x}_k) \tag{6.88}$$

Now take the following Cholesky factorization: $Q = CC^T$ (so that $C$ is $O(\Delta t^{\frac{1}{2}})$. Heuristically we anticipate that, in the limit of infinitesimal time steps, we will have the following dynamics with respect to $\mathfrak{I}_z$:

$$\begin{aligned} dz &= HA\hat{x}dt + HCdw \\ d\hat{x} &= A\hat{x}dt + K_t HCdw \end{aligned} \tag{6.89}$$

where we have altered the notation by reabsorbing factors of $\Delta t \to dt$. Note $dw$ is a standard Brownian motion with $dwdw^T = I$. (Note also that the dynamics of observables and estimators are perfectly correlated.)

As indicated in (6.89), the Kalman gain matrix $K$ is time dependent and $O(1)$. In fact, we have that

$$\tilde{P} = FP_k F^T + Q = P_k + AP_k + P_k A^T + AP_k A^T + Q \tag{6.90}$$

and recalling the fact that $HP_k = 0$, we get that

$$\begin{aligned} \tilde{P}H^T &= (P_k A^T + AP_k A^T + Q)H^T \\ H\tilde{P}H^T &= H(AP_k A^T + Q)H^T \end{aligned} \tag{6.91}$$

so in the small time-step limit we have (again rescaling to leading order)

$$K = \tilde{P}H^T(H\tilde{P}H^T)^{-1} = (P_k A + Q)\Delta t H^T (HQH^T)^{-1} \Delta t^{-1} \tag{6.92}$$

with the time-step scalings obviously canceling out. Thus, the (infinitesimal) dynamics of the covariance matrix are found to be[32]

$$dP_{k+1} = \begin{bmatrix} Q - QH^T(HQH^T)^{-1}HQ + \\ (I - QH^T(HQH^T)^{-1}H)AP_k + \\ P_k A^T(I - H^T(HQH^T)^{-1}HQ) \end{bmatrix} dt \tag{6.93}$$

#### 6.2.2.6 *Trading strategies and observables*

Obviously, after all of this, it is fair to ask: why do we even care? We have repeatedly stressed the importance of (variance) scaling laws in commodity markets, and for our purposes here we will associate such scaling with some form of mean reversion in the underlying commodity. Of course, when there is mean reversion, there are opportunities to trade: buy when the commodity is sufficiently below its (long-term) mean, sell when it is sufficiently above (with "sufficient" referring to a likelihood of return being high enough). This certainly sounds nice, but there is a small problem: we rarely know what the (true) mean is, and therefore we rarely know when we have identified a trading opportunity! Leaving this minor complication aside for the moment, a very interesting mathematical problem for discerning the underlying structure of such strategies can be crafted in terms of stochastic control and dynamic programming, topics which have been investigated in Section 3.3 (and will be returned to in Section 7.6).

To make this context a bit clearer, consider the results of Boguslavsky and Boguslavskaya (2004), introduced in Section 3.3.1. A standard OU process is studied, and a trader is assumed to make transactions with the objective of maximizing terminal wealth.[33] The resulting stochastic control problem proves tractable (via the well-known HJB formulation), yielding results for value function dynamics and optimality criteria for trading decisions. This work is based, of course, on the unrealistic assumption that the true mean reversion level is known. It would be highly useful to extend these results to the case where the mean is *not* known, and decisions must be made based on some projection of that mean, as inferred from observable prices. This is precisely where the representation in (6.89) and (6.93) could prove fruitful. This system should allow for application of the HJB-type procedures discussed in Chapter 3 under the incomplete information filtration (which of course is the actual situation we encounter) by taking a value function of the form $V(z, \hat{x}, w, t)$ wrt. to the (Markovian) dynamics in (6.89). (Here $w$ denotes wealth, *e.g.*, for the trading control problem.) It is not immediately clear how tractable the resulting PDE will be, as it lacks the affine structure possessed by the complete information case. (Another complication is that under the observation filtration, the [instantaneous] covariance matrix in the dynamics of (6.89) becomes time dependent.) We cannot hope to address this topic here, and so leave it as a potentially interesting avenue of future research.

#### 6.2.2.7 *Performance*

A natural question finally arises: how well does all this work in practice? Sadly, the answer is, not very well. We will illustrate with a familiar example, a

mean-reverting process with stochastic reversion level. The continuous-time form and its discrete-time counterpart are:

$$\begin{aligned} dx &= \kappa_x(y-x)dt + \sigma_x dw_x \\ dy &= -\kappa_y y dt + \sigma_y dw_y \\ x_{n+1} &= (1-\kappa_x \Delta t)x_n + \kappa_x \Delta t \cdot y_n + \varepsilon_n^x \\ y_{n+1} &= -\kappa_y \Delta t \cdot y_n + \varepsilon_n^y \end{aligned} \tag{6.94}$$

with (instantaneous) correlation $\rho$. In the notation of (6.63), we will consider the following numerical example:

$$F = \begin{pmatrix} 0.25 & 0.75 \\ 0 & 0.05 \end{pmatrix}, Q = \begin{pmatrix} 0.3^2 & -0.6 \cdot 0.3 \cdot 0.02 \\ -0.6 \cdot 0.3 \cdot 0.02 & 0.02^2 \end{pmatrix} dt \tag{6.95}$$

with $dt = 1/365$. In other words, the process $x$ is essentially a standard mean-reverting process; the mean-reversion level is stochastic but very slowly varying (and very nearly white noise). We first consider the full-information case, which in the notation of (6.64) implies $H = I$. In this case the usual machinery of VAR econometrics can be employed, but it is useful to see the output of a numerical, nonlinear optimization. This is shown in Figure 6.3 for the variance of process $x$. As can be seen, the estimator, although convergent (the average across paths gives a volatility of 30%, the true value), is definitely not Gaussian. In contrast, exact solution of the MLE conditions for the VAR *does* produce an asymptotically normal distribution. Clearly, there are effects due to the numerical optimization at play. We will see more on this kind of problem in Section 6.5.4.

Next we consider the partial information case, where only process $x$ is observed, so that we take $H = (1 \;\; 0)$. Results are shown in Figure 6.4.

Now, in addition to non-normality, the convergence properties are worse (the sample average volatility is 27%). In fact, the estimators of the (unobserved) mean and correlation perform terribly, giving sample averages of 200% and −4% (respectively)! It almost goes without saying that the estimator of the mean reversion coefficients (the matrix $F$ in (6.95)) is poor; we have already seen examples of this behavior with the standard VAR. What is striking here is that the covariance estimator is also poor; typically, covariance estimators are comparatively robust. We have attempted to put forward the best case for the filtering estimator here by focusing on the variance; however, we actually have produced a cautionary tale about filtering.

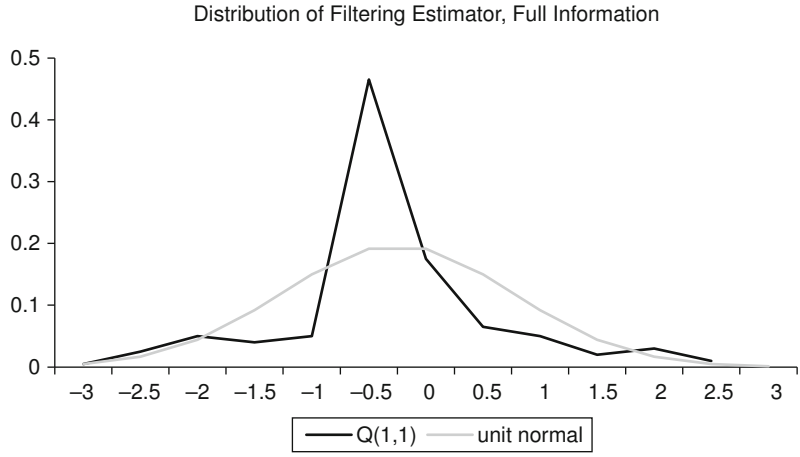



**Figure 6.3** Standardized filtering distribution, full information case. The estimator is unbiased and consistent, but due to numerical issues is clearly non-normal (200 simulations)

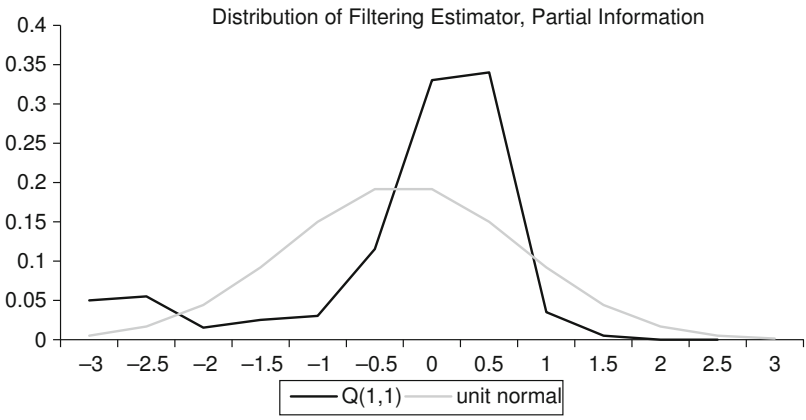



**Figure 6.4** Standardized filtering distribution, partial information case

### 6.2.3 Heston vs. generalized autoregressive conditional heteroskedasticity (GARCH)

*6.2.3.1 Overview and commentary*

As we have just seen, the possibility that the dynamics of interest may depend on unobserved entities such as stochastic volatility makes it necessary to resort to computationally intensive procedures such as filtering. In addition to the formidable challenges of the actual calculations, there is the ubiquitous problem of non-robustness in finite samples, especially for nonlinear problems (even those without

a great deal of structure). There is thus some interest in dealing *only* with observable entities. The so-called autoregressive conditional heteroskedasticity (ARCH) and GARCH classes of models[34] effectively model returns, or more accurately residuals, both of which are directly observable. To illustrate, consider the AR($p$) process

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots \phi_p y_p + u_t \tag{6.96}$$

The random error/shock/innovation $u$ is now assumed to have the following form:

$$u_t = \sqrt{h_t}\varepsilon_t \tag{6.97}$$

where $\varepsilon_t$ is standard white noise and $h$ satisfies

$$h_t = \omega + \beta_1 h_{t-1} + \cdots + \beta_r h_r + \alpha_1 u_{t-1}^2 + \cdots + \alpha_q u_{t-q}^2 \tag{6.98}$$

Eq. (6.98) is termed a GARCH($r, q$) model. (For $r = 0$ it becomes simply an ARCH($q$) model.) Note that, at time $t$, the "amplitude," so to speak, of the randomness in (6.97) is known via (6.98), although of course the white noise component is not. (As can be seen from (6.98), one of the purposes of ARCH models is to capture the well-known effect of volatility clustering where periods of relatively greater turbulence are followed by periods of relatively greater tranquility.) The system (6.96)–(6.98) can equivalently be written as

$$\begin{aligned} y_t &= \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \sqrt{h_t}\varepsilon_t, \;\; \varepsilon_t \sim N(0,1) \\ h_t &= \omega + \sum_{i=1}^{r} \beta_i h_{t-i} + \sum_{i=1}^{q} \alpha_i \left( y_{t-i} - \phi_0 - \sum_{j=1}^{p} \phi_j y_{t-i-j} \right)^2 \end{aligned} \tag{6.99}$$

The system in (6.99) is jointly estimated via maximum likelihood, using the conditional density $\exp(-\frac{1}{2}(y_t - \phi_0 - \phi_1 y_{t-1} - \cdots - \phi_p y_p)^2/h_t)/\sqrt{2\pi h_t}$.[35,36] This nonlinear problem ordinarily must be solved via numerical optimization. The primary point here is that, unlike stochastic volatility models that are similar (but not equivalent, see below), all of the analysis takes place in terms of observables (*i.e.*, no filtering is required). This particular issue does not greatly concern us here, as there is quite a voluminous literature on GARCH models that the reader can refer to if interested (see Endnote 34). We are instead concerned with tractable GARCH models that relate to the canonical stochastic volatility processes previously considered in Chapter 2. It is a somewhat common misconception that, in the small-time limit, GARCH models converge to a diffusive, stochastic volatility process. This is not necessarily the case, as certain technical conditions must be further imposed

(see Alexander and Lazar [2004]). Thus, GARCH models should be thought of as an *alternative* to stochastic volatility models, and not a discrete-time counterpart.

This nonequivalence is in fact not surprising. Precisely the point of these models is to separate unobservable state dynamics from their observable manifestations. In other words, the operative viewpoint is that it is not so important *why* returns/residuals are, say, autocorrelated with nondeterministic volatility, only *that* they are so. The main concern is then with modeling these observable features directly, as opposed to indirectly via some mechanism responsible for their manifestation, as in stochastic volatility models. One obvious concern here is that these mechanisms may, in fact, be important, and as a consequence it may be good to try to understand them as opposed to simply (formally) capturing their effects in a model of price returns. For example, a GARCH model that mimics the behavior of a Heston process may not correspond to any discretization of the Heston SDE and may not provide any insight on the underlying stochastic variance driver. This may not be important for many applications, but that is a separate issue.

Of larger concern is the fact that these models in some sense replace one problem with another one, while avoiding the main challenges. Without doubt, the filtering problem posed by stochastic volatility models (or any other model with unobserved state variables) is often quite formidable, as we have seen. Avoiding this problem is definitely beneficial. However, the GARCH paradigm accomplishes this by constructing an essentially *ad hoc* modeling framework that possesses just as much (if not more) structure as the stochastic volatility models it seeks to overturn. (This *ad hoc* nature is why, to be honest, so much GARCH modeling gives the impression of an everything-*and*-the-kitchen-sink approach; again refer to Endnote 34.) Thus, the concerns we have raised here, namely the lack of robustness of complex models in the face of limited data and information, apply just as much to the GARCH repertoire as they do to more standard, continuous-time models. It is simply not clear what is gained by the alternative paradigm put forth by GARCH.

#### *6.2.3.2 Time scales: they are everywhere*

There is in fact a stronger case to be made against GARCH modeling, and it relates to one of our central themes: GARCH models are not scale invariant. That is to say, they are not consistent across iterations of the (purported) dynamics. To see what we mean, consider a standard AR(1) process:

$$x_n = \phi x_{n-1} + \sigma \varepsilon_n \tag{6.100}$$

with $\varepsilon \sim N(0,1)$. Now, iterating (6.100), we see that

$$x_n = \phi^2 x_{n-2} + \sigma \varepsilon_n + \phi\sigma \varepsilon_{n-1} = \phi^2 x_{n-2} + \sigma\sqrt{1+\phi^2}\tilde{\varepsilon}_n \tag{6.101}$$

with $\tilde{\varepsilon} \sim N(0,1)$. It can be seen that (6.101) is *still* an AR(1) process, with of course different parameters. The relation between the parameters as the process is iterated reflects the scaling behavior of the underlying process. The process remains AR(1) whether one considers the dynamics at a daily resolution or a monthly resolution (say). There is an isomorphism, so to speak, across time horizons.

In contrast, GARCH models do not display this consistency. For example, consider the GARCH(1, 1) model

$$\begin{aligned} x_n &= \phi x_{n-1} + \sqrt{h_n}\varepsilon_n \\ h_n &= \omega + \beta_1 h_{n-1} + \alpha_1 (x_{n-1} - \phi x_{n-2})^2 \end{aligned} \tag{6.102}$$

We see that

$$\begin{aligned} x_n &= \phi^2 x_{n-2} + \phi\sqrt{h_{n-1}}\varepsilon_{n-1} + \sqrt{h_n}\varepsilon_n \\ &= \phi^2 x_{n-2} + \sqrt{h_n + \phi^2 h_{n-1}}\tilde{\varepsilon}_n = \phi^2 x_{n-2} + \sqrt{\tilde{h}_n}\tilde{\varepsilon}_n \end{aligned} \tag{6.103}$$

Now, it is not hard to see that $\tilde{h}_n \equiv h_n + \phi^2 h_{n-1}$ does *not* follow the GARCH recipe in (6.102) for a two-step time horizon. In other words, we cannot speak of a GARCH model as such across time scales/horizons. Rather, we would have separate GARCH models for different time horizons under consideration. This fact highlights yet again the essentially *ad hoc* nature of GARCH. But perhaps more importantly, it demonstrates the basic inability of GARCH to shed light on the dynamics that manifest themselves in the scaling laws whose importance in actual valuation problems cannot be understated.

We have stepped onto our soap box here because GARCH modeling *is* extremely popular, so we feel the need to explain why we devote such little time to it in this volume. As we have already mentioned, we do wish to focus on a special case where there is a continuous-time connection to a member of the canonical class of affine jump diffusions, namely the Heston stochastic volatility model.

#### 6.2.3.3 *A continuous-time limit*

Heston and Nandi (2000) propose the following GARCH-type model:[37]

$$\begin{aligned} z_t &= z_{t-1} + \lambda h_t + \sqrt{h_t}\varepsilon_t \\ h_t &= \omega + \beta h_{t-1} + \alpha\left(\varepsilon_{t-1} - \gamma\sqrt{h_{t-1}}\right)^2 \end{aligned} \tag{6.104}$$

with $\varepsilon \sim N(0,1)$ and independent of $h$. This model differs from the more familiar VGARCH and NGARCH models with which it bears a passing similarity, specifically through the presence of conditional variance as a return premium in the

(log-)price dynamics (compare also with Duan [1995]). Note in (6.104) that $h_t$ is observable at time $t-1$, while of course $z_t$ is not. By making the scaling $h_t = \Delta \cdot v_t$ where $\Delta$ is the time step, and introducing a suitable reparameterization, Heston and Nandi show that in the small step limit

$$\begin{aligned} &E_{t-\Delta}(v_{t+\Delta} - v_t) = k(\theta - v_t)\Delta + \mathrm{O}(\Delta^2) \\ &\mathrm{var}_{t-\Delta}(v_{t+\Delta}) = \sigma^2 v_t \Delta + \mathrm{O}(\Delta^2) \\ &\mathrm{Corr}_{t-\Delta}(v_{t+\Delta}, z_t) \to \pm 1 \end{aligned} \tag{6.105}$$

so that the limiting dynamics take the familiar Heston form

$$\begin{aligned} dz &= \lambda v dt + \sqrt{v} dw \\ dv &= \kappa(\theta - v)dt + \sigma\sqrt{v} dw \end{aligned} \tag{6.106}$$

So we see that the Heston-Nandi GARCH model does indeed have a continuous-time limit corresponding to a (well-known) stochastic volatility model, albeit one with rather stringent requirements on the instantaneous joint dynamics of log price and variance (*i.e.*, perfect correlation or anticorrelation). The Heston-Nandi model, like its continuous time counterpart, permits tractable computation of the conditional characteristic function (and hence facilitates the calculation of option values). To see this, consider the following iterated expectations:

$$\begin{aligned} E_t e^{i\phi z_{t+n}} &= E_t E_{t+n-1} e^{i\phi z_{t+n}} = E_t e^{i\phi(z_{t+n-1} + \lambda h_{t+n})} E_{t+n-1} e^{i\phi\sqrt{h_{t+n}}\varepsilon_{t+n}} \\ &= E_t e^{i\phi(z_{t+n-1} + \lambda h_{t+n}) - \phi^2 h_{t+n}/2} = E_t E_{t+n-2} e^{i\phi z_{t+n-1} + (i\phi\lambda - \phi^2/2)h_{t+n}} \\ &= E_t e^{i\phi(z_{t+n-2} + \lambda h_{t+n-1})} E_{t+n-2} \\ &\quad e^{i\phi\sqrt{h_{t+n-1}}\varepsilon_{t+n-1} + (i\phi\lambda - \phi^2/2)\left(\omega + \beta h_{t+n-1} + \alpha\left(\varepsilon_{t+n-1} - \gamma\sqrt{h_{t+n-1}}\right)^2\right)} \end{aligned} \tag{6.107}$$

Now, we will spare the reader the gory details (Heston and Nandi [2000] may be consulted if so desired), but it should be clear from (6.107) that as this iteration is continued, an essentially affine exponential form will be produced. In fact, as Heston and Nandi show, the coefficients in this form can be calculated recursively as a system of difference equations, clearly an analogue to the system of ODEs that prevails in the continuous time case. It is interesting to note that a subsidiary result is the calculation of expectations such as $Ee^{aw+bw^2}$, with $w$ a standard normal. These expectations will play an important part in the discussion of Wishart processes to be analyzed in Subsections 6.3 and 8.1.5.

One facet of almost all the problems we encounter in energy markets is the question of joint dependence. This is an extremely important topic that we must now turn to, and we will focus on the particular dependency notion of copulas.

## 6.3 Sampling distributions

### 6.3.1 The reality of small samples

At the risk of repeating ourselves, limited sample size is a pervasive feature of energy market data. In fact, the risk of redundancy is worth taking because the impact of small samples on econometric investigations cannot be overstated. We simply cannot caution enough about the potential for producing seriously misleading results when this issue is not handled with sufficient care. (There is, however, little need to elaborate on the sources of this state of affairs: relatively [in comparison to financial markets] rapid changes in underlying market structure rendering older price history irrelevant, high volatilities/jumps presenting a barrier to robust estimation, *etc.*[38])

Ultimately, the central question comes down to one of stability of an estimator. It is important to know, when actual numerical values are churned out by an estimator operating on a particular set of data, how sensitive these values are to this specific realization. In principle, the underlying population could have produced a different realization, and it would be good to know if the estimator will produce wildly different values for a new sample, or whether the new output is reasonably well behaved. Recall that our objective is the establishment of *actual* portfolio positions to facilitate the extraction of a suitably identified value driver. It is thus imperative that whatever operational tools we adopt, they avoid simply optimizing to noise. This goal leads us necessarily to the question of estimator performance in finite samples, as opposed to some operationally meaningless asymptotic sense.

#### *6.3.1.1 Estimation in small samples*

We introduced the concept of sampling distributions in Section 2.2.2. There, the objective was to discuss entities (some might call them, wrongly, test statistics) that *indirectly* reflect a population characteristic of interest, while at the same time prove amenable to analysis of their distributional properties (in essence, an understanding of their robustness). A paradigmatic example is variance scaling laws as indicators of mean reversion. As is well known, variances are typically more robustly estimated than mean-reversion rates. An associated issue is how to ascertain the behavior of a variance estimator in an actual, finite sample.

A class of matrix-valued processes called Wishart processes prove suitable as a means of investigating the *finite* sample properties of variance estimators. (We will revisit such distributions in Section 8.1.5 as representations of matrix-valued joint dependency.) It will turn out that we can say quite a bit about these distributions in fairly general Gaussian cases (including the case of non-i.i.d. returns characterized by mean reversion), and derive more limited (but still useful) results for non-Gaussian cases (specifically concerning the mean of the estimator, thus representing an expansion of the results in Section 2.2.3).

### 6.3.2 Wishart distribution and more general sampling distributions

*6.3.2.1 Standard case*

Consider a $p$–vector $x$ of (zero mean) normal deviates with covariance matrix $\Sigma$. Assume further that we have $n$ independent realizations of this random vector, which we arrange as a $p \times n$ matrix:

$$X = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^n \\ x_2^1 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \cdots & \vdots \\ x_p^1 & x_p^2 & \cdots & x_p^n \end{pmatrix} \tag{6.108}$$

Now, note that the elements of $XX^T$ correspond to the usual ensemble estimates of the covariance matrix (*i.e.*, the sample covariance). For example, $(XX^T)_{ii} = x_i^j x_i^j$ which is of course a common estimator of $n\Sigma_{ii}$. We can thus begin to understand the properties of this estimator by analyzing the (random) matrix $XX^T$.

The $p \times p$ matrix $S \equiv XX^T$ is said to possess a Wishart distribution, denoted as $S \sim W_p(\Sigma, n)$. Here the positive integer $n$ is called the number of degrees of freedom. This law can be thought of as a generalization of the familiar chi-squared distribution (for the same degree of freedom):[39] $\chi^2(n) \stackrel{\text{law}}{=} W_1(1, n)$. The Wishart distribution is the distribution of the sample covariance matrix of a joint normal $N_p(0, \Sigma)$; *i.e.*, it is the sampling distribution for the MLE estimator of the covariance matrix.

Not surprisingly, we will examine the characteristic function of $S$ as a means to determining its probability distribution. For matrix-valued random variables we consider

$$f(\Theta) = Ee^{i\mathrm{Tr}(XX^T\Theta)} = Ee^{ix_k^j x_l^j \Theta_{lk}} = Ee^{ix_k^j \Theta_{kl} x_l^j} = (Ee^{ix_k^1 \Theta_{kl} x_l^1})^n \tag{6.109}$$

for (symmetric) matrix-valued Fourier variables $\Theta$. In (6.109) we have used the fact that the columns of $X$ (essentially, the sample of $x$) in (6.108) are independent. The problem is thus reduced to the calculation of an expectation of the form

$$g = E_t e^{i\Theta_{ij} z_i(T) z_j(T)} \tag{6.110}$$

with $z$ a correlated, driftless Brownian motion (with covariance structure represented by $\Sigma$). We have that

$$g_t + \frac{1}{2}\Sigma_{ij} g_{z_i z_j} = 0 \tag{6.111}$$

Owing to the quadratic structure of the exponent in (6.110), we seek a solution to (6.111) of the form

$$g = e^{\omega_0 + \frac{1}{2}\omega_{ij} z_i z_j} \tag{6.112}$$

with the coefficients $\omega$ being functions of time only (really $T - t$). Substituting (6.112) into (6.111) (and as usual making the transformation $t \to \tau \equiv T - t$), we get the following system of (matrix) Riccati ODEs:

$$\begin{aligned} \dot{\omega}_0 &= \frac{1}{2}\sum_{ij}\omega_{ij} = \frac{1}{2}\mathrm{Tr}(\omega\Sigma) \\ \dot{\omega} &= \omega\Sigma\omega \end{aligned} \tag{6.113}$$

with initial conditions $\omega_0(0) = 0$, $\omega(0) = 2i\Theta$. We will see such ODEs again in Section 8.1.5, *e.g.*, (8.68). Adopting the solution technique there, we write $\omega = u^{-1}$ so that $\dot{u} = -\Sigma$, from which we get $u = -\frac{i}{2}\Theta^{-1}(I - 2i\Theta\Sigma\tau)$ and finally

$$\omega = 2i(I - 2i\Theta\Sigma\tau)^{-1}\Theta \tag{6.114}$$

Turning attention to the first equation in (6.113), we have that[40]

$$\begin{aligned} \dot{\omega}_0 &= \mathrm{Tr}((I - 2i\Theta\Sigma\tau)^{-1} i\Theta\Sigma) \Rightarrow \\ \omega_0 &= -\frac{1}{2}\mathrm{Tr}\log(I - 2i\Theta\Sigma\tau) \Rightarrow \\ e^{\omega_0} &= \frac{1}{\sqrt{\det(I - 2i\Theta\Sigma\tau)}} \end{aligned} \tag{6.115}$$

Putting all of this together, by taking $z = 0$ and $\tau = 1$ we find that the characteristic function of the general Wishart RV is

$$Ee^{i\mathrm{Tr}(XX^T\Theta)} = \det(I - 2i\Theta\Sigma)^{-n/2} \tag{6.116}$$

From this result, the probability density of a Wishart variable is given by

$$\Pr(S) = \frac{\det(S)^{\frac{n-p-1}{2}} e^{-\frac{1}{2}\mathrm{Tr}(\Sigma^{-1}S)}}{2^{\frac{np}{2}} \det(\Sigma)^{n/2} \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{n-j+1}{2}\right)} \tag{6.117}$$

See Adhikari (2007). Note from (6.116) that

$$Ee^{i\mathrm{Tr}(\frac{1}{n}XX^T\Theta)} \to e^{i\mathrm{Tr}(\Theta\Sigma)} \tag{6.118}$$

as $n \to \infty$, indicating convergence to the true covariance matrix in the large sample limit (owing to the delta function form of the inverse transform of (6.118)). (In fact, it can also be seen from (6.116) that $E\frac{1}{n}XX^T = \Sigma$, so as an estimator the sample covariance matrix is unbiased [for i.i.d. Gaussian variables].)

#### 6.3.2.2 *Extensions I: non-I.I.D. Gaussian*

We will now indicate how the results for the Wishart distribution in (6.116) can be extended to more general situations. Space constraints will prevent us from providing full details, but by now the reader should be sufficiently well equipped with the necessary tools to embark on the project himself.

Recall the canonical affine diffusion model (*sans* jumps) from Section 5.2.3:

$$dz_i = (A_{ij}z_j + b_i)dt + \sigma_i^k \sqrt{z_k} dw_i^k + \sigma_i^0 dw_i^0 \tag{6.119}$$

where selected elements of $z$ can be thought of as log prices. We remain concerned with the distributional properties of the sample covariance estimator over a (finite) set of $N$ returns:[41]

$$\hat{X}_N = \frac{1}{N}\sum_{i=1}^{N} \Delta z_i \Delta z_i^T \tag{6.120}$$

where $\Delta z_i \equiv z(t_i) - z(t_{i-1})$ for some set of observation times $t_i$. Proceeding as in the i.i.d. case, we can discern the distribution of the sample covariance matrix $\hat{X}_N$ from its characteristic function. However, the problem in the general case is that returns are not necessarily i.i.d., so we cannot invoke an argument like that used in (6.109). However, we apply an iterative procedure that allows us to build up the sample-wise characteristic function from the (in principle) known conditional characteristic functions. We will outline the idea here; we will return to it in Section 6.5.5 when we discuss spectral methods in econometrics.

First we treat the purely Gaussian case, where $\sigma^k = 0 \cdot \forall k \geq 1$. Consider the characteristic function of $\hat{X}_N$:

$$E_{t0} \exp(i\mathrm{Tr}(\Phi \hat{X}_N)) = E_{t_0} \exp\left( i\frac{1}{N}\sum_{i=1}^{N} \mathrm{Tr}(\Phi \Delta z_i \Delta z_i^T) \right) \tag{6.121}$$

Now, unlike the i.i.d. case, we cannot reduce (6.121) to the $N$-th power of the single time step characteristic function (as in (6.109)). However, using iterated expectations, we can write (6.121) as

$$E_{t_0}\exp\left(i\frac{1}{N}\sum_{i=1}^{N}\mathrm{Tr}(\Phi\Delta z_i\Delta z_i^T)\right)$$
$$=E_{t_0}\left[\exp\left(i\frac{1}{N}\sum_{i=1}^{N-1}\mathrm{Tr}(\Phi\Delta z_i\Delta z_i^T)\right)E_{t_{N-1}}\exp\left(i\frac{1}{N}\mathrm{Tr}(\Phi\Delta z_N\Delta z_N^T)\right)\right] \tag{6.122}$$

The innermost expectation in (6.122) can be written as

$$\exp\left(i\frac{1}{N}z_{N-1}^T\Phi z_{N-1}\right)E_{t_{N-1}}\exp\left(i\frac{1}{N}(z_N^T\Phi z_N-z_N^T\Phi_{N-1}-z_{N-1}^T\Phi z_N)\right) \tag{6.123}$$

A difficulty presented by the expectation in (6.123) is the presence of the time $t_{N-1}$ state variables, which means that the conditional expectation is *not* a martingale (nor is it amenable to the discounting adjustment used in the derivation of the familiar Feynman-Kac formula). However, the following expectation

$$E_{t_{N-1}}\exp\left(i\frac{1}{N}(z_N^T\Phi z_N-2z_N^T\Phi x)\right) \tag{6.124}$$

for arbitrary $x$ *is* amenable to standard techniques, whereby the expectation satisfies a PDE (an evolution equation involving the process infinitesimal generator) whose solution can be reduced, owing to the special form of the underlying dynamics, to a system of ODEs. Specifically, we seek a solution to (6.124) of the form

$$\exp(z_{N-1}^T\alpha z_{N-1}+z_{N-1}^T\beta+\gamma) \tag{6.125}$$

where $\alpha$, $\beta$, and $\gamma$ are resp. matrix-, vector-, and scalar-valued functions (of time to maturity). Using (by now) standard techniques, it is straightforward to show that these coefficients satisfy the following (matrix) Riccati system of ODEs:

$$\begin{aligned}\dot{\alpha}&=2A^T\alpha+2\alpha X^\circ\alpha\\ \dot{\beta}&=(A^T+2\alpha X^\circ)\beta+2\alpha b\\ \dot{\gamma}&=b^T\beta+\frac{1}{2}\beta^T X^\circ\beta+\mathrm{Tr}(\mathrm{X}^\circ\alpha)\end{aligned} \tag{6.126}$$

where we have the initial conditions $\alpha(0)=\Theta$, $\beta(0)=-2\Theta x$, and $\gamma(0)=0$.[42] With this (formal) solution in hand, the conditional expectation in (6.123) can be obtained by setting $x=z_{N-1}$.

We are not concerned here with the solution of the system in (6.126), as we have already briefly discussed matrix Riccati ODEs. Rather, our interest lies in exploiting

the tractable nature of the calculation of the one-step conditional expectation (via reduction to a system of ODEs) to iteratively construct the characteristic function of the finite sample estimator in (6.120). To see how this is possible, consider conditioning on time $t_{N-2}$ in (6.122). Using the expression in (6.125), we are confronted with the following expectation:

$$E_{t_{N-2}} \exp\left( i\frac{1}{N}\Delta z_{N-1}^T \Phi \Delta z_{N-1} + z_{N-1}^T \alpha z_{N-1} + z_{N-1}^T \beta + \gamma \right) \tag{6.127}$$

Now, owing to the quadratic form of the exponent in (6.127), we can compute the conditional expectation in the same way as done with the construction in (6.125) and (6.126), with of course different initial conditions for the ODEs. This approach is facilitated by the following important fact. The system in (6.126), owing to its structure and initial conditions, is affine-quadratic in the state variable. (Specifically, $\alpha$ does not depend on the state variable, $\beta$ is affine, and $\gamma$ is affine-quadratic.) Thus, the exponent in (6.127) retains the necessary affine-quadratic form. It should then be clear that continued iterations of the expectation operator (conditional on the decreasing series of observation times) will produce the desired characteristic function. We leave as an open question whether any tractable results akin to those in Section 2.2.3 can be obtained.

#### 6.3.2.3 *Extensions II: non-Gaussian*

In the non-Gaussian case (*e.g.*, generalized versions of Heston stochastic volatility), we unfortunately cannot employ the methods used in the previous section, because of the presence of expressions involving $z$ in the Hessian (second-derivative) term. While the characteristic function may be out of reach (as far as those tractable techniques are concerned), the expected value of the finite size sample variance *is* obtainable. In other words, we can evaluate

$$E_{t_0} \hat{X}_N = \frac{1}{N} E_{t_0} \sum_{i=1}^{N} \Delta z_i \Delta z_i^T \tag{6.128}$$

Ultimately, the problem amounts to the calculation of expectations of the following form:

$$E_{t_0} z_m(t_i) z_n(t_i) \tag{6.129}$$

(Note that we have switched notation so that subscripts now refer to a vector component, not a vector evaluated at some time index.) We can seek solutions to (6.129) of the form $z^T \alpha z + z^T \beta + \gamma$, with the coefficients (as functions of time to maturity)

satisfying the following system of ODEs:

$$\begin{aligned}\dot{\alpha} &= 2A^T\alpha \\ \dot{\beta} &= A^T\beta + 2\alpha b + H \\ \dot{\gamma} &= b^T\beta + H_0\end{aligned} \tag{6.130}$$

where the vector $H$ has coefficients given by $H_k = \mathrm{Tr}(X^k\alpha)$ and the initial conditions are $\alpha(0) = \delta_{mn}$, $\beta(0) = \gamma(0) = 0$. Again, we eschew detailed analysis of this system, and reaffirm the main point that expressions for the finite sample ensemble covariance can be derived for the very general class of affine diffusions.

Having given some attention to the analytics of finite sample estimation, we now turn to some of the more practical matters.

## 6.4 Resampling and robustness

In this section we will address issues pertaining to the information that can be extracted from a sample *itself*, while making minimal resort to assumptions about the underlying DGP and theoretical results that depend on large sample sizes.

### 6.4.1 Basic concepts

It will prove useful to revise somewhat our abstract definition of an estimator from Section 2.1.1. There, we defined an estimator to be a map from some (finite) sample to some set of parameters, which stand in some relation to a set of model parameters defined through an underlying probability distribution. Here, we wish to retain that sense of abstraction while being a bit less explicit about the underlying probabilistic nature.

To this end, let us define an econometric model to be a relation of the following form:

$$\mathfrak{M}(z_i, \varepsilon_i; \theta) = 0, \quad \forall i \in \{1, \ldots, S\} \tag{6.131}$$

where $z_i$ denotes the $i^{\text{th}}$ element of some vector-valued sample (of size $S$) of observable data, $\varepsilon_i$ is an unobservable vector (which can be thought of as corresponding to some random driver or noise with generic distributional/dynamic properties [although we will typically assume independence across realizations]), and $\theta$ is some vector of model parameters. Equivalently, we can view the model as a parameter-dependent map from the set of unobservables to the set of observables: $\mathfrak{M}_\theta : \varepsilon \Rightarrow z$. An estimator associated with a model is defined to be a map from the set of observables to an estimate of the model parameters:

$$\mathfrak{E}_{\mathfrak{M}}(\{z_i\}) = \hat{\theta} \tag{6.132}$$

Finally, we specify a mechanism for extracting estimates of the unobservables from the parameter estimates and the observables:[43]

$$\hat{\varepsilon}_i = \mathfrak{M}^{-1}(z_i;\hat{\theta}) \tag{6.133}$$

We term these the (realized) residuals.

A simple example would be the familiar AR(1) process from (2.17). Here the observables are the primary variable $x$ and its lagged value: $z_i = (x_i \;\; x_{i-1})^T$. The model parameter is simply the autoregressive coefficient: $\theta = \phi$. It is clear what the (unobservable) random term is ($\mathcal{E}$ is a Gaussian driver). We note here, however, that unlike the conventional AR formulation, the abstract definitions in (6.131)–(6.133) make no specifications about the nature of $\mathcal{E}$ (*e.g.*, there is no necessary assumption of normality). Indeed, we have not even claimed that the *model* in (6.131) is a DGP (or inquired as to the consequences if it is, *e.g.*, hypothesis testing). We thus make no (explicit) reference to a relation between estimator and true parameter value as in Section 2.1.1. Consequently, if we *formally* invoke a method such as MLE to estimate $\phi$ (as in (2.18)), we do not attempt to connect this construct to any kind of moment or other population condition. (See (2.16), and Section 6.5 below). However, we can still extract residuals via $\hat{\varepsilon}_i = x_i - \hat{\phi}x_{i-1}$.[44]

Since we decline to specify any sense in which the estimator establishes a connection to the "true" parameter values (or even what such a connection would mean), it may well be asked: how useful is this revised formalism? We will now attempt to address this concern.

### 6.4.2 Information conditioning

#### *6.4.2.1 Econometric challenges*

What are the primary challenges facing any econometric analysis? At the risk of painting with broad brush strokes, there are two central issues:[45]

1. Sample sizes are not large enough to permit robust analysis.
2. The DGP that generated the actual data is not (and cannot be) known in entirety, and indeed, may not be constant over time.

We will mention again that both of these issues are particularly acute in energy markets. As we have already pointed out (and will consider in great detail in Section 6.5), results for the properties of many popular estimators are often asymptotic in nature; that is, they apply only for large samples. But, it is never known in practice whether a given sample is large enough for the asymptotic results to be valid or useful, and we will see an example in Section 6.5 where a (simulated) data set that is large by the standards of energy-market data (*i.e.*, more than 1,000 points) can yield extremely poor results in comparison to the asymptotic theory. This example is particularly striking because the DGP is actually known. This brings us to point

number 2 above: we never really know what the DGP is! It is true that we may have various, good reasons for believing that volatility is stochastic or that a given price series has jumps. But it remains the case that we simply never know for sure, and at any rate energy markets are particularly characterized by rather rapid evolution in their structure (see EW). It is simply inconceivable that we would ever have a model that is both reasonably tractable and acceptably complex. This situation is in turn exacerbated by the small sample issue brought up in point number 1.

In (empirical) modeling, a general principle applies: there is a definite trade-off between robustness and structure. It is *very* easy to add complexity to a model sufficient to destroy its robustness *given* the sample size at our disposal. Given these problems, it is desirable to have an estimation technique that is both accommodative of small sample sizes *and* makes as few explicit modeling assumptions as possible. We can now turn to the abstract estimator framework introduced in Section 6.4.1.

#### *6.4.2.2 Models, estimators, and stability*

Consider the following situation. We have a set of variables $\{(x_i, y_i)\}$ and we want to investigate the relationship (if any) between $x$ and $y$. There are a few possibilities that immediately suggest themselves: (1) there is a linear relationship between the two, apart from Gaussian noise; (2) there is a nonlinear relationship (say, polynomial in $x$) with Gaussian noise; and (3) there is a linear relationship with non-Gaussian noise. Assume further the sample is small, say 25 elements. Which model, based on one of these three possibilities, is best in this situation? The answer is actually not obvious. Depending on the nature of $x$ and $y$, it can be very difficult (if not practically impossible) to statistically distinguish between the three models in a small sample (recall the example in (2.23), a stationary series with very weak mean reversion). Put differently, in this scenario could a wrong model perform better than a model based on the true relationship?

We will investigate this latter question, in the context of standard econometric techniques, in Section 6.5. Here we ask a different question: *given* a particular choice of model and an associated estimator (*e.g.*, a linear model along with OLS), what information can we extract from the available data themselves about this model-estimator pair? The issue can be characterized as one of stability. Recall that our estimation procedure produces estimates of the (unobservable) noise terms (see (6.133)). As noise *qua* noise (assuming independence for the moment), any particular realization of these terms in the sample is statistically equivalent to any other realization for the sample in question. This latter caveat is important: we are always concerned with estimation based on a sample of a particular size, and not in some operationally empty limiting case. Thus we see a potential criterion for assessing model robustness: how sensitive are the estimated parameters to the particular residuals (estimated errors) that happened to appear in the sample in question?

Put differently, if these errors changed, how much would the estimated parameters change?

Our concern here is to avoid optimizing to noise, which is very easy to do in models with much structure in small samples. The question then becomes: for our sample (which is the only data we have available), just how much structure is possessed by the model in question? It is important to understand that this question does not stand in an abstract sense, but only in relation to the data being modeled *and* estimated. Put this way, it is natural to investigate the impact of the residuals implied by the model and estimator. We would expect the effects of model structure, as manifested in the estimation procedure, to be reflected by the resulting residuals, which have the advantage of being easier to analyze. In fact, these residuals offer a natural means of hypothesis testing, so to speak. If the model in question characterizes the particular sample, *and* the associated estimator performs effectively for this sample, then the (implied) residuals should be good representatives of noise and not structure. In principle, then, these residuals could be used to generate (in some as-yet unspecified manner) *new* noise terms and thus new (synthetic, really) samples (incorporating supposed structure and noise). These synthetic samples can in turn be estimated (in a manner of speaking), and the output assessed.

#### 6.4.2.3 *Residuals and conditioning*

In other words, *conditional on the actual sample*, the model/estimator pair produces residuals whose (supposed) status *as* noise can be tested, at least in terms of stability. Visually this procedure can be expressed as

$$\begin{aligned} &\mathfrak{M}(z_i, \varepsilon_i; \theta) = 0, \ \mathfrak{E}_{\mathfrak{M}}(\{z_i\}) = \hat{\theta} \rightarrow \\ &\hat{\varepsilon}_i = \mathfrak{M}^{-1}(z_i; \hat{\theta}) \rightarrow \\ &\mathfrak{M}_{\hat{\theta}} : \hat{\varepsilon}'_i \Rightarrow z'_i \Rightarrow \mathfrak{E}_{\mathfrak{M}}(\{z'_i\}) = \hat{\theta}' \end{aligned} \tag{6.134}$$

In the event that this picture is worth less than a thousand words, let us explain what we mean here. From an econometric model that expresses a relationship between structure and noise, we have an estimator mapping observables to estimates of underlying (structural) parameter values. The original model then allows us to derive residuals (*i.e.*, estimates of the [unobservable] noise), as a function of the sample and estimated parameter values. In turn, these residuals can be used to construct synthetic samples from which *reestimated* parameter values can be obtained. Finally, the stability of the estimator and the robustness of the model, always in the context of a particular sample, can be assessed.

Note that this procedure can *always* be carried out, and it should be. The standard approach that views econometrics in terms of hypothesis testing of models and statistical significance of parameters is critically dependent on large samples and (largely) untestable assumptions about the underlying DGP.[46] By contrast, viewing

the problem as one of stability analysis is much broader in nature. For large enough samples and with sufficient knowledge of the DGP, stability analysis can of course be crafted as familiar testing of statistical significance. The reverse, however, is generally *not* true, and of course the usual situation is precisely one of small samples and insufficient information about the true DGP.

We are emphasizing two, although ultimately related, concepts in (6.134). First, we have the fact that *all* relevant information flows from the actual sample that we possess. (Obviously, if we have good prior reason to believe that this sample should be truncated or augmented, this should be done, but when the econometric analysis commences, we can *only* use that data for drawing conclusions.) Related to points we will raise when we discuss simulation in Chapter 7, imposing structure (through formal modeling) *cannot* create information not already present in the original sample. Furthermore, the residuals that arise from econometric analysis on this data are dependent on the model and associated estimator that are being employed. Second, as these residuals are resampled (in some model-dependent but otherwise still unspecified way) to produce synthetic samples, new estimates of the model parameter are produced. These new estimates can be used to assess, in light of the estimates derived from the original sample, the stability of the underlying estimator.

Thus, we have the following synergy:

- The sample as the central source of conditional information
- The residuals as dependent on sample, model, and estimator
- Resampled residuals as a means of assessing the robustness of the estimator given the model.

We can now turn attention to the operational question of how the generation of synthetic samples in (6.134) can be carried out, and shed more light on this synergy.

### 6.4.3 Bootstrapping

The idea behind the standard bootstrap, which can be characterized as resampling with replacement,[47] is actually quite simple and can best be conveyed with a concrete example. Consider estimation of a VAR process:

$$x_n = \Phi x_{n-1} + \varepsilon_n \tag{6.135}$$

Standard MLE (say) produces an estimate $\hat{\Phi}$ of the (regression) coefficient matrix and a corresponding set of residuals $\hat{\varepsilon}_n (\equiv x_n - \hat{\Phi} x_{n-1})$. As we have seen, in small samples (or more accurately, samples small in relation to the characteristic time scales of the underlying dynamics[48]) this estimator can perform very poorly. By this we mean that the (asymptotic) results that provide an idea of the distribution of the estimator in large samples are of little practical use in finite samples.

Bootstrapping eschews the question of distributional properties of the estimators and instead addresses the question of estimator sensitivity (of course, in large samples these two issues are basically equivalent). It does this by constructing synthetic time series on which re-estimation can be conducted. The basis of this re-estimation is the original estimated coefficient matrix and original residuals. A *new* set of residuals is created by resampling the original set, uniformly with replacement. This approach leads to a synthetic time series $x^i$ via

$$x_n^i = \hat{\Phi} x_{n-1}^i + \hat{\varepsilon}_{n^i} \tag{6.136}$$

where $n^i$ denotes the index of the $n$th element of the residuals for $i$th resampling.[49] This procedure may be repeated as many times as necessary, producing a set of (resampled) estimates $\hat{\Phi}^i$ . (An occasionally useful variant is so-called parametric bootstrapping, where some suitable distributional form is assumed for/fitted to the [finite sample] residuals and used as the basis of regenerating samples.) Note of course an implicit, but important, assumption underlying the method: independence of the population disturbances. This is a limitation of the method's applicability, and must be balanced against its undoubted simplicity. (So-called block bootstrapping, whereby blocks of data are resampled, is means of preserving some serial dependence.) For more on bootstrapping, see MacKinnon (2006).

Although it is common to present bootstrapping as a facilitator for conducting hypothesis testing and constructing confidence intervals, this is not how we would view it. As we discussed in Section 3.1.1, however useful the language and formalism of statistical inference may be, it is ultimately tangential to our ultimate objective of forming *definite* portfolios for the purpose of extracting a *specific* value associated with some structured product. As such, we cannot be content with knowing (in whatever sense) that some model parameter value is statistically significant or not; we have to put on actual values/hedges. It is therefore imperative that we have some idea of how sensitive these values are to the particular sample used in the econometrics. We are always receiving new information (data), and we need to know how important this information is for the task at hand. Bootstrapping is precisely such a tool for addressing these kinds of questions. The formalism of the previous subsection can now be seen as a necessary precursor for establishing the conceptual framework behind what is, ultimately, a very simple methodology (precisely the appeal).

Just in case the point is not sufficiently clear yet, we will close this chapter with a detailed discussion of the shortcomings of several popular econometric techniques in finite samples.

## 6.5 Estimation in finite samples

As a final lesson, we discuss the econometric ramifications of the reality of finite samples. We will focus on the results of an important study by Zhou (2001).[50] In this work, estimation of a Cox-Ingersoll-Ross (CIR) square-root process is considered:

$$dr = \kappa(\theta - r)dt + \sigma\sqrt{r}dw \tag{6.137}$$

(The study takes place in the context of interest rates, but (6.137) can be immediately recognized as the stochastic variance process of the Heston model.) This work provides a cautionary tale on the use of asymptotic results for drawing statistical inference, and nicely illustrates the trade-off between robustness and structure in estimation. The central lesson is that, even in situations where we have realized paths of length much greater than commonly encountered in energy markets and where the underlying DGP is known (Zhou applies simulation to generate samples paths of (6.137) of size 500–1,500), the asymptotic diagnostics (which are the typically the only theoretical/analytical results available for many common econometric techniques) can give extremely misleading (if not incorrect) results regarding the model parameter estimates.

Several standard techniques are applied, some of which have already been introduced; in the interest of being reasonably complete, we will provide brief overviews.

### 6.5.1 Basic concepts

#### *6.5.1.1 Central limit theorem*

We start with a foundational result that is, often only implicitly, at the heart of much econometric analysis: the celebrated Central Limit Theorem (CLT)[51]. This result states that for a realized sample $\{X_i\}$ of size $N$ of i.i.d. variables with mean $\mu$ and *finite* variance $\sigma^2$, the sample mean and population mean are related via

$$\frac{\sqrt{N}}{\sigma}\left(\frac{X_1 + \cdots X_N}{N} - \mu\right)^d \sim N(0,1) \tag{6.138}$$

as $N \to \infty$ . In other words, the entity on the LHS of (6.138) is distributionally a standard normal as the sample size tends to infinity. We know the reader would be

disappointed if we did not invoke characteristic functions to establish this result, so we have that (WLOG we take $\mu = 0$):

$$E\exp\left(i\phi N^{-1/2}\sum_i X_i\right) = (Ee^{i\phi N^{-1/2}X_1})^N = f_X(\phi N^{-1/2})^N$$
$$= \left(1 - \frac{\sigma^2\phi^2}{2N} + O(N^{-2})\right)^N \to e^{-\sigma^2\phi^2/2} \qquad (6.139)$$

The CLT thus provides a connection between sample average and population mean. Let us now see how the CLT is used in econometric diagnosis. We first turn to the actual estimators themselves.

*6.5.1.2 Maximum likelihood*

The standard workhorse in a great many econometric techniques is MLE, which we have already introduced in Section 2.1.1, and in fact used throughout. To facilitate the fluidity of the exposition we repeat the relevant results here. In MLE we seek those values of a model that, if true, are most consistent with a given set of realized data. (By consistency we mean the plain language meaning: with alternative values, there is a lower probability that the model would have generated the observed data.) This goal is achieved by maximizing the so-called log-likelihood function over the data set:

$$\hat{\theta} = \arg\max_\theta \mathcal{L}(z;\theta) \qquad (6.140)$$

where $\theta$ are the model parameters, the hat denotes estimated entities, $z$ is the data set, and the log-likelihood function is given by

$$\mathcal{L}(z;\theta) = \frac{1}{T}\log\prod_t f(z_t;\theta) = \frac{1}{T}\sum_t \log f(z_t;\theta) \qquad (6.141)$$

where $f(z_t;\theta)$ is the *functional form* of the (parameter-dependent) density of observation $z_t$ and $T$ the number of observations. (The implicit assumption in (6.141), as we shall see, is that the observations are independent and identically distributed.) A simple application would be to the standard linear model in (2.5), where the density is Gaussian and the usual OLS estimators can be recovered.[52] Note that, evaluated at the true parameter value $\theta^*$, the first-order optimality condition satisfies (assuming certain regularity conditions; see Hamilton [1994])

$$E\frac{\partial}{\partial\theta}\mathcal{L}(z;\theta)\Big|_{\theta^*} = \frac{1}{T}\sum_t E\frac{\partial}{\partial\theta}\log f(z_t;\theta)\Big|_{\theta^*}$$
$$= \frac{1}{T}\sum_t \frac{\partial}{\partial\theta}\int dz f(z;\theta)\Big|_{\theta^*} = 0 \qquad (6.142)$$

since the density integrates to 1. This essentially demonstrates the (asymptotic) consistency of MLE.

To support this latter claim, we need to investigate what, exactly, the formal mechanics of optimizing (6.141) means. In other words, how does the output of the optimization procedure relate to the actual model parameters? It is here we invoke the CLT. *Under i.i.d. assumptions*, the *sample* average

$$\frac{1}{T}\sum_t \frac{\partial}{\partial\theta}\log f(z_t;\theta) \tag{6.143}$$

is asymptotically distributed (normally) around the *population* expectation

$$E\frac{\partial}{\partial\theta}\log f(z;\theta) \tag{6.144}$$

with the variance of the distribution decreasing like the reciprocal of the sample size. Since the expectation in (6.144) is zero for the true model parameter, the first-order optimization condition in MLE corresponds to the sample analogue of a particular population condition. (We will see in Section 6.5.3 how this condition can be effectively inverted to provide [asymptotic] distributional information about the estimators themselves.)

Now, the i.i.d. assumption is fairly restrictive, and it is common to apply the formal optimization in (6.140) in situations outside the range where the CLT is rigorously valid. Specifically, when applied to the realizations of processes such as (6.137), the constituent terms in the log-likelihood function (6.141) are *conditional* densities, not unconditional densities as required by the CLT. Our chief concern here is to outline what can go wrong in such cases. Of course, a formal procedure can always be used mechanistically; however, extreme care must be employed when interpreting the output. The question ultimately remains: what is the connection between the sample condition derived from (6.143) and the population condition derived from (6.144)? Let us frame this question in an appropriate form.

#### *6.5.1.3 Averages: population vs. sample vs. time*

The typical approach for applying MLE to processes such as (6.137) is to optimize the following log-likelihood function:

$$\mathcal{L}(z;\theta) = \frac{1}{T}\sum_t \log \Pr_\theta(z_t|z_{t-1}) \tag{6.145}$$

where $\Pr_\theta(z_t|z_{t-1})$ denotes the parameter-dependent (conditional) transitional density for the process. Clearly, as random variables, each constituent term in the summation in (6.145) has a different distribution, hence the i.i.d. assumption *cannot* be valid, and we cannot appeal to the standard CLT in associating the first-order

condition from optimizing (6.145) with a population condition as we did in (6.143) and (6.144). However, *conditionally* this connection of course still exists. That is, for a *given x* we can write

$$\frac{1}{T_z}\sum_t \frac{\partial}{\partial\theta}\log \Pr_\theta(z_t|x) = 0 \Leftrightarrow E_x \frac{\partial}{\partial\theta}\log \Pr_\theta(z|x) = 0 \tag{6.146}$$

as the conditional sample size $T_z \to \infty$. Now, thinking of $x$ as drawn from some population with an unconditional density $\Pr_\theta(x)$, we can invoke iterated expectations and write the following:

$$\frac{1}{T_z T_x}\sum_{t,u} \frac{\partial}{\partial\theta}\log \Pr_\theta(z_t|x_u) = 0 \Leftrightarrow E \frac{\partial}{\partial\theta}\log \Pr_\theta(z|x) = 0 \tag{6.147}$$

with the usual CLT asymptotics applying (for the limit of large $T_x$). However, it is difficult (if not impossible) to implement (6.147) in practice because we typically do not have "cross-sectional"[53] (so to speak) realizations of the DGP; we typically have only a *single* path on which to base the estimation.[54] (This is precisely why time averages across conditional densities [as in (6.145)] are universally resorted to in practice.)

Nonetheless, we anticipate that, *for sufficiently large sample sizes*, time averages such as (6.145)[55] will well approximate the unconditional expectations needed for invoking standard limiting results, so long as the underlying stochastic drivers are sufficiently "nice." To be a bit more precise, what is required is the property of *ergodicity*, which means that large-sample time averages of a (stationary) stochastic process converge (in the appropriate distributional sense) to the unconditional expectation of the process. (Intuitively, ergodicity implies that points sufficiently far apart on a path are independent.) In fact, the standard CLT can be generalized to the case of stationary, ergodic martingale differences[56] (Billingsley, 1961). To outline the idea, consider a stochastic process $X_t$ (wrt. some filtration) satisfying $E_{t-1}X_t = 0$, with state-independent conditional variance $E_{t-1}X_t^2 = \sigma_{t-1}^2 < \infty$, and possessing an unconditional mean: $EX_t = 0$. Proceeding as in (6.139), we see that

$$\begin{aligned}
E\exp\left(i\phi N^{-1/2}\sum_{i=1}^{N} X_i\right) &= E\left(\exp\left(i\phi N^{-1/2}\sum_{i=1}^{N-1} X_i\right) E_{N-1} e^{i\phi N^{-1/2}X_N}\right) \\
&= E\left(\exp\left(i\phi N^{-1/2}\sum_{i=1}^{N-1} X_i\right) f(i\phi N^{-1/2}; X_{N-1})\right) \\
&= E\left(\exp\left(i\phi N^{-1/2}\sum_{i=1}^{N-2} X_i\right) E_{N-2}\left(e^{i\phi N^{-1/2}X_{N-1}}\left(1 - \frac{\sigma_{N-1}^2\phi^2}{2N} + O(N^{-2})\right)\right)\right)
\end{aligned} \tag{6.148}$$

where $f$ is the (conditional) characteristic function of the process $X$. Continuing with the iterated expectations, we see that (6.148) behaves asymptotically like

$$\begin{aligned} & E\exp\left(I\phi N^{-1/2}\sum_{i=1}^{N}X_i\right) \\ & \quad =\prod_{i=1}^{N}E\left(1-\frac{\sigma_{i-1}^2\phi^2}{2N}+O(N^{-2})\right)\sim\exp\left(-\frac{\bar{\sigma}^2\phi^2}{2}\right) \end{aligned} \tag{6.149}$$

as $N\to\infty$, and where $\bar{\sigma}^2=\lim_{N\to\infty}N^{-1}\sum_{i=1}^{N}\sigma_{i-1}^2$. (See Proposition 7.8 in Hamilton [1994].) The asymptotic normality of the large-sample time average is thus established by (6.149).[57]

Keep in mind that the relevant characterization of stationarity/ergodicity applies to the process $X$, which may itself be a function of other process, which is the actual objective of the econometrics. *E.g.*, in MLE we would be concerned with time averages such as (6.145), from which we extract estimates of the model parameters for the underlying DGP. With the obvious extension of (6.149) to higher dimensions, we calculate the conditional covariance in the case of MLE to be (recall the argument used in (6.142))

$$\begin{aligned} & E_t\frac{\partial}{\partial\theta}\log\mathrm{pr}_\theta(z_{t+1}|z_t)\frac{\partial}{\partial\theta^T}\log\mathrm{Pr}_\theta(z_{t+1}|z_t) \\ & \quad =E_t\frac{1}{\mathrm{Pr}_\theta}\frac{\partial^2}{\partial\theta\partial\theta^T}\mathrm{Pr}_\theta(z_{t+1}|z_t)-E_t\frac{\partial^2}{\partial\theta\partial\theta^T}\log\mathrm{Pr}_\theta(z_{t+1}|z_t) \\ & \quad =-E_t\frac{\partial^2}{\partial\theta\partial\theta^T}\log\mathrm{Pr}_\theta(z_{t+1}|z_t) \end{aligned} \tag{6.150}$$

yielding the familiar result that the asymptotic covariance of the MLE estimator is $-T^{-1}\mathfrak{I}^{-1}$, where $\mathfrak{I}$ is the well-known Fisher information matrix (the Hessian of the log density). A more casual derivation will be provided in Section 6.5.2.

#### *6.5.1.4 Asymptotics: a sense of false security*

We have engaged in this somewhat lengthy exposition for several reasons. The formal mechanics of popular tools such as MLE (and related techniques to be considered shortly) are not terribly interesting as such, as they are ultimately just cranks for producing numbers. What *is* of importance is the ability to interpret those numbers in light of some model representing a stochastic process of interest. Let us stress again: our primary concern is the construction of good valuations for structured products. This means relating the structure in question to some set of available hedging instruments (or more accurately relating the respective payoffs), and accounting for residual risk. Obviously, any estimation procedure produces estimates

of both a particular relation and of an associated residual. Uncertainty around the estimated relation is the flip side of uncertainty regarding the residual, which drives the valuation (since the penultimate deal price is the cost of hedging plus a charge for remaining risk). As we have endeavored to show here, the justification for many standard econometric techniques (we have focused on MLE, but the point applies more generally) only holds asymptotically, *i.e.*, for large sample sizes.

The unfortunate fact is that the formal theory offers almost no guidance as to whether a particular set of data is large enough for these asymptotic results to be appealed to. Since stationarity/ergodicity is at the heart of these asymptotic results, we see once again the role that time scales play in the empirical analysis of the problem. The sample size is not large or small as such, but only in relation to the operative time scales of the underlying DGP.

With the relevant concepts and rationales laid out, we can turn more briefly to some important, related techniques before turning to a concrete example.

### 6.5.2 MLE and QMLE

We discussed MLE in a fair amount of detail in Section 6.5.1. We now turn to a related technique, for which many of the asymptotic arguments of that section can be readily adapted.

#### *6.5.2.1 Quasi-Maximum Likelihood Estimation (QMLE) and informational requirements*

The feasibility of MLE requires that the densities underlying the objective function in (6.141) be both known and sufficiently tractable to facilitate the ensuing optimization (which typically requires some sort of numerical procedure). To overcome these potentially serious difficulties, an alternative, related approach known as QMLE is often resorted to. Under QMLE, the log-likelihood function that is maximized is given by

$$\mathcal{L}(\theta) = \frac{1}{T}\sum_t \log f'(z_t;\theta) \tag{6.151}$$

where $f'$ is *not* the true density of the underlying DGP. Rather, it is an auxiliary or proxy density that (hopefully) bears some similarity to the true density (call it $f$) while being more tractable. Thus, there is not necessarily a "true" value of the model parameter as such, although the auxiliary density could be (and often is) taken as an approximation to the true density, in which case the QMLE parameters do correspond to the parameters of the DGP. To understand how this relationship manifests itself, note that

$$\mathcal{L}(\theta) = -\frac{1}{T}\sum_t \log\frac{f(z_t)}{f'(z_t;\theta)} + \frac{1}{T}\sum_t \log f(z_t) \tag{6.152}$$

Thus, maximizing the (quasi) log-likelihood function is equivalent to minimizing the so-called Kullback-Leibler (KL) divergence (defined as $E\log(f/f')$), well known from information theory. (The KL divergence is obviously related to the relative entropy measures considered in Section 5.2.9.) In other words, for a given class of auxiliary models, QMLE selects the one that (effectively) minimizes relative entropy/informational discrepancy under the DGP.

*6.5.2.2 Convergence and consistency*

To understand the sense in which the QMLE estimator is consistent (or inconsistent), we turn to the question of the asymptotic behavior. To leading order, the first order condition for maximization of the (quasi) log-likelihood function (6.151) becomes

$$0 = \frac{1}{T}\sum_t \frac{\partial}{\partial\theta}\log f'(z_t;\theta^*) + \frac{1}{T}\sum_t \frac{\partial^2}{\partial\theta\partial\theta^T}\log f'(z_t;\theta^*)(\hat{\theta}-\theta^*) + \cdots \tag{6.153}$$

where $\theta^*$ is the "true" parameter value, $\hat{\theta}$ is the QMLE estimator. What is meant by "true" here? Recalling the caveats from Section 6.5.1 and assuming sufficient stationarity/ergodicity conditions, we see that the first term on the RHS of (6.153) converges to

$$E_f \frac{\partial}{\partial\theta}\log f'(z;\theta^*) = \frac{\partial}{\partial\theta}\left(E_f \log\frac{f'}{f} - E_f \log f\right)\Big|_{\theta^*} \tag{6.154}$$

(Here $E_f$ denotes expectation wrt. the true density $f$.) Thus, *if $\theta^*$ is taken to be the minimizer of the KL divergence*, (6.154) is zero (first-order condition) and we see from (6.153) that

$$\hat{\theta}-\theta^* \sim N(0, \frac{1}{T}H^{-1}JH^{-1}) \tag{6.155}$$

and

$$\begin{aligned} H &= -E_f\left(\frac{\partial^2\log f'}{\partial\theta\partial\theta^T}\Big|_{\theta^*}\right)^{-1} \\ J &= E_f\left(\frac{\partial\log f'}{\partial\theta}\Big|_{\theta^*}\right)\left(\frac{\partial\log f'}{\partial\theta}\Big|_{\theta^*}\right)^T \end{aligned} \tag{6.156}$$

giving rise to the familiar Huber "sandwich" for the parameter covariance. Note that when $f' = f$ (*i.e.*, regular MLE), the parameter covariance reduces (along with $T^{-1}$ scaling) to the usual inverse of the Fisher information matrix (the Hessian of the log density), by the well-known matrix information equality.[58]

Having established the sense in which the QMLE estimator can properly be considered consistent, let us now consider the case where the auxiliary density $f'$ is

actually an approximation to the true density $f$ of the underlying DGP. As we have just seen, the QMLE estimator is asymptotically the minimizer of the KL divergence. It also stands in some relation to the true parameter of the DGP:

$$\theta^*_{QMLE} = K(\theta^*_{DGP}) \tag{6.157}$$

It will not in general be true that the map $K$ in (6.157) is the identity transformation. Consequently, the QMLE estimator is inconsistent (although it may well be consistent under certain conditions; see Newey and Steigerwald [1997]).[59] However, we would anticipate that if the auxiliary density is a good approximation of the true density, then the KL divergence will be small and the map $K$ is near identity. In such a case, QMLE may be inconsistent, but only moderately so. But there is a deeper point, as we shall see: in small samples, QMLE, though *asymptotically* inconsistent, may actually outperform MLE based on the true density!

Having covered likelihood-based methods, we can now (briefly) turn to a more encompassing, expectation-based technique (and its offshoots).

### 6.5.3 GMM, EMM, and their offshoots

#### *6.5.3.1 Moment conditions*

MLE is actually a special case of a more general technique, which we generically refer to as moment matching. As the name suggests, for a given model we may have some known expression for expectations of the form $Eg(z;\theta) = 0$ for some vector $g$, which is solved for the parameter $\theta$ by replacing population entities by their sample analogues; compare with (6.142). A very simple (and common) example is estimating a Gaussian variable by matching sample means and standard deviations to their population counterparts. An extremely influential extension of this (classical) idea, due to Hansen (1982) is known as the Generalized Method of Moments (GMM). Suppose we start with some vector-valued set of (population) expectations and its corresponding set of sample averages:[60]

$$Eg(z;\theta) \Leftrightarrow \bar{g}(Z;\theta) \equiv \frac{1}{T}\sum_t g(z_t;\theta) \tag{6.158}$$

In general, the number of moment conditions may be different from the number of parameters being estimated (*e.g.*, $Eg(z;\theta) = 0 \in \Re^d$ and $\theta \in \Re^p$ with $p \neq d$), making classical moment matching ill-determined. The idea behind GMM is to instead craft the sample analogue of the moment conditions as a minimization problem, for a suitable weighting matrix across conditions. In other words, for some symmetric matrix $W$ we take

$$\hat{\theta} = \arg\min_\theta \bar{g}(Z;\theta)^T \cdot W \cdot \bar{g}(Z;\theta) \tag{6.159}$$

We do not here go into detail about the optimal choice of weighting matrix (see chapter 14 of Hamilton [1994]), although we will note that in practice, GMM is often implemented iteratively, first using $W = I$ (the identity matrix) and then updating with the sample covariance matrix (for the conditions (6.158); recall the estimator of the noise covariance in a VAR, *e.g.*, (6.29)).

As its name suggests, GMM is quite general and includes as special cases a number of well-known techniques. These include OLS where the moment conditions amount to orthogonality between regressors and residual (see (2.5)): $E(y - \alpha x - \beta) = 0$ and $Ex(y - \alpha x - \beta)$. MLE also clearly falls in this category, as the relevant moment condition can be seen in (6.142): $E\frac{\partial}{\partial\theta}\log \Pr_\theta(z) = 0$. (In these cases, there are as many moment conditions as unknown parameters, so there is no need for a weighting matrix; *i.e.*, GMM is just identifiable.)

*6.5.3.2 Asymptotics*

The first order optimality condition associated with (6.159) is

$$\frac{\bar{g}(Z;\theta)^T}{\partial\theta} \cdot W \cdot \bar{g}(Z;\theta) = 0 \tag{6.160}$$

Proceeding similarly to the case of QMLE (see (6.153)), the asymptotic consistency and normality of the GMM estimator can be established as follows. Denoting the true parameter value by $\theta^*$, to leading order we have

$$\begin{aligned} 0 &= \frac{1}{T}\sum_t g_\theta(z_t;\theta^* + \hat{\theta} - \theta^*)^T \cdot W \cdot \frac{1}{T}\sum_t g(z_t;\theta^* + \hat{\theta} - \theta^*) \\ &\approx \frac{1}{T}\sum_t g_\theta(z_t;\theta^*)^T \cdot W \cdot \left(\frac{1}{T}\sum_t g(z_t;\theta^*) + \frac{1}{T}\sum_t g_\theta(z_t;\theta^*)(\hat{\theta} - \theta^*)\right) \end{aligned} \tag{6.161}$$

Thus, by smuggling in the appropriate assumptions regarding stationarity/ergodicity, we see (via CLT) that

$$\hat{\theta} - \theta^* \sim N(0, \frac{1}{T}(D^T W D)^{-1}\ (D^T W C W D)(D^T W D)^{-1}) \tag{6.162}$$

as $T \to \infty$, where $D \equiv Eg_\theta(z;\theta)$ and $C \equiv Eg(z;\theta)g(z;\theta)^T$. (The optimal weighting matrix is given by $W = C^{-1}$, in which case the estimator covariance matrix reduces to $\frac{1}{T}(D^T C^{-1} D)^{-1}$; see Hamilton [1994].) It should be emphasized that moment conditions are often based on some process dynamics, and take a *conditional* form, *e.g.*,

$$E_t g(z_{t+1}) = \tilde{g}(z_t;\theta) \tag{6.163}$$

for some parameter-dependent process $z$. However, by invoking arguments similar to those used in the discussion of MLE (*e.g.*, (6.149)), these *asymptotic* results still

hold. (Hansen [1982, 2007] is quite careful to emphasize these important qualifications.) However, this is precisely the point that must be stressed: the concerns raised in Section 6.5.1. about the appropriateness of relying on asymptotic results hold just as true here.

For completeness we briefly mention some related techniques. A drawback of GMM is that the choice of moment conditions is virtually unlimited and, in large part, arbitrary. This is offset by the fact that GMM is quite flexible and (usually) very tractable. It thus contrasts with MLE with its comparatively greater efficiency (it can attain the information-theoretical Cramér-Rao lower bound on estimator [co]variance; essentially the inverse of the Fisher information matrix) but frequently intractable nature. It is thus desirable to seek a middle ground, which is the objective of the Efficient Method of Moments (EMM). EMM combines aspects of QMLE and so-called *indirect inference*. We have already discussed QMLE, so we will give a quick overview of indirect inference.

#### 6.5.3.3 *Indirect inference*

The idea behind indirect inference is to estimate a model for which, say, MLE is intractable by introducing an auxiliary model (as in QMLE) which *is* tractable. The auxiliary model typically possesses some set of parameters $\upsilon$ distinct from (in fact, generally unrelated to) the actual model parameters $\theta$. Assume we have an estimator $\hat{\upsilon}(Y)$ of the auxiliary parameters for the sample in question, $Y = \{y_t\}$. Now, for a particular value $\theta$ of the primary model we construct a set of simulated time series $Y_s^\theta$ according the DGP/model that is being estimated. For each such path, we find a set of estimators from the auxiliary model $\hat{\upsilon}(Y_s^\theta)$. The central idea behind indirect inference is to construct an estimator $\hat{\theta}$ of the model in question by making the set of (simulation-derived) auxiliary parameters as close as possible to the original estimator (based on the actual sample), in the appropriate sense. For example, we could match the original estimator to the average estimator (across the $S$ paths):

$$\hat{\theta}(Y) = \arg\min_\theta \left\| \hat{\upsilon}(Y) - \frac{1}{S}\sum_s \hat{\upsilon}(Y_s^\theta) \right\|^2 \tag{6.164}$$

Alternatively, the estimator across an average auxiliary log-likelihood function could be used:

$$\hat{\theta}(Y) = \arg\min_\theta \left\| \hat{\upsilon}(Y) - \arg\max_\upsilon \frac{1}{S}\sum_s \mathfrak{L}_{\text{aux}}(Y_s^\vartheta; \upsilon) \right\|^2 \tag{6.165}$$

Of course, a weighting matrix could be introduced in the minimization criteria, as in GMM. Indirect inference is particularly useful when the underlying DGP has unobservable components such as stochastic volatility. (Thus avoiding resort

to filtering.) For more on indirect inference, see Smith (2008) and the references therein.

*6.5.3.4 EMM*

The idea behind EMM can now be better understood. First, an auxiliary model, parameterized by $\upsilon$, is introduced, and a corresponding estimate $\hat{\upsilon}$ is obtained via QMLE. This initial step establishes the following moment conditions, deduced from the sample analogue of the first order conditions:

$$\frac{1}{T}\sum_t \frac{\partial}{\partial \upsilon}\log f_{\text{aux}}(y_t;\hat{\upsilon}) = 0 \Leftrightarrow E\frac{\partial}{\partial \upsilon}\log f_{\text{aux}}(y;\upsilon) = 0 \tag{6.166}$$

where $y_t$ is the sample in question and $f_{\text{aux}}$ is the auxiliary density. (In (6.166), the (population) expectation is with respect to the underlying DGP.) Next, GMM is applied, using the moment conditions in (6.166) as the minimization criteria. This typically entails, as in the case of indirect inference, the use of simulation. That is, we construct pathwise averages of the form:[61]

$$m(\theta;\hat{\upsilon}) = \frac{1}{S}\sum_s \frac{\partial}{\partial \upsilon}\log f_{\text{aux}}(y_s^{\theta};\hat{\upsilon}) \tag{6.167}$$

where $\theta$ refers to the parameter of interest of the underlying DGP. The estimator of this parameter is then taken (in usual GMM fashion) to be

$$\hat{\theta} = \arg\min_{\theta} m(\theta;\hat{\upsilon})^T \cdot W \cdot m(\theta;\hat{\upsilon}) \tag{6.168}$$

for suitable (symmetric) weighting matrix $W$. Due to its simulation-based nature, EMM is well suited to problems such as stochastic volatility, where the volatility is not observable.[62] For more on EMM, see Andersen *et al.* (1999) or Gallant and Tauchen (2010). Needless to say, the small-sample robustness of methods such as EMM is every bit dependent on underlying stationarity/ergodicity as the previous methods we have discussed.

We are now finally in a position to discuss Zhou's (2001) findings.

### 6.5.4 A study of estimators in small samples

*6.5.4.1 The setup*

The CIR model in (6.137) proves a nice econometric testing ground precisely because it adds a bit of structure to standard GBM (mean-reversion and non-Gaussian deviations) while retaining an analytical form for the transition density:

$$\Pr(r_{t+\Delta t}|r_t) = ce^{-u-v}\left(\frac{v}{u}\right)^{q/2} I_q(2\sqrt{uv}) \tag{6.169}$$

where $q = \frac{2\kappa\theta}{\sigma^2} - 1$, $c = \frac{2\kappa}{\sigma^2(1-e^{-\kappa\Delta t})}$, $u = ce^{-\kappa\Delta t} r_t$, $v = cr_{t+\Delta t}$, and $I_q$ denotes the modified Bessel function of the first kind of order $q$. Note that it is straightforward to obtain the (conditional) mean and variance:[63]

$$\begin{aligned}\tilde{\mu}_{t,t\Delta t} &\equiv E_t r_{t+\Delta t} = r_t e^{-\kappa\Delta t} + \theta(1 - e^{-\kappa\Delta t}) \\ \tilde{\sigma}^2_{t,t+\Delta t} &\equiv E_t r^2_{rt+\Delta t} - (E_t r_{t+\Delta t})^2 = \frac{\sigma^2}{\kappa}(1 - e^{-\kappa\Delta t})(r_t e^{-\kappa\Delta t} + \frac{\theta}{2}(1 - e^{-\kappa\Delta t}))\end{aligned} \tag{6.170}$$

giving rise to a Gaussian approximation of the true density via

$$\Pr(r_{t+\Delta t}|r_t) \approx \frac{1}{\sqrt{2\pi\,\tilde{\sigma}^2_{t,t+\Delta t}}} \exp\left(-\frac{(r_{t+\Delta t} - \tilde{\mu}_{t,t+\Delta t})^2}{2\tilde{\sigma}_{t,t+\Delta t}}\right) \tag{6.171}$$

The (conditional) density in (6.169) can plainly be used in standard MLE, *e.g.*, (6.140) and (6.141). Obviously the main impediment is calculation of the Bessel function, but there exist efficient algorithms for carrying this out (see Press *et al.* [2007]). Additionally, the very convenient form of the (conditional) moment expressions in (6.170) make possible the application of techniques such as QMLE[64] or GMM. What could be better? As we shall see, an embarrassment of riches is not all it is cracked up to be.

#### *6.5.4.2 Simulation tests*

Another nice aspect of the CIR model is that it is quite amenable to simulation. Zhou (2001) exploits the mixture of Poisson and Gamma characterization to generate random samples; more recent approaches (in the context of Heston) include Broadie and Kaya (2006) and Andersen (2006). Thus it is easy to create many realizations of a given sample size. Zhou considers several scenarios, each with two different sample sizes: 500 and 1,500. The scenarios essentially run the gamut from low mean reversion-low conditional variance to high mean reversion-high conditional variance. Both of these extreme cases are computationally and econometrically challenging. With a step size of seven days (one week), these cases have parameter values given below:

| Low Mean Reversion, Low conditional Variance | High Mean Reversion, High conditional variance |
|---|---|
| $\kappa = 0.15$ | $\kappa = 15$ |
| $\theta = 8.7$ | $\theta = 87$ |
| $\sigma = 0.2$ | $\sigma = 7.9$ |
| $E_t r_{t+\Delta t=0.997 r_t + 0.025}$ | $E_t r_{t+\Delta t} = 0.75 r_t + 2.17$ |
| $Var_t r_{t+\Delta t} = 7.5 \cdot 10^{-4} r_t + 9.4 \cdot 10^{-6}$ | $\mathrm{Var}_t r_{t+\Delta t} = 0.79 r_t + 1.14$ |

(6.172)

Zhou finds that the distributions of the estimators from these popular techniques (he also considers EMM) are decidedly non-normal and that their biases decrease very slowly with increased sample size. It is a striking result precisely because the true model is known (and its structure, though non-Gaussian, is not terribly complex) and the sample sizes quite large in relation to actual market data. It should be further stressed that Zhou's study took place in the context of interest-rate modeling; the sample sizes he used are especially large by energy-market standards. His conclusions, broadly stated, are that while MLE has the highest efficiency (in terms of variance), QMLE and EMM perform best in terms of inference. QMLE in particular offers great benefits in terms of computational feasibility and ease of implementation. His results strongly caution against trying to seek out full-information estimators as a first resort.

However, while we do not doubt the prescriptions that follow from this study, we do believe there are numerical issues that amplify the starkness of the results (*e.g.*, figures 1 and 2 in Zhou [2001]). We will therefore illustrate these points with our own computational analysis.

#### *6.5.4.3 In revision of Zhou*

The primary challenge in implementing full-information MLE for the CIR model is the calculation of the modified Bessel function in the transition density in (6.169). Zhou (2001) employs the aforementioned Poisson-Gamma mixing series solution. However, this series has regions of non-convergence, especially in the low mean reversion case (Zhou's scenario 1). In this case, the argument $q$ of the Bessel function is very large and the series solution breaks down. (There is a gamma function calculation involved that becomes highly problematic in this regime; consult any text on special functions.) Zhou thus resorts to a standard asymptotic expansion in this case for evaluation of the Bessel function. We believe this distorts his results, although his conclusions are largely left intact.

We employ an approach based on a continued fraction expansion from Press *et al.* (2007). (For the optimization we use standard iterated Newton/gradient search, constraining the mean reversion rate $\kappa$ to be positive.) We generate 500 realizations of time series of length 500 each (keeping the weekly time step used by Zhou [2001] and the mixing approach for drawing a random deviate at each time step). We "normalize" the estimates by subtracting the corresponding true parameter value and dividing by the standard error (which we take as the standard deviation across paths); *i.e.*, we look at the $t$-statistics essentially. We in fact analyze an alternative formulation of QMLE, based not on moment matching but instead on the discretized (Euler) approximation to the underlying SDE, so that we still take a Gaussian transition density but with conditional mean and variance given by $v_t + \kappa(\theta - v_t)\Delta t$ and $\sigma^2 v \Delta t$; compare with the small step limit of (6.170). (This approach is commonly termed discretized maximum likelihood estimation [DMLE].) The results for the low mean reversion case are presented in Figures 6.5 through 6.7.

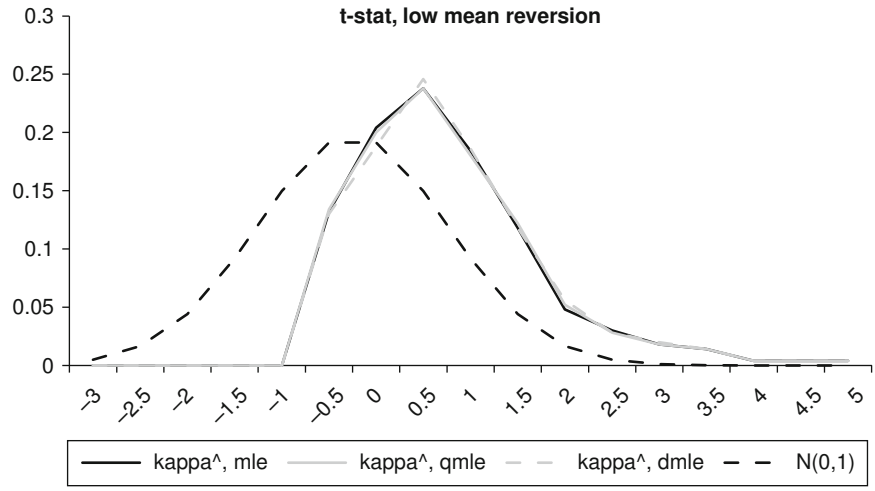



**Figure 6.5** Distribution of *t*-statistic, mean reversion rate

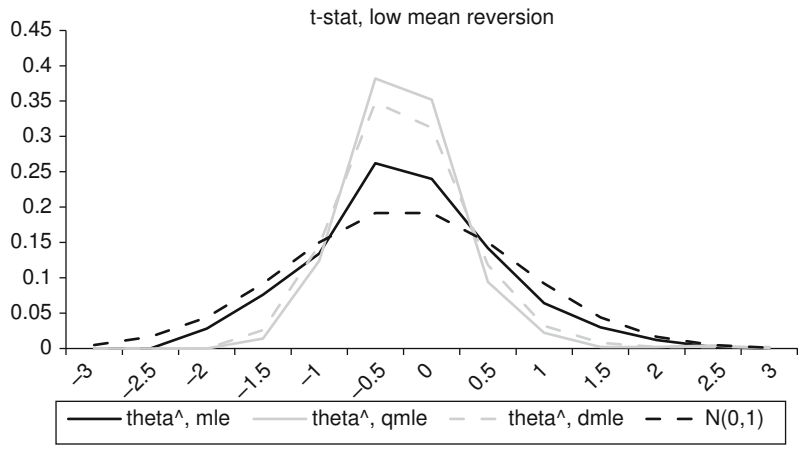



**Figure 6.6** Distribution of *t*-statistic, mean reversion level

The following points stand out:

1. The three estimators are generally very similarly distributed, and largely *not* asymptotically normal.
2. The estimator of the mean reversion rate is highly biased and skewed right.
3. The estimator of the mean reversion level is not very biased, but very fat-tailed (somewhat symmetrically).
4. The estimator of the volatility is asymptotically normal about the true value.

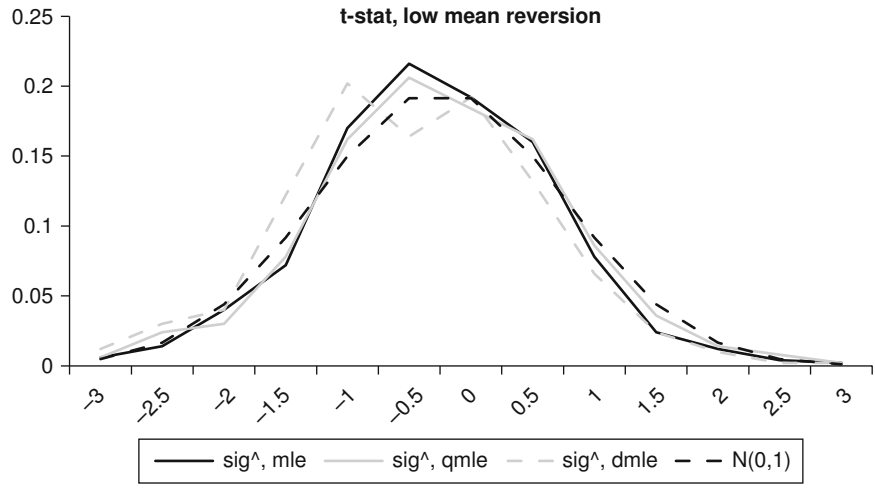



**Figure 6.7** Distribution of *t*-statistic, volatility

Point number 2 should come as no surprise. It is well known that in finite samples it is very difficult to identify mean reversion in nearly nonstationary time series. We have already seen this in a very simple example in (2.23). Point number 4 also is not particularly shocking. It highlights the general comparatively superior robustness of volatility estimation. Again, we indicated why this should be expected to be the case in the example in (2.24). The results here clearly conform to these familiar (stylized) facts.

Point number 1 represents the biggest departure between our results and those of Zhou (2001). *Contra* Zhou, we see little statistical discrepancy between the full-information MLE and various approximations based on retaining second-order moment information only, with the exception of mean reversion level (although even here we do not see the biases in MLE as reported by Zhou). (However, it is clear that these estimators are *not* asymptotically normal, except in the case of volatility estimation.) Nonetheless, these results point to an important fact: full information is either *not* necessary for estimation of certain parameters (volatility) or is *not* helpful for the estimation of others (mean reversion rate). It goes without saying that computationally QMLE is much faster than MLE (since it avoids the burden of the Bessel function calculations). Put differently, even when we know the true DGP (and so resort to approximation is not necessary), it may be better to simply ignore this information and apply an approximation anyway. We thus affirm Zhou's basic conclusion as to the superiority of QMLE over MLE.

Note that it remains the case that some approximations are better than others. While in the low mean reversion/low conditional variance case, DMLE is basically

indistinguishable from QMLE (and MLE), this is not true in the other scenario under consideration, namely high mean reversion/high conditional variance (Zhou's [2001] scenario 8). We present our results for this case in Figures 6.8 through 6.10.

First, note that, unlike the previous scenario, the MLE and QMLE *do* conform asymptotically to normal behavior.[65] (This again represents a point of departure from Zhou's [2001] results, which actually suggests problems with his Bessel function calculation that go beyond the breakdown of the series representation.) Second, except in the case of the mean reversion level, DMLE is clearly strongly

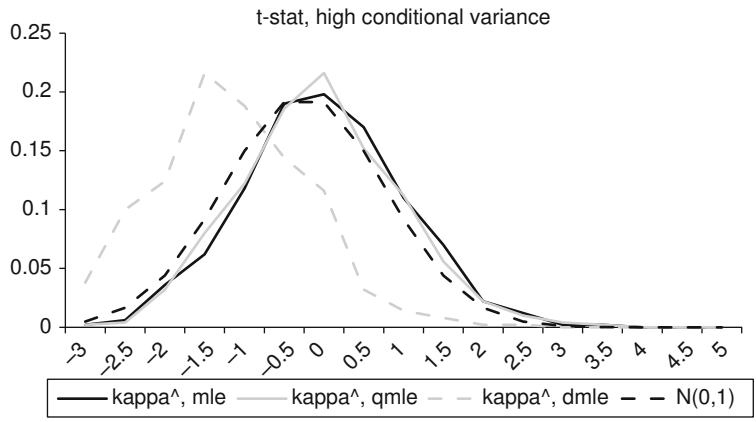



**Figure 6.8** Distribution of *t*-statistic, mean reversion rate

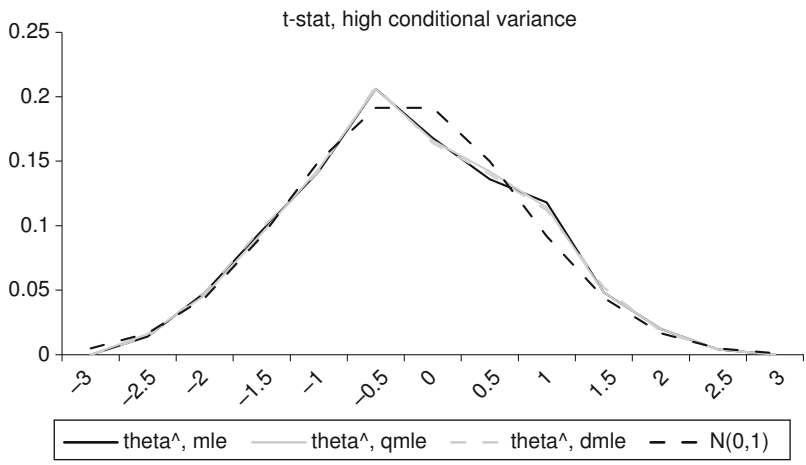



**Figure 6.9** Distribution of *t*-statistic, mean reversion level

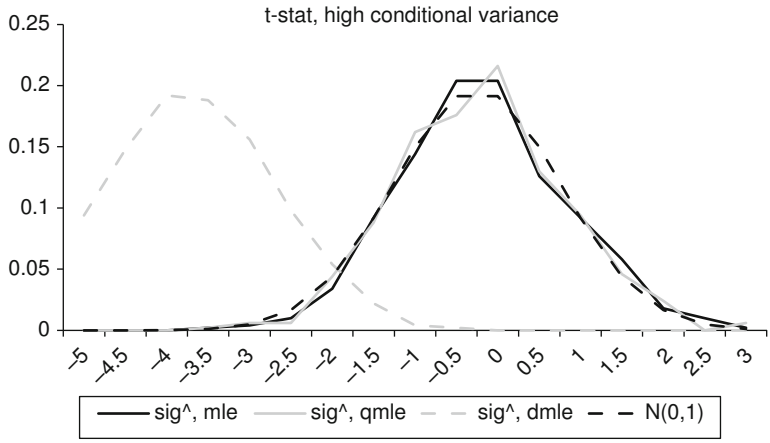



**Figure 6.10** Distribution of *t*-statistic, volatility

biased, even for the volatility. This is not terribly surprising: in the case of high conditional variance, the Euler discretization should not be expected to be a very good approximation to the underlying DGP, and these results bear that out. So, while this scenario again affirms the basic point of preferring QMLE over MLE, it also highlights the critical point that the issue in estimation invariably comes down to the use of *conditional* information. Actually, more specifically it is a question of time scales (once again). The divergence (of the approximation) comes down to the size of $\kappa \Delta t$, which precisely reflects the relevant time scale of the process.

We can now provide an assessment.

#### *6.5.4.4 The lessons learned*

First, we see the basic irrelevance of asymptotic results for estimation diagnostics. For a time series of length 500 (which corresponds to 500 weekly observations, which greatly exceeds the data availability in many [if not most] energy markets) the distribution of several popular estimators are clearly non-Gaussian, which is the predicted form from standard asymptotic analysis. (See, again, Sections 6.5.2 and 6.5.3.) Thus, in assessing the stability of an estimator for an actual sample (recall the discussion in Section 6.4) one must exercise great caution in employing asymptotic results. (It would probably not be an overstatement to say that one should never rely on asymptotics.) Now, there are some qualifications to this pessimism. Volatility estimators generally *do* conform to their asymptotic distribution, *if* conditional information is correctly specified in light of the prevailing time scales. (Recall the poor performance of DMLE in the case of high conditional variance.) But as a general observation, asymptotics are fairly useless in practice, and what appears to be a large sample size may in reality be small, again in comparison to the underlying time scales. In particular, direct estimation of mean reversion rates is extremely

problematic if the process in question is not sufficiently stationary (again, in the operational sense of a sample and not in the abstract sense of a population), and effort is better spent on indirect estimation as revealed by variance scaling laws (as in Section 2.2).

Second, using full information in an estimator often provides little benefit and can in fact be needlessly costly. We saw in two econometrically challenging cases (weak mean reversion, strong conditional variance) that partial information QMLE was distributed almost identically to full information MLE. When we speak of partial information here, we mean that in QMLE we (correctly) specified the first two (conditional) moments, as opposed to using essentially all the moments as entailed in MLE. Now, as we showed in Section 6.5.2, QMLE identifies that proxy density that minimizes the relative entropy measure (KL distance) between the proxy and true density. Our results here indicate that the correctly specified conditional Gaussian is indeed a very good approximation to the non-Gaussian CIR (square root) process (in the sense of relative entropy). (In fact, for the nearly non-stationary case, even the incorrectly specified DMLE approximation is a good one.) This is an important point, as it is not *a priori* obvious that an estimator based on the Gaussian approximation should produce results similar to that of an estimator based on the true density. (We in fact observe that the estimators behave similarly *pathwise*, not just distributionally.) MLE thus proves needlessly costly from a computational standpoint[66] and in fact provides little real benefit, as can be seen in estimation of the mean reversion parameter. Mean reversion rates are simply extremely hard to estimate in finite samples, with or without full information.[67]

It is worth stressing that in this study, we *know* the actual DGP. Obviously in actual applications we will not have this knowledge. The analysis here suggests that in fact we are better off *not* worrying about what the actual DGP is, if we can instead get better *conditional* information about the first two moments only. As always, the main econometric challenge here is overcoming the informational constraints imposed by the sample in question. Generally speaking, the smaller the sample (and sample size is in truth a function of the operative time scales), the less structure one can introduce and expect to achieve robust results. Indeed, while we have presented QMLE as superior alternative to MLE (which it clearly is in many ways), it, too, suffers from many of the same problems. (Which is why, *e.g.*, we generally advise against trying to directly estimate mean reversion.) We are thus led to present the following important corollary: *knowing the qualitative features of a process is not the same thing econometrically as knowing the actual DGP.*[68]

The upshot here is clear: while it is important to understand the econometric techniques presented here, one needs to be acutely aware that they have rather severe limitations and, if they are to be employed, one must adopt an extremely critical eye when assessing their output.

### 6.5.5 Spectral methods

We know the reader would be disappointed if we did not close out our little (econometric) adventure with an application of characteristic functions to the problem of estimation. For those models where the characteristic function is known (or at least is amenable to numerical determination, *e.g.*, the class of affine jump diffusions), the characteristic function can be used as a matching criterion. (In some sense, the characteristic function can be thought of as providing *all* of the moments to be used in a moment-matching procedure.) We start with the basic relationship

$$f(\phi; z_t) = E_t e^{i\phi^T z_{t+1}} \tag{6.173}$$

where $z$ is some $d$–dimensional process, $\phi \in \Re^d$, and $f$ is the conditional characteristic function. (Recall the examples from Section 5.2.5, such as standard Black-Scholes, Merton's jump diffusion, and Heston's stochastic volatility model.) There are several econometric techniques based on the expression in (6.173), which we proceed to discuss now.

#### *6.5.5.1 Direct ML and GMM estimation*

Singleton (2001) proposes inverting the relationship (6.173) to obtain (via the Fourier inversion theorem) the (conditional) density of the underlying process, which can then be used in standard MLE. The obvious problem here is the computational burden, which grows exponentially with the dimension of the state space, and of course the well-known challenges presented by oscillatory integrands (to say nothing of truncation error of the infinite integration region). Possible remedies here include partial information estimations based on suitable linear combinations of the state variables. For example, (6.173) can obviously be used to calculate the characteristic function of entities such as $t_l^T z$, where $t_l$ has elements $\delta_{kl}$. Characteristic function inversion then amounts to one-dimensional integrations for any index $l$, producing (conditional) densities for any single-state variable given *all* previous-time state variables. Another possibility is to extract moment conditions from the derivatives of the characteristic function (evaluated at $\phi = 0$) for use in standard GMM.

Apart from the computational issues, further difficulties are presented when unobservable variables (such as stochastic volatility) are introduced. Singleton (2001) considers simulation-based methods for essentially filtering out such latent variables; recall the discussion in Section 6.5.3. (In addition, if possible, option market data can be exploited in the estimation, since as we know that the affine framework permits a reasonably efficient computational framework for option pricing.) To consider alternative approaches in this case, we must present some prefatory results.

*6.5.5.2 Unconditional entities and latent variables*

For exposition we will confine attention to the affine case, and for the moment ignore jumps. We will also assume time homogeneity, so that only the sample time step (further assumed uniform) matters. Broadly, we will characterize the state variables as observable (log) prices and unobservable stochastic volatility (the paradigmatic example of course being Heston). We will indicate these (multidimensional) categories by $z$ and $v$, respectively. Then, using the results from Section 5.2, we have that

$$\begin{aligned} f(\phi,0;z_{t-1},v_{t-1}) &= E_{t-1}e^{i\phi^T z_t + i0^T v_t} \\ &= \exp(\alpha^T(\Delta t;\phi,0)z_{t-1} + \beta^T(\Delta t;\phi,0)v_{t-1} + \gamma(\Delta t;\phi,0)) \end{aligned} \tag{6.174}$$

where the coefficients $(\alpha^T,\beta^T,\gamma)^T$ solve a system of ODEs (as in, say, (5.84)) with initial conditions $(\alpha^T,\beta^T,\gamma)^T(0) = (i\phi^T,0^T,0)^T$, and $\Delta t$ is the time step between points in the sample.

Now, we are interested in the (joint) characteristic function of the observables, so introducing another Fourier variable, employing iterated expectations in (6.174), and again exploiting affinity, we see that

$$\begin{aligned} E_{t-2}e^{i\phi_0^T z_{t-1} + i\phi_1^T z_t} &= E_{t-2}e^{i\phi_0^T z_{t-1}}\exp(\alpha^T(\Delta t;\phi_1,0)z_{t-1} + \beta^{T(\Delta t;\phi_1,0)v_{t-1}+\gamma(\Delta t;\phi_1,0)}) \\ &= \exp\begin{pmatrix} \alpha^T(\Delta t;\phi_0 - i\alpha(\Delta t;\phi_1,0),-i\beta(\Delta t;\phi_1,0))z_{t-2}+ \\ \beta^T(\Delta t;\phi_0 - i\alpha(\Delta t;\phi_1,0),-i\beta(\Delta t;\phi_1,0))v_{t-2}+ \\ \gamma(\Delta t;\phi_1,0) + \gamma(\Delta t;\phi_0 - i\alpha(\Delta t;\phi_1,0),-i\beta(\Delta t;\phi_1,0)) \end{pmatrix} \\ &\equiv \exp(\alpha^T(\Delta t;\phi_0^*,\psi_0^*)z_{t-2} + \beta^T(\Delta t;\phi_0^*,\psi_0^*)v_{t-2} + \gamma(\Delta t;\phi_1,0) + \gamma(\Delta t;\phi_0^*,\psi_0^*)) \end{aligned} \tag{6.175}$$

Repeating this process, we get the following recursive result (see Jiang and Knight [2002] and Rockinger and Semenova [2005]):

$$\begin{aligned} &E_{t-p-1}e^{i\phi_0^T z_{t-p} + \cdots + i\phi_p^T z_t} \\ &\quad = \exp\left(\sum_{k=0}^{p}\gamma(\Delta t;\phi_k^*,\psi_k^*) + \alpha^T(\Delta t;\phi_0^*,\psi_0^*)z_{t-p-1} + \beta^T(\Delta t;\phi_0^*,\psi_0^*)v_{t-p-1}\right) \end{aligned} \tag{6.176}$$

with

$$\begin{aligned} &\phi_p^* = \phi_p, \psi_p^* = 0 \\ &\phi_k^* = \phi_k - i\alpha(\Delta t;\phi_{k+1}^*,\psi_{k+1}^*), \psi_k^* = -i\beta(\Delta t;\phi_{k+1}^*,\psi_{k+1}^*) \end{aligned} \tag{6.177}$$

Assuming sufficient stationarity in all the variables (such that unconditional expectations can be meaningfully defined), we get from (6.176) that

$$Ee^{i\phi_0^T z_{t-p}+\cdots+i\phi_p^T z_t} = \exp\left(\sum_{k=0}^{p}\gamma\left(\Delta t;\phi_k^*,\psi_k^*\right)\right)$$
$$\cdot E\exp(\alpha^T(\Delta t;\phi_0^*,\psi_0^*)z_{t-p-1}+\beta^T(\Delta t;\phi_0^*,\psi_0^*)v_{t-p-1}) \tag{6.178}$$

Note that in the special case where the observables are not mean reverting (in other words, in the affine dynamics there are no drift terms proportional to $z$) we have that $\alpha = i\phi$. Then, in terms of (purely stationary) log returns $r_t \equiv z_t - z_{t-1}$, (6.177) and (6.178) can be used to show that

$$Ee^{i\phi_0^T r_{t-p}+\cdots+i\phi_p^T r_t} = Ee^{-i\phi_0^T z_{t-p-1}+(\phi_0-\phi_1)^T z_{t-p}\cdots+(\phi_{p-1}-\phi_p)^T z_{t-1}+i\phi_p^T z_t}$$
$$= \exp\left(\sum_{k=0}^{p}\gamma\left(\Delta t;\phi_k,\psi_k^*\right)\right)\cdot E\exp(\beta^T(\Delta t;\phi_0,\psi_0^*)v_{t-p-1}) \tag{6.179}$$

with $\psi_p^* = 0$, $\psi_k^* = -i\beta(\Delta t,\phi_{k+1},\psi_{k+1}^*)$. In the customary example of a Heston-driven stochastic variance, the unconditional expectation can be obtained as a long-term limit of the conditional expectation, yielding (see (5.123) to review the parametric notation)

$$E\exp(\beta^T(\Delta t;\phi_0,\psi_0^*)v_{t-p-1}) = \left(1-\frac{\beta(\Delta t;\phi_0,\psi_0^*)\sigma^2}{2\kappa}\right)^{2\kappa\theta/\sigma^2} \tag{6.180}$$

Recall from Section 5.2 that in many cases (including those with jumps), the coefficients $\alpha$, $\beta$, and $\gamma$ are known analytically. In general, however, numerical solution of the underlying system of ODEs will be necessary.

It is worth asking at this stage: what is our objective here? Latent variables such as stochastic volatility (which are by definition not observable) manifest themselves in certain observable characteristics, prominent among these being heteroskedasticity and autocorrelation (of returns). We cannot condition on something that we cannot observe, so to capture the joint structure of (observable) prices or returns, we must work in terms of unconditional entities such as (6.178), with the contribution from conditional latent dynamics effectively integrated out. This approach can thus be thought of as an alternative to explicit or formal filtering. We are now led to consider estimation based on matching to the so-called empirical characteristic function, which can be extracted from the (observable) data for any block length $p+1$.

*6.5.5.3 Empirical characteristic function estimation*

Consider a sample of size $T$, denoted by

$$X = \begin{pmatrix} z_1 & \cdots & z_T \\ v_1 & \cdots & v_T \end{pmatrix} \tag{6.181}$$

Now, divide this sample into $T - p$ overlapping blocks of size $p + 1$ and define the vector $y$ via $y_j = (z_j^T, \ldots, z_{j+p}^T)^T$. The empirical characteristic function (ECF) is then given by

$$g(\phi; y) = \frac{1}{T-p} \sum_{j=1}^{T-p} e^{i\phi^T y_j} \tag{6.182}$$

Denoting the (unconditional) characteristic function in (6.178) (or (6.179)) by $f(\phi)$ and the underlying vector of model parameters by $\theta$. An ECF estimator can then be crafted à la GMM as

$$\hat{\theta} = \arg\min_{\theta} \int_{-\infty}^{\infty} |f(\phi) - g(\phi; y)|^2 w(\phi) d\phi \tag{6.183}$$

for an appropriate weighting function $w$ (*e.g.*, Gaussian). The role of the block length $p+1$ can be understood intuitively as a trade-off between cost and efficiency: larger values of $p$ (*i.e.*, longer block lengths) may be needed to better incorporate the effects of ergodicity arising from the (non-i.i.d.) stochastic variances, but this also of course increases the dimension of the integration in (6.183). In general, the estimator in (6.183) will display asymptotically normal consistency, a result whose utility we know must be viewed with a bit of trepidation. See Jiang and Knight (2002) and Rockinger and Semenova (2005) for a fuller discussion.[69]

## 6.6 Appendix

### 6.6.1 Continuous vs. discrete time

Up to this stage, the bulk of our discussion has concerned continuous time processes, *e.g.*, the canonical class of affine jump diffusions in Chapter 5. There is no need to delve into any deep philosophical issues to recognize the appeal of continuous time modeling. (We have already seen in Chapter 5 the great utility continuous time modeling provides for our central objective of identifying relevant value drivers arising from portfolio dynamics.) At the same time, there is a natural (unavoidable?) tendency to disregard information that may arise in the time intervals between discretely observed data, if only as a (reasonable) approximation. The

issue here is not so much the fact that data, and their subsequent econometric analysis, are (perhaps necessarily) reckoned as discrete entities. Rather, the question concerns the nature of the process that generated the data, and in what ways the passage of time matters. As we will see in due course, the importance to which we assign the role of time scales in characterizing the relevant behavior of a process is considerable. The horizons of interest are typically not impacted by any distinction between continuous time reality and discrete time approximation. Nonetheless, we should comment here on the role that small-time discretization effects play in econometric analysis.

Consider a generic diffusive process governed by

$$dx = \mu(x)dt + \sigma(x)dw$$

for some vector-valued process $x$. We typically have a set of observations denoted by

$$x(t_0), x(t_1), \cdots, x(t_N)$$

for some set of $N$ times, usually equally spaced, *e.g.*, $t_i = ih$ for some time step $h$. Following Yu (2014), we distinguish the following two limiting cases:

$$\begin{aligned} &N \to \infty, h \text{ fixed} \\ &h \to 0, N \text{ fixed} \end{aligned} \tag{6.184}$$

These two cases correspond to increasing sample size for a fixed resolution and increasing resolution for a fixed sample size, respectively.[70] To illustrate, consider the standard mean-reverting process

$$dz = \kappa(\theta - z)dt + \sigma\, dw \tag{6.185}$$

and its discrete-time analogue

$$z_n = c + \phi z_{n-1} + \varepsilon_n \tag{6.186}$$

with $\varepsilon_n \sim N(0, \sigma^2 \Delta t)$, $c = \kappa\theta\Delta t$, and $\phi = 1 - \kappa\Delta t$ for some (small) time step $\Delta t$. The conditional densities are both Gaussian, respectively:

$$\begin{aligned} &z_T | z \sim N(ze^{-\kappa(T-t)} + \theta(1 - e^{-\kappa(T-t)}), \sigma^2 \frac{1 - e^{-2\kappa(T-t)}}{2\kappa} \\ &z_{n+1} | z_n \sim N(\phi z_n + c, \sigma^2 \Delta t) \end{aligned} \tag{6.187}$$

Even simple econometric ramifications can immediately be seen from (6.187). The optimization problem posed by MLE, say, is considerably simpler in the discrete

time case. Note also that (6.186) can of course be thought of as a standard Euler discretization/approximation to (6.185). Note that this particular scheme is not the only possible one, and we see right away another challenge in analyzing continuous time models, namely the integration of underlying SDEs. Certainly, in the small time-step limit, the densities in (6.187) correspond to one another. It is thus reasonable to wonder how important the continuous/discrete dichotomy really is. As is so often the case, it depends. We are, in fact, more interested in a third alternative to the categories in (6.184), namely the role time horizons (for whatever class of dynamics) play in characterizing the flow of information, and ultimately their impact on valuation problems. (We are of course also interested in the basic inapplicability of large sample results for crafting robust estimators, and how to deal with the reality of small samples.) So, we will have little more to say on this issue, and will primarily operate in the widely used discrete-time framework, shifting to continuous time when suitable for extracting certain effects of interest.

### 6.6.2 Estimation issues for variance scaling laws

#### *6.6.2.1 Multidimensional diffusive processes*

In an important paper, Grzywacz and Wolyniec (2011) discuss some of the ramifications of the fact that commodity markets exhibit volatility whose effects are manifested (econometrically and otherwise) over differing time scales. We discuss these results in the context of a multidimensional setting. So consider a general Gaussian process with dynamics

$$dz_i = (A_{ij}z_j + b_i)dt + \sigma_i dw_i \tag{6.188}$$

for $i = 1,\ldots,N$ and (instantaneous) covariance structure given by $dw_i dw_j = \rho_{ij}\sigma_i\sigma_j dt \equiv X_{ij}dt$. Consider now the sample covariance of $T+1$ discretely observed realizations of (6.188):

$$\bar{\nu} = \frac{1}{T}\sum_{i=1}^{T} \Delta z_i \Delta z_i^T - \bar{\mu}\,\bar{\mu}^T \tag{6.189}$$

where $\Delta z_i \equiv z_i - z_{i-1}$ and the sample mean $\bar{\mu}$ is given simply by

$$\bar{\mu} = \frac{1}{T}\sum_{i=1}^{T} \Delta z_i = \frac{1}{T}(z_T - z_0)^{71} \tag{6.190}$$

*6.6.2.2 A familiar story: sample vs. population*

We are interested in knowing what, if any, relationship the sample variance in (6.189) has to the properties of the underlying process in (6.188). This is not as trivial a question as it may seem. Consider the case where $A = 0$, so the underlying process is simply a (correlated) random walk. Then clearly, in expectation (6.189) converges to the population covariance, and sensibly we can associate the sample variance with the population variance. However, in the presence of mean reversion (say), it is not immediately obvious what the sample variance would converge to. Note that differences are not i.i.d., so a standard premise commonly appealed to (if only implicitly) in statistical applications is no longer valid. It is precisely one of the contributions of Grzywacz and Wolyniec (2011) to address this question in the one-dimensional case. We now extend those results to higher dimensions.

*6.6.2.3 Another familiar story: affine analysis*

First, we need to understand the distributional properties of $z$. Using the characteristic function methods developed in Chapter 2, we see that $f = E_t e^{i\phi^T z(T')} = e^{\alpha^T z + \alpha_0}$ where the coefficients satisfy the following system of ODEs:

$$\begin{aligned} \dot{\alpha} &= A^T \alpha, \quad \alpha(0) = i\phi \\ \dot{\alpha}_0 &= b^T \alpha + \frac{1}{2} \alpha^T X \alpha, \quad \alpha_0(0) = 0 \end{aligned} \tag{6.191}$$

where we have made the transformation $\tau = T' - t$ (for some time horizon $T'$). Now, it is not hard to see from (6.191) that the overall structure of the process will be Gaussian (*i.e.*, the characteristic function will be an exponential of a linear-quadratic form in $\phi$), but it remains necessary to understand, at some level, the precise form. To this end, we introduce the diagonal factorization previously considered in Section 6.1.4 (again postponing consideration of the complications introduced by multiple eigenvalues): $A = V \Lambda V^{-1}$ where $V$ are the eigenvectors of $A$ and $\Lambda$ is a diagonal matrix of the eigenvalues $\lambda_i$. Thus the natural substitution $\beta = V^T \alpha$ transforms the system (6.191) into

$$\begin{aligned} \dot{\beta} &= \Lambda \beta, \quad \beta(0) = iV^T \phi \\ \dot{\alpha}_0 &= \tilde{b}^T \beta + \frac{1}{2} \beta^T \tilde{X} \beta, \quad \alpha_0(0) = 0 \end{aligned} \tag{6.192}$$

where $\tilde{b} = V^{-1} b$ and $\tilde{X} = V^{-1} X V^{-T}$. The transformed system (6.192) is easy to solve (the system now decouples owning to the diagonal nature), and upon transforming back we get

$$\begin{aligned}\alpha &= iV^{-T}L_\tau V^T\phi = i\tilde{L}_\tau^T\phi \\ \alpha_0 &= ib^T V^{-T}G_\tau V^T\phi - \frac{1}{2}\phi^T V\left(\int_0^\tau dsL_s\tilde{X}L_s\right)V^T\phi \\ &= ib^T\tilde{G}_\tau^T\phi - \frac{1}{2}\phi^T\int_0^\tau ds\tilde{L}_s X\tilde{L}_s^T\phi \end{aligned} \tag{6.193}$$

where $L$ and $G$ are diagonal matrices given by

$$\begin{aligned}L_\tau &= \text{diag}(\exp(\lambda_i\tau)) \\ G_\tau &= \int_0^\tau dsL_s = \text{diag}\left(\frac{\exp(\lambda_i\tau)-1}{\lambda_i}\right)\end{aligned} \tag{6.194}$$

and where $\tilde{L} \equiv VLV^{-1}$ (with a similar expression for $G$). Thus we see that $z$ is indeed normally distributed, with

$$z(\tau) \sim N_N\left(\tilde{L}_\tau z + \tilde{G}_\tau b, \int_0^\tau ds\tilde{L}_s X\tilde{L}_s^T\right)^{72} \tag{6.195}$$

or alternatively

$$z(\tau) - z \sim N_N\left(\widehat{L}_\tau z + \tilde{G}_\tau b, \int_0^\tau ds\tilde{L}_s X\tilde{L}_s^T\right) \tag{6.196}$$

where $\widehat{L} \equiv \tilde{L} - I$.

With the necessary dynamics in hand, we can now ask what the expectation of the sample variance is. We have

$$\begin{aligned}E_0\bar{\nu} &= \frac{1}{T}\sum_{i=1}^T E_0\Delta z_i\Delta z_i^\tau - E_0\bar{\mu}\,\bar{\mu}^T \\ &= \frac{1}{T}\sum_{i=1}^T E_0\Delta z_i\Delta z_i^T - \frac{1}{T^2}E_0(z_T - z_0)(z_T - z_0)^T\end{aligned} \tag{6.197}$$

As the covariance of the sample mean is fairly straightforward to evaluate, we will focus on the terms in the summation. We write these as

$$\frac{1}{T}\sum_{i=1}^{T} E_0 \Delta z_i \Delta z_i^T = \frac{1}{T}\sum_{i=1}^{T} E_0 E_{i-1}(z_i - z_{i-i})(z_i - z_{i-i})^T \tag{6.198}$$

where we have invoked the telescoping property of conditional expectations in the last equation. Now, from the process dynamics we have

$$\begin{aligned} &E_{i-i}(z_i - z_{i-i})(z_i - z_{i-i})^T \\ &= (\widehat{L}_{\Delta t} z_{i-1} + \tilde{G}_{\Delta t} b)(\widehat{L}_{\Delta t} z_{i-1} + \tilde{G}_{\Delta t} b)^T + \int_0^{\Delta t} ds \tilde{L}_s X \tilde{L}_s^T \\ &= \widehat{L}_{\Delta t} z_{i-1} z_{i-1}^T \widehat{L}_{\Delta t}^T + 2\mathrm{sym}(\widehat{L}_{\Delta t} z_{i-1} b^T \tilde{G}_{\Delta t}^T) + \tilde{G}_{\Delta t} b b^T \tilde{G}_{\Delta t}^T \int_0^{\Delta t} ds \tilde{L}_s X \tilde{L}_s^T \end{aligned} \tag{6.199}$$

where $\Delta t$ is the time step between observations, so that the time index $i$ will represent $\tau_i \equiv i\Delta t$ (so that $\Delta t = T'/T$), and sym denotes the symmetrization of a matrix: $\mathrm{SYM}(A) \equiv \frac{1}{2}(A + A^T)$. Also,

$$\begin{aligned} E_0 z_{i-1} &= d_{i-1} \\ E_0 z_{i-1} z_{i-1}^T &= d_{i-1} d_{i-1}^T + \int_0^{\tau_{i-1}} ds \tilde{L}_s X \tilde{L}_s^T \end{aligned} \tag{6.200}$$

where $d_{t-1} = \tilde{L}_{\tau_{i-1}} z_0 + \tilde{G}_{\tau_{i-1}} b$. From this we get

$$\begin{aligned} E_0(z_i - z_{i-1})(z_i - z_{i-1})^T &= \widehat{L}_{\Delta t} \left( d_{i-1} d_{i-1}^T + \int_0^{\tau_{i-1}} ds \tilde{L}_s X \tilde{L}_s^T \right) \widehat{L}_{\Delta t}^T \\ &+ 2\mathrm{sym}(\widehat{L}_{\Delta t} d_{i-1} b^T \tilde{G}_{\Delta t}^T) + \tilde{G}_{\Delta t} b b^T \tilde{G}_{\Delta t}^T + \int_0^{\Delta t} ds \tilde{L}_s X \tilde{L}_s^T \end{aligned} \tag{6.201}$$

*6.6.2.4 Putting it all together, and a one-dimensional example*

We can now start to put these results together. The expectation (conditional on initial time) of the sample variance is

$$E_0\bar{\nu} = \frac{1}{T}\sum_{i=1}^{T}\left(\begin{array}{l} \widehat{L}_{\Delta t}\left(d_{i-1}d_{i-1}^T + \int_0^{\tau_{i-1}} ds\tilde{L}_s X\tilde{L}_s^T\right)\widehat{L}_{\Delta t}^T + \\ 2\mathrm{sym}(\widehat{L}_{\Delta t}d_{i-1}b^T\tilde{G}_{\Delta t}^T) + \\ \tilde{G}_{\Delta t}bb^T\tilde{G}_{\Delta t}^T + \int_0^{\Delta t} ds\tilde{L}_s X\tilde{L}_s^T \end{array}\right) - \frac{1}{T^2}E_0(z_T - z_0)(z_T - z_0)^T \tag{6.202}$$

It is easy to see that if the process is purely diffusive (so that $L = I$ and for convenience we ignore drift), then this expectation is just the population covariance $X\Delta t$.

In the one-dimensional (standard mean-reverting) case, we see that

$$\begin{aligned} E_0\bar{\nu} &= \frac{1}{T}\sum_{i=1}^{T}\left(\begin{array}{l} (1-e^{-\kappa\Delta t})^2\left(\begin{array}{l}(e^{-\kappa\tau_{i-1}}z_0 + \theta(1-e^{-\kappa\tau_{i-1}}))^2 \\ +\frac{\sigma^2}{2\kappa}(1-e^{-2\kappa\tau_{i-1}})\end{array}\right) - \\ 2(1-e^{-\kappa\Delta t})^2\theta(e^{-\kappa\tau_{i-1}}z_0 + \theta(1-e^{-\kappa\tau_{i-1}})) + \\ (1-e^{-\kappa\Delta t})^2\theta^2 + \frac{\sigma^2}{2\kappa}(1-e^{-2\kappa\Delta t}) \end{array}\right) \\ &\quad - \frac{1}{T^2}E_0(z_T - z_0)(z_T - z_0)^T \\ &= \frac{1}{T}\sum_{i=1}^{T}\left(\begin{array}{l} (1-e^{-\kappa\Delta t})^2(z_0-\theta)^2e^{-2\kappa\tau_{i-1}} + \\ \frac{\sigma^2}{2\kappa}(1-e^{-\kappa\Delta t})^2(1-e^{-2\kappa\tau_{i-1}}) + \\ \frac{\sigma^2}{2\kappa}(1-e^{-2\kappa\Delta t}) \end{array}\right) - \frac{1}{T^2}E_0(z_T - z_0)(z_T - z_0)^T \end{aligned} \tag{6.203}$$

Now, ignoring the last term as asymptotically small and assuming for convenience that the process starts at the long-term mean, (6.203) can be rewritten as

$$\begin{aligned} E_0\bar{\nu} = \frac{\sigma^2}{2\kappa} &= (1-e^{-\kappa\Delta t})\frac{1}{T}\sum_{i=1}^{T}((1-e^{-\kappa\Delta\tau})(1-e^{-2\kappa\tau_{i-1}}) + 1 + e^{-\kappa\Delta t} \\ &= \frac{\sigma^2}{\kappa}(1-e^{-\kappa\Delta t})\frac{1}{T}\sum_{i=1}^{T}(1 - \frac{1}{2}(1-e^{-\kappa\Delta t})e^{-2\kappa\tau_{i-1}} \end{aligned} \tag{6.204}$$

The contribution of the exponentials to the summation will tend to be small for larger sample sizes and/or time horizons,[73] so we arrive at the result

$$E_0\bar{\nu} \approx \frac{\sigma^2}{\kappa}(1-e^{-\kappa\Delta t}) \tag{6.205}$$

which is essentially the result of Grzywacz and Wolyniec (2011), namely that in expectation the sample variance displays evidence of mean reversion at *half* the actual rate.

#### 6.6.2.5 *Some general asymptotics*

Let us continue to consider the special case where the process starts at the reversion level, which we take to be zero. This amounts to requiring that $z_0 = b = 0$. Then, the result in (6.202) can be written conveniently as

$$E_0\bar{\nu} = \frac{1}{T}\sum_{i=1}^{T} V\left((L_{\Delta t} - I)\int_0^{\tau_{i-1}} ds L_s \tilde{X} L_s (L_{\Delta t} - I) + \int_0^{\Delta t} ds L_s \tilde{X} L_s\right) V^T$$
$$- \frac{1}{T^2} E_0(z_T - z_0)(z_T - z_0)T \tag{6.206}$$

Now, the integrands in (6.206) have the form

$$\int_0^{\tau} ds \tilde{X}_{ij} e^{(\lambda_i+\lambda_j)s} = \tilde{X}_{ij}\frac{e^{(\lambda_i+\lambda_j)s} - 1}{\lambda_i + \lambda_j} \tag{6.207}$$

and thus can be written as $\tilde{X} \circ K_\tau$, where the denotes Hadamard (element-by-element) matrix multiplication and the elements of $K$ are given by[74] $K_{ij}(\tau) = (\exp(\lambda_i + \lambda_j)\tau - 1)/(\lambda_i + \lambda_j)$. Ignoring again the second term in (6.206) (as it will vanish asymptotically), we are led to consider

$$E_0\bar{\nu} = V((L_{\Delta t} - I)(\tilde{X} \circ \frac{1}{T}\langle K_{\tau_{i-1}}\rangle)(L_{\Delta t} - I) + \tilde{X} \circ K_{\Delta t})V^T \tag{6.208}$$

where brackets denote ensemble summation. Now, assuming all eigenvalues are (strictly) negative (so, roughly speaking, the process can be thought of as stationary over long enough time horizons), we see (as in the one dimensional case, *e.g.*, (6.204)) we see that the long-term sample average of the matrix $K$ is simply the matrix $M_{ij} = -(\lambda_i + \lambda_j)^{-1}$ and the asymptotic behavior of the estimator (the sample variance) is given by

$$E_0\bar{\nu} \sim V(-(L_{\Delta t} - I)(\bar{X} \circ M)(L_{\Delta t} - I) + \tilde{X} \circ K_{\Delta t})V^T \tag{6.209}$$

*6.6.2.6 A two-dimensional example*

It would now be helpful to consider an actual example, and we will use again the "cascading" extension of the standard mean-reverting model, with a stochastic mean-reversion level:[75]

$$\begin{aligned} dx &= \kappa_x(y-x)dt + \sigma_x dw_x \\ dy &= \kappa_y(\theta-y)dt + \sigma_y dw_y \end{aligned} \tag{6.210}$$

The associated eigenvalues and eigenvector matrix are, respectively, given by $(-\kappa_x, -\kappa_y)$ and $\begin{pmatrix} 1 & r \\ 0 & 1 \end{pmatrix}$, where $r = \frac{\kappa_x}{\kappa_x - \kappa_y}$.[76] For further convenience we will assume the processes in (6.210) are independent, so that $X = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$ and $\tilde{X} = \begin{pmatrix} \sigma_x^2 + r^2\sigma_y^2 & -r\sigma_y^2 \\ -r\sigma_y^2 & \sigma_y^2 \end{pmatrix}$. The asymptotic result (6.209) then becomes

$$\begin{aligned} &\begin{pmatrix} 1 & r \\ 0 & 1 \end{pmatrix} \left[ \begin{pmatrix} (\sigma_x^2 + r^2\sigma_y^2)\frac{(1-e^{-\kappa_x\Delta t})^2}{2\kappa_x} & -r\sigma_y^2\frac{(1-e^{-\kappa_x\Delta t})(1-e^{-\kappa_y\Delta t})}{\kappa_x+\kappa_y} \\ -r\sigma_y^2\frac{(1-e^{-\kappa_x\Delta t})(1-e^{-\kappa_y\Delta t})}{\kappa_x+\kappa_y} & \sigma_y^2\frac{(1-e^{-\kappa_y\Delta t})^2}{2\kappa_y} \end{pmatrix} + \right. \\ &\left. \begin{pmatrix} (\sigma_x^2 + r^2\sigma_y^2)\frac{1-e^{-2\kappa_x\Delta t}}{2\kappa_x} & -r\sigma_y^2\frac{1-e^{-(\kappa_x+\kappa_y)\Delta t}}{\kappa_x+\kappa_y} \\ -r\sigma_y^2\frac{1-e^{-(\kappa_x+\kappa_y)\Delta t}}{\kappa_x+\kappa_y} & \sigma_y^2\frac{1-e^{-2\kappa_y\Delta t}}{2\kappa_y} \end{pmatrix} \right] \\ &\begin{pmatrix} 1 & 0 \\ r & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & r \\ 0 & 1 \end{pmatrix} \left[ \begin{pmatrix} (\sigma_x^2 + r^2\sigma_y^2)\frac{1-e^{-\kappa_x\Delta t}}{\kappa_x} & -r\sigma_y^2\frac{2-e^{-\kappa_x\Delta t}-e^{-\kappa_y\Delta t}}{\kappa_x+\kappa_y} \\ -r\sigma_y^2\frac{2-e^{-\kappa_x\Delta t}-e^{-\kappa_y\Delta t}}{\kappa_x+\kappa_y} & \sigma_y^2\frac{1-e^{-\kappa_y\Delta t}}{\kappa_y} \end{pmatrix} \right] \\ &\begin{pmatrix} 1 & 0 \\ r & 1 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_x^2\frac{1-e^{-\kappa_x\Delta t}}{\kappa_x} + r^2\sigma_y^2\left(\frac{1-e^{-\kappa_x\Delta t}}{\kappa_x} - 2\frac{2-e^{-\kappa_x\Delta t}-e^{-\kappa_y\Delta t}}{\kappa_x+\kappa_y} + \frac{1-e^{-\kappa_y\Delta t}}{\kappa_y}\right) & \\ r\sigma_y^2\left(\frac{1-e^{-\kappa_y\Delta t}}{\kappa_y} - \frac{2-e^{-\kappa_y\Delta t}-e^{-\kappa_y\Delta t}}{\kappa_x+\kappa_y}\right) & \\ r\sigma_y^2\left(\frac{1-e^{-\kappa_y\Delta t}}{\kappa_y} - \frac{2-e^{-\kappa_x\Delta t}-e^{-\kappa_y\Delta t}}{\kappa_x+\kappa_y}\right) & \\ \sigma_y^2\frac{1-e^{-\kappa_y\Delta t}}{\kappa_y} & \end{pmatrix} \end{aligned} \tag{6.211}$$

The result in (6.211) must be compared with the corresponding analytical results for the process covariance matrix over a time interval $\Delta t$, which is given by

$$\begin{pmatrix} \sigma_x^2 \frac{1-e^{-2\kappa_x \Delta t}}{2\kappa_x} + r^2\sigma_y^2 \left( \frac{1-e^{-2\kappa_x \Delta t}}{2\kappa_x} - 2\frac{1-e^{-(\kappa_x+\kappa_y)\Delta t}}{\kappa_x+\kappa_y} + \frac{1-e^{-2\kappa_y \Delta t}}{2\kappa_y} \right) & & \\ r\sigma_y^2 \left( \frac{1-e^{-2\kappa_y \Delta t}}{2\kappa_y} - \frac{1-e^{-(\kappa_x+\kappa_y)\Delta t}}{\kappa_x+\kappa_y} \right) & & \\ r\sigma_y^2 \left( \frac{1-e^{-2\kappa_y \Delta t}}{2\kappa_y} - \frac{1-e^{-(\kappa_x+\kappa_y)\Delta t}}{\kappa_x+\kappa_y} \right) & & \\ \sigma_y^2 \frac{1-e^{-2\kappa_y \Delta t}}{2\kappa y} & & \end{pmatrix} \tag{6.212}$$

We thus see that the result of Grzywacz and Wolyniec (2011) has an analogue in higher dimensions, specifically the convergence of the sample covariance to a form reflecting *half* of the true mean-reversion rate.

However, as interesting as this generalization may be, there is actually a more important observation to be made. Recall that the applicability of the asymptotic result in (6.211) depends on the sample size being large enough that the ensemble (matrix) averages[77] $\frac{1}{T}\langle \exp((\lambda_i + \lambda_j)\tau_k) \rangle$ tend to zero. With sufficient algebraic effort, it can be shown that these correction terms for sample variance of process $x$ are given by

$$-T^{-1} \begin{pmatrix} \sigma_x^2 \frac{1-e^{-\kappa_x \Delta t}}{1+e^{-\kappa_x \Delta t}} \frac{1-e^{-2\kappa_x T\Delta t}}{2\kappa_x} + r^2\sigma_y^2 \left( \frac{1-e^{-\kappa_x \Delta t}}{1+e^{-\kappa_x \Delta t}} \frac{1-e^{-2\kappa_x T\Delta t}}{2\kappa_x} \right. \\ \left. -2\frac{(1-e^{-\kappa_x \Delta t})(1-e^{-\kappa_y \Delta t})}{1-e^{-(\kappa_x+\kappa_y)\Delta t}} \frac{1+e^{-(\kappa_x+\kappa_y)T\Delta t}}{\kappa_x+\kappa_y} + \frac{1-e^{-\kappa_y \Delta t}}{1+e^{-\kappa_y \Delta t}} \frac{1+e^{-2\kappa_y T\Delta t}}{2\kappa_y} + \right) \end{pmatrix} \tag{6.213}$$

(Compare with (6.204).) Of course, for stationary processes, the expression in (6.213) tends to zero as $T \to \infty$. However, the issue becomes a bit more subtle for finite sample sizes. It can be seen from (6.213) that the most slowly varying terms are those involving the smallest mean-reversion rates. In the typical case, the stochastic mean ($y$) in (6.210) exhibits much less mean reversion than the primary asset ($x$). That is to say, deviations of the stochastic mean from *its* equilibrium level represent a more slowly varying transient that deviations of the heat rate from the mean itself. A common example is a heat rate, where the stochastic mean reflects capital effects (such as stack growth), which are typically "more" non-stationary than heat rates themselves, which are also driven by factors such as weather. As a result there is typically a balance between longer-term, relatively non-stationary (supply) effects, and shorter-term, relatively stationary (demand) effects. (This is not just a theoretical nicety: we will shortly see evidence in the actual data, as manifested in variance scaling laws.) Again we see the importance of time scales in a problem, where the impact of information from particular sources depends on the time scale over which such sources operate.

Thus, the effects of small samples (specifically, their corruption of asymptotic convergence of estimators and the attendant instability) on the standard

estimator (6.189) are most acutely channeled by the most non-stationary factors in the process dynamics. As the correction terms in (6.213) indicate, even for simple estimators such as the sample variance, reducing the impact of slower varying time scales on the presumed relation between sample and population requires that the sample size be large *in relation to* the characteristic time scale of the most non-stationary underlying driver. *A fortiori* we would expect more sophisticated methods that impose much greater structure on the problem to be even more susceptible to these time-scaling effects. Let us further stress: this is not simply an issue of being unable to extract useful information regarding long-term effects. *Even the estimation of short-term effects is affected by the presence of (relatively) more non-stationary factors!*

### 6.6.3 High-frequency scaling

With the advent of technological innovations in information processing and data collection, there has been much interest in trading over very small time horizons.[78] This interest applies to commodity markets no less than equity markets. While much econometric work concerning high-frequency data is well beyond the scope of the present volume (a good recent source is Aït-Sahalia and Jacod [2014]), we will mention some tools that represent generalizations of the concepts that apply over more conventional (*e.g.*, daily) time horizons. Specifically, we are interested in modeling realized variance, which as we have already seen (*e.g.*, Chapter 3) plays a central role in valuation. Furthermore, the relative importance of time scales in commodity markets means that valuation is critically dependent on the choice of hedging strategy and portfolio formation. Here, we will take a somewhat different approach from prior sections and look into the behavior of variance over decreasingly small time horizons. Of interest here will be the role played by jumps.

#### *6.6.3.1 Quadratic variation*

First we start with the notion of quadratic variation, which should be very familiar to most readers. The basic foundational concept is the well-known class of semimartingales, which are stochastic processes $X_t$ that can be decomposed into two càdlàg,[79] adapted[80] processes as

$$X_t = M_t + A_t \tag{6.214}$$

where $M_t$ is a local martingale[81] and $A_t$ is a finite variation process. For our purposes this latter characteristic means that, over any finite time interval $[0, T]$,

$$\operatorname*{plim}_{h\to 0+} \sum_{i=1}^{\mathrm{int}(T/h)} |A_{ih} - A_{(i-1)h}| < \infty \tag{6.215}$$

where int($x$) denotes the integer part of $x$. Then, we can define the quadratic variation (QV)[82] of $X$:

$$[X]_T = \operatorname*{plim}_{h\to 0+} \sum_{i=1}^{\text{int}(T/h)} (X_{ih} - X_{(i-1)h})^2 \tag{6.216}$$

giving rise to the familiar isometry $d[X]_t = (dX_t)^2$. Plainly QV relates to realized variance (RV), although we do not elaborate on this fact here.[83] (Recall the standard result that Brownian motion has infinite variation but finite quadratic variation, while a continuous finite variation process has zero quadratic variation; it is not hard to see that $[w]_T = T$ for a standard Brownian motion.)

Before going further, we should recall why we would care about things like quadratic variation. We have actually already seen the main ideas laid out in Chapter 3 on dynamic optimality and informational efficiency for portfolio construction. In fact, eq. (3.25) makes clear the role quadratic variation plays as a value driver in terms of which residual exposure must be reckoned. That discussion took place in the context of continuous (diffusion) processes and continuous portfolio rebalancing. To this latter point, we can never actually re-hedge in continuous time, so we are obviously interested in the limiting behavior in discrete time of entities such as (6.216). In many markets, intraday rebalancing is possible (and as a result, potentially significant incremental value over re-hedging based on daily settlement only), which raises the question of how value drivers behave over higher (time) resolutions. In addition, it should be fairly clear that sums-of-squares are rather easy to implement in practice and so possess some inherent econometric appeal.

#### *6.6.3.2 Jumps and such*

As just noted, much of our previous discussion has neglected jump effects. There is no doubting the existence of jump effects in energy markets, so it is worth investigating the impact jumps have on entities of interest such as quadratic variation. For simplicity, we will confine attention to the case where $A$ is a continuous process (and thus has zero quadratic variation) so that any jump effects are associated with the (local) martingale component $M$. We thus write (6.214) as

$$X_t = M_t^c + M_t^d + A_t \tag{6.217}$$

where the superscripts $c$ and $d$ denote continuous and discontinuous components, respectively. We will further assume that $A$ and $M^{c,d}$ are independent. With these assumptions, it is reasonably straightforward to show that

$$[X]_T = \left[M^c\right]_T + \left[M^d\right]_T \Rightarrow$$

$$E_0[X]_T = E_0 \int_0^T \text{var}_u(dM_u^c) + E_0 \sum_{u=1}^{N(T)} E_{u-}(\Delta M_u^d)^2 \tag{6.218}$$

where $N(T)$ is the number of jumps in the time interval in question and where $\Delta M_u^d = M_u^d - M_{u-}^d$ is the jump size at the jump time $u$. In other words, in the limit of infinitesimal resolution, the sum-of-squares of the process differences converges at least in expectation to the integrated (conditional, instantaneous) variance of the continuous martingale term plus the quadratic variation of the discontinuous (jump) component. (In fact, stronger results are possible, such as convergence in probability for specified dynamics; for more on the relationship between QV, RV, and integrated variance [IV], see Barndorff-Nielsen and Shephard [2002] or McAleer and Medeiros [2008].)

As (6.218) indicates, the realized variance/quadratic variation will in general include contributions from both the (continuous) diffusion components and the jump components. It is of interest to be able to distinguish the two effects, so we consider a variant of standard quadratic variation, originally due to Barndorff-Nielsen and Shephard (2004a, b).[84]

#### 6.6.3.3 *Power/bipower variation*

A natural generalization of (6.216) is the $r$th order power variation, defined as

$$[X]_T^r = \underset{h\to 0+}{\text{plim}}\, h^{1-r/2} \sum_{i=1}^{\text{int}(T/h)} |X_{ih} - X_{(i-1)h}|^r \tag{6.219}$$

Note that for a Brownian motion with time-inhomogeneous volatility, *i.e.*, with dynamics $dX_t = \sigma_t dw_t$, we have that

$$[X]_T^r = E|\phi|^r \int_0^T \sigma_t^r dt \tag{6.220}$$

with $\phi$ a unit normal. A further extension is the $(r,s)$ -order bipower variation, defined by

$$[X]_T^{r,s} = \underset{h\to 0+}{\text{plim}}\, h^{1-(r+s)/2} \sum_{i=2}^{\text{int}(T/h)} |X_{ih} - X_{(i-1)h}|^r |X_{(i-1)h} - X_{(i-2)h}|^s \tag{6.221}$$

For the aforementioned inhomogeneous Brownian motion, we would have that

$$[X]_T^{r,s} = E|\phi|^r \cdot E|\phi|^s \int_0^T \sigma_t^{r+s} dt \tag{6.222}$$

While interesting in their own right, power and bipower variation find their greatest utility in identifying jumps in a process. We will confine attention to the class of finite activity jump processes (so essentially compound Poisson processes). We write such processes generically as

$$X_t = \int_0^t \mu_s ds + \int_0^t \sigma_s dw_s + \sum_{i=1}^{N_t} J_i \tag{6.223}$$

for suitably adapted, càdlàg drifts, volatilities, and jumps (with diffusions and jumps independent). Then, it is not hard to see that the following hold:[85,86]

$$\begin{aligned} [X]_t &= \int_0^t \sigma_s^2 ds + \sum_{i=1}^{N_t} J_i^2 \\ [X]_t^{1,1} &= \tfrac{2}{\pi} \int_0^t \sigma_s^2 ds \end{aligned} \tag{6.224}$$

Thus, the difference between RV and $\frac{\pi}{2}$ times the realized (1, 1) bipower variation can serve as consistent estimator of the QV of the underlying jump component. This simple procedure can serve as a simple means for detecting, if not the actual presence of jumps, then the magnitude of their impact. As always, the question we ask regarding any effect is: how much does it matter?

As we illustrated with examples in (2.43) and (2.45), the impact of jumps depends (to a very large extent) on the time horizon over which particular jumps operate. Specifically, the issue concerns how jumps in spot processes manifest themselves as jumps in expectations that form the basis of trading (*i.e.*, in forwards). For example, we saw how the standard mean-reverting plus jump model of heat rates with spikes (basically, Merton plus mean reversion) gives rise to forward dynamics where the impact of jumps is primarily close to expiration. It is well known that delta hedging in the presence of jumps can be difficult (because of the nature of the hedging error). As we have noted, this difficulting is reflected in the instability of estimates of value drivers in the range where jumps are operative. It is thus important that one recognize where these effects can occur, and adjust accordingly (*e.g.*, by not over-attributing collectable value to jumps). Techniques such as bipower variation can be useful in identifying jumps, but only if one knows the right place to look.[87]

# 7 Numerical Methods

## 7.1 Basics of spread option pricing

Spread options are pervasive in energy markets. While we must refer the reader to EW for a complete treatment of the various structures that entail such optionality, for context we will outline a few examples here, stressing that other, more mathematical, issues are the central focus here. To recap, examples include:

- Tolling/heat rate options: a right to buy fuel and sell power
- Transport: a right to flow natural gas from one location to another
- Storage: a right to buy natural gas in summer and sell it in winter.

Indeed, as we have seen, phenomena in energy markets such as mean reversion that offer a striking contrast to equity markets can be well understood (in terms of how they impact valuation) from a study of spread options, making the techniques developed here very relevant for understanding the valuation of other kinds of products. Such options can be either financially or physically settled, and we have discussed in great detail the relevant factors impacting both kinds of valuation. Here, we simply note that the essential characteristic of such optionality is a payoff of the form

$$(S_2 - S_1 - K)^+ \tag{7.1}$$

This payoff has some definite similarities with the payoff of a single asset option, and one might suppose that at least a few of the well-developed methods for pricing the latter kind of option can be applied to the former. This intuition is true to an extent, and we will discuss in this section both the commonalities and differences.

### 7.1.1 Measure changes

First, we examine to what extent the standard BS results carry over when an additional asset is introduced. Assume the two asset prices follow GBM:

$$\frac{dS_1}{S_1} = \mu_1 dt + \sigma_1 dw_1$$
$$\frac{dS_2}{S_2} = \mu_2 dt + \sigma_2 dw_2 \tag{7.2}$$

with some (instantaneous) correlation between them ($dw_1 dw_2 = \rho dt$ in the Itô isometry). Now, following the usual heuristic approach adopted in derivations of the BS equations, assume we hold a portfolio $\Pi$ of an option $V$ and some positions $\Delta_i$ in the underlyings, and demand that the resulting portfolio is instantaneously riskless (we again neglect any effects due to discounting, and hence omit any kind of cash/bond term in the portfolio). The portfolio dynamics are given by

$$\begin{aligned} d\Pi &= dV + \Delta_1 dS_1 + \Delta_2 dS_2 \\ &= (V_t + \tfrac{1}{2}(\sigma_1^2 S_1^2 V_{S_1 S_1} + 2\rho\sigma_1\sigma_2 S_1 S_2 V_{S_1 S_2} + \sigma_2^2 S_2^2 V_{S_2 S_2}))dt \\ &\quad + (V_{S_1} + \Delta_1)dS_1 + (V_{S_2} + \Delta_2)dS_2 \end{aligned} \tag{7.3}$$

Now, the usual argument (which, as we have seen in the previous sections of this chapter, is rather problematic even if it does formally reproduce the correct result) holds that if we take $\Delta_i = -V_{S_i}$, then the option value must satisfy (to avoid arbitrage opportunities that are assumed to not exist) the following PDE:

$$V_t + \tfrac{1}{2}(\sigma_1^2 S_1^2 V_{S_1 S_1} + 2\rho\sigma_1\sigma_2 S_1 S_2 V_{S_1 S_2} + \sigma_2^2 S_2^2 V_{S_2 S_2}) = 0 \tag{7.4}$$

with terminal conditions given by (7.1): $V(S_1, S_2, T) = (S_2 - S_1 - K)^+$. There are clearly similarities with the usual BS equation here, as we are still dealing with a second-order, parabolic PDE. The obvious question is: can it be solved, as it can in the BS case?

It turns out for the special case $K = 0$ (*i.e.*, no fixed strike) that it can. With the similarity transformation $S_2 = SS_1$, $V = S_1 U(S)$, the PDE becomes

$$U_t + \tfrac{1}{2}\sigma^2 S^2 U_{SS} = 0 \tag{7.5}$$

with $U(S, T) = (S-1)^+$. The entity $\sigma$ is given by

$$\sigma^2 = \sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2 \tag{7.6}$$

and can be termed, for obvious reasons, the *ratio volatility*.[1] The solution for $U$ is clearly of BS type, and the general solution for the option price can be written as

$$V(S_1, S_2) = S_1 V_{BS}\left(\frac{S_2}{S_1}, 1, \sigma\right) \tag{7.7}$$

or explicitly as

$$V = S_2 N\left(\frac{\log(S_2/S_1) + \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}}\right) - S_1 N\left(\frac{\log(S_2/S_1) - \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}}\right) \tag{7.8}$$

Where, as usual, time to maturity is given by $\tau \equiv T - t$. A straightforward calculation shows that the deltas are simply the coefficients of $S_1$ and $S_2$ in the expression (7.8).[2] The evident correspondence of the asset $S_1$ with the fixed strike in the standard BS setup is in fact not a coincidence, as will be made clear in the next section when change-of-measure techniques will be used to derive this result much more compactly.

The result in (7.8) is commonly referred to as the Margrabe formula, and provides a very useful means for valuing spread-type structures. Note in particular that, after rescaling in terms of $S_1$, the option value depends *only* on the ratio of the assets, time to maturity, and the ratio volatility. That is, the individual leg volatilities and correlation do not enter into the calculation directly, but only through the specific form given by (7.6). This point has been seen to have great importance for reducing the scope of identifying the relevant valuation parameters for a given pricing problem.[3]

Before turning attention to the complications arising from the case of a nonzero fixed strike $K$ in the payoff (7.1), it is worth noting another approach to modeling spread-type structures. Note that this payoff depends only on the difference between the two assets (hence the term "spread"). Now, it is certainly plausible that in some situations it is more appropriate to model this difference directly, that is, without separate reference to the constituent legs comprising this difference. For example, instead of viewing this spread as the difference of lognormal assets, we could view the spread itself as normally distributed.[4] This is the so-called Bachelier model of option pricing. An application we will briefly note here is a natural gas transport option (recall Section 1.2.2). In these markets, locational prices (say, a Rockies price) do not trade as stand-alone commodities. Rather, they trade as differentials to some primary hub or backbone price (say, Henry Hub gas).[5] Now, in general such differentials can (and often are) negative, so a lognormal representation of the process relevant for valuation might not be appropriate.[6] In such cases the spread is often modeled directly (as opposed to separate modeling of the two legs and their joint behavior), and an obvious choice is to treat the spread as normally distributed. A suitable model of the process under the pricing measure (recall that as tradeables the spreads/differentials are martingales under such a measure) is

$$dz = \sigma\, dw \tag{7.9}$$

A straightforward calculation yields

$$E_t^Q(z_T - k)^+ = (z-k)N\left(\frac{z-k}{\sigma}\right) + \sigma \cdot \Phi\left(\frac{z-k}{\sigma}\right) \tag{7.10}$$

We will have more to say on the value drivers appropriate for using such models later. The point that must be stressed here is that the choice of model is driven by the structure of the underlying market. In markets where the basis is liquidly traded on a forward basis, these products are the appropriate hedging instrument, and valuation (which we stress again is the necessary counterpart of hedging) must take place in terms of these (and not the legs as such).

We now turn attention to the question of how to price spread options when there is a nonzero strike term.

### 7.1.2 Approximations

It should not be terribly hard to see that the similarity solution used in (7.7) to derive the Margrabe formula cannot be used when there is a nonzero strike (that is, $K \neq 0$). Although we will discuss in great detail in this chapter various numerical methods (primarily quadrature) for effectively computing the option price in this case, there is great utility in having analytical expressions for the price, even if only approximately. We would expect the Margrabe formula to still provide some range of validity, by expanding the payoff function for small $K$:

$$(S_2 - S_1 - K)^+ \approx (S_2 - S_1)^+ - K \cdot H(S_2 - S_1) \tag{7.11}$$

where $H$ denotes the Heaviside step function. Thus the full spread option is, at least for sufficiently small strikes, approximately a combination of a Margrabe-type option and a digital/binary option. An important conclusion, which we will expand upon later, is that even in the general case we would expect the ratio volatility to play an important role in the valuation of the option. This has major ramifications for identifying and estimating the appropriate value driver in a given valuation problem.

However, even if we can establish the relevance of the ratio volatility for a particular valuation problem, it may be desirable to account for "higher order," so to speak, terms in the expansion in (7.11). (Note that even a stand-alone digital spread option will have formal dependence on the correlation between the two legs.) A number of approximations have been developed to this end. A veritable menagerie is presented in Venkatramanan and Alexander (2011). We will focus on the most well-known techniques here.

By far the most widely used approach is originally due to Kirk (1996). The basic idea can be understood by treating the "aggregate strike" $\tilde{S}_1 = S_1 + K$ as approximately lognormal,[7] in which case the standard Margrabe result can be applied. The necessary (covariance) parameters are obtained via moment matching:

$$\begin{aligned} E_t \tilde{S}_1(T) S_2(T) &= S_2(S_1 e^{\rho\sigma_1\sigma_2\tau} + K) \to \tilde{S}_1 S_2 e^{\tilde{\rho}\tilde{\sigma}_1\sigma_2\tau} \\ E_t \tilde{S}_1(T)^2 &= S_1^2 e^{\sigma^2\tau} + 2S_1 K + K^2 \to \tilde{S}_1^2 e^{\tilde{\sigma}_1^2\tau} \end{aligned} \tag{7.12}$$

where the tildes denote the equivalent lognormal entities (we anticipate that the volatility of $S_2$ will remain unchanged). Of course, (7.12) can be readily solved, but an extremely useful form comes from approximating the exponentials by a first-order Taylor series, in which case we find that

$$\tilde{\sigma}_1 \tilde{S}_1 = S_1 \sigma_1, \quad \tilde{\rho} = \rho \tag{7.13}$$

or more commonly

$$\tilde{\sigma}_1 = \frac{S_1}{S_1 + K} \sigma_1 \tag{7.14}$$

The approximation thus becomes

$$V(S_1, S_2, \sigma_1, \sigma_2, \rho) \approx \tilde{S}_1 V_{BS}\left(\frac{S_2}{\tilde{S}_1}, 1, \tilde{\sigma}\right) \tag{7.15}$$

where the adjusted ratio volatility is given by

$$\tilde{\sigma}^2 = \tilde{\sigma}_1^2 - 2\rho\tilde{\sigma}_1\sigma_2 + \sigma_2^2 \tag{7.16}$$

Observe that to leading order in the relative strike $(K/S_1)$ the adjusted volatility is approximately

$$\tilde{\sigma}^2 = \sigma^2 + 2\sigma_1(\rho\sigma_2 - \sigma_1)\frac{K}{S_1} \tag{7.17}$$

which highlights not only the relevance of the ratio volatility even in problems with nonzero strike, but also how the correction term is related to a vega hedge in one of the leg assets (see Endnote 3 above).[8] A point that is sometimes overlooked is that the delta with respect to leg 1 will include a vega contribution due to the dependence of $\sigma_1$ on $S_1$ in (7.14).

Now, an obvious question is: how good is the approximation? The short answer is, surprisingly good. Consider the following set of parameters:

$$S_1 = S_2 = 1, \quad \sigma_1 = \sigma_2 = 0.5, \quad \rho = 0.8, \quad \tau = 0.75$$

Figure 7.1 shows a comparison of the extrinsic value (defined as option value minus intrinsic value [current payoff]) for a range of strikes, between the Kirk approximation and the exact value (here obtained from quadrature, to be discussed in greater detail in Section 7.4). As can be seen, Kirk provides an excellent approximation for a very wide range of strikes (in this case, up to 50% of either underlying leg), especially for positive strikes. The poor agreement for negative strikes reflects the fact that the underlying equivalent lognormal approximation in (7.12) breaks down for negative values of the summands. (The obvious remedy in such cases is to treat $S_2 - K$ as lognormal, but we will not pursue such an investigation here.) But for the deep OTM strikes (where the equivalent lognormal approximation should be expected to be reasonable but *a priori* not necessarily fantastic), it is quite striking how well Kirk performs.

There are of course limits to any approximation, and we show the results (for the OTM case) for very high correlation ($\rho = 0.99$) in Figure 7.2.

So we can see in such cases that the approximation is less impressive. (*A fortiori* the Kirk deltas and other greeks will likewise be less accurate, as well.) In fact, the accuracy of Kirk depends, as well, on the moneyness between the two legs and time to maturity, and we do not propose to catalog the extent of its range of validity here. The point is that Kirk is a very effective (and not surprisingly, very popular) technique for getting good spread-option values in a wide range of situations, without much more pain than a standard European option calculation.

Although we shall not delve into any details here, we will note that the equivalent lognormal approximation (via moment matching) in (7.12) can be effectively applied to higher dimensional options, such as those with payoffs such as

$$(S_3 - \min(S_1, S_2) - K)^+ \tag{7.18}$$

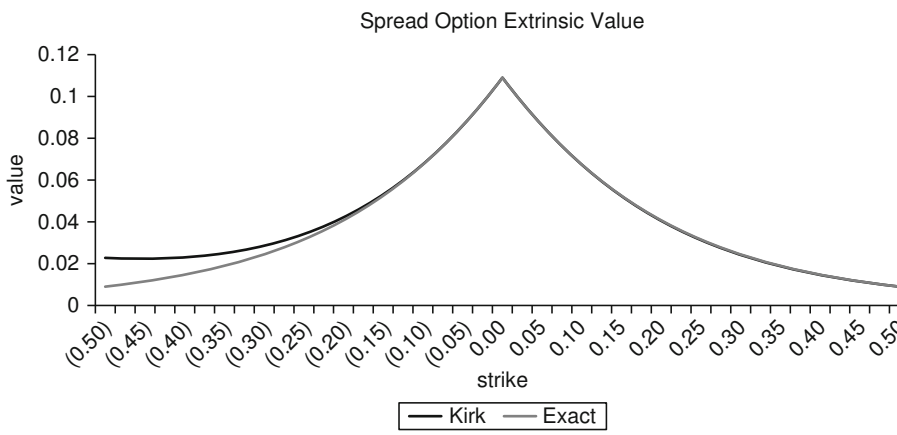



**Figure 7.1** Comparison of spread option extrinsic value as a function of strike. $\rho = 0.8$

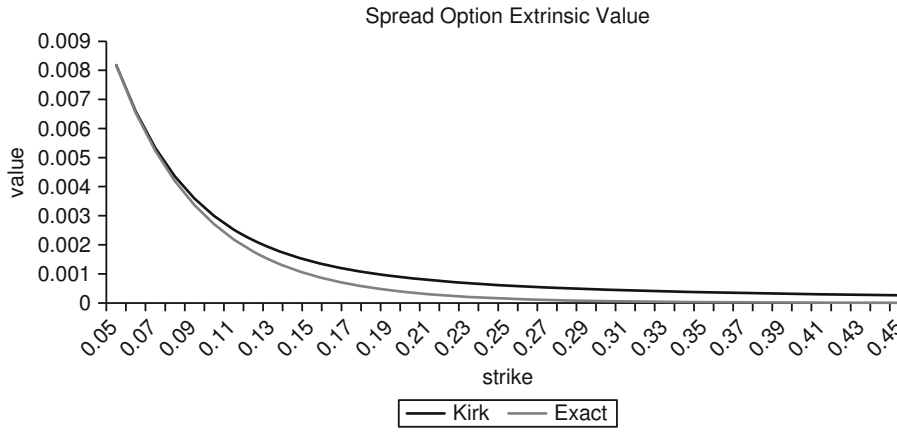



**Figure 7.2** Comparison of spread option extrinsic value as a function of strike. $\rho = 0.99$, OTM only

which might be encountered in a tolling deal with fuel-switching optionality. Here, the Kirk-type transformation would be $\tilde{S}_i = S_i + K$ for $i = 1, 2$ and a corresponding adjustment for volatilities. The resulting option (*sans* fixed strike) was considered (in very general form) in Section 5.1 via change-of-measure techniques.

A final approximation to be (briefly) considered is due to Bjerksund and Stensland (2006). The main idea here is to approximate the nonlinear (in terms of log prices) exercise boundary by an effectively linear region, over which the resulting option can be readily evaluated. That is, the payoff $(e^{z_2} - e^{z_1} - K)^+$ is replaced by

$$(e^{z_2} - e^{z_1} - K)1(z_2 > az_1 + b) \tag{7.19}$$

Note that this essentially represents valuation under a suboptimal exercise policy.[9] Thus, tight (hopefully) lower bounds can be obtained by choosing the parameters $a$ and $b$ to maximize the expectation of (7.19). The natural question is, how tight? Bjerksund and Stensland report excellent lower bounds for levels of ratio volatility typically encountered in equity markets. Our experience with volatilities more characteristic of energy markets suggest good, but not overwhelming, results. We do not propose to undertake a full critique/analysis here. We merely wish to note that the approach is very interesting, and can be readily extended to higher dimensions. The important point to note is that the resulting expectation of payoffs such as (7.19) can again be handled well by change-of-measure techniques, as we will see.

## 7.2 Conditional expectation as a representation of value

As we have stressed throughout, the conditional expectation of the terminal payoff of some structured product only has meaning as the value of this product to the extent that it is related to some portfolio constructed around that product. Typically this will involve some sort of (dynamic) hedge in the underlying commodities. The classic example is the BS paradigm, where the value of an (European) option on some (lognormal) asset is given by the expectation of the terminal payoff under a (unique, in this case) martingale measure. Of course, in the given context this value *means* that the structure in question can be perfectly replicated by a particular hedging strategy:

$$\Pi = (S_T - K)^+ - V_t - \int_t^T \Delta_s dS_s = 0 \tag{7.20}$$

*if* the value *and* hedges are taken to be the corresponding BS values. Now, we have gone into great detail elsewhere (principally, Chapter 3) regarding the sense in which we can continue to write $V_t = E_t^Q(S_T - K)^+$ for an appropriate measure $Q$, stressing the critical connection of this value process to the hedging process $\Delta_t$ and the overall construction of a portfolio process that creates exposure to a particular entity (the value drivers). We do not intend any further review here, except to remind the reader that our objective in evaluating such conditional expectations is never as a free-floating calculation, but rather as a means to a specific end, namely producing good, robust valuations of actual structured products. Still, the actual calculations are by no means trivial (the basic formulation in (7.20) becomes all the more challenging with the introduction of multiple assets, operational constraints, non-Gaussian dynamics, and so on), so the remainder of our journey here will focus on numerical techniques appropriate for valuation problems in energy markets.

## 7.3 Interpolation and basis function expansions

We have seen in Chapter 5 how change-of-measure techniques can greatly ameliorate many of the computational challenges that arise in pricing problems, even vanilla ones. This facilitation largely takes place through dimensional reduction, effectively "removing" one asset from consideration by treating it as a numeraire (roughly speaking, the units in which the other assets are expressed). Unfortunately, in most applications it is not possible to fully exploit this reduction, primarily due to the existence of certain costs, both variable and fixed. (In truth what this means is that there is in fact an extra dimension to any pricing problem, specifically a bond-type asset.) Simple examples include variable operation and maintenance (VOM)

costs in tolling and commodity charges in gas transport and storage. The basic payoff in such cases can generically be written as

$$(S_T^2 - S_T^1 - K)^+$$

with the associated pricing problem becoming the evaluation of conditional expectations (under an appropriate measure) of the following form:

$$E_t^Q(S_T^2 - S_T^1 - K)^+$$

We have discussed in Section 7.1 various approximation methods (*e.g.*, Kirk) that allow change–of-measure techniques to be employed. While some of these techniques are surprisingly effective, it is of course important to ascertain just how effective they are, and to have methods applicable outside the range of the approximations' domain of viability. In this section we will explore some approaches, broadly categorized as interpolation based, for systematically decomposing such problems.

### 7.3.1 Pearson and related approaches

*7.3.1.1 Conditioning and dimension reduction*

We consider a technique originally due to Pearson (1995). The underlying stochastic process is jointly lognormal, and it will prove convenient to work in terms of log prices. We take the terminal payoff to be

$$(e^{\mu_2+\sigma_2 z_2} - e^{\mu_1+\sigma_1 z_1} - K)^+ \tag{7.21}$$

For further convenience we have absorbed factors of square root of time to maturity in the volatilities.[10] The joint density is given by

$$\Pr(z_1, z_2) = \frac{1}{\sqrt{(2\pi)^2\rho_s^2}} \exp\left(-\frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{2\rho_s^2}\right) \tag{7.22}$$

where $\rho$ is the correlation (between the asset returns) and $\rho_s \equiv \sqrt{1-\rho^2}$. The density in (7.22) can be rewritten (either by simple algebra or from standard results for conditional normals) as

$$\Pr(z_1, z_2) = \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \frac{1}{\sqrt{2\pi\rho_s^2}} \exp(-(z_2 - \rho z_1)^2/2\rho_s^2) \tag{7.23}$$

This fact proves critical to the method, for the expectation of the payoff (7.21) can be written

$$E(e^{\mu_2+\sigma_2 z_2} - e^{\mu_1+\sigma_1 z_1} - K)^+ = \int_{-\infty}^{\infty} dz_1 \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \int_{-\infty}^{\infty} dz_2 (e^{\mu_2+\sigma_2 z_2} - e^{\mu_1+\sigma_1 z_1} - K)^+ \Pr(z_2|z_1) \quad (7.24)$$

Now, making the substitution $z_2 = \rho_s \xi + \rho z_1$ in (7.24), the "inner" integral can be written as

$$\int_{-\infty}^{\infty} d\xi (e^{\mu_2+\sigma_2\rho z_1+\sigma_2\rho_s\xi} - e^{\mu_1+\sigma_1 z_1} - K)^+ \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} = e^{\mu_2+\sigma_2\rho z_1+\frac{1}{2}\sigma_2^2\rho_s^2} N(d+\sigma_2\rho_s) - (e^{\mu_1+\sigma_1 z_1} + K)N(d) \quad (7.25)$$

with $d = \frac{\mu_2+\sigma_2\rho z_1 - \log(e^{\mu_1+\sigma_1 z_1}+K)}{\sigma_2\rho_s}$. Consequently, (7.24) becomes

$$\begin{aligned} E(e^{\mu_2+\sigma_2 z_2} - e^{\mu_1+\sigma_1 z_1} - K)^+ &= \int_{-\infty}^{\infty} dz_1 \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} e^{-\mu_2+\sigma_2\rho z_1+\frac{1}{2}\sigma_2^2\rho_s^2} N(d+\sigma_2\rho_s) \\ &\quad - \int_{-\infty}^{\infty} dz_1 \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} e^{-\mu_1+\sigma_1 z_1} N(d) \\ &\quad - K \int_{-\infty}^{\infty} dz_1 \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} N(d) \end{aligned} \quad (7.26)$$

*7.3.1.2 Evaluation via interpolation*

Now, the first thing to note is that the two-dimensional integration in (7.24) has been reduced to a one-dimensional integration.[11] This in itself is quite beneficial, as there are a host of well-known, effective techniques available for quadrature in one dimension (see, *e.g.*, Press *et al.* [2007]). The approach adopted by Pearson was to interpolate the normal cumulative distribution functions (CDF) in (7.26) by a piecewise exponential-affine approximation along some discrete grid along the real $z_1$–axis. That is, a set of grid points

$$z_1^i = (i - N/2)h, \; i = 0, \ldots N \quad (7.27)$$

is introduced (so $\pm Nh/2$ are very large in absolute value), and between neighboring grid points the normal CDF is approximated by[12]

$$N(d(z_1)) \approx \alpha_i e^{z_1} + \beta_i, z_1^i < z_1 < z_1^{i+1} \quad (7.28)$$

This is achieved by matching end-point values:

$$\begin{aligned}\alpha_i &= \frac{N(d(z_1^{i+1})) - N(d(z_1^i))}{e^{z_1^{i+1}} - e^{z_1^i}} \\ \beta_i &= N(d(z_1^i)) - \alpha_i e^{z_1^i}\end{aligned} \tag{7.29}$$

For example, the third integral in (7.26) is approximated by

$$\sum_i \int_{z_1^i}^{z_1^{i+1}} dz_1 \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2}(\alpha_i e^{z_1} + \beta_i) \tag{7.30}$$

Similar expressions arise for the other integrals.[13] Thus, the problem becomes one of evaluating integrals of the form

$$\int_{z_1^i}^{z_1^{i+1}} dz_1 \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2}(\alpha_i e^{c+bz_1} + \beta_i) \tag{7.31}$$

The terms proportional to $\beta$ are plainly just evaluations of the standard normal CDF, *e.g.*, $N(z_1^{i+1}) - N(z_1^i)$. For the terms proportional to $\alpha$, completing the square in the exponent also yields a standard evaluation, *e.g.*, $N(z_1^{i+1} - b) - N(z_1^i - b)$. Note that the numerical error in this case is driven by the extent to which the normal CDFs in (7.26) are well approximated by piecewise affine exponentials in (7.28). Since this interpolation is a good approximation in this case (note that this method is easily tailored to adaptability depending on local steepness), and since the underlying interpolation and attendant standard normal CDF evaluations are quite easy to carry out (see Press *et al.* [2007] for a standard algorithm), it can be seen that Pearson provides an extremely useful method for numerically computing spread-option values. As we will see in the next subsection, the underlying idea here can be greatly generalized.

#### *7.3.1.3 Extension to higher dimensions*

As a brief example of this latter point, consider an option with payoff $(e^{z_3} - \min(e^{z_1}, e^{z_2}) - K)^+$. The general expectation of this payoff for nonzero fixed strike $K$ will require a three-dimensional integration. However, by factoring the underlying probability density via $\Pr(z_1, z_2, z_3) = \Pr(z_3|z_1, z_2)$ and using the fact that normality is retained under conditioning,[14] an "inner" integral (in $z_3$) can be formed that is analytically tractable for given values of $z_1$ and $z_2$ (recall the approaches developed in Chapter 5 for max/min options). Then, adopting a "bilinear" approximation of the form

$$\alpha_{ij} e^{z_1+z_2} + \beta_{ij} e^{z_1} + \gamma_{ij} e^{z_2} + \delta_{ij} \tag{7.32}$$

within a two-dimensional set of grid points (in the $z_1-$ and $z_2-$dimensions), we can use efficient algorithms for the two-dimensional normal CDF (again, to be presented in the subsection on quadrature), to obtain accurate valuations.

Note also that the payoff function in (7.21) can be written in terms of prices as

$$S_2 \cdot 1(S_2 - S_1 - K > 0) - S_1 \cdot 1(S_2 - S_1 - K > 0) - K \cdot 1(S_2 - S_1 - K > 0) \quad (7.33)$$

So, using change-of-measure results (and Euler's theorem), we see that the deltas are essentially binary option values under the appropriate numeraire. In practice they can simply be read off from the coefficients of the underlying price in the option value formula. Thus the Pearson ensemble in (7.26) automatically provides the deltas of the spread option, and with a bit of algebra *all* of the greeks (*e.g.*, gammas and vegas) can be obtained, *without* recourse to numerical differentiation (*e.g.*, finite difference).

In fact, we can see here a useful application of the change-of-measure techniques presented in Chapter 5. Equation (7.26) can be written as

$$\begin{aligned} & E(e^{\mu_2+\sigma_2 z_2} - e^{\mu_1+\sigma_1 z_1} - K)^+ \\ & \quad = E^n(e^{\mu_2+\sigma_2\rho z_1+\frac{1}{2}\sigma_2^2\rho_s^2} N(d+\sigma_2\rho_s)) \\ & \qquad - E^n(e^{\mu_1+\sigma_1 z_1} N(d)) - KE^n(N(d)) \end{aligned} \quad (7.34)$$

where $E^n$ denotes expectation wrt. a standard unit normal. Now, we have seen how exponential factors inside expectations lend themselves to convenient measure changes, *e.g.*,

$$Ee^{az} f(z) = Ee^{az} \cdot E^a f(z) \quad (7.35)$$

where under the measure change effected by the RN derivative $e^{az}/Ee^{az}$ a unit normal becomes distributed as $N(a,1)$ , *i.e.*, the mean is simply shifted. Thus, expectations of the form (7.35) are reducible to expectations wrt. a normal of general mean. In the context of Pearson's approach, if the function $f$ in (7.35) is approximated as

$$f(z) \approx \sum_i (\alpha_i e^z + \beta_i) 1(z_i < z < z_{i+1}) \quad (7.36)$$

for some set of grid points $z_i$. Hence evaluation of (7.36) entails evaluation of

$$\begin{aligned} & E^a \sum_i (\alpha_i e^z + \beta_i) 1(z_i < z < z_{i+1}) \\ & \quad = \sum_i \alpha_i E^{a+1} 1(z_i < z < z_{i+1}) + \sum_i \beta_i E^a 1(z_i < z < z_{i+1}) \end{aligned} \quad (7.37)$$

Now, any of these expectations can be readily evaluated in terms of the standard normal CDF:

$$E^a 1(z_i < z < z_{i+1}) = N(z_{i+1} - a) - N(z_i - a) \tag{7.38}$$

So again we see that if the approximation in (7.36) is good, then the resulting numerical algorithm will be quite effective.

#### *7.3.1.4 Spread options under affine processes*

As an extension of the result in (7.37), consider the problem of spread-option valuation when the underlying legs belong to the class of affine jump diffusions studied in Section 5.2. As we saw there, the characteristic function of the underlying process option pricing can be obtained from a system of ODEs (which can often be solved analytically). In turn, the characteristic function can be used to price options via quadrature. In fact, more basic (binary) structures with indicator payoffs whose values can be expressed as

$$E_t^Q 1(z_1(T) < \alpha_1, z_2(T) < \alpha_2) \tag{7.39}$$

are amenable to this technique. This fact can be used to price spread options under general (affine) processes by applying the above interpolation techniques. The basic valuation problem can be written as follows:

$$\begin{aligned} &E_t^Q (e^{z_2(T)} - e^{z_1(T)} - K)^+ \\ &\quad = E_t^Q e^{z_2(T)} \cdot E_t^{Q_2} 1(e^{z_2(T)} \\ &\qquad - e^{z_1(T)} - K > 0) - E_t^Q e^{z_1(T)} \cdot E_t^{Q_1} 1(e^{z_2(T)} - e^{z_1(T)} - K > 0) \\ &\qquad - K \cdot E_t^Q 1(e^{z_2(T)} - e^{z_1(T)} - K > 0) \end{aligned} \tag{7.40}$$

using the measure change $\frac{dQ_k}{dQ} = \frac{e^{z_k(T)}}{E_t^Q e^{z_k(T)}}$.

So, the first thing to note is the primary valuation problem reduces to the evaluation of the expectation of certain indicator functions. The problem, again, is that the underlying region of integration is nonlinear for nonzero strike $K$. We circumvent this problem by adopting a discretization in $z_1$ and a piecewise linear approximation to the exercise region in the $(z_1, z_2)$ plane. Specifically, we approximate the constituent expectations in (7.40) by the following ensemble:

$$\sum_i E_t^Q 1(z_1^i < z_1(T) < z_1^{i+1}) 1(z_2(T) > \alpha_i z_1(T) + \beta_i) \tag{7.41}$$

The issue amounts to the calculation of the joint characteristic function of $(z_1, \tilde{z}_2^i)$, where $\tilde{z}_2^i \equiv \alpha_i z_1 + \beta_i$. But this is straightforward, as we have

$$\begin{aligned} E_t^Q e^{i\phi_1 z_1(T) + i\phi_2 \tilde{z}_2(T)} &= E_t^Q e^{i(\phi_1 - \alpha\phi_2) z_1(T) + i\phi_2 z_2(T)} \\ &= f(\phi_1 - \alpha\phi_2, \phi_2) \end{aligned} \tag{7.42}$$

We will see more on expectations like these in Section 7.7.

*7.3.1.5 Further generalizations*

As a final note, using the law of iterated expectations, we have

$$\begin{aligned} E(e^{\mu_2+\sigma_2 z_2} - e^{\mu_1+\sigma_1 z_1} - K)^+ &= E\left[E((e^{\mu_2+\sigma_2 z_2} - e^{\mu_1+\sigma_1 z_1} - K)^+ | z_1)\right] \\ &= E\left[E\left(e^{\mu_2+\sigma_2 z_2} 1\left(z_2 > \frac{\log X - \mu_2}{\sigma_2}\right) \middle| z_1\right)\right] \\ &\quad - E\left[X \cdot E\left(1\left(z_2 > \frac{\log X - \mu_2}{\sigma_2}\right)\right) \middle| z_1\right)\Bigg] \end{aligned} \tag{7.43}$$

with $X = e^{\mu_1+\sigma_1 z_1} + K$. Now, in any case where the conditional expectation is known, the inner expectation can (at least in principle) be obtained as a function of $z_1$, at which point the interpolation-based approach described above (*e.g.*, in (7.37)) can be applied to the outer expectation. As we will see in the discussion of copulas in Chapter 8, there is a class of distributions for which this is (conceivably) possible, namely the so-called elliptical distributions (which include, as a special case, joint normality). This class is in fact reasonably rich and enjoys a fairly wide range of applications.[15]

The basic idea underlying Pearson's approach can thus be seen to be quite general and powerful. In fact, the essence of the idea can be greatly expanded to include features such as early exercise optionality, which in general present considerable computational challenges. Before delving into that problem in detail, we outline a highly efficient algorithm for interpolation-based valuation.

### 7.3.2 The grid model[16]

*7.3.2.1 Main idea*

We will discuss here a method originally developed by Eydeland (1994, 1996); the implementation considered here is synopsized in Eydeland and Mahoney (2002). Consider the following expectation:

$$V(z,t) = E_t U(z_t) = \int_{-\infty}^{\infty} U(z_t) \Pr(z_T | z) dz \tag{7.44}$$

(The function $U$ may be thought of as an option payoff function, with $T$ denoting expiry.) For the time being we will assume the random variable $z$ to be (conditionally) normal, so that (7.44) can be written

$$V(z,t) = \int_{-\infty}^{\infty} U(z_T) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(z_T-\mu)^2/2\sigma^2} dz \tag{7.45}$$

To specify some structure, we take the (conditional) variance to be state independent (we will subsume any explicit time dependence in the parameterization) and take the (conditional) mean to be affine in the state variable $z$: $\mu = az + b$. For example, for GBM we would have

$$\mu = z + \mu' - \tfrac{1}{2}\sigma'^2, \quad \sigma = \sigma'$$

while for a mean-reverting process we would have

$$\mu = \theta(1 - e^{-\kappa\tau}) + ze^{-\kappa\tau}, \quad \sigma = \sigma' \sqrt{\frac{1 - \exp(-2\kappa\tau)}{2\kappa\tau}}$$

using standard notation.[17]

Now, in contrast to the Pearson approach, we introduce *two* sets of grids/discretizations, one for the current time $t$ and one for the terminal time $T$. Specifically, we take

$$\begin{aligned} z^i &= \mu_t + (i - N/2)h_t \\ z_T^i &= \mu_T + (i - N/2)h_T \end{aligned} \tag{7.46}$$

for $i = 0, \ldots, N$ (compare with (7.27)). Note that in general the (discretization) step size will differ between the two times. The "means" about which the grids are centered may also differ. We now proceed as in Pearson. The payoff function is interpolated between the time-$T$ grid points quasi-linearly as in (7.28), so evaluating (7.45) at $z^i$ gives (we drop explicit dependence on the current time)

$$V(z^i) = \sum_j \int_{z_T^j}^{z_T^{j+1}} (\alpha_j e^{z_T} + \beta_j) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(z_T - az^i - b)^2/2\sigma^2} dz \tag{7.47}$$

Upon making the substitution $z_T = \sigma\xi + az^i + b$, (7.47) becomes

$$V(z^i) = \sum_j \int_{(z_T^j - az^i - b)/\sigma}^{(z_T^{j+1} - az^i - b)/\sigma} (\alpha_j e^{az^i + b + \sigma\xi} + \beta_j) \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} d\xi \tag{7.48}$$

Of course, integrals of this form can be evaluated easily, as we have seen with Pearson. However, we will now see that there is a rationale for the discretization chosen in (7.46). Note that the lower (say) limit of integration in (7.48) can be written (apart from the divisor $\sigma$) as

$$\mu_T + (j - N/2)h_T - a\mu_t - (i - N/2)ah_t - b \tag{7.49}$$

So, if we make the choice

$$h_T = ah_t \tag{7.50}$$

then (7.49) depends only on the *difference* between the indices $i$ and $j$. For convenience we can take $\mu_T = a\mu_t + b$. (For further convenience we could take $\mu_t = z$, so that the center of the time $t$ grid corresponds to the current log price.)

*7.3.2.2 Computational feasibility*

Consequently (7.48) becomes

$$\begin{aligned} V(z^i) = e^{z_T^i + \sigma^2/2} \sum_j \left( N\left( \frac{j - i + 1}{\sigma} h_T - \sigma \right) - N\left( \frac{j - i}{\sigma} h_T - \sigma \right) \right) \alpha_j \\ + \sum_j \left( N\left( \frac{j - i + 1}{\sigma} h_T \right) - N\left( \frac{j - i}{\sigma} h_T \right) \right) \beta_j \end{aligned} \tag{7.51}$$

Thus, the calculations in (7.51) entail a pair of matrix multiplications of the form

$$\sum_j T_{i-j} x_j \tag{7.52}$$

The matrix $T$ has a special form: it is constant along subdiagonals. As such, matrix multiplications such as (3.64) can be carried out efficiently via the Fast Fourier Transform (FFT). We have already encountered many fruitful applications of Fourier analysis and general transform methods in Chapter 5. Our intent here is not to provide a full-blown treatment of numerical methods that are well covered in standard texts (*e.g.*, Dahlquist and Björck [1974] or Press *et al.* [2007]). We will assume that methods such as the FFT are sufficiently familiar to the reader that we may merely emphasize the relevant points.

As is well known, the FFT permits matrix multiplications that would ordinarily require a computationally costly $O(N^2)$ operation count (with $N$ representing the size of the matrix in question) to be done in a much more manageable $O(N \log N)$ operations. In particular, Toeplitz matrix multiplications such as (3.64) can be performed very quickly.[18,19] Thus, valuation of the expectation in (7.44) can be efficiently obtained for a *vector* of initial (log) prices, not just a scalar as in Pearson. The significance of this point will be made clear in the next section when we begin to consider early exercise-type options (*e.g.*, American options), when it is necessary to have valuations across a range of underlying (log) prices. Note again that, like Pearson, the main driver of the discretization error is the accuracy with which the piecewise affine function used in (7.28) approximates the payoff function in (7.44). We would expect this approximation to be good in a wide range of applications. In fact, the grid model outlined here will be exact for a European call, so long as the (log) strike is taken as one of the discretization points. Another advantage of this method that should be emphasized is that, by incorporating the asymptotic behavior of the payoff function (usually known in many applications), problems created by the truncation of the grid at finite end points will be considerably muted, as these correction terms are easily obtained in terms of standard normal CDFs and added on to (7.51).

In other words, so long as there is no error incurred in representing the transition density in the time step between $t$ and $T$, interpolation error is the *only* source of error in the calculation. Note that the problem being solved here is essentially an integration of a PDE of the form[20]

$$V_t + (\chi - \kappa z) V_z + \tfrac{1}{2}\sigma^2 V_{zz} = 0 \tag{7.53}$$

with terminal condition $V(z, T) = U(z)$. Now, there are standard approaches for integrating such PDEs numerically (such as Crank-Nicolson [finite difference]; in the context of financial applications, see Wilmott *et al.* [1997]). The main issue of interest here is the fact that, generally speaking, such numerical evaluation cannot, for stability reasons, treat the temporal and spatial discretizations independently. One of the big advantages of the grid model is that these two aspects of the problem (space and time, so to speak), *can* be treated independently. There is no need to constrain the time discretization to obtain a desired spatial discretization.

#### *7.3.2.3 Extraction of greeks*

Since the value function is obtained across a grid of points (centered about the current value of the underlying log price), greeks such as deltas and gammas can easily be obtained numerically via finite difference. Generally speaking it is preferable to avoid recourse to finite differences, but since the necessary inputs are automatically produced by the grid model, differencing is certainly a natural choice in practice. However, there is also an interpretation similar to the one appealed to with Pearson,

as in (7.33) via Euler. Using the relation between the grids in (7.46) and (7.50), the factor multiplying the terms involving $\alpha$ in (7.51) becomes

$$e^{z_T^i+\sigma^2/2} = e^{az^i+b+\sigma^2/2} \tag{7.54}$$

and we anticipate that the $z$–derivative (from which the actual delta is obtained via $S\partial_S = \partial_z$ ) will be given by

$$ae^{z_T^i+\sigma^2/2}\sum_j \left(N\left(\frac{j-i+1}{\sigma}h_T-\sigma\right)-N\left(\frac{j-i}{\sigma}h_T-\sigma\right)\right)\alpha_j \tag{7.55}$$

This is in fact correct, as we can see from the following formulation (which we will revisit when we consider simulation-based approaches):

$$\frac{\partial V}{\partial z} = a\int_{-\infty}^{\infty} U(z_T)\frac{(z_T-\mu)}{\sigma^2}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(z_T-\mu)^2/2\sigma^2}dz \tag{7.56}$$

Now, if we perform the usual interpolation, (7.56) becomes

$$\left.\frac{\partial V}{\partial z}\right|_{z^i} = a\sum_j \int_{(z_T^j-az^i-b)/\sigma}^{(z_T^{j+1}-az^i-b)/\sigma} (\alpha_j e^{az^i+b+\sigma\xi}+\beta_j)\frac{\xi}{\sigma}\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}d\xi \tag{7.57}$$

which further reduces to

$$\begin{aligned}
\left.\frac{\partial V}{\partial z}\right|_{z^i} &= ae^{z_T^i+\sigma^2/2}\sum_j \left(N\left(\frac{j-i+1}{\sigma}h_T-\sigma\right)-N\left(\frac{j-i}{\sigma}h_T-\sigma\right)\right)\alpha_j \\
&\quad \frac{a}{z}e^{z_T^i+\sigma^2/2}\sum_j \alpha_j\left(\varphi\left(\frac{j-i+1}{\sigma}h_T-\sigma\right)-\varphi\left(\frac{j-i}{\sigma}h_T-\sigma\right)\right) \\
&\quad + \frac{a}{\sigma}\sum_j \beta_j\left(\varphi\left(\frac{j-i+1}{\sigma}h_T\right)-\varphi\left(\frac{j-i}{\sigma}h_T\right)\right) \\
&= ae^{z_T^i+\sigma^2/2}\sum_j \left(N\left(\frac{j-i+1}{\sigma}h_T-\sigma\right)-N\left(\frac{j-i}{\sigma}h_T-\sigma\right)\right)\alpha_j \\
&\quad + \frac{a}{\sigma}\sum_j \left(U(z_T^{j+1})\varphi\left(\frac{j-i+1}{\sigma}h_T\right)-U(z_T^j)\varphi\left(\frac{j-i}{\sigma}h_T\right)\right)
\end{aligned} \tag{7.58}$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is the standard normal density, and in the last equation in (7.58) we have used the fact that at each grid point $U(z) = \alpha e^z + \beta$.[21] As this latter telescoping sum is only evaluated at the extremities of the grid, it vanishes and result (7.55) is established. (Note that homogeneity/Euler does not apply, in general, to the class of problems to be considered in the remainder of the chapter, as the presence of mean reversion means the problem is *not* scale-independent, so to speak.)

#### *7.3.2.4 Higher dimensions*

Finally, as with Pearson, extensions to higher dimensions are possible. However, it turns out there is a twist. Consider the following generalizations of (7.44) and (7.45):

$$V(x,y,t) = E_t U(x_T, y_T) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_T, y_T)\Pr(x_T, y_T|x,y)\,dx_T\,dy_T \tag{7.59}$$

and

$$V(x,y,t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_T, y_T)\frac{1}{\sqrt{(2\pi)^2\sigma_x^2\sigma_y^2(1-\rho^2)}} e^{-\frac{1}{2}\left(\frac{(x_T-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x_T-\mu_x)(y_T-\mu_y)}{\sigma_x\sigma_y} + \frac{(y_T-\mu_y)^2}{\sigma_y^2}\right)} dx_T\,dy_T \tag{7.60}$$

Now, as in the one-dimensional case we confine attention to (Gaussian) affine cases where the conditional covariance structure is state-independent and the conditional means take the form

$$\begin{aligned} \mu_x &= a_{xx}x + a_{xy}y + b_x \\ \mu_y &= a_{yx}x + a_{yy}y + b_y \end{aligned} \tag{7.61}$$

An example would be a mean-reverting process with a stochastic drift:

$$\begin{aligned} dz &= \kappa(\theta - z)dt + \sigma_z dw_z \\ d\theta &= \mu\,dt + \sigma_\theta\,dw_\theta \end{aligned} \tag{7.62}$$

in which case $a_{xx} = e^{-\kappa\tau}$, $a_{xy} = 1 - e^{-\kappa\tau}$, $a_{yx} = 0$, $a_{yy} = 1$.

We can now see where the aforementioned twist arises. Continuing with the generalization, we introduce (spatial) grids at each time:

$$\begin{aligned} x^i &= \mu_t^x + (i - N_x/2)h_t^x,\ y^j = \mu_t^y + (j - N_y/2)h_t^y \\ x_T^i &= \mu_T^x + (i - N_x/2)h_T^x,\ y_T^j = \mu_T^y + (j - N_y/2)h_T^y \end{aligned} \tag{7.63}$$

Again proceeding similarly to the one-dimensional case, we introduce an interpolation of the ("bilinear") form (7.32). Then (allowing for shifts arising from suitable changes of measure), we find that we are led to consider integrals of the form [22]

$$\sum_{m,n} \int_{\xi_{ij}^m}^{\xi_{ij}^{m+1}} \int_{\zeta_{ij}^n}^{\zeta_{ij}^{n+1}} \frac{1}{\sqrt{(2\pi)^2(1-\rho^2)}} e^{-(\xi^2 - 2\rho\xi\zeta + \zeta^2)/2(1-\rho^2)} d\xi\, d\zeta \tag{7.64}$$

where, for example,

$$\xi_{ij}^m = \frac{\mu_T^x + (m - N_x/2)h_T^x - a_{xx}(\mu_i^x + (i - N_x/2)h_t^x) - a_{xy}(\mu_t^y + (j - N_y/2)h_t^y)}{\sigma_x} \tag{7.65}$$

Now, our objective is to get the underlying interpolation in (7.64) into a form similar to (3.64) so that the FFT (in two dimensions) can be applied. That is, we need the underlying matrix multiplication to be of the form[23]

$$\sum_j T_{i-m,j-n} x_{mn} \tag{7.66}$$

Unfortunately, due to its two-dimensional structure, (7.65) does not permit (7.64) to be expressed in the form (7.66). In the one-dimensional case, suitably adjusting the grid size between times as in (7.50) was sufficient to yield the Toeplitz structure (3.64) in the valuation algorithm. However, note that if the two-dimensional price/state dynamics were such that the means "decoupled" (in the sense that a given asset's [instantaneous] mean depends only on it and not the other assets[24]), then the same approach used in the one-dimensional case would be applicable here.

This observation leads us to introduce an (evenly spaced) *auxiliary* grid given by

$$\begin{aligned} \tilde{x}^i &= \tilde{\mu}_t^x + (i - N_x/2)\tilde{h}_t^x \\ \tilde{y}^j &= \tilde{\mu}_t^y + (i - N_y/2)\tilde{h}_t^y \end{aligned} \tag{7.67}$$

and related to the time–$t$ grid via

$$\begin{aligned} \tilde{x} &= a_{xx}x + a_{xy}y + b_x \\ \tilde{y} &= a_{yx}x + a_{yy}y + b_y \end{aligned} \tag{7.68}$$

In matrix form, $z = a^{-1}(\tilde{z} - b)$. Consequently, this relation induces an irregularly spaced time–$t$ grid,[25] and by also introducing an auxiliary value function the approximation to the basic valuation expression in (7.59) and (7.60) becomes

$$\begin{aligned} \tilde{V}(\tilde{x}_i, \tilde{y}_j, t) &\equiv V(x_{ij}, y_{ij}, t) \\ &= \sum_{m,n} \int_{\mu_T^x + (m-N_x/2)h_T^x}^{\mu_T^x + (m+1-N_x/2)h_T^x} \int_{\mu_T^y + (n-N_y/2)h_T^y}^{\mu_T^y + (n+1-N_y/2)h_T^y} U_{mn}(x_T, y_T)\varphi_2 \\ &\quad \left( \frac{x_T - \mu_x}{\sigma_x}, \frac{y_T - \mu_y}{\sigma_y}; \rho \right) dx_T dy_T \end{aligned} \tag{7.69}$$

with $U_{mn}$ denoting an interpolant of the form (7.32), and where $\varphi_2$ is the standard two-dimensional Gaussian density. Now, upon making the obvious substation $(z \to \mu + \sigma z)$), by construction (see (7.67) and (7.61)) the limits of integration (allowing for the usual shifts arising from measure changes) will have the form (upon using the inverse form of (7.68))

$$\begin{aligned} &\frac{\mu_T^x + (m - N_x/2)h_T^x - a_{xx}x_{ij} - a_{xy}y_{ij} - b_x}{\sigma_x} \\ &\quad = \frac{\mu_T^x + (m - N_x/2)h_T^x - \tilde{\mu}^x - (i - N_x/2)\tilde{h}^x}{\sigma_x} \end{aligned} \tag{7.70}$$

with similar results for the $y$–dimension limits. Thus, if we take the auxiliary time–$t$ grid to be *identical* to the time–$T$ grid (*i.e.*, $h^T = \tilde{h}$) then the various constituent matrix multiplications inherent in (7.69) take the form (7.69). As stated, this is effectively a two-dimensional convolution,[26] for which the FFT can be applied, yielding a huge computational savings, namely $O(N^2 \log N)$ as opposed to $O(N^4)$.

Of course, a disadvantage in this case is the fact that we often need the time–$t$ value function at a regularly shaped grid such as in (7.63) and not the irregular grid induced by the transformation in (7.68). This is the case for the kinds of early exercise and control problems to be considered in the next section, when the basic valuation procedure in (7.59) must be carried out recursively. What must be done in this case is to perform *another* interpolation, between the irregular grid where the value function is known, and the desired regular grid. This is not a particularly taxing task, as the regular grid can be mapped into the new coordinates

represented by (7.68), at which stage every mapped point can be identified as being inside a rectangle, and then interpolated via (7.32). This is not a particularly excessive computational burden (either in terms of pure cost or additional algorithmic architecture), although it does introduce an additional source of error. However, this must be balanced by the significant gains in speed via the FFT.

We emphasize these points because there are oftentimes a need to introduce auxiliary grids. For example, in the 1D case this is not necessary because we can still fully exploit the efficiencies of the FFT by suitably adjusting the relationship between grids via (7.50). However, we will see in the next section that this adjustment might not always be the most advisable approach, as there are cases where the grid size adjustment across time steps can manifest itself in numerical overflow/underflow in certain situations. The question of time stepping is an important one, and the real power of the grid model comes in applications involving early exercise or control-type features, which are ubiquitous in energy markets. We now consider this very rich class of problems, which were already introduced in Section 3.3; the reader may want to review that discussion.

#### *7.3.2.5 Early exercise options*

Problems involving timing decisions and their optimization are very commonplace in energy markets. The basic structure of such problems is that a decision must be made now that will alter the future state of the system under consideration, and hence the value that can be extracted from that system is determined by the particular decision policies. More accurately, the value is characterized by the optimization of such policies. The basic optionality usually encountered in energy markets is operational in nature. Typical examples are storage and tolling. In storage, an injection (say) changes the state of the facility by increasing the inventory available for future use (*e.g.*, future withdrawals and sale into the market). In tolling, a decision to turn the plant on may entail operation during off-peak periods during which margins are low, in order to collect high margins during on-peak periods and avoid start-up costs. We will see detailed representations and analysis of both kinds of deals later. The main point here is that these kinds of deals entail *optimal stopping times*.

To quickly review, stopping times are random times that are measureable with respect to the filtration of some stochastic process.[27] What this means practically is that a decision depending on a stopping time can be made on the basis of information available at the present time. An example might be the first time a price crosses a certain price from above. Again, we refer the reader to sources such as Etheridge (2002) or Shreve (2004a) for greater exposition. The main result that concerns us here is that many of these problems of interest can be crafted (in a general sense) as an optimal stopping-time problem. That is, they take the form

$$V(S,t) = \sup_{\tau} E_t^Q F(S_\tau) \tag{7.71}$$

where $\tau$ represents a stopping time, $F$ is some payoff function, and "sup" denotes supremum. The expression in (7.71) says that the value of the early exercise structure is an expectation of the payoff function, optimized with respect to measureable exercise policies. Not surprisingly, such problems present considerable computational challenges, as determination of the optimal exercise/control/decision policy is usually extremely difficult (even apart from evaluation of the associated expectation). This is true for generic financial problems (there is no known analytical solution of the standard, American option problem), and the challenges are compounded even further in physical energy structures where various operational constraints must be accounted for. Effective means of approaching these problems will be our focus for the bulk of this chapter.

We will necessarily limit our concern to discrete-time versions of (7.71). First consider an American option.[28] The standard approach to the valuation is to note that, at any point in time, one can compare the value from exercising the option immediately, to the expected value of the option given that it is held (*i.e.*, not exercised). The problem is solved recursively. First, at expiration (say, time $T$), there is no early exercise value obviously and the value is simply intrinsic:

$$V(S,T) = (S-K)^+ \tag{7.72}$$

At previous time steps (prior to maturity), the value is the greater of either intrinsic or the expected value of the next-step value:

$$V_n(S) = \max((S-K)^+, e^{-r\Delta t} E_n^Q V_{n+1}) \tag{7.73}$$

where we adopt the notation $V_n(S) \equiv V(S, t+n\Delta t)$ and $\Delta t$ is the time-step size. For $N$ time steps, $n=0$ corresponds to the initial time, and $n=N$ corresponds to the terminal time, where the value function is known (eq. (7.72)). It is important to understand that the conditional expectation in (7.73) (commonly referred to as the continuation value) requires, for its valuation, knowledge of the value function along the time grid at the *next* time step. Here we see the essence of the computational challenge posed by early exercise problems. Because the expectation in (7.73) is essentially a projection of the next-time value function onto the current-time state space,[29] this calculation must be carried out along *each* point of the current-time grid. This is in general a computationally intensive task. Indeed, it renders simulation-based approaches very hard to apply (we will discuss this issue in great detail later). In general, there is a "computation-within-a-computation" aspect to the problem.

#### *7.3.2.6 Comparison with trees*

A common approach to the problem is the binomial tree model (see, *e.g.*, Shreve [2004a]), where the underlying is modeled as taking one of two possible future

states, say, up or down. Since the relation along a time step is specified by the up/down probabilities of the representation of the underlying, we have[30]

$$E_n^Q V_{n+1} = pV_{n+1}(uS) + qV_{n+1}(dS) \tag{7.74}$$

Since, by construction of the tree, a node corresponding to a particular value of the underlying is connected to the nodes corresponding to up/down moves on the next-time tree, the calculation in (7.74) can be readily carried out. In general, however, we would like to be able to specify a bit more detail about the underlying transition probability. This is precisely where the grid model comes in. It should be clear from the analysis leading to (7.51) that the grid model will produce the expectation of the next-time value function along the entire current-time grid, as required to evaluate (7.73). Furthermore, it does so quite efficiently, exploiting the $O(N \log N)$ speed of the FFT. Eydeland and Mahoney (2003) present a battery of tests comparing the grid method to standard binomial trees, demonstrating both the algorithmic efficiency and superior convergence properties. We reproduce a very small subset here.

It is not hard to see that the binomial tree involves $N(N+1)/2$ multiplications, while the grid model requires $O(T \cdot N \log N)$ operations, where $T$ is the number of time steps. Thus, in terms of operation count, there can in fact be a trade-off between the two methods when the number of times steps is high enough (say, daily time steps for a year). On the other hand, precisely one of the strengths of the grid is the ability to separate time discretization from spatial discretization, so it can be well tailored to problems where the early-exercise optionality is on discrete time scales (*e.g.*, Bermudan options). We will emphasize this fact by comparing results from the grid model and binomial tree for the case of a European option, where a single time step is sufficient (for the grid model). We show the runtimes in Table 7.1.

Of course, the actual comparative runtimes do not conform to their theoretical ratios, reflecting various overhead issues. However, the relevant pattern is clear:

**Table 7.1** Runtimes, grid vs. binomial. 110% OTM, one year maturity, 50% volatility.

| OTM European Option | | |
|---|---|---|
| | Runtimes (cs) | |
| N | Grid | Binomial |
| 128 | 0.026 | 0.025 |
| 256 | 0.063 | 0.062 |
| 512 | 0.074 | 0.148 |
| 1024 | 0.132 | 0.462 |
| 2048 | 0.238 | 1.694 |
| 4096 | 0.526 | 7.373 |
| 8192 | 0.983 | 30.541 |

binomial runtime grows quadratically with spatial resolution, while grid runtime grows (approximately) linearly. Figure 7.3 shows the convergence rates of the two methods, again for increasing spatial resolution. The grid model is clearly superior in this case, reflecting the fact that error arises *solely* from the spatial discretization (and none from temporal discretization).

*7.3.2.7 Example: gas storage*

A concrete energy-related illustration is in order. We consider an example from gas storage, which should also serve as a preview of the dynamic programming methods to be considered later. We assume that, under the pricing measure $Q$, gas spot prices are given by

$$G = e^{g'+\chi} \tag{7.75}$$

where $\chi$ is a (deterministic) seasonality factor and $g'$ is a mean-reverting process:

$$dg' = \kappa(\theta - g')dt + \sigma\, dw \tag{7.76}$$

so that the log-price $g$ satisfies[31]

$$dg = (\kappa(\theta - g) + \dot{\chi} + \kappa\chi)dt + \sigma\, dw \tag{7.77}$$

Note that the dynamics of the forward price

$$F_{t,T} = E_t^Q e^{g_T} \tag{7.78}$$

are given by

$$\frac{dF_{t,T}}{F_{t,T}} = \sigma\, e^{-\kappa(T-t)}\, dw \tag{7.79}$$

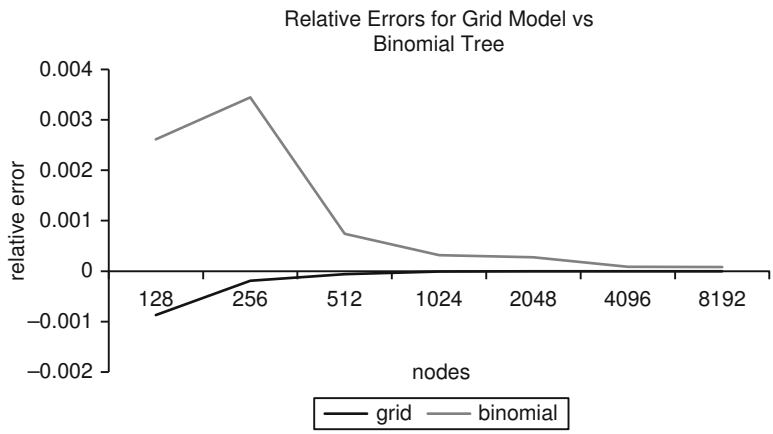



**Figure 7.3** Convergence rates, grid vs. binomial

(Not surprisingly, the forward price is a $Q$–martingale.) Explicitly, we have the following relation between spot and forward:

$$F_{t,T} = \exp\left( g e^{-\kappa(T-t)} + \theta(1 - e^{-\kappa(T-t)}) + \chi_T - \chi_t e^{-\kappa(T-t)} + \frac{\sigma^2}{4\kappa}(1 - e^{-2\kappa(T-t)}) \right) \tag{7.80}$$

These results can be obtained by various means, but the characteristic functions methods developed in Chapter 5 are particularly convenient and should prove a useful exercise to the reader.

Now, the valuation of storage depends not only on the current gas price, but also on the current level of inventory (*i.e.*, the state of the system, to use the language of dynamic programming). Depending on both price and inventory, the optimal decision may be to inject (increase inventory by buying gas from the market) or to withdraw (decrease inventory by selling gas into the market). Algorithmically, we must discretize *both* price and inventory, making the problem inherently two-dimensional. The basic valuation problem can be expressed as

$$V_n(G_i; Q_j) = \max_q \{ -qP(G_i; q) + e^{-rdt} E_n V_{n+1}(S; Q_j + q) \} \tag{7.81}$$

where the "payoff" $P$ is given by

$$P(G; q) = \begin{cases} G/(1 - f_{inj}) + k_{inj}, & q > 0 (\text{inj}) \\ G/(1 - f_{wdr}) - k_{wdr}, & q > 0 (\text{wdr}) \end{cases} \tag{7.82}$$

where $f$ and $k$ denote fuel losses[32] and commodity charges, respectively. A typical constraint on the optimization in (7.82) is of the form

$$q_{\min} \leq q \leq q_{\max} \tag{7.83}$$

where $q_{\min} < 0$ is the maximum withdrawal rate and $q_{\max} > 0$ is the maximum injection rate. These constraints will obviously be binding (in the plain language sense of not being able to inject/withdraw at the fastest possible rate) when the inventory is large enough or small enough, but also when ratchets are present, so that the maximum allowable flow rates are explicitly inventory dependent. Typically, there are also constraints on the terminal inventory (*e.g.*, the facility must be empty at the end of the deal). Note that it is often more convenient to write (7.81) in terms of the state (inventory) rather than the control:

$$V_n(G_i; Q_j) = \max_Q \{ -(Q - Q_j) P(G_j; Q - Q_j) + e^{-rdt} E_n V_{n+1}(S; Q) \} \tag{7.84}$$

Now, for a given inventory level, the grid model can be used to efficiently evaluate the conditional expectation in (7.81) or (7.84). Thus, if we discretize the allowable state space (inventory level) we can determine the optimal control conditional

on information at time $n$.[33] An issue that arises is that, in some applications, the mean-reversion rate $\kappa$ in (7.77) might be sufficiently large that numerical overflows/underflows can occur when the grid-matching condition (7.50) is imposed. In this, the relation between grid sizes is given by

$$h_{n+1} = e^{-\kappa \Delta t} h_n \tag{7.85}$$

where $\Delta t$ is the temporal step size, typically one day (*i.e.*, 1/365). For example, Bjerksund *et al.* (2008) consider a case with a very high mean reversion rate of 18.25. (We will have more to say about this particular work later.) Thus, the grid step size gets reduced by about 5% for each time step. This means that the grid step size at the initial time step is insignificant compared to the grid step size at the terminal time step. (For a one-year deal the ratio would be about $10^{-8}$). Equivalently, the terminal time grid step size is enormous compared to the initial grid step size. Ideally, the set of grid points represents a sampling of the underlying distribution across most of its range, say, five standard deviations. A tiny step size means this sampling is severely truncated, whereas a huge step size can lead to numerical overflow from too big a range.

Fortunately we can correct for this effect and still retain the power of the underlying method. To do this, we introduce a "parallel" grid at each time step, akin to the auxiliary grid employed in the two-dimensional case in the previous section:

$$\tilde{z}_n^i = \tilde{\mu}_n + (i - N/2)\tilde{h}_n \tag{7.86}$$

We proceed as before. The time $n+1$ value function is again interpolated between the grid points on this auxiliary grid:

$$V_n(z_n^i) = \sum_j \int_{\tilde{z}_{n+1}^j}^{\tilde{z}_{n+1}^{j+1}} dz(\alpha_{n+1}^j e^z + \beta_{n+1}^j) \frac{1}{\sqrt{2\pi\sigma_{n,n+1}^2}} e^{-(z-\mu_{n,n+1}^i)^2/\sigma_{n,n+1}^2} \tag{7.87}$$

where $\sigma_{n,n+1}$ denotes the step volatility (*i.e.*, the conditional volatility between times $n$ and $n+1$). We now impose the same condition as before to create the Toeplitz form, as well as some conditions for convenience:

$$\begin{aligned} \tilde{h}_{n+1} &= a_{n,n+1} h_n \\ h_{n+1} &= h_n \\ \tilde{\mu}_{n+1} &= a_{n,n+1}\mu_n + b_{n,n+1} \\ \tilde{\mu}_n &= \mu_n \end{aligned} \tag{7.88}$$

In other words, we obtain the value at the primary grid by integrating the next step value function along the auxiliary grid. For a given time step, both primary and auxiliary grids are "centered" at the same mean, but of course their grid step sizes differ. (We take a constant step grid size for each grid across time steps.) See Figure 7.4.

Note that there is essentially no need for an additional interpolation due to the nesting of the grids. For example, we only need to identify, for each auxiliary grid point, the primary grid points that contain it:

$$z_{n+1}^{j'} < \tilde{z}_{n+1}^{j} < \tilde{z}_{n+1}^{j+1} < z_{n+1}^{j''} \tag{7.89}$$

This amounts to finding the index such that $(j' - N/2)h_{n+1} < (j - N/2)\tilde{h}_{n+1}$. Thus, we can perform the usual exponential interpolation of the time $n+1$ value function between the primary grid points $j'$ and $j''$ and confine the integration interval to the relevant auxiliary grid points.

#### *7.3.2.8 Example: swings/recalls*

Two popular products in energy markets involve structures with multiple exercise rights, either in terms of the ordinal number of exercises or the allocation of some total volume across some time period. For example, a swing option usually involves the transaction of a fixed volume (*e.g.*, natural gas or electricity) over a definite period of time (*e.g.*, the upcoming winter or summer). One of the parties is obliged to acquire this volume (either physically or in terms of financial settlement) by contract end, but they have the option of *when* they transact specific blocks (typically, there are minimum and maximum blocks limiting the sizes in which they can exercise). This product can be crafted exactly like (7.81), (usually) without complications arising from physical operation (*e.g.*, ratchets).

In contrast, recalls are essentially generalized American (or Bermudan) options with a specified number of exercise rights in a given time period. (Thus, there are limiting valuations: a single recall right amounts to an ordinary American option,

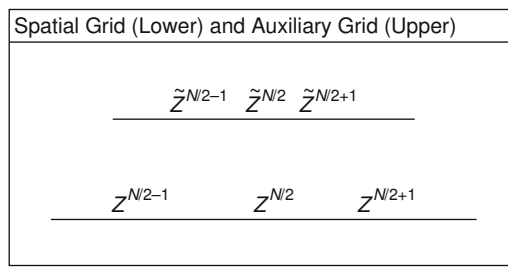



**Figure 7.4** Grid alignment

while the number of recalls equal to the number of days in the period in question amounts to an ordinary daily European option.) In the intermediate case, this structure can be valued by the following generalization of (7.73):

$$V_n^k(S) = \max((S-K)^+ + e^{-r\Delta t}E_n^Q V_{n+1}^{k-1}, e^{-r\Delta t}E_n^Q V_{n+1}^k) \tag{7.90}$$

where the superscript $k$ denotes the number of available recalls (exercise rights), with the convention $V^0 \equiv 0$. (In the context of the familiar binomial tree model (7.74), one might think of a binomial "forest," so to speak.) Note that the exercise boundary (separating regions of exercise from holding/non-exercise) now becomes three-dimensional in nature.

While the general overview of the grid model has been confined to the standard (log-) normal case, the method can be extended to a much richer class of processes, as we now see.

### 7.3.3 Further applications of characteristic functions

*7.3.3.1 Grid model, generalized*

Let us write again the main valuation problem from (7.44), noting explicit dependence on a particular measure:

$$V(z,t) = E_t^Q U(z_T) = \int_{-\infty}^{\infty} U(z_T)\Pr(z_T|z)dz \tag{7.91}$$

We again employ the usual ("quasi-linear") interpolation used in (7.28), but it proves convenient here to write (7.91) as

$$V(z,t) = E_t^Q \sum_j (\alpha_j e^{z_T} + \beta_j)1(z_T^j < z_T < z_T^{j+1}) \tag{7.92}$$

To evaluate the expectations in (7.92), we will employ the characteristic function methods presented in Section 5.2. Denoting the (conditional) characteristic function by $f$, the Fourier inversion formula implies that

$$\Pr(z_T|z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\phi f(\phi;z)e^{-i\phi z_T} = \frac{1}{2\pi}\int_{\Gamma} d\phi f(\phi;z)e^{-i\phi z_T} \tag{7.93}$$

where, as in Section 5.2, we allow for contour changes (indicated by $\Gamma$) that will justify reversing orders of integration implicit in (7.92). We are thus led to consider expressions of the form

$$E_t^Q 1(z_T^j < z_T < z_T^{j+1}) = \int_{-\infty}^{\infty} dz_T 1(z_T^j < z_T < z_T^{j+1}) \Pr(z_T|z)$$

$$= \int_{-\infty}^{\infty} dz_T 1(z_T^j < z_T < z_T^{j+1}) \frac{1}{2\pi} \int_{\Gamma} d\phi f(\phi; z) e^{-i\phi z_T} \tag{7.94}$$

Now, by choosing the contour $\Gamma$ such that the orders of integration in (7.94) may be reversed, (7.94) becomes

$$\frac{1}{2\pi} \int_{\Gamma} d\phi f(\phi; z) \frac{e^{-i\phi z_T^{j+1}} - e^{-i\phi z_T^j}}{-i\phi} \tag{7.95}$$

In the cases considered in Section 5.2 for a standard call, one of the grid points was at $-\infty$ and the contour was taken to be parallel to but below the real $\phi$–axis (*i.e.*, of the form $\phi_r - i\varepsilon$ with $\varepsilon > 0$ and $\phi_r \in (-\infty, \infty)$), ensuring convergence under the interchanged order of integration. We note here that this standard choice is in fact not necessary, and by retaining flexibility in the choice of contour, there can be great numerical benefits, as we shall see later in this chapter. Our main concern here is with employing fast computational algorithms for more general processes with known characteristic functions.

#### *7.3.3.2 Application to affine processes*

To this end, consider the class of affine jump diffusions studied in Chapter 5. These will have characteristic functions of the form[34]

$$f(\phi; z) = e^{\alpha(\tau;\phi)z + \beta(\tau;\phi)} \tag{7.96}$$

Specifically, we consider in the one-dimensional case a process of the form

$$dz = (\chi - \kappa z)dt + \sigma\, dw + j dq \tag{7.97}$$

in which case the coefficient of $z$ in (7.96) is separable: $\alpha = i\phi a(\tau)$ where $a = e^{-\kappa\tau}$. Thus, (7.95) takes the form

$$\frac{1}{2\pi} \int_{\Gamma} d\phi e^{\beta(\tau;\phi)} \frac{e^{-i\phi(z_T^{j+1} - a(\tau)z)} - e^{-i\phi(z_T^j - a(\tau)z)}}{-i\phi} \tag{7.98}$$

So, if we construct spatial grids as in (7.46) with the step sizes related by $h_T = a(\tau)h_t$, the expression in (7.98) will depend only on the difference in indices, just

as in the Gaussian case previously considered. Consequently, as long as we can numerically evaluate integrals such as those in (7.98) without great burden, we have reduced the valuation problem to an efficient, FFT-based matrix multiplication. (See Eydeland and Mahoney [2003], and related work by Lord *et al.* [2008].)

*7.3.3.3 Issues with American structures*

Another, non-standard example would be an American variance option under Heston.[35] The payoff at expiry is $(V_T - V - K)^+$ where $V$ is realized (integrated) variance with $Q$–dynamics given by

$$\begin{aligned} dv &= \kappa(\theta - v)dt + \sigma\sqrt{v}dw \\ dV &= v\,dt \end{aligned} \tag{7.99}$$

so that $V_T - V = \int_t^T v_s ds$. Conventionally, we shall take $V = 0$. The general backward induction algorithm for valuation can be written

$$U_n(v, V) = \max((V - K)^+, E_n^Q U_{n+1}(v', V')) \tag{7.100}$$

As before, we require an efficient means of evaluating the conditional expectation within the max operator in (7.100). In this case, it makes more sense to employ a true linear interpolation (*i.e.*, not in terms of exponentials) and so, assuming a discretized grid in $(v, V)$–space,[36] we are led to consider expectations of the following form:

$$\begin{aligned} g(v_n^k, V_n^l) &\equiv E_n^Q 1(v_{n+1}^i < v' < v_{n+1}^{i+1}) 1(V_{n+1}^j < V' < V_{n+1}^{j+1}) \\ h(v_n^k, V_n^l) &\equiv E_n^Q v' V' 1(v_{n+1}^i < v' < v_{n+1}^{i+1}) 1(V_{n+1}^j < V' < V_{n+1}^{j+1}) \end{aligned} \tag{7.101}$$

(as well as one involving a factor of just $v$ and one involving a factor of just $V$). For the first expectation we have, using characteristic functions, that

$$\begin{aligned} &E_n^Q 1(v_{n+1}^i < v' < v_{n+1}^{i+1}) 1(V_{n+1}^j < V' < V_{n+1}^{j+1}) \\ &\quad = \frac{1}{(2\pi)^2} \int_{\Gamma_1} \int_{\Gamma_2} d\phi_1\, d\phi_2\, e^{i\phi_2 V_n^i + \beta(\phi_1,\phi_2) v_n^k + \gamma(\phi_1,\phi_2)} \\ &\qquad \frac{e^{-i\phi_1 v_{n+1}^{i+1}} - e^{-i\phi_1 v_{n+1}^i}}{-i\phi_1} \frac{e^{-i\phi_2 V_{n+1}^{j+1}} - e^{-i\phi_2 V_{n+1}^j}}{-i\phi_2} \end{aligned} \tag{7.102}$$

for appropriately chosen contours $\Gamma_{1,2}$ and where $\beta$ and $\gamma$ satisfy

$$\begin{aligned} \dot{\beta} - \kappa\beta + \tfrac{1}{2}\sigma^2\beta^2 + i\phi_2 &= 0 \\ \dot{\gamma} + \kappa\theta\beta &= 0 \end{aligned} \tag{7.103}$$

with terminal[37] conditions $\beta = i\phi_1$ and $\gamma = 0$. Of course, the system in (7.103) can be solved analytically as it can in the regular Heston case. However, there is little point here in writing out the explicit solution. The main point we want to emphasize here is that the solution is *nonlinear* in the Fourier variables $\phi_{1,2}$. This fact presents some challenges for crafting the backward induction in (7.100) in a form amenable to fast transform techniques.

Because $\beta$ is nonlinear in $\phi_{1,2}$, the combination $\beta(\phi_1,\phi_2)v_n^k - i\phi_1 v_{n+1}^i$ (say) in (7.102) does not depend (and cannot be made to depend) only on the difference between grid indices $k-i$ as it does in the (linear) Gaussian case. However, since in the typical case the time discretization is small (*e.g.*, a single day so $1/365 \approx 0.003$ in annualized terms), we can exploit the fact that $\beta$ is approximately $i\phi_1$. We can write

$$\beta = i\phi_1 + \beta' \tag{7.104}$$

where $\beta'$ is small over the time interval in question. Thus, using a Taylor series expansion, we can write(7.102) as

$$\begin{aligned} &\frac{1}{(2\pi)^2}\int_{\Gamma_1}\int_{\Gamma_2} d\phi_1 d\phi_2 \\ &e^{\beta'(\phi_1,\phi_2)v_n^k + \gamma(\phi_1,\phi_2)} \frac{e^{-i\phi_1(v_{n+1}^{i+1} - v_n^k)} - e^{-i\phi_1(v_{n+1}^{i} - v_n^k)}}{-i\phi_1} \frac{e^{-i\phi_2(V_{n+1}^{j+1} - V_n^l)} - e^{-i\phi_2(V_{n+1}^{j} - V_n^l)}}{-i\phi_2} \\ &= \frac{1}{(2\pi)^2}\int_{\Gamma_1}\int_{\Gamma_2} d\phi_1 d\phi_2 \\ &e^{\gamma(\phi_1,\phi_2)} \frac{e^{-i\phi_1(v_{n+1}^{i+1} - v_n^k)} - e^{-i\phi_1(v_{n+1}^{i} - v_n^k)}}{-i\phi_1} \frac{e^{-i\phi_2(V_{n+1}^{j+1} - V_n^l)} - e^{-i\phi_2(V_{n+1}^{j} - V_n^l)}}{-i\phi_2} \\ &+ \frac{1}{(2\pi)^2} v_n^k \int_{\Gamma_1}\int_{\Gamma_2} d\phi_1 d\phi_2 \beta'(\phi_1,\phi_2) \\ &e^{\gamma(\phi_1,\phi_2)} \frac{e^{-i\phi_1(v_{n+1}^{i+1} - v_n^k)} - e^{-i\phi_1(v_{n+1}^{i} - v_n^k)}}{-i\phi_1} \frac{e^{-i\phi_2(V_{n+1}^{j+1} - V_n^l)} - e^{-i\phi_2(V_{n+1}^{j} - V_n^l)}}{-i\phi_2} + \cdots \end{aligned} \tag{7.105}$$

Note that each integration in the expansion can now be made to depend (for a suitable grid discretization) only on the difference of respective indices. Thus, the

problem can be treated as a series of problems where each term is amenable to FFT techniques. Plainly, American options on an asset driven by Heston dynamics can be similarly treated.[38]

To illustrate the plausibility of this approach, let us consider the standard mean-reverting process

$$dz = \kappa(\theta - z)dt + \sigma\, dw \tag{7.106}$$

with characteristic function $f = \exp(i\phi(ze^{-\kappa\tau} + \theta(1 - e^{-\kappa\tau})) - \phi^2\sigma^2(1 - e^{-2\kappa\tau})/(4\kappa))$ over a time interval $\tau = T - t$, which will be assumed small in the subsequent analysis. We are concerned with evaluating integrals of the form

$$\Pr(z_T < \gamma) = \frac{1}{2\pi}\int_{\Gamma} d\phi f(\phi)\frac{e^{-i\phi\gamma}}{-i\phi} \tag{7.107}$$

Now, using the form of the characteristic function, (7.107) can be written as

$$\begin{aligned}&\frac{1}{2\pi}\int_{\Gamma} d\phi \frac{e^{i\phi(\gamma - z)+\alpha_0(\phi)}}{-i\phi} e^{i\phi z(e^{-\kappa\tau}-1)} \\ &\quad = \frac{1}{2\pi}\sum_n \frac{z^n(e^{-\kappa t}-1)^n}{n!}\int_{\Gamma} d\phi (i\phi)^n \frac{e^{-i\phi(\gamma - z)+\alpha_0(\phi)}}{-i\phi}\end{aligned} \tag{7.108}$$

where it should be clear what $\alpha_0$ represents. From the basic result $\frac{1}{2\pi}\int_{\Gamma} d\phi \frac{e^{-i\alpha\phi-\beta^2\phi^2/2}}{-i\phi} = N(\frac{\alpha}{\beta})$, it can be seen through successive differentiations wrt. $\alpha$ that the terms in (7.108) will decay like $\frac{(1-e^{-\kappa\tau})^n}{(1-e^{-2\kappa\tau})^{n-\frac{1}{2}}}$. Thus, although the integral in (7.107) can of course be evaluated exactly, in the regime of small $\tau$ it can be well approximated by a series of integrals involving only the difference $\gamma - z$, which is the necessary condition for being able to employ fast convolution methods for early exercise options (*e.g.*, American variance swaps under Heston in (7.100)).

We have established to this point a quite general framework for evaluating early-exercise options, which will see further applications in certain control problems of relevance, such as tolling and storage. In general, however, even problems without such features are of interest (and challenging), so we will now turn attention to more general quadrature techniques.

## 7.4 Quadrature

In this section we will discuss rather generic methods for numerically evaluating integrals that arise as expectations in problems of interest.

### 7.4.1 Gaussian

*7.4.1.1 Basic formulas*

As mentioned in Section 5.2, the integrals that arise in characteristic function-based approaches (*e.g.*, option valuation) are often well handled by Gaussian quadrature. While it is not our intent here to provide a self-contained (much less full) exposition of such methods here (standard texts such as Press *et al.* [2007] should be consulted for that purpose), it will prove useful to give a cursory overview as the basic notions will be useful for later discussions.

The essential idea is that integrals of the form

$$\int_a^b dx \cdot w(x) f(x) \tag{7.109}$$

can be well approximated by the sum

$$\sum_{i=1}^{N} w_i f(x_i) \tag{7.110}$$

for reasonably small values of $N$ *if* the function $f$ can be well approximated by a polynomial with respect to the weighting function $w$ (in the sense of orthogonality across the interval in question). In fact, the approximation will be exact for functions within that class of basis polynomials. (Generally speaking, $N$ quadrature points give exact results for polynomials of degree $2N-1$, for the schemes we will be interested in here.) Computation of the weights $w$ and abscissas/collocation points $x$ is fairly standard provided the set of orthogonal polynomials with respect to the weighting function can be obtained without great difficulty. Common examples are Gauss-Laguerre, Gauss-Hermite, and Gauss-Legendre, with respective weight functions and interval of integration given by[39]

$$w(x) = x^{\alpha} e^{-x}, \quad 0 < x < \infty \tag{7.111}$$

and

$$w(x) = e^{-x^2}, \quad -\infty < x < \infty \tag{7.112}$$

and

$$w(x) = 1, \quad -1 < x < 1 \tag{7.113}$$

Of course, we would be remiss if we failed to mention that old warhorse, the trapezoidal rule, even though it does not technically fall under the rubric of Gaussian quadrature. The trapezoidal rule is exact for 1st degree ("linear") polynomials and

is related to the previously discussed interpolation/basis function approaches, with expansions in terms of "hat functions" of the form

$$\varphi_i(x) = \left\{ \begin{array}{ll} 1 - \frac{1}{h}|x - ih|, & |x - ih| < h \\ 0, & |x - ih| > h \end{array} \right\} \tag{7.114}$$

Assume we are concerned with integration over the interval $[0,1]$. Then, the quadrature points are given by $x_i = ih$ where $h = 1/N$ for $N+1$ quadrature points and the weights are given by $w_i = h \, \forall i$ except $x_0 = x_N = h/2$.

Not surprisingly, the relevant set of orthogonal polynomials (which determine the class of functions for which the approximation (7.110) will be accurate) are, respectively, the classical Laguerre and Hermite polynomials. The Hermite case is obviously relevant for the valuation of expectations under Gaussian densities, as we will shortly explore. Gauss-Laguerre is quite useful for general valuation of integrals in the expression (2.6.21). As an example, we can see the convergence in the case of Heston (for typical parameter values) in Figure 7.5.

#### *7.4.1.2 Pearson, revisited*

We can also compare quadrature-based approaches to Pearson's technique for evaluating spread options with nonzero strikes. Recall from (7.26) that the basic valuation problem can be reduced (via use of conditional densities) to an integral of the form

$$E(e^{\mu_2 + \sigma_2 z_2} - e^{\mu_1 + \sigma_1 z_1} - K)^+ = \int_{-\infty}^{\infty} dz_1 \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} F(z_1) \tag{7.115}$$

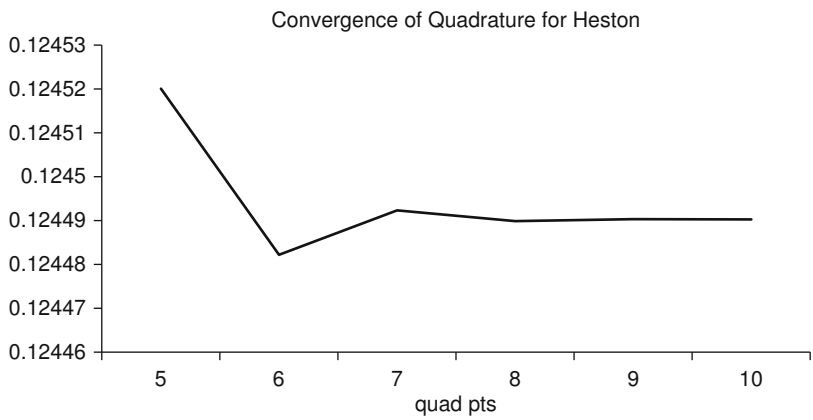



**Figure 7.5** Convergence of Gauss-Laguerre quadrature for Heston. Parameter values are $S = 1$, $K = 1$, $v = 0.1$, $\tau = 1$, $\kappa = 6$, $\theta = 0.1$, $\rho = -0.6$, $\sigma = 0.3$

for some nonlinear function $F$ of $z_1$ (known in terms of the standard normal CDF). In Pearson's approach, this function was interpolated in terms of piecewise affine exponentials that could be evaluated exactly (within the constituent intervals of interpolation). Alternatively, (7.115) lends itself well to Gauss-Hermite quadrature. In fact, from (7.33) we see that the deltas can likewise be alternatively obtained via quadrature. (Indeed, as has been noticed, *all* relevant greeks [*i.e.*, gammas and vegas] are obtainable via integration.) Recall that

$$E^Q(e^{z_2} - e^{z_1} - K)^+ = E^Q e^{z_2} \cdot E^{Q_2} 1(e^{z_2} - e^{z_1} - K > 0) \\ - E^Q e^{z_1} \cdot E^{Q_1} 1(e^{z_2} - e^{z_1} - K > 0) - K \cdot E^Q 1(e^{z_2} - e^{z_1} - K > 0) \tag{7.116}$$

where the measure changes are given by $\frac{dQ_i}{dQ} = \frac{e^{z_i}}{E^Q e^{z_i}}$. Now for homogeneous problems (*e.g.*, as occur with martingale pricing), Euler's theorem implies that the expected values of the indicator functions in (7.116) are the option deltas. Thus, an expression such as (7.115) can be used to compute these deltas. In particular, the deltas will involve an integral of the form

$$\int_{-\infty}^{\infty} dz_1 \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} N\left( \frac{\mu_2 + \sigma_2 \rho z_1 - \log(e^{\mu_1 + \sigma_1 z_1} + K)}{\sigma_2 \rho_s} \right) \tag{7.117}$$

where the means $\mu_i$ will entail some adjustment through the mean change. Thus, we see that the evaluations of the deltas amount to the same kind of quadrature problem. (With a bit of algebra the other greeks, *e.g.*, gammas and vegas, can likewise be extracted; see Section 7.5.2 for likelihood-based methods in simulation.)

#### *7.4.1.3 Generalizations (and limitations)*

More generally, consider max/min options with nonzero strikes. In energy markets such structures appear naturally in tolling problems with fuel-switching units, natural gas storage with multiple injection and/or withdrawal points, or natural gas transport with segmented paths (such that gas may flow from the cheaper of multiple receipt points to the more expensive of multiple delivery points). In most cases there are variable (non-stochastic) charges which prevent change-of-measure techniques from being (directly) used. A suitable payoff to illustrate how to approach such problems is

$$(\max(e^{z_2}, e^{z_3}) - e^{z_1} - K)^+ \tag{7.118}$$

with $K \neq 0$. Following Pearson, we write the joint (three-dimensional) density as $\Pr(z_1, z_2, z_3) = \Pr(z_1 | z_2, z_3) \Pr(z_2, z_3)$, where the conditional density can be readily obtained by standard results (*e.g.*, see (7.128)). Then we write

$$V = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dz_2 dz_3 \Pr(z_2, z_3) \int_{-\infty}^{\infty} dz_1 \Pr(z_1|z_2, z_3)(X(z_2, z_3) - e^{z_1})^+ \qquad (7.119)$$

where $X(z_2, z_3) \equiv \max(e^{z_2}, e^{z_3}) - K$. Now, the conditional density $\Pr(z_1|z_2, z_3)$ remains normal, therefore the "inner" integral in (7.119) over $z_1$ can be evaluated analytically for *given* $z_{2,3}$ (it is just a standard BS put problem). It remains now to handle the "outer" integral over $z_{2,3}$. As we have done before, we shall jump ahead to topics that will be considered in greater detail later, in this case multidimensional quadrature. Note that the density $\Pr(z_2, z_3)$ is jointly normal, so it might seem plausible that some kind of generalized Gauss-Hermite quadrature can be applied here. In fact, in the special case of zero correlation between assets 2 and 3, the joint density is just the product of two separate normal densities, in which case the usual quadrature can be applied across each dimension. We will see in the next section that this idea can indeed be generalized to arbitrary correlation structures.[40]

The idea is to "diagonalize" the correlation structure by a suitable linear transformation (in terms of eigenvectors). This approach is quite effective for computing the option value. An important issue obviously concerns the deltas (and other greeks). Here, though, we see some limitations to quadrature, namely situations where the underlying functional is *not* well approximated by polynomials. As can be seen from change-of-measure results, this calculation essentially entails the expectation of an exponential factor over the exercise region. For the delta wrt. asset 1 in (7.118), this is not problematic, because the resulting integrand remains continuous and hence amenable to Gaussian quadrature. However, for the other two deltas, due to the presence of the max( ) function between assets 2 and 3, a discontinuity is introduced which renders Gaussian quadrature ineffective. A feasible way around this problem is to further decompose the "outer" probability kernel in (7.119) via $\Pr(z_2, z_3) = \Pr(z_2|z_3)\Pr(z_3)$ and perform a Pearson-style interpolation in terms of $z_2$, as a function of $z_3$. Finally, Gaussian quadrature can be carried out wrt. $z_3$. This latter integration can typically be carried out with a fairly small number of points. We are thus effectively employing an adaptive sort of quadrature, where much work is devoted to one dimension, and much less to the other (recall that one dimension is analytically accounted for).

#### *7.4.1.4 Application: tolling*

Yet another problem of interest is encountered in tolling,[41] where the optionality to run a unit can extend across an entire day, yielding a weighted average of on- and off-peak power prices at minimum generation levels, plus incremental optimality in both blocks to ramp up to maximum capacity. The basic payoff is

$$\left[\begin{array}{l} C_{\min}(16P_{on}+8P_{off}-24HR_{\min}G-24\cdot VOM)-X-F\cdot G+zP_{off}+ \\ (C_{\max}-C_{\min})(16(P_{on}-HR_{inc}G-VOM)^{+}+8(P_{off}-HR_{inc}G-VOM)^{+}) \end{array}\right]^{+} \tag{7.120}$$

so that, on a given (week) day, one can start the unit for the entire day, running for 16 peak hours and 8 off-peak hours at minimum operating capacity, incurring variable fuel costs (at minimum heat-rate conversion) and operational costs, as well as fixed start-up costs and start-up fuel amounts. (There can also be off-peak power revenue for a start-up.) Once up at the min level, one can ramp up to max level in each temporal block, which incurs only variable costs.[42] By normalizing the problem by the total daily megawatt hours (*i.e.*, $24C_{\max}$), (7.120) can be written (in terms of log prices) as

$$\left[\begin{array}{l} a_1e^{z_1}+a_2e^{z_2}-g_0e^{z_3}-K_0+ \\ (b_1e^{z_1}-g_1e^{z_3}-K_1)^{+}+(b_2e^{z_2}-g_2e^{z_3}-K_2)^{+} \end{array}\right]^{+} \tag{7.121}$$

where we will not bother to explicitly identify the parameters in (7.121) with the various unit parameters, as it should be clear enough what they comprise. The expectation of the terminal payoff given by (7.121) in general entails a three-dimensional integration. However, following the approach of Pearson adopted for max/min options, we write the three-dimensional density as $\Pr(z_1,z_2,z_3) = \Pr(z_1,z_2)\Pr(z_3|z_1,z_2)$, and try to attack the problem as a two-dimensional integration over $z_1$ and $z_2$ wrt. to a function with $z_3$ integrated out (so to speak). The complication here is that, as a function of $z_{1,2}$, the expression in (7.121) does not appear reducible to a form that is analytically tractable (*e.g.*, a standard put payoff). However, note that for *fixed* $z_{1,2}$ (*e.g.*, at the "outer" quadrature abscissas), (7.121) has the form (recall the notation from Section 4.2)

$$(Z_0-g_0e^{z_3}+(Z_1-g_1e^{z_3})^{+}+(Z_2-g_2e^{z_3})^{+})^{+} \tag{7.122}$$

For $z_3$ sufficiently large, the expression in (7.122) equals zero; for $z_3$, sufficiently small, is a payoff for a standard in-the-money put. Plainly, there is some critical value $z_3^*$ (possibly at $-\infty$, in which case (7.122) makes no contribution for the outer quadrature points in question) for which the "inner" $z_3-$ integration can be written as $\int_{-\infty}^{z_3^*} dz_3(Z'-g'e^{z_3})\Pr(z_3)$, which *is* expressible as a standard BS put value. Determining this cutoff point is a simple problem of trial and error; there are only a few points that could qualify (*e.g.*, it could be where all the constituent terms in (7.122) are positive, *etc.*) and these just need to be tested to find the largest. Consequently, the extension of Pearson's method used for max/min options can also be applied here.

*7.4.1.5 Application: normal CDFs*

Another example is computation of standard normal distribution functions so ubiquitous in mathematical finance. Consider first the two-dimensional case:

$$N_2(a,b;\rho) = \frac{1}{2\pi}\int_{-\infty}^{a}\int_{-\infty}^{b} dxdy \cdot e^{-(x^2-2\rho xy+y^2)/2(1-\rho^2)} \tag{7.123}$$

Following the approach used in Pearson's method in Section 7.3.1, we decompose the density in the integrand via conditioning to get

$$\begin{aligned} N_2(a,b;\rho) &= \frac{1}{2\pi}\int_{-\infty}^{a} dx \cdot e^{-x^2/2}\int_{-\infty}^{b} dy \cdot e^{-(y-\rho x)^2/2(1-\rho^2)} \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} dx \cdot e^{-x^2/2} N\left(\frac{b-\rho x}{\sqrt{1-\rho^2}}\right) \end{aligned} \tag{7.124}$$

Now, it might seem tempting to employ Gauss-Hermite quadrature here, by including a factor of $1(x < a)$ in the integrand and extending the integration range to $+\infty$. However, it would quickly be revealed, as has already been pointed out, that the success of Gaussian quadrature depends critically on the integrand being well behaved in the appropriate sense. The discontinuity of the indicator function renders the method ill-suited in this case. (A similar phenomenon will be seen when we consider the calculation of various greeks in option-pricing problems, where Gaussian quadrature works extremely well for the option values, but very poorly for certain greeks.)

Fortunately, there is an alternate tack that allows effective use of Gaussian quadrature. Taking the derivative of (7.124) wrt. $\rho$ (and assuming differentiation can be taken under the integral sign), we get

$$\begin{aligned} \frac{\partial N_2}{\partial \rho} &= \frac{1}{2\pi}\int_{-\infty}^{a} dx \cdot e^{-x^2/2} e^{-(b-\rho x)^2/2(1-\rho^2)}\left(\frac{b\rho - x}{(1-\rho^2)^{3/2}}\right) \\ &= \frac{1}{2\pi} e^{-b^2/2}\int_{-\infty}^{a} dx \cdot e^{-(x-\rho b)^2/2(1-\rho^2)}\left(\frac{b\rho - x}{(1-\rho^2)^{3/2}}\right) \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-(a^2-2\rho ab+b^2)/2(1-\rho^2)} \end{aligned} \tag{7.125}$$

Using the known result for $\rho = 0$, we get

$$N_2(a, b; \rho) = \frac{1}{2\pi} \int_0^{\rho} dr \frac{1}{\sqrt{1-r^2}} e^{-(a^2 - 2rab + b^2)/2(1-r^2)} + N(a)N(b) \tag{7.126}$$

Finally, the substitution $r = \sin\theta$ in (7.126) gives the computationally useful result (originally due to Sheppard; see Amos [1969])

$$N_2(a, b; \rho) = \frac{1}{2\pi} \int_0^{\sin^{-1}\rho} d\theta e^{-(a^2 - 2\sin\theta ab + b^2)/2\cos^2\theta} + N(a)N(b) \tag{7.127}$$

which can be handled easily and efficiently with Gauss-Legendre quadrature.[43] Note that the originally two-dimensional problem in (7.123) has been reduced to a one-dimensional one, which can in fact be accurately evaluated with not much computational cost. An example is shown in Figure 7.6.

As we saw in Section 5.1.4 on the quintessential option-pricing formula, many applications of interest can be transformed to calculations involving the cumulative normal distribution function in higher dimensions. It can be shown that the kinds of conditioning carried out in the evaluation of the two-dimensional

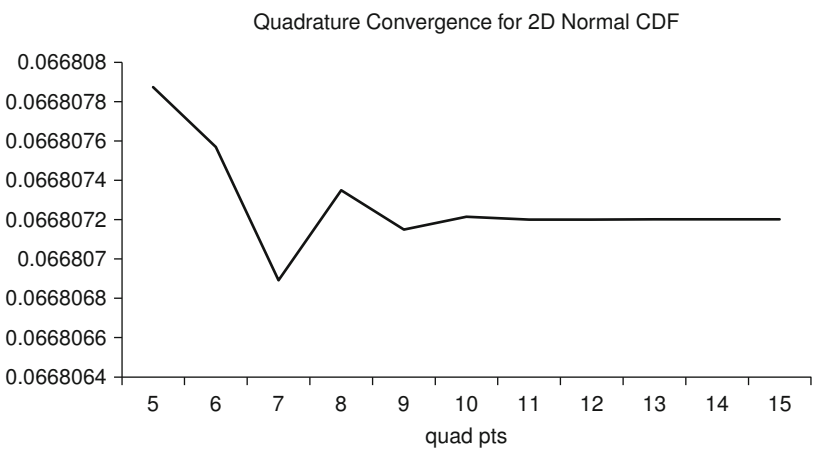



**Figure 7.6** Convergence results for 2-dimensional normal CDF. Parameter values are $a = 1.5$, $b = 1.5$, $\rho = -0.99$

problem can be extended to higher dimensions, allowing the calculation of interest to be obtained recursively. Recalling the basic result for conditional normals:

$$\begin{aligned}\mu_{y|x} &= \mu_y + \Sigma_{yz}\Sigma_{xx}^{-1}(x-\mu_x)\\ \Sigma_{y|x} &= \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\end{aligned} \tag{7.128}$$

in terms of the unconditional means and covariances ($\mu$ and $\Sigma$, respectively), we proceed as follows. First, for the three dimensional case, we have

$$N_3(\gamma_1,\gamma_2,\gamma_3;\rho_{12},\rho_{13},\rho_{23}) = \int_{-\infty}^{\gamma_1} dx_1 \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} N_2\left(\frac{\gamma_2-\rho_{12}^s x_1}{\rho_{12}^s}, \frac{\gamma_3-\rho_{13}^s x_1}{\rho_{13}^s}; \tilde{\rho}_{23}\right) \tag{7.129}$$

where $\rho_{1i}^s = \sqrt{1-\rho_{1i}^2}$ and $\tilde{\rho}_{23} = \frac{\rho_{23}-\rho_{12}\rho_{13}}{\rho_{12}^s\rho_{13}^s}$. Next, the substitution $y = N(x_1)$ allows the integral to be transformed into a form suitable for Gauss-Legendre quadrature, using the algorithm previously derived for $N_2$ and accurate approximations for $N^{-1}$ in terms of rational functions or Halley's third-order iterative scheme.[44] Note that exploiting the conditional structure again achieves dimensional reduction, from three to two, and in fact high accuracy can be obtained with relatively few quadrature points in each dimension.

Generalizations to higher dimensions proceed similarly. We have

$$\begin{aligned}N_n(\gamma;\Sigma) &= \int_{-\infty}^{\gamma} dx \frac{1}{\sqrt{(2\pi)^n \det\Sigma}} e^{-x^T\Sigma^{-1}x/2}\\ &= \int_{-\infty}^{\gamma_1} dx_1 e^{-x_1^2/2} N_{n-1}\left(\frac{\gamma_{i+1}-\rho_{1,i+1}^s x_1}{\rho_{1,i+1}^s}; \tilde{\Sigma}\right)\end{aligned} \tag{7.130}$$

where the index $i$ in (7.130) denotes dependence on elements 2 through $n$ of the $n$–dimensional vector $\gamma$ and $\rho_{1,i+1}^s = \sqrt{1-\rho_{1,i+1}^2}$ for $i = 1,\ldots,n-1$. The conditional covariance correlation $\tilde{\Sigma}$ has elements given by

$$\tilde{\Sigma}_{ij} = \frac{\rho_{i+1,j+1}-\rho_{1,i+1}\rho_{1,j+1}}{\rho_{1,i+1}^s\rho_{1,j+1}^s} \tag{7.131}$$

The inverse normal CDF substitution again allows use of Gauss-Legendre. The procedure is iterative, working backward (for the quadrature calculation in each dimension) until, say, a calculation of $N_2$ is required.

Obviously this nesting of the problem makes the overall procedure rather burdensome as the number of dimensions increases. There is reason to hope that the

efficiency of Gaussian quadrature, namely the need for a relatively small number of quadrature points (and hence computations) in each dimension will ameliorate this problem, but it certainly will not eliminate it, so we will consider some of these issues now.

### 7.4.2 High dimensions

*7.4.2.1 Generic setup*

A common class of problems encountered in financial applications (including commodity markets) takes the following form:

$$\int_{-\infty}^{\infty} dx H(x) \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{7.132}$$

for an $n$–dimensional vector $x$. Now, as $\Sigma$ will typically represent a (symmetric, positive-definite) covariance matrix, it permits a factorization of the form $\Sigma = LL^T$, in which case the substitution $x = Ly + \mu$ transforms (7.132) to (note that the Jacobian of the transformation is $L = \sqrt{\det \Sigma}$)

$$\int_{-\infty}^{\infty} dy H(Ly+\mu) \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}y^T y} \tag{7.133}$$

Given the form of the integrand in (7.133), a natural approach to is look for a multi-sum version of (7.110) for the Gauss-Hermite case (7.112).[45] The integral in (7.133) would be approximated as

$$\sum_{i,j,\ldots} w_i w_j \cdots \tilde{H}(y_i, y_j, \ldots) \tag{7.134}$$

where $\tilde{H}(y) \equiv H(Ly+\mu)$ is evaluated at the Gauss-Hermite abscissas. This latter computation can (and obviously should) be performed outside the multiplication loops in (7.134). For example, the arguments of $H$ must be evaluated at

$$x_k^{ij\ldots} = \mu_k + L_{k1} y_1^i + L_{k2} y_2^i + \cdots \tag{7.135}$$

where $y^i$ denote the quadrature points. The element-by-element multiplication in (7.135) needs to only be computed for a *single* set of quadrature points, across each column in $L$ for each row/dimension. Then, the appropriate contribution to each argument in (7.134) can be invoked as needed.

Now, there are several possible choices for the underlying factorization $\Sigma = LL^T$. An obvious candidate is the standard Cholesky factorization where $L$ is a

lower triangular matrix.[46] Another possible choice is in terms of the eigenvalues/eigenvectors of the covariance matrix:[47]

$$\Sigma = V \Lambda V^T \tag{7.136}$$

where $V$ is a matrix whose columns are the eigenvectors of $\Sigma$ and $\Lambda$ is a diagonal matrix of the eigenvalues. Consequently, we could take $L = V\Lambda^{1/2}$ (where the exponent $1/2$ represents element-by-element square root). In fact, Cholesky can also be put in this tri-factor form by taking

$$LL^T = ADA^T \tag{7.137}$$

where $D$ is diagonal and $A$ is lower triangular with ones across the diagonal.[48] In general, for any possible factorization $L$, another possible factorization is given by $LO$ where $O$ is an orthogonal matrix (that is, $OO^T = I$). To the best of our knowledge, it does not appear possible to definitively say which factorization is best for a given problem; some experimentation is probably necessary. Note that an advantage of representations such as (7.136) or (7.137) is in the identification of certain "characteristic" scales in a given problem, which can facilitate various kinds of adaptive quadrature. (This is akin to the notion of effective dimension reduction that will be seen later in conjunction with certain simulation-based techniques.) We will only demonstrate here the feasibility of this technique by comparing it to some test cases.[49]

So, consider a payoff function $\max(e^{z_1}, e^{z_2}, e^{z_3}, e^{z_4})$ for jointly normal variables, and a correlation structure given by $\text{corr}(z_i, z_j) = \rho^{|i-j|}$. We show typical convergence results (against the "exact" quintessential formula) for different values of $\rho$ in Figures 7.7 and 7.8. As can be seen, depending on the particular correlation structure, convergence can be extremely rapid for one correlation matrix decomposition, but not necessarily for another. As noted, some experimentation will be necessary in practice. However, we can clearly see the basic effectiveness of the approach here.

While (7.134) is certainly straightforward, it obviously suffers from the fact that its computational cost grows exponentially with dimension, rendering it infeasible for high-dimensional problems.[50] Still, given the nice convergence properties of Gaussian quadrature in general, it is worth investigating the extent to which (7.134) can be employed. If a small number of quadrature points in each dimension is sufficient for high accuracy, then the dimensionality of tractable problems may be reasonably high.[51] In our experience the number dimensions for which quadrature is feasible is five, perhaps six. It remains the case, then, that the dimension of problems for which the above "full-grid" approach (we will understand this choice of words shortly) is feasible is not terribly large. An alternative approach in higher dimensions is still required.

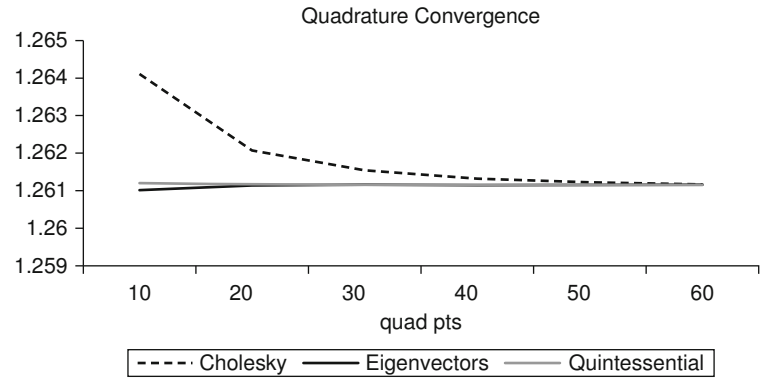



**Figure 7.7** Convergence of Gaussian quadrature. Comparison of different decompositions of correlation matrix for (effective) diagonalization. Payoff is max across 4 assets, each with expected value 1, time to maturity 1, volatilities 0.4, 0.5, 0.6, and 0.7. Correlation structure is given by $\text{corr}(z_i, z_j) = \rho^{|i-j|}$ with $\rho = 0.9$

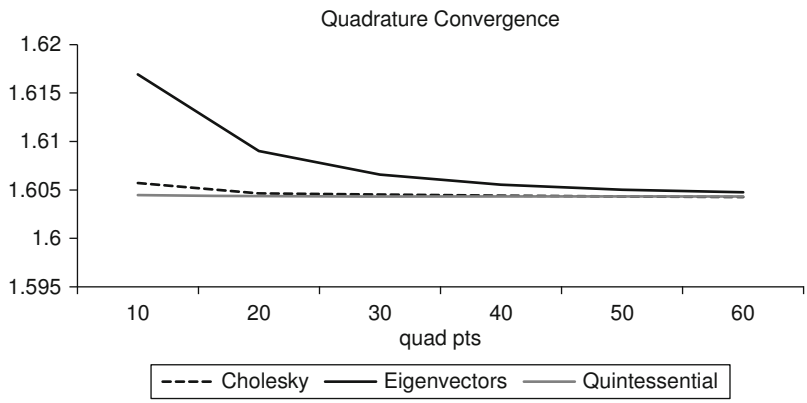



**Figure 7.8** Convergence of Gaussian quadrature. Same parameters as Figure 7.7, but $\rho = -0.9$

*7.4.2.2 Sparse grid quadrature*

We begin by introducing some notation. First, we consider quadrature schemes that are "embedded," that is to say, as we go to finer levels of resolutions/discretization, the quadrature points used at the previous level are included in subsequent levels.[52] A simple example would be the familiar trapezoidal rule in (7.114), where the resolution is increased from $2^l + 1$ quadrature points to $2^{l+1} + 1$ points (*i.e.*, the step size is halved). We define a (one-dimensional) level $l$ quadrature rule $Q_l$ as an operator on some function $f$ via

$$Q_l f \equiv \sum_{i=0}^{N_l} w_i^l f(x_i^l) \tag{7.138}$$

for some set of (level-dependent) weights $w_i^l$ and quadrature points $x_i^l$ (with $N_l + 1$ weights and points, respectively). Let us also denote by $\aleph^l$ the class of functions for which the scheme in (7.138) is exact. (We will assume that this class is nested; *i.e.* $\aleph^l \subset \aleph^{l+1}$.[53]) Note that multidimensional quadrature can be constructed via tensor products as

$$(Q_{l_1} \otimes Q_{l_2} \otimes \cdots \otimes Q_{l_D}) f = \sum_{i_1=0}^{N_{l_1}} \sum_{i_2=0}^{N_{l_2}} \cdots \sum_{i_D=0}^{N_{l_D}} w_{i_1}^{l_1} w_{i_2}^{l_2} \cdots w_{i_D}^{l_D} f(x_{i_1}^{l_1}, x_{i_2}^{l_2}, \cdots, x_{i_D}^{l_D}) \tag{7.139}$$

where $D$ is the number of dimensions and in general the level of resolution can differ across dimensions. Note we see from (7.139) the motivation for the characterization "full grid": *all* of the (one-dimensional) quadrature points are used across *each* dimension.

We now note a rather trivial (but ultimately important) point. Using telescoping sums, we have that

$$Q_l f = \sum_{i=0}^{l} \Delta Q_i f \tag{7.140}$$

where $\Delta Q_i \equiv Q_i - Q_{i-1}$ (with $Q_{-1} \equiv 0$). The implication of (7.140) is the following: for any function $f \in \aleph^l$ (or more generally functions well approximated by members of this class), the *incremental* value of going to a higher-level resolution in the quadrature scheme (7.138) is *zero* (or more generally tends to zero).[54] Consequently, the additional workload of going to this higher level is largely unnecessary and inefficient. This idea, namely that quadrature points should be added only if they are needed for a given level of accuracy, can be exploited to create quadrature schemes in higher dimensions that greatly ameliorate the so-called curse of dimensionality (*i.e.*, the fact that the computational burden grows exponentially with dimension).

The idea (originally due to Smolyak [1963]; see also Holtz [2011]) starts with the following representation. Invoking (7.140), we write (7.139) as

$$(Q_{l_1} \otimes Q_{l_2} \otimes \cdots \otimes Q_{l_D}) f = \sum_{i_1=0}^{l_1} \sum_{i_2=0}^{l_2} \cdots \sum_{i_D=0}^{l_D} (\Delta Q_{i_1} \otimes \Delta Q_{i_2} \otimes \cdots \otimes \Delta Q_{i_D}) f \tag{7.141}$$

which follows from telescoping sums and the bilinearity of the tensor product. We now note the following: for quadrature rules that are exact for certain categories of

functions, the full-grid ensemble (7.141) will involve redundancies across quadrature points. For example, consider a scheme that is exact for polynomials of (total) degree $d$. Introduce monomials of the form

$$x_1^{k_1} x_2^{k_2} \cdots x_D^{k_D} \tag{7.142}$$

with $k_1 + k_2 + \cdots k_D = d$. Then it can be seen that (7.141) is inefficient because many of the terms vanish for monomials such as (7.142). This point leads us to seek multidimensional schemes that are exact for monomials of a given degree, an obvious generalization of one-dimensional Gaussian quadrature. However, unlike the one-dimensional case, it is in general not possible to constructively prescribe the minimal number of quadrature points necessary to achieve a given (polynomial) degree of accuracy (see Cools [2002]).[55]

Smolyak's very clever idea was to consider a compromise by taking the tensor product in (7.141) only over those terms contributing to the discretization error as the degree of accuracy is increased. The approach is highly reminiscent of modern, multi-resolution techniques such as wavelet analysis (in the context of quantitative finance, see Dempster *et al.* [2000] or de Wiart and Dempster [2011]). That is, we write the quadrature scheme as

$$I_l(f) = \sum_{|i|=0}^{l} (\Delta Q_{i_i} \otimes \Delta Q_{i_2} \otimes \cdots \otimes \Delta Q_{i_D}) f \tag{7.143}$$

where the index $i$ is a $D$–dimensional vector of non-negative integers with "norm" given by $|i| \equiv i_1 + \cdots i_D$. In other words, each term in the summation in (7.143) is taken over all indices with a given total degree (or more broadly conceived, a given level of resolution). The expression (7.143) can also be written as (see Wasilkowski and Woźniakowski [1995])

$$I_l(f) = \sum_{|i|=l-D+1}^{l} (-)^{l-|i|} \binom{D-1}{l-|i|} (Q_{i_i} \otimes Q_{i_2} \otimes \cdots \otimes Q_{i_D}) f \tag{7.144}$$

We may think of the one-dimensional schemes $Q_k$ as being exact for polynomials of degree $2k-1$ (as in the standard Gaussian case), in which case the multidimensional scheme in (7.143) is exact for multinomials of total degree $l$. More generally, we can think of (7.143) as a device for introducing, for increasing levels of resolution, only those quadrature points (and associated weights) necessary for the level of exactness required for that level. Thus, for a given dimension, as the resolution increases the amount of new points (and hence the resulting computational burden) grows much more slowly than the corresponding full-grid case. (We refer the reader to table 1 in Heiss and Winschel (2006).)

**Table 7.2** Comparison of quadrature techniques. Probability mass of 4D unit hypercube for standard multi-normal. "Exact" value given by Gauss-Legendre quadrature via (7.130) and 10 quadrature points in each dimension. Full grid value based on trapezoidal rule and $65 = (2^7 + 1)$ points per dimension (redundant multiplications in the tensor product (7.139) are taken into account by appropriate nesting). Sparse grid result also based on trapezoidal rule with resolution levels 2 through 8 in (7.144).

| | Value | Operation count |
|---|---|---|
| Full | 0.058622 | 18,129,540 |
| Sparse | 0.058566 | 117,532 |
| GauLeg | 0.058642 | 1,110 |

To illustrate, we compute the expected value of $1(0 \leq x \leq 1)$ with $x$ a four-dimensional unit normal with correlation matrix with elements $\rho_{ij} = 0.8^{|i-j|}$. We use the trapezoidal rule, and compare with Gaussian quadrature via (7.130). The results are shown in Table 7.2. Obviously, we would never use the trapezoidal rule in practice for evaluating such an integral. However, this example serves to make clear the great computational feasibility of sparse grid quadrature.

## 7.5 Simulation

*7.5.0.1 Preamble*

In this section we will provide an overview of simulation-based methods relevant for problems arising in energy markets. As usual, no claims to completeness are made, and we assume some basic familiarity on the reader's part. We recommend the excellent text by Glasserman (2004) for greater detail than that provided here. We do feel it necessary to emphasize here that simulation methods are inherently *computational* techniques. That is, they are simply means to an end. They can be extremely useful means (and in many cases, they are the only feasible means), but they are ultimately just tools for calculating a certain entity of interest. Like any other tool, they can be appropriate or inappropriate for a given problem, and cleverly applied or completely misapplied. It must be said bluntly: simulation does *not* add to one's knowledge.[56] Put differently, there is nothing that can be produced by a simulation that was not already present in the initial assumptions. This is not to say that the simulation cannot produce concrete numbers (*e.g.*, prices, hedge ratios, *etc.*) that could not otherwise be obtained. The point is that the resulting numbers are only as good as the input numbers used to generate the output.

We engage in this bit of editorializing here because we have commonly encountered a tendency to believe that, the more complex a problem is, the quicker one must resort to simulation in order to arrive at some kind of solution, and that the

less need there is to understand the essential features of a problem. This viewpoint is flawed for a number of reasons. First, as already noted, bad input almost certainly guarantees bad output, and in energy markets with rather sparse data series and many fundamental changes in the recent history, input data will often be of mixed (at best) quality. It is simply an illusion of security that churning out a great number of scenarios based on spotty underlying data will produce useful results. Second, reliance on simulation can encourage laziness of thought in tackling a problem (to be quite frank about it), as the relative ease with which many simulations can be implemented tends to allow one to overlook the need to identify the relevant features of a given problem, features that could (indeed, should) be used to produce robust approximations. Worse: a tempting belief that scenario generation can compensate for poorly estimated model parameters can become hard to resist. (As a general rule, robust approximations are to be preferred to precise nonsense.) Of course, overreliance on a particular methodology is not unique to simulation, but in our experience there is greater danger here than with other methods. This danger is especially enhanced in energy markets, where various operational constraints present numerous challenges. Finally, we repeat a recurring theme of this work: valuation *means* replication, and any valuation technique that does not produce hedges whose precise meaning refers to a trading strategy around which the stated value can be extracted/collected (with residual risk properly identified and accounted for) is simply rubbish.

The simple fact is that the great complexity of many structured products is of little relevance in producing *good* valuations. While we have often encountered great surprise when we argue that, say, hourly electricity prices should not (necessarily) be modeled when valuing tolling or full requirement deals, it remains the case that (as of this writing) there is no kind of traded product that gives direct exposure to prices at that level of resolution.[57] It is thus irrelevant whether one has a good model of, say, hourly price-load correlation[58] (and in reality, one never has such a model), if this is not the entity one has exposure to in light of available hedging instruments. There is little reason to believe that trying to infer the desired relationship between exposure and available portfolios via elaborate auxiliary models is superior to directly ascertaining that relationship from the start. The former approach is at best unnecessary, at worst seriously misleading.

With this little lecture out of the way, we now turn attention to some specific techniques related to simulation.

### 7.5.1 Monte Carlo

#### *7.5.1.1 Integration as expectation*

The basic idea here needs little elaboration. The central problem of interest is the evaluation of integrals of the form

$$V = \int_{\Omega} dx \cdot f(x) = \int_0^1 dx \cdot f(x) \cdot 1_{\Omega}(x) \tag{7.145}$$

for some region $\Omega$ inside the unit hyper-cube of $R^n$. The expression in (7.145) clearly can be expressed as an expectation:

$$V = E\left[f(x) \cdot 1_{\Omega}(x)\right] \tag{7.146}$$

with $x$ a uniform deviate in $R^n$. The natural intuition is to approximate (7.145) by a *sample average*

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) 1_{\Omega}(x_i) \tag{7.147}$$

where $x_i$ are drawn identically and independently (IID) from a uniform deviate in the unit hypercube. Now, intuition is fine as far it goes, but it is worth asking how (if at all) the arithmetic entity in (7.147) relates the integral of interest in (7.145). To answer this question, first note that the expectation of each of term in (7.147) is the same (via the assumption of identity in distribution), hence the expectation of (7.147) equals (7.146) and thus integral of interest in (7.145). This equality in expectation implies that the approximation in (7.147) is *unbiased*. Second, since the entity in (7.147) is itself a random variable, it is useful to ask how far any particular realization of it can be from its expected value. Here the assumption of independence of distribution, along with the Central Limit Theorem,[59] tells us that the variance of the approximation/estimate in (7.147) is $O(\frac{1}{N})$. Hence, as more simulations are used, the probability that any particular realization of (7.147) is far (in terms of some arbitrary tolerance) from the true value in (7.145) is correspondingly small. These two points convey the sense in which the entity (7.147) has a connection to the integral of interest in (7.145).

#### *7.5.1.2 Generation of random deviates*

We belabor these rather basic points because we wish to emphasize the crucial distinction between sample averages and population means, a point that was given much attention in Chapters 2 and 6 in the context of estimation. As an arithmetic expression (7.147) is merely a crank; for the output to be meaningful it must be understood what the purpose of the tool is and how it relates to the objective in question. We should note here some standard techniques for actually generating the random variables used in simulation-based calculations. Of course, the idea of algorithmically generating random variables is a contradiction in terms, hence the actual methodology is termed pseudorandom number generation. The basic idea, used in the vast majority of popular applications and algorithms, involves so-called

linear congruential generation. For some initial value $X_0$ (termed the seed) and a modulus $m$, the following sequence is generated:

$$X_{n+1} = (aX_n + b) \bmod m \tag{7.148}$$

for suitable choices of $a$ and $b$. Obviously, the recursion in (7.148) will only produce $m$ distinct values, and will eventually repeat itself. However, for "good" choices of the underlying parameters, the maximal period ($m$) can be attained, and for sufficiently long periods the resulting sequence will resemble "true" randomness (in the sense that it will pass statistical hypothesis tests). Press *et al.* (2007) provide a solid overview, as well as algorithms of good practical use; see also Park and Miller (1988).[60] The sequence in (7.148) thus effectively gives a uniformly distributed set of integers, which obviously can be normalized to yield a set of (independent) uniforms on $[0, 1]$.

*7.5.1.3 Non-uniform deviates*

Of course, in a great many applications it is more suitable to employ simulations in terms of normally distributed random variables, rather than uniform deviates. As is well-known, for continuously valued random variables $X$ with known distribution function $F(x) \equiv \Pr(X \leq x)$, $F(X)$ is uniformly distributed. Thus $F^{-1}(U)$ is distributed as $X$ for $U$ uniformly distributed on $(0, 1)$. Thus, by simulating standard uniforms, general random variables can be generated by inverting the distribution function. Clearly, the feasibility of this method depends on the distribution function being sufficiently tractable that the resulting inversion is not too numerically taxing. A quite simple example, useful for modeling jump diffusions, is the exponential density, where $\Pr(x) = \lambda e^{-\lambda x}$.[61] In practice, there are superior methods, such as acceptance/rejection (to be outlined shortly; see also Press *et al.*, 2007), which actually form the basis of other important applications such as Markov Chain Monte Carlo (which were mentioned in Chapter 6 in the context of stochastic filtering). Ideally, a method should be tailored to the nature of the distribution to be simulated/sampled. A very useful example is the simulation of standard normals. Of course, there exist very efficient techniques for inverting the standard normal CDF, but the algorithm of choice is the Box-Muller method, which is a transformation of two independent uniform deviates $(x_1, x_2)$ via

$$\begin{aligned} x_1 &= e^{-\frac{1}{2}(y_1^2 + y_2^2)} \\ x_2 &= \frac{1}{2\pi} \tan^{-1} \frac{y_2}{y_1} \end{aligned} \tag{7.149}$$

To see this, we argue in reverse: if $(y_1, y_2)$ are independent unit normals, then upon transforming to polar coordinates in the CDF it is easy to see that (7.149) implies that $(x_1, x_2)$ are independent uniforms. [62] Note that the deviates are produced in pairs, a fact that can be exploited for efficiency.

*7.5.1.4 Acceptance/rejection*

The method just outlined requires that the inverse of the CDF in question be feasibly calculated (or that the underlying structure permits suitable tricks as in the Gaussian/Box-Muller case). This requirement may be restrictive in many cases, so we note an alternative here. Assume we have a CDF $G$ which is not only easily invertible[63] but also "close" to the desired CDF $F$. In fact, we require the ratio $F/G$ to be bounded by a (positive) constant $c$, as close to 1 as possible. Now assume that $Y$ is distributed according to $G$: $Y \sim G$. Denote the respective densities by $f$ and $g$. Now draw independently a uniform RV $U$. The acceptance/rejection method works by keeping (accepting) $Y$ if $U \leq \frac{f(Y)}{cg(Y)}$ and repeating the process otherwise (rejecting). The claim is that the accepted deviates (again, drawn from $G$ but only conditionally kept) are distributed according to $F$.

To see this, consider the following conditional probability: $\Pr(Y \leq y | U \leq \frac{f(Y)}{cg(Y)})$. Since

$$\Pr\left(U \leq \frac{f(Y)}{cg(Y)}\right) = \int_{\Omega}\int_{[0,1]} dudyg(y)1\left(u \leq \frac{f(y)}{cg(y)}\right) = \int_{\Omega} dyg(y)\frac{f(y)}{cg(y)} = c^{-1}$$

$$\Pr\left(Y \leq y, U \leq \frac{f(Y)}{cg(Y)}\right) = \int_{\Omega}\int_{[0,1]} dudy'g(y')1(y' \leq y)1\left(u \leq \frac{f(y')}{cg(y')}\right)$$

$$= \int_{\Omega} dy'g(y')1(y' \leq y)\frac{f(y')}{cg(y')} = c^{-1}F(y) \tag{7.150}$$

Thus, by Bayes's rule, we see that $Y$ conditioned on $U$ being within the acceptance region is indeed distributed as $F$.[64]

*7.5.1.5 Joint dependency*

Here, we note a very standard approach for generating multidimensional Gaussian deviates with a given covariance structure. (We will consider multidimensional aspects of simulation for copulas in Section 8.1.1.) We are referring of course to Cholesky decomposition, which is a factorization of a symmetric matrix into a lower triangular form: $\Sigma = LL^T$. If $\Sigma$ is a correlation matrix and $z$ is a vector of independent unit normals, then $w = Lz$ is clearly a vector of correlated unit normals (with correlation matrix $\Sigma$ ). As already pointed out in the subsection on high dimensional quadrature, this factorization is not unique. Obviously, Cholesky is a simple and convenient choice, and in fact alternative choices (*e.g.*, diagonalization via [orthogonal] eigenvectors) do not offer any substantial advantages (in terms of convergence, *etc.*). The simulation error depends on the underlying covariance structure, which obviously cannot be affected by one particular factorization as opposed to another. However, we will point out cases where this is not true, when we

consider low-discrepancy alternatives to Monte Carlo simulation (*i.e.*, quasi-Monte Carlo).

As already indicated, the variance of any particular simulation is of great importance in determining the usefulness of the simulation's output. We will now discuss various standard techniques for improving simulation efficiency (*i.e.*, reducing the variance of the simulation).

## 7.5.2 Variance reduction

### *7.5.2.1 Antithetics*

There are a number of techniques that are fairly easy to incorporate into any simulation scheme that can aid nontrivially in variance reduction. One such method involves so-called antithetic variates. The idea here is very simple. In many applications the random deviates in question are identically distributed under simple, say, affine, transformations. For example, if $x$ is $N(0,1)$, then so is $-x$. Similarly, if $x$ is $U(0,1)$ then $1-x$ is also. Thus, if a random sample is generated, it is very simple to create another random sample under this transformation. Apart from reducing the computational cost of having to generate another random set from whatever congruence generator one is using, the resulting variance of the estimator is reduced. This can be seen as follows. For the Gaussian case we have the following (unbiased) estimator:

$$\hat{V} = \frac{1}{2}\left(\frac{1}{N}\sum_i f(x_i) + \frac{1}{N}\sum_i f(-x_i)\right) \tag{7.151}$$

Using the independence of each draw, we see that the variance of the estimator is given by

$$\mathrm{var}(\hat{V}) = \tfrac{1}{2N}\left(\mathrm{var}(f(x)) + \tfrac{1}{2}\mathrm{cov}(f(x), f(-x))\right) \tag{7.152}$$

Therefore, in those cases where the objective function is negatively correlated under the transformation, the variance in (7.152) will be less than the constituent variances, in particular the variance of the base case estimator (*i.e.*, without the antithetic variate).[65]

### *7.5.2.2 Control variates*

A generally more effective technique is the use of so-called control variates, often referred to as a regression-based technique, for reasons that will become obvious. Consider the following estimator:

$$\hat{V} = \frac{1}{N}\sum_i f(x_i) + \Delta\left(\frac{1}{N}\sum_i g(x_i) - g_0\right) \tag{7.153}$$

where $g$ is some function whose expectation is known, with $Eg(x) = g_0$. Clearly this estimator is unbiased, for any value of $\Delta$. With this freedom, we can choose $\Delta$

to minimize the variance of $\hat{V}$. Straightforward calculation shows that this optimal value is given by

$$\widehat{\Delta} = -\frac{\text{cov}(f(x), g(x))}{\text{var}(g(x))} \tag{7.154}$$

The connection between control variates and regression should be clear now.

Of course, the actual covariance in (7.154) is not known (the expectation of $f$ is precisely what is being sought), so in practice we use the sample covariance (and variance). It is easy to see that for $f = g$, $\hat{\Delta} = -1$ and the estimator becomes $\hat{V} = g_0$ with variance (trivially) zero. A classic example of this technique is to evaluate an Asian option using a European option (with known BS value) as the control variate. Finally, we note an obvious generalization of (7.153) and (7.154). We consider a multidimensional problem, with multiple control variates

$$\hat{V} = \frac{1}{N}\sum_i f(x_i) + \Delta_1\left(\frac{1}{N}\sum_i g_1(x_i) - g_1^0\right) + \cdots \Delta_K\left(\frac{1}{N}\sum_i g_K(x_i) - g_K^0\right) \tag{7.155}$$

where $x$ now denotes a multidimensional random variable, and $g_k$ are functions with known expectations $g_k^0$. Variance minimization of the estimator $\hat{V}$ requires that the following linear system be satisfied:

$$\sum_k \text{cov}(g_i, g_k)\Delta_k = \text{cov}(g_i, f) \tag{7.156}$$

again illustrating the connection to regressions.

#### 7.5.2.3 *Importance sampling and its connection to measure change*

Since the purpose of these variance reduction techniques is to effectively increase the convergence properties of a simulation, we point out a method that is closely related to the change-of-measure techniques described in Chapter 5. A simple example will suffice here, as we will further develop these concepts later in the chapter in conjunction with contour integration. Assume we are interested in the calculating following tail probability via simulation:

$$\Pr(z > \alpha) = E[1(z > \alpha)] \tag{7.157}$$

where $z$ is a unit normal and $\alpha$ is large (*e.g.*, more than two standard deviations). Clearly, simulation will converge very slowly because an extremely large number of paths will be required to get nonzero samples. (The same issue arises in trying to value deep OTM options by simulation.) A standard textbook trick is to note that, by completing the square, the underlying probability density in (7.157) can be rewritten as $\frac{1}{\sqrt{2\pi}}e^{-z^2/2} = \frac{1}{\sqrt{2\pi}}e^{-(z-\alpha)^2/2}e^{-\alpha z+\alpha^2/2}$, in which case the desired expectation can be re-crafted as a *new* expectation wrt. a Gaussian density with mean

centered about the tail point of interest,[66] which will consequently converge more rapidly. For obvious reasons, this approach is termed *importance sampling*. There is in fact a deeper aspect to the approach beyond algebraic manipulations. Note that the additional exponential factor is reminiscent of a Girsanov-type transformation, which naturally leads us to consider applications of change of measure. Explicitly introducing an underlying probability measure $P$, (7.157) can be written as

$$\begin{aligned} E^P 1(z > \alpha) &= E^P \frac{\xi(z)}{E^P \xi(z)} \frac{E^P \xi(z)}{\xi(z)} 1(z > \alpha) \\ &= E^P \xi(z) \cdot E^{P_\xi} \xi(z)^{-1} 1(z > \alpha) \end{aligned} \tag{7.158}$$

where $\xi(z)$ is some (suitably chosen) function of $z$ and the measure $P_\xi$ is defined via $\frac{dP_\xi}{dP} = \frac{\xi(z)}{E^P \xi(z)}$. Note that (7.158) is completely general and the only real consideration in applying this approach is that the ensuing change of measure be feasible/tractable. Maintaining this generality, a possible consideration in choosing $\xi(z)$ is that, under the new measure, the expectation of $z$ is centered about the tail quantile $\alpha$:

$$E^{P_\xi} z = \frac{E^P z \xi(z)}{E^P \xi(z)} = \alpha \tag{7.159}$$

Plainly, the familiar class of affine jump diffusions we have considered in great detail throughout provides such feasibility, in particular with adjustment factors of the form $\xi = e^{\gamma z}$. It should be clear how the standard textbook example can be recovered from this framework. Importance sampling is thus seen to be an important and quite general technique for facilitating the convergence of standard Monte Carlo schemes, when done properly (see Endnote 66). We will consider other connections to complex variable theory later in this chapter.

#### *7.5.2.4 Brownian bridge construction*

Finally, it is worth noting here a technique, although not strictly directed toward variance reduction or convergence as such, is nonetheless quite useful in facilitating certain calculations based on simulation. In many applications, we are interested not simply in the distribution of some entity at a terminal time (*e.g.*, in the case of a vanilla or European option), but in the distribution of an entire *path* (that is to say, the distribution of a process through time). Familiar examples from financial markets are Asian and lookback options. Obviously, path dependency is quite important in energy markets, due to various physical/operational constraints that impact valuation. Examples include natural gas storage and tolling. Commonly, we are interested in discrete time versions of a (vector) Brownian motion with drift, *e.g.*,

$$\Delta z_i^j = \mu_i \Delta t + \sigma_i \sqrt{\Delta t} \phi_i^j \tag{7.160}$$

where $\phi$ are unit normals with correlation structure $E\phi_i^j\phi_{i'}^{j'} = \delta_{jj'}\rho_{ii'}$ (so $j$ is a time-step index and $i$ is a commodity index). While increments are independent, (log) prices have the following correlation structure: $\text{corr}(z^t, z^{t'}) = \sqrt{\min(t,t')/\max(t,t')}$. The aggregate correlation structure (across time and asset) can be conveniently represented via the Kronecker (matrix) product:

$$(\sqrt{\min(t,t')/\max(t,t')}) \otimes (\rho_{ij}) \tag{7.161}$$

(In other words, the matrix structure has the form of blocks of cross-commodity correlations for a particular pair of time horizons.) It is not hard to show that the Cholesky decomposition of a Kronecker product is the Kronecker product of the constituent factors.[67] (The temporal correlation matrix has a particularly simply Cholesky factorization: $(1/\sqrt{\max(t,t')})$ for the lower triangular form.) Thus, paths can be simulated either by directly iterating (7.160) or by imposing the aggregate correlation structure in (7.161). As a general rule, there is no compelling advantage of one approach over the other. However, there is a case for preferring the former approach to the latter. It often proves useful for facilitating the convergence of a simulation procedure by adjusting the simulated paths to match certain known properties of the underlying distribution. Typically, this involves moment matching of some sort. A suitable approach is an affine transformation at the level of the normal deviates, *e.g.*, $z \to A(z - \langle z \rangle)$, where brackets denote sample averages, and the matrix $A$ is chosen so that the sample covariance matrix matches a specified covariance matrix $\Sigma = CC^T$. This is achieved by taking $A = C\hat{C}^{-1}$ where $\hat{C}$ is the Cholesky factor of the sample covariance matrix $\langle (z - \langle z \rangle)(z - \langle z \rangle)^T \rangle$. (By construction the sample average is zero.)

Another approach that proves useful in many applications involves the so-called Brownian bridge construction. The Brownian bridge is a Brownian motion conditioned to be 0 at some terminal endpoint. There are several different constructions of the Brownian bridge that are well known in the literature; the construction of interest here entails a linear interpolation. For the general case this can be represented as

$$B_t^0 \sim N\left(B_{t_1} + \frac{t - t_1}{t_2 - t_1}(B_{t_2} - B_{t_1}), \frac{(t_2 - t)(t - t_1)}{t_2 - t_1}\right) \tag{7.162}$$

where $B$ is a Brownian motion, $B^\circ$ is a Brownian bridge, and $t_1 \le t \le t_2$ The result in (7.162) provides a means of simulating paths of a Brownian motion. The problem is broken up into increasingly finer levels of resolution. Consider a terminal time $T$. We have $B_T = \sqrt{T} \cdot z$ with $z \sim N(0,1)$. Then we note that by interpolating between $B_0 = 0$ and $B_T$ via (7.162), an entity with the proper distribution is obtained. In

fact, the necessary covariance structure is also obtained. This observation leads to the construction of the following sequence:

$$B_{t_{2k+1}^{n+1}} = \frac{t_{k+1}^n - t_{2k+1}^{n+1}}{t_{k+1}^n - t_k^n} B_{t_k^n} + \frac{t_{2k+1}^{n+1} - t_k^n}{t_{k+1}^n - t_k^n} B_{t_{k+1}^n} + \sqrt{\frac{(t_{k+1}^n - t_{2k+1}^{n+1})(t_{2k+1}^{n+1} - t_k^n)}{t_{k+1}^n - t_k^n}} \cdot z$$

$$B_{t_{2k}^{n+1}} = B_{t_k^{n+1}}, B_{t_{2k+2}^{n+1}} = B_{t_{k+1}^n} \tag{7.163}$$

for some sequence of points satisfying $t_{2k}^{n+1} = t_k^n < t_{2k+1}^{n+1} < t_{k+1}^n = t_{2k+1}^{n+1}$, and with $z$ drawn from a sequence of independent unit normals. The result in (7.163) can be readily derived by induction, by first calculating the characteristic function of $B_{2k+1}^{n+1}$ conditional on information at resolution level $n$, then further conditioning on $B_{t_k^n}$. The most convenient approach to take in practice is to successively bisect the time horizon in question, so that first $B_T$ is simulated, then $B_{T/2}$ is constructed, then $B_{T/4}$ and $B_{3T/4}$, *etc.*; see the algorithm in Glasserman (2004). As with the prior approaches for path generation, there is no real computational or statistical advantage to employing (7.163). The main advantage comes when considering alternatives to Monte Carlo simulation, and in fact relates to our ubiquitous theme of time scales. The multi-resolution aspect of the Brownian bridge involves the construction of components operating on coarser/broader time scales, with increasingly finer time scales being filled out each step, in contrast to typical time-stepping methods. This multilevel structure can often be exploited to facilitate variance reduction, since the most important drivers can often be more readily identified, if not isolated. Note that in higher dimensions the algorithm in (7.163) can be suitably modified by taking $z \to C_z$ where $C$ is some factorization satisfying $CC^T = \Sigma$ for the desired correlation structure. We have already seen in the subsection on quadrature that different choices for this factorization can produce different convergence results. This is also the case here, where the dimension of the main drivers, so to speak, of a problem can be best extracted from specific representations. (An example would be familiar PCA, for transforming data such that most of the underlying variance is concentrated in a few drivers.)

We will further consider these issues in the section on quasi-Monte Carlo. Our next topic continues with a central theme of this work, namely, the connection of the so-called greeks (delta, gamma, *etc.*) to valuation, specifically as a means of both extracting value (through hedging) and quantifying the (residual) exposure induced by the value function. Simulation as a computational tool is only useful for a particular valuation problem to the extent that the necessary greeks can also be obtained. We now demonstrate effective techniques for carrying out these calculations.

*7.5.2.5 Application: likelihood ratio for greeks*

As we have emphasized throughout, the greeks/sensitivities for any valuation problem are not just theoretical niceties, but are critical to the meaning and extraction of value in the first place. Computation of these greeks is thus of central concern to any computational methodology. A common approach to determining greeks is simply finite differencing: the sensitivity to a given parameter (*e.g.*, volatility) is obtained by central differences, by first reevaluating the problem with the parameter bumped up by some small value (either additively or multiplicatively), then reevaluating again with the parameter bumped down by the same amount. The numerical derivative is then the approximation to the desired greek. This method is certainly straightforward, and there is often no other feasible approach to the problem. However, it suffers from a number of drawbacks that would preferably be avoided.

First, inasmuch as any non-analytical (*i.e.*, numerical) result is an approximation (however good), numerical differentiation necessarily adds another layer of approximation (*i.e.*, error) to the problem. For some higher-order greeks (such as gamma), taking too small a perturbation size can result in numerical instabilities, especially when simulation is used. Second, for simulation-based methods great care must be exercised in ensuring that *only* the entity of interest is varied between paths (although often this is simply a case of ensuring that the same set of paths are used for each bump). Finally, the underlying simulation may be computationally costly, and having to calculate multiple perturbations (*e.g.*, for each price along a given forward curve) may become rather burdensome. In general it is useful to have a means of calculating important greeks without having to make recourse to finite differences.

The so-called likelihood ratio method (so named because it involves expressions often encountered in common likelihood tests) is one such method. We will consider the general exposition in the Gaussian case. The basic valuation integral (in $N$–dimensions) is

$$V = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \int_{-\infty}^{\infty} dx \cdot e^{-x^T \Sigma^{-1} x/2} H(\exp(\mu + \sigma \circ x)) \tag{7.164}$$

where $\circ$ denotes the Hadamard product (element-by-element matrix/vector multiplication). We have baked time dependence into the volatilities $\sigma$, and $\mu = \log F - \frac{1}{2}\sigma \circ \sigma$ in terms of the inception prices $F$. Let $C = \Sigma^{-1}$ be the inverse of the correlation matrix and denote any normalization constant by $D$. Introduce the variable $\xi = \mu + \sigma \circ x$ and write the valuation problem as

$$V = \frac{D^{-1}}{\sigma_1 \cdots \sigma_N} \int_{-\infty}^{\infty} d\xi \exp\left( -\frac{1}{2} \sum_{ij} \frac{C_{ij}(\xi_i - \mu_i)(\xi_j - \mu_j)}{\sigma_i \sigma_j} \right) H(e^{\xi}) \tag{7.165}$$

Thus we have (using $\frac{\partial}{\partial F} = \frac{1}{F}\frac{\partial}{\partial \mu}$ and symmetry of $C$)

$$\frac{\partial V}{\partial F_k} = \frac{D^{-1}}{F_k \sigma_1 \cdots \sigma_N} \int_{-\infty}^{\infty} d\xi \exp\left(-\frac{1}{2}\sum_{ij} \frac{C_{ij}(\xi_i - \mu_i)(\xi_j - \mu_j)}{\sigma_i \sigma_j}\right) H(e^{\xi}) \sum_j \frac{C_{kj}(\xi_j - \mu_j)}{\sigma_k \sigma_j} \tag{7.166}$$

so that upon re-substituting for $x$ we get

$$\frac{\partial V}{\partial F_k} = \frac{D^{-1}}{F_k \sigma_k} \int_{-\infty}^{\infty} dx \exp(-\tfrac{1}{2} x^T C x) H(\exp(\mu + \sigma \circ x))(Cx)_k \tag{7.167}$$

In other words, the vector of deltas are obtained by taking the expectation of the payoff function times the vector $Cx$. Similarly, gamma is given by

$$\frac{\partial^2 V}{\partial F_k \partial F_{k'}} = -\frac{\Delta_k}{F_k}\delta_{kk'} - \frac{C_{kk'} V}{F_k F_{k'} \sigma_k \sigma_{k'}} + \frac{D^{-1}}{F_k F_{k'} \sigma_k \sigma_{k'}} \int_{-\infty}^{\infty} dx \exp\left(-\frac{1}{2} x^T C x\right) H(\exp(\mu + \sigma \circ x))(Cx)_k (Cx)_{k'} \tag{7.168}$$

For the (forward) vega[68] we have

$$\frac{\partial V}{\partial \sigma_k} = -\frac{D^{-1}}{\sigma_k \sigma_1 \cdots \sigma_N} \int_{-\infty}^{\infty} d\xi \exp\left(-\frac{1}{2}\sum_{ij} \frac{C_{ij}(\xi_i - \mu_i)(\xi_j - \mu_j)}{\sigma_i \sigma_j} H(e^{\xi}) + \frac{D^{-1}}{\sigma_1 \cdots \sigma_N} \int_{-\infty}^{\infty} d\xi \exp(-\Xi) H(e^{\xi}) \left(\sum_j \frac{C_{kj}(\xi_k - \mu_k)(\xi_j - \mu_j)}{\sigma_k^2 \sigma_j} - \sum_j \frac{C_{kj}(\xi_j - \mu_j)}{\sigma_j}\right)\right) \tag{7.169}$$

(where the meaning of the variable $\Xi$ should be clear) so that

$$\frac{\partial V}{\partial \sigma_k} = -\frac{D^{-1}}{\sigma_k} \int_{-\infty}^{\infty} d\xi \exp\left(-\frac{1}{2}x^T Cx\right) H(\exp(\mu + \sigma \circ x))(-1 + (x_k - \sigma_k)(Cx)_k) \tag{7.170}$$

Now, in general we will need the correlation greeks as well (because there may be dependencies fixed through a heat rate volatility that will manifest themselves through correlation sensitivities). To get these, note that

$$\begin{aligned} C\Sigma &= I \\ C_\theta \Sigma + C\Sigma_\theta &= 0 \\ C_\theta &= -C\Sigma_\theta C \end{aligned} \tag{7.171}$$

and

$$\begin{aligned} \det(\Sigma + d\theta\, \Sigma_\theta) &= \det(\Sigma)\det(I + d\theta\ C\Sigma_\theta) \\ &= \det(\Sigma)(1 + d\theta\ \mathrm{Tr}(C\Sigma_\theta)) + o(d\theta) \end{aligned} \tag{7.172}$$

so that

$$\begin{aligned} \frac{\partial \det(\Sigma)}{\partial \theta} &= \det(\Sigma)\mathrm{Tr}(C\Sigma_\theta) \\ \frac{\partial \det(\Sigma)^{-1/2}}{\partial \theta} &= -\frac{1}{2}\det(\Sigma)^{-1/2}\mathrm{Tr}(C\Sigma_\theta) \end{aligned} \tag{7.173}$$

Finally, we see that

$$\frac{\partial V}{\partial \rho_{kl}} = D^{-1} \int_{-\infty}^{\infty} dx \exp\left(-\frac{1}{2}x^T Cx\right) H(\exp(\mu + \sigma \circ x))\frac{1}{2}(x^T C\Sigma_{\rho kl} Cx - \mathrm{Tr}(C\Sigma_{\rho kl})) \tag{7.174}$$

To understand the effect of the sensitivity correlation matrix, note that (using the 3 asset case for illustrative purposes)

$$\Sigma_{\rho_{12}} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{7.175}$$

and it can be readily seen that the effect of multiplication is simply to filter out any elements of the multiplied matrix or vector except those with indices 1 or 2. Thus,

$$\frac{\partial V}{\partial \rho_{kl}} = D^{-1} \int_{-\infty}^{\infty} dx \exp\left(-\frac{1}{2} x^T C x\right) H(\exp(\mu + \sigma \circ x))((Cx)_k (Cx)_l - C_{kl}) \quad (7.176)$$

Note that there is in fact a relationship between the gammas and vegas.[69] Recall (3.56), which reproduce here with suitably modified notation:

$$\begin{aligned} V_{\rho_{kl}} &= F_k F_l \sigma_k \sigma_l \Gamma_{kl} \\ V_{\sigma_k} &= F_k \sum_l \rho_{kl} \sigma_l F_l \Gamma_{kl} \end{aligned} \quad (7.177)$$

The first equation in (7.177) follows immediately from (7.176) and (7.168). The second equation is straightforward to derive using the result for gamma (simply cross multiply by the correlation matrix), and we leave it as an exercise for the reader. The obvious point here is the computational benefit offered by (7.177): once gammas are in hand, vegas (including correlation sensitivities) fall out at once.

Now, although these results are formally rather nice, they suffer from a serious problem in practice: they have poor convergence properties, precisely due to the various factors $Cx$ introduced in the calculation. The reason is as follows. The likelihood ratio method essentially takes a derivative of the logarithm of the underlying (parameterized) probability density, call it $\Pr_\theta$. The resulting derivative is the weighting factor introduced in the various expressions above. Since by definition $\int dx \Pr_\theta(x) = 1$, it follows that $\int dx \frac{\partial}{\partial \theta} \Pr_\theta(x) = \int dx \Pr_\theta(x) \frac{\partial}{\partial \theta} \log \Pr_\theta(x) = 0$ and thus the expectation of the weighting factor is 0 (see Capriotti, 2008)[70]. (Note that the vector factor $Cx$ in, *e.g.*, (7.167) clearly has zero mean.) What this means is that ensemble-based approximations will (potentially) involve *both* positive and negative terms in the (sample) average, contributing cancelations and hence noise (*i.e.*, additional variance) to the calculation. Figure 7.9 compares likelihood ratio results against straightforward finite differencing, for a standard (ATM) call. As can be seen, finite difference clearly outperforms the likelihood ratio.

#### *7.5.2.6 Illustration: spread option calculations*

This defect highlights the need for a good control variate. We will illustrate with a standard spread option with nonzero strike. A suitable control variate is the corresponding zero strike structure, which of course can be evaluated analytically via the Margrabe formula. Figure 7.10 compares the results of the delta calculation using likelihood ratio only, likelihood in conjunction with the (Margrabe) control variate, and finite differencing.

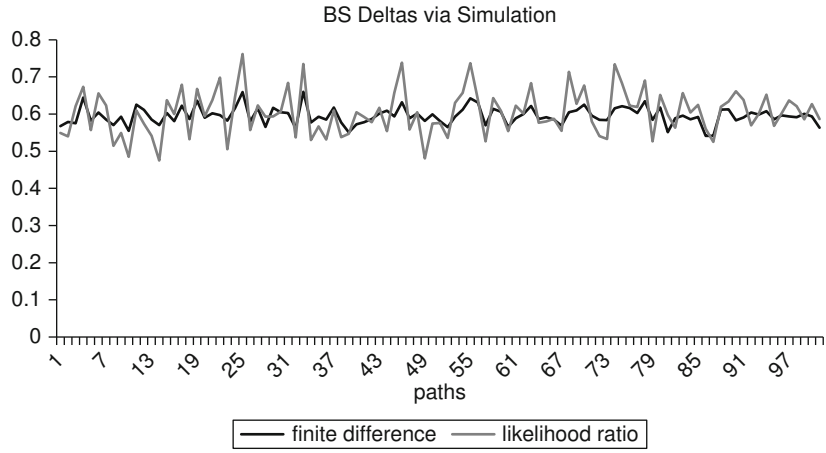



**Figure 7.9** Delta calculations

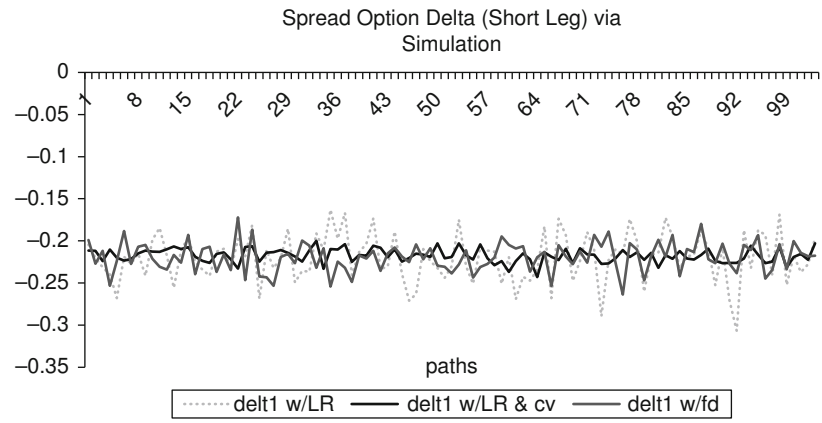



**Figure 7.10** Comparison of greek calculations via simulation

Visually it is clear that the likelihood ratio combined with a control variate yields superior results. (Although it is not immediately clear from the graph, finite differencing outperforms likelihood ratio only, just as it did in the Black-Scholes example shown in Figure 7.9.) These examples serve to drive home a larger point: simulation is *only* a computational device, and like any tool it can be misappropriated, or made more efficient in collaboration with other computational foils. Ultimately it remains just a means to an end. The idea that good greeks can be produced from raw simulations is doubly problematic. First, the value that those greeks (as actual hedge volumes) are meant to extract must be determined by a valuation framework *prior* to any simulation. The simulation itself provides no information regarding

the essence of this value (as related to a replication strategy). Second, if simulation is the chosen means for conducting the calculation in question (and no doubt, oftentimes simulations is the only feasible means), great care must be employed in the implementation. Noisy greeks are worse than useless.

This question of efficiency and convergence brings us to our next topic.

### 7.5.3 Quasi-Monte Carlo

We know from appeals to the Central Limit Theorem that the rate of convergence of Monte Carlo schemes (at heart, pseudo-random number generation) is $O(\frac{1}{\sqrt{N}})$, where $N$ is the number of simulations. Not only is this rate rather slow, but standard Monte Carlo suffers from an additional inefficiency. The essence of any quadrature scheme is to discretize, so to speak, the unit hypercube.[71] By construction, Monte Carlo schemes fill out the unit hypercube independently (memoryless would probably be a better characterization). That is, as more points are generated (thus presumably facilitating the accuracy of the computation), there is no relationship to the points previously generated. In particular, it is possible that a newly generated point is nearby previously generated points. This effect manifests itself in the well-known clustering[72] of simulated points, which can easily be seen in a two-dimensional scatterplot. This can lead to inefficiencies (for a particular set of simulated points) in the sense that the integrand is effectively and needlessly resampled. Conversely, gaps or empty regions left unfilled lead to undersampling. Standard quadrature schemes avoid these problems by uniformly discretizing the integration region. However, such schemes are prohibitively expensive due to the exponentially growing cost with the number of dimensions. There is also no way of adding a new quadrature point without completely reconstructing the quadrature grid (and of course completely resampling the integrand). It is desirable to have a method that retains the uniformity of standard discretization while permitting easy addition of new points.

#### *7.5.3.1 Low-discrepancy sequences*

This objective is accomplished, to a large degree, with so-called quasi-Monte Carlo schemes. These schemes generate a (multidimensional) sequence that fills out the unit hypercube more uniformly than standard (pseudo-random) Monte Carlo. The basis of such methods is so-called low-discrepancy sequences, which (not surprisingly) are sequences for which any finite subsequence has low discrepancy. In the event that this description is not helpful, we provide the following definition of the discrepancy of a $d$-dimensional set $S = \{x_1, \ldots, x_N\}$ with $x_i \in [0,1]^d$:

$$D_N(S) = \sup_{B \in I} \left| \frac{\#(x_i \in B)}{N} - V(B) \right| \tag{7.178}$$

where $I$ is the set of $d$-dimensional boxes within the unit hypercube and $V$ is the volume of a box. Intuitively, (7.178) provides a measure of how uniformly well (or poorly) the set fills any region of the unit hypercube. The discrepancy (or more accurately, its close cousin the star discrepancy) of a set is used (via the Koksma-Hlawka inequality) to bound the error of approximations such as (7.147) to integrals of the form (7.145), along with a measure of the variation of the integrand (roughly speaking, its smoothness). These bounds are often not very useful in practice, because apart from being difficult to actually calculate, they can be infinite, as well. For many popular choices of low-discrepancy sequences (to be listed below), the discrepancy has the following order of magnitude:

$$O\left(\frac{\log^d N}{N}\right) \tag{7.179}$$

Although $\log N$ grows much slower than $N$, the seemingly impressive rate of convergence in (7.179) (in comparison to $O(N^{-1/2})$ for standard Monte Carlo) must be tempered by the fact that as the dimension $d$ grows, quasi-Monte Carlo loses its effectiveness. Furthermore, the ability of low-discrepancy sequences to fill in the unit hypersphere more uniformly than (pseudo-) Monte Carlo is a result of the fact that there *is* an effective correlation (or memory, for lack of a better term) between successive iterates in the sequence. (Put differently, pseudo-random schemes fill the unit hypersphere indirectly, while quasi-Monte Carlo methods fill it directly.) This fact makes quasi-Monte Carlo unsuitable for problems requiring path generation; it is essentially a quadrature scheme at heart. However, since a great many problems of interest *can* be crafted as expectations/quadratures, quasi-Monte Carlo is a viable option in many cases. This is especially true when used in conjunction with so-called scrambling techniques that overcome the aforementioned deficiencies in higher dimensions.

#### *7.5.3.2 Specific examples and constructs*

Very useful expositions on low discrepancy sequences and their applications can be found in Caflisch (1998) and Glasserman (2004). Here we will briefly outline the construction of some of the more popular sequences, and give examples of their effectiveness in energy market applications. We start with the van der Corput sequence. This sequence simply reverses the representation of the integers under some base $b$:

$$n = \sum_{j=0}^{m} a_j(n) b^j \Rightarrow \Phi_b(n) = \sum_{j=0}^{m} a_j(n) b^{-j-1} \tag{7.180}$$

The Halton sequence is a natural generalization of the van der Corput sequence to higher dimensions:

$$h(n) = (\Phi_{b_1}(n), \ldots, \Phi_{b_d}(n)) \tag{7.181}$$

with $b_i$ relatively prime. Most modern approaches are based on the work of Niederreiter (1988), who introduced the terms $(t, m, s)$ -nets and $(t, s)$ -sequences. A very popular special case is the so-called Sobol' sequence, which is constructed as follows (see Joe and Kuo, 2008b). In $d$ dimensions, the $j^{\text{th}}$ component of a Sobol' sequence is generated by starting with a primitive polynomial:

$$x^{s_j} + a_{1,j}x^{s_j-1} + a_{2,j}x^{s_j-2} + \cdots a_{s_j-1,j}x + 1 \tag{7.182}$$

where the coefficients $a_{k,j}$ are either 0 or 1. Next, introduce the following recursion relationship:

$$\begin{aligned} m_{k,j} &= 2a_{1,j}m_{k-1,j} \otimes 2^2 a_{2,j}m_{k-2,j} \otimes \cdots \otimes \\ & 2^{s_j-1}a_{s_j-1,j}m_{k-s_j+1,j} \otimes 2^{s_j} m_{k-s_j,j} \otimes m_{k-s_j,j} \end{aligned} \tag{7.183}$$

for some positive integers $m_{k,j}$, and where $\oplus$ is the bitwise exclusive-or (XOR) operator.[73] Then, a set of *direction numbers* is constructed from $v_{k,j} = \frac{m_{k,j}}{2^k}$, from which the desired Sobol' component is given by

$$x_{i,j} = i_1 v_{1,j} \otimes i_2 v_{2,j} \otimes \cdots \tag{7.184}$$

where the binary form of $i$ is given by $i = (\cdots i_3 i_2 i_1)_2$. This discussion should serve as an outline of the main idea; actual implementations (via Gray code in C/C++) can be found in Press *et al.* 2007 or Joe and Kuo (2008b).

#### *7.5.3.3 Issues in high dimensions*

One rather serious drawback to using low-discrepancy sequences is their tendency to display strange patterns in higher dimensions; effectively, they start clustering along hyperplanes and so begin to fill the entire space very inefficiently. Again, this is an inherent feature of the construction of these sequences, namely an effective autocorrelation. We illustrate this phenomenon in Figure 7.11, using a straightforward implementation (specifically, the algorithm from Press *et al.* [2007]).

There are a few ways of addressing this issue. So-called scrambling techniques essentially randomize the binary representation of the Sobol' iterates; see Chi *et al.* (2005). Alternatively, a judicious choice of initial seed,[74] so to speak, can produce Sobol' sets without clustering. Figure 7.12 shows the results from the algorithm of Joe and Kuo (2008b).

It bears repeating that simulation techniques are ultimately quadrature methods. Their chief utility comes in applications to high-dimensional problems. We provide an illustration in Figure 7.13. We price again an option with payoff equal to the maximum of four lognormal assets, which are taken as independent with unit variances. (An "exact" numerical value is obtained from the methods used in Section

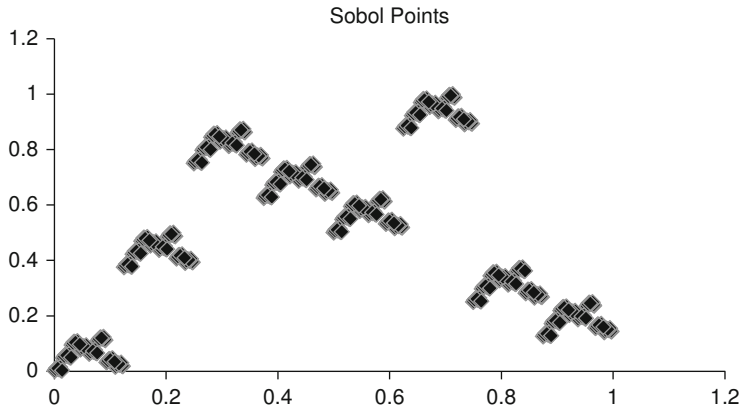



**Figure 7.11** Clustering of Sobol' points

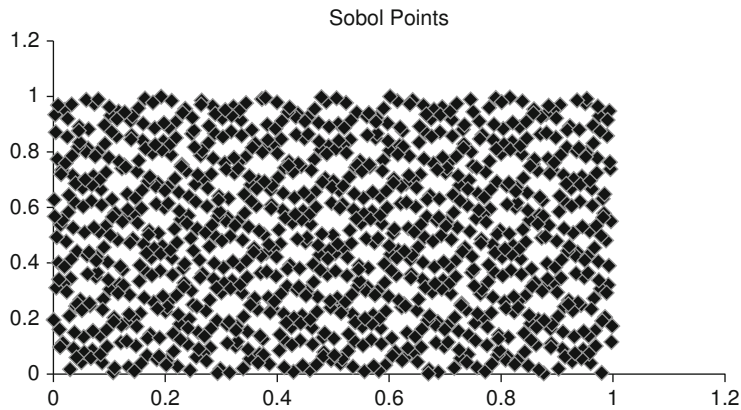



**Figure 7.12** Sobol' points with suitably chosen seed

7.4.2.) It can be seen that, while both quasi- and pseudo-Monte Carlo techniques converge for sufficiently large numbers of realizations, quasi-Monte Carlo (Sobol' in this case) does so more uniformly. This is of course a reflection of the nature of the method, to fill out the unit hypercube more regularly than pseudo-Monte Carlo (*i.e.*, random number generation). Both approaches, however, inherently perform the same task: quadrature.

We now turn attention to an important topic in energy markets with these features, namely control problems.

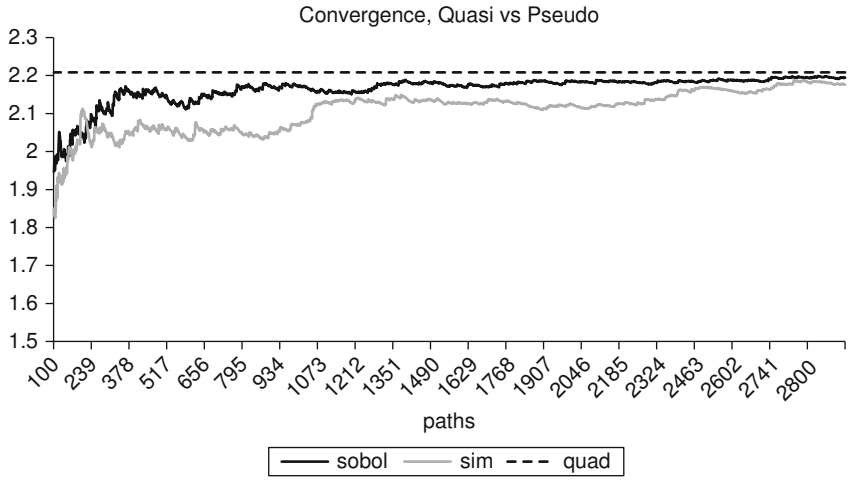



**Figure 7.13** Convergence of quasi- and pseudo-Monte Carlo

## 7.6 Stochastic control and dynamic programming

An important class of problems encountered in energy markets involves optimal control of some stochastic process. This situation arises because there are many structures whose value depends on some (operational) state of the underlying (physical) system, and the decision to induce a change in this state depends on the dynamic, stochastic behavior of some underlying price system. Typical examples (which have been discussed elsewhere in this work) are tolling and natural gas storage. In the former, the decision to start up or shut down a power plant depends on spark spreads now, plus the expected value of the future toll given a change in the system state (*i.e.*, the plant turned on or off). In the latter, a decision to inject or withdraw fuel depends on the price of gas now plus the expected value of future storage given a system change (*i.e.*, capacity increased or reduced). In these problems, there is an apparent paradox in that the value depends on the decision taken now, but the decision to take now depends on the value!

Of course, there is no paradox, and we have in fact discussed these kinds of problems already in several places, most notably Section 3.3. In fact, even in this chapter we have presented methods of solution (*e.g.*, Section 7.3.2). We will briefly review the main ideas before focusing on another numerical approach that ties in with the simulation-based techniques that we have just presented. The major theme of this section should be stated quite clearly: control problems are quite challenging computationally. It is unavoidable that some sort of approximate solution method must be resorted to, which by nature provides a lower bound on value. The imperative

question then becomes: how good is this value? Put differently, how much *more* value is there? As we will see, a powerful set of techniques already encountered in Chapter 3 (duality) provide a feasible means of answering these questions.

### 7.6.1 Hamilton-Jacobi-Bellman equation

As already noted, the relevant material here has already been covered in Section 3.3.1, so we will only present the pertinent results. A general representation of the kinds of problems of interest is captured by (3.97):

$$V(S_t, A_t, t) = \sup_a E_t \left( \int_t^T e^{-r(s-t)} g(S_s, A_s; a_s) ds + e^{-r(T-t)} f(S_T, A_T) \right),$$
$$A \in \mathfrak{A}, a \in \mathfrak{a} \tag{7.185}$$

over some finite time horizon. The meaning of (7.185) is that the value $V$ (dependent on prices $S$ and some system state $A$) is the supremum taken over all (adapted) actions ($a$) of the expected value accruing from some terminal payoff $f$ and accumulated gains/losses $g$. Bellman's principle of optimality states the following. For Markov decision processes (almost always the typical case), the optimal present decision consists of the action that optimizes any immediate payoff plus the expected value of the objective function over the remaining horizon, *given* the present action. This principle leads to the following ensembles for numerically solving (7.185) (recall again Section 3.3.1):

$$V(S_t, A_t, t) = \max_a (g(S_t, A; a)\Delta t + e^{-r\Delta t} E_t V(S_{t+\Delta t}, A + a\Delta t, t + \Delta t))$$
$$V_t + \tfrac{1}{2}\mathfrak{L}V + \max_a (g(S, A; a) + \mu^T V_S + aV_A) - rV = 0 \tag{7.186}$$

The second equation in (7.186) is commonly referred to as the Hamilton-Jacobi-Bellman equation, although the first equation (a precursor in discrete time) is probably more commonly employed as a solution technique.

In fact, we have already seen implementations of this discrete-time calculation, namely, the Grid Model of Section 7.3.2. The feasibility of this approach hinges critically on the ability to efficiently carry out the underlying expectations in (7.186).

### 7.6.2 Dual approaches

We can now contrast direct methods for solving (7.186) with so-called dual methods. Again, we will be brief, as this material has already appeared in Section 3.3.2. As noted above, direct methods seek solutions to (7.186) by optimizing across

the space of actions/decisions/controls (or equivalently, across the space of state changes). An alternative approach is to frame the problem in such a manner that optimization is carried out over the space of candidate value functions. To see this, recall the analysis in Section 3.3.2. The standard American option problem (in terms of stopping times) satisfies the following relations:

$$\begin{aligned} V(S_t,t) &= \sup_{t\le\tau\le T} E_t(e^{-r(\tau-t)}f(S_\tau)) = \sup_{t\le\tau\le T} E_t(e^{-r(\tau-t)}f(S_\tau)-\pi_\tau+\pi_\tau) \\ &\le \sup_{t\le\tau\le T} E_t(e^{-r(\tau-t)}f(S_\tau)-\pi_\tau)+\pi_t \le E_t \max_{t\le\tau\le T}(e^{-r(\tau-t)}f(S_\tau)-\pi_\tau) \\ &\quad +\pi_t \Rightarrow \cdot V(S_t,t) \le \inf_{\pi}\left\{E_t \max_{t\le\tau\le T}(e^{-r(\tau-t)}f(S_\tau)-\pi_\tau)+\pi_t\right\} \end{aligned} \tag{7.187}$$

where $\pi$ represents a supermartingale (*i.e.*, a process satisfying $\pi_t \ge E_t\pi_T$). In particular, since the (discounted) value function is itself a martingale, (7.187) implies that the closer a candidate supermartingale is to the true value process, the tighter the upper bound is. (The difference between an upper- and lower-bound valuation via (7.187) is commonly referred to as the *duality gap*.) In fact, this latter point is worth emphasizing: (7.187) provides an *upper* bound, in contrast to many other methods that, by their nature, can only provide lower bounds. Let us now consider some of these methods, before turning to how duality-based upper bounds can be used in conjunction with simulation-based valuation methods.

### 7.6.3 LSQ

#### *7.6.3.1 Challenges for dynamic programming*

Let us consider the standard backward induction formulation of an American derivative in discrete time (so that it is really Bermudan in nature), in terms of the comparison between immediate payoff and the so-called continuation value (*i.e.*, the value of holding or non-exercise):[75]

$$V_n(S_i) = \max(P(S_i), E_{n,i}V_{n+1}(S)) \tag{7.188}$$

where $P$ is some payoff function, and at some terminal time $T$ (the starting point of the recursion) we have that $V_T(S) = P(S)$. In (7.188), the subscript $n, i$ in the conditional expectation operator is meant to emphasize the dependence on the time $n$ asset value $S_i$. This latter point highlights the difficulty in applying simulation-based methods to problems of this sort. Simulations typically involve the generation of paths from some initial point. But, to carry out the backward induction in (7.188), conditional expectations (of the next-step value function) must be calculated at *each* point on a path. These expectations could similarly be computed via simulation, however, it is precisely *this* necessary information that the typical simulation does

not provide (because they are not generated from intermediate points, only from the initial point). What is needed, in essence, is a "simulation within a simulation," so to speak. While grid- or tree-based methods are (reasonably) well suited for this task (because the connection between grids/trees of neighboring time steps via transition densities can often be efficiently exploited), simulation is computationally infeasible here.

*7.6.3.2 Regression-based approximations*

An alternative approach is the following, broadly known as Least Squares Monte Carlo (LSQ). First, recall the modern notion of conditional expectation as a projection onto a sub-sigma algebra of the underlying probability space. This framework clearly has affinities with the regression techniques encountered in Chapter 2, and we are accordingly led to represent conditional expectation as an appropriate expansion in terms of some set of regressors.[76] That is, we look for an approximation (to the continuation value) of the following form:

$$C_n(z_n^i) \equiv E_{n,i} V_{n+1}(z) \approx \sum_j \beta_j \beta_j(z_n^i) \equiv \hat{C}_n(z_n^i) \tag{7.189}$$

for some set of basis functions $B_j$. In (7.189), superscripts on $z$ denote path indices and subscripts denote time steps. The coefficients are computed from a standard regression across paths of the time $n+1$ value function on the basis functions evaluated along the time $n$ path variables. There is typically little *a priori* guidance for how to best choose these basis functions for a given problem, and in practice the choice is fairly *ad hoc*; typical examples include:

- Monomials: $1, z, z^2, \ldots$
- Indicator and related functions: $H(z), z^+, \ldots$, and powers thereof
- Special functions, *e.g.*, Hermite or Laguerre polynomials
- Tensor products of the above choices in higher dimensions.

The pioneering work on this technique was done by Longstaff and Schwartz (2001) and Tstitsiklis and van Roy (2001).[77] (These will henceforth be denoted by LS and TR, respectively.) It is worth briefly noting the subtle difference between the approaches adopted in these two works. The essential point concerns how the continuation value is propagated in the backward induction. First, note that (7.188) can be written as a recursion on continuation value via

$$\begin{aligned} C_n(z_n) &= E_n V_{n+1}(z_{n+1}) = E_n \max(P(z_{n+1}), E_{n+1} V_{n+2}(z_{n+2})) \\ &= E_n \max(p(z_{n+1}), C_{n+1}(z_{n+1})) \end{aligned} \tag{7.190}$$

with the convention $C_T = 0$. (The setup in (7.190) is sometimes referred to as $Q$ value iteration.) In the TR approach, the backward induction is implemented as

$$V_n(Z_n) = \max(P(z_n), \hat{C}_n(z_n)) \tag{7.191}$$

whereas in LS it is performed as

$$V_n(z_n) = P(z_n)1(P(z_n) > \hat{C}_n(z_n)) + V_{n+1}(z_n)1(\hat{C}_n(z_n) > P(z_n)) \tag{7.192}$$

It can thus be seen that TR and LS can be broadly characterized as value function approximations and stopping time approximations, respectively. It is probably not inaccurate to say that LS is the more popular technique in practice, although there are certainly arguments in favor of using TR. The TR formulation is continuous, whereas LS is discontinuous, which may impact pathwise sensitivities for calculations of greeks. However, TR is potentially more sensitive to the propagation of error across time steps, which is generally less of an issue for LS (since the approximation to the continuation value is only used to decide on the exercise boundary). See Stentoft (2012) for a recent numerical study.

*7.6.3.3 A suitable choice of regressors*

In truth, the choice of the particular methodology for employing the continuation value is of less importance than having good bounds on whatever particular value is produced, and we will investigate this problem shortly. (We tend to find TR to be the more convenient framework for this objective, but this preference is largely of little impact.) With this in mind, we will now indicate a particular choice of indicator functions in (7.189) that have an intuitive interpretation and are simple to implement. First note how the coefficients are calculated:

$$\sum_i B_k(z_n^i) V(z_{n+1}^i) = \sum_j \beta_j \sum_i B_j(z_n^i) B_k(z_n^i) \tag{7.193}$$

In matrix notation, (7.193) can be written as $\langle BB^T \rangle \beta = \langle BV \rangle$, where brackets denote pathwise summations. A very convenient choice here would be an "orthonormal" basis, *i.e.*, $\forall z$ there is only *one* index $j$ such that

$$B_j(z) = 1 \tag{7.194}$$

and zero otherwise. That is, if (7.194) holds for some value of $j$, then we have that $B_k(z) = 0$ for $k \neq j$. The coefficients are then given by

$$\beta_k = \frac{\sum_i B_k(z_n^i) V(z_{n+1}^i)}{\sum_i B_k(z_n^i)} \tag{7.195}$$

Note that this expression has the intuitive interpretation of an average of the value function across those paths that are in some sense "similar" to the path upon which the conditioning is based. It is basically equivalent to kernel-based estimation.[78] Heuristically, we can write (7.189) as

$$E_{z_n^j} V(z_{n+1}) \approx \frac{\sum_i V(z_{n+1}^i) 1(z_n^i \sim z_n^j)}{\sum_i 1(z_n^i \sim z_n^j)} \tag{7.196}$$

where the symbol $\sim$ is meant to convey some notion of "similarity" between paths (*e.g.*, akin to a common node in a multinomial tree, albeit a "fuzzy" node, so to speak). Operationally, this notion could (in practice does) simply mean something like "approximately equal to."

For example, since we will often be considering Gaussian problems, the path variables will take the form

$$z_n^i = z_0 - \frac{1}{2}\sigma^2 t_n + \sigma\sqrt{t_n}\phi_n^i \tag{7.197}$$

with $\phi_n^i \sim N(0,1)$. Thus, there is a simple transformation relating the path variables to standard normal deviates, and the notion of "similarity" driving the basis functions can be applied at the level of the generated random deviates. The exercise can thus be treated as an issue of indexing paths appropriately and averaging over paths with a common index. Since this indexing can be handled outside of any loops that handle the backward induction central to control problems, the actual calculation of the expected value is greatly facilitated.

Continuing the example, consider the following choice:

$$E_{n,i} V(z_{n+1}) \approx \beta_1 B_1(z_n^i) + \beta_2 B_2(z_n^i) \tag{7.198}$$

where

$$\begin{aligned} B_1(z) &= 1\left(z > z_0 - \tfrac{1}{2}\sigma^2 t_n\right) \\ B_2(z) &= 1\left(z < z_0 - \tfrac{1}{2}\sigma^2 t_n\right) \end{aligned} \tag{7.199}$$

In other words, the basis functions depend on whether the underlying random deviate is positive or negative. Thus, the conditional expectation is an average across paths for which the log price (say) is above (below) its unconditional mean. Note that extensions to higher dimensions can be implemented straightforwardly via tensor products, *e.g.*,

$$B_{11}(z) \equiv B_1(z_1) \otimes B_1(z_2) = 1\left(z_1 > z_{10} - \tfrac{1}{2}\sigma_1^2 t_n, z_2 > z_{20} - \tfrac{1}{2}\sigma_2^2 t_n\right) \tag{7.200}$$

In general, there will be no definite relationship (in an ordinal sense) between regression-based approximations such as (7.189) and the true value function. Ideally, we would like to have some sense of a lower-bound valuation that can be used (we will see) in conjunction with duality-based approaches to derive an upper-bound valuation. To attain lower bounds, LSQ typically proceeds in two stages. First, a preliminary set of (price) paths is generated that serves as the basis for the regression calculation in (7.189). Then, a secondary set of paths is generated that is used, in conjunction with the basis function coefficients computed in the preliminary stage, to carry out the backward induction of (7.188) (as implemented either via LS or TR). This two-stage process effectively means that a suboptimal exercise policy is used in the (final) procedure, and hence that the resulting valuation will be a lower bound to the true value.

For example, using (7.195) for a set of paths denoted by $\tilde{z}_t^i$, we would calculate for *another* set of paths:

$$E_{z_t} V(z_{t+1}) \approx \sum_k \beta_k B_k(z_t) = \beta_{k_z}$$
$$= \sum_i V(\tilde{z}_{n+1}^i) \frac{B_{k_z}(\tilde{z}_n^i)}{\sum_{i'} B_{k_z}(\tilde{z}_n^{i'})} = \sum_i V(\tilde{z}_{n+1}^i) \frac{1(i \sim k_z)}{\sum_{i'} 1(i' \sim k_z)} \tag{7.201}$$

where $k_z$ is the index $k$ st. $B_k(z_t) = 1$ and 0 otherwise. The notation in the last equation in (7.201) indicates that the conditional expectation of $V$ using the *new* set of paths is a weighted average of $V$ using the *old* set of paths, which are similar (in the sense of sharing the same non-zero basis function index) to the (new) point on which the conditioning is based.

#### *7.6.3.4 Assessing the basis functions/lower bound*

So we again turn to the all-important question of just how good these lower bounds are. Let us stress that our primary objective in employing techniques such as LSQ is to have an approximation of the conditional expectation in order to create a proxy value function to be used in upper-bound duality approaches. Keep in mind that, in addition to being (by construction) upper bounds, expressions such as (7.187) replace stopping time/control problems by pathwise expectations (of a look-back nature). This is an enormous advantage because, as we have seen, the former class of problems is very difficult to solve via simulation, while the latter are quite amenable to simulation. More generally, control problems often suffer from the computational burden associated with high dimensions (hence the need for approximate valuations), something for which simulation-based methods are much better suited. Duality approaches thus provide a ready means for ascertaining the

quality of a lower-bound valuation (obtained from simulation, policy approximation, *etc.*): we simply check how good or bad the proxy is by seeing how big the resulting duality gap is. We will now turn to this topic.

### 7.6.4 Duality (again)

#### *7.6.4.1 Multiple stopping times/exercise rights*

Since the kinds of problems we are typically confronted with in energy markets entail multiple exercise rights (*e.g.*, the ability to change some operational state of the system in question more than once in some time period), it is worth examining some generalizations of the basic duality result in (7.187). We first turn to work by Meinshausen and Hambly (2004) (hereafter referred to as MH), which crafts the problem in terms of *incremental* exercise value. (For context, it may help to review the formulation in (7.90).) To begin the exposition, assume the structure has $k$ exercise rights and denote the payoff upon exercise by $H_t$, a stochastic process that may have some dependence on other, more basic processes.[79] Define a policy $\pi_t$ to be a sequence of (strictly) ordered stopping times $t \leq \tau_1 < \tau_2 < \cdots < \tau_n \leq T$ over some time horizon $[t, T]$. Then the value of the structure is given by

$$V_t^k = \sup_{\pi_t} E_t^Q \sum_{n=1}^{k} H_{\tau_n} \tag{7.202}$$

under a suitable pricing measure $Q$. Confining attention to the discrete-time case and appealing to Bellman, (7.202) can be written as

$$V_t^k = \sup_{\mathrm{t} \leq \tau_1 \leq T} E_1^Q (H_{\tau_1} + V_{\tau_1+1}^{k-1}) \tag{7.203}$$

where the "+1" in the subscript refers to the minimum time between exercises. In the more familiar operational form (see again (7.90)), (7.203) can be written as

$$V_t^k = \max(H_t + E_t^Q V_{t+1}^{k-1}, E_t^Q V_{t+1}^k) \tag{7.204}$$

which has the intuitive interpretation of comparing exercise now plus the expected value of one less right with the expected value of all rights.

#### *7.6.4.2 Marginal valuation*

To further facilitate the analysis, we state the familiar Doob-Meyer decomposition (in discrete time; see Etheridge [2002]): under suitable technical conditions, a stochastic process $X$ adapted to some filtration $\mathfrak{F}$ can be written as $X_t = M_t + A_t$, where $M$ is a martingale wrt. $\mathfrak{F}$ and $A$ is a previsible process (*i.e.*, $A_{t+1}$ is

$\mathfrak{F}_t$-measureable) with $A_0 = 0$ (conventionally). If $X$ is a supermartingale, then $A$ is nonincreasing (*i.e.*, $A_{t+1} \leq A_t$). Constructively we have

$$\begin{aligned} A_{t+1} &= A_t + E_t X_{t+1} - X_t \\ M_{t+1} &= M_t + X_{t+1} - E_t X_{t+1} \end{aligned} \tag{7.205}$$

We can apply this result as follows.[80] We first introduce the *marginal* value function defined by $\Delta V_t^k \equiv V_t^k - V_t^{k-1}$.[81] Then from (7.203) we see that

$$\begin{aligned} \Delta V_t^k &= \sup_{t \leq \tau_1 < T} E_t^Q (H_{\tau_1} + V_{\tau_1+1}^{k-1}) - V_t^{k-1} \\ &= \sup_{t \leq \tau_1 < T} E_t^Q (H_{\tau_1} + V_{\tau_1}^{k-1} + E_t^Q V_{\tau_1+1}^{k-1} - V_{\tau_1}^{k-1}) - V_t^{k-1} \\ &= \sup_{t \leq \tau_1 < T} E_t^Q (H_{\tau_1} + V_{\tau_1}^{k-1} + A_{\tau_1+1}^{k-1} - A_{\tau_1}^{k-1}) - V_t^{k-1} \\ &= \sup_{t \leq \tau_1 < T} E_t^Q (H_{\tau_1} + M_{\tau_1}^{k-1} + A_{\tau_1+1}^{k-1}) - V_t^{k-1} = \sup_{t \leq \tau_1 < T} E_t^Q (H_{\tau_1} + A_{\tau_1+1}^{k-1}) - A_{\tau_1}^{k-1} \end{aligned} \tag{7.206}$$

where we introduce obvious notation for rights-dependent $A$ and $M$ from Doob-Meyer and optional stopping is invoked in the last equation in (7.206). (Note from (7.204) that $A_{t+1}^k - A_t^k = -(H_t - E_t^Q \Delta V_{t+1}^k)^+$, from which it can be proved that the optimal exercise time of the $k^{\text{th}}$ right is identical to the first time where $A_t^k$ is strictly negative.) Using induction, (7.206) can be used to prove the intuitive result that the marginal value decreases with the number of exercise rights. In fact, the marginal value is also a supermartingale (however, it is dominated by the incremental continuation value for *one less* right; *i.e.*, $E\Delta V_{t+1}^k \leq \Delta V_t^k \leq E\Delta V_{t+1}^{k-1}$).

Now, being essentially a nested series of single-stopping time problems, (7.206) forms the cornerstone of a generalization of the basic Kogan-Haugh duality result in (7.187). We defer the bulk of the details to MH and here only outline the main idea of the argument. By splitting the stopping times and using the fact that $A$ is nonincreasing, we see that

$$\begin{aligned} \Delta V_0^k &= \sup_{0 \leq \tau \leq \tau_1} E_0^Q ((H_\tau + A_{\tau+1}^{k-1}) 1_{\{\tau < \tau_1\}} + \sup_{\tau_1 \leq \tau' \leq T} E_{\tau_1}^Q (H_{\tau'} + A_{\tau'+1}^{k-1}) 1_{\{\tau = \tau_1\}}) \\ &\leq \sup_{0 \leq \tau \leq \tau_1} E_0^Q (H_\tau 1_{\{\tau < \tau_1\}} + (\sup_{\tau_1 \leq \tau' \leq T} E_{\tau_1}^Q (H_{\tau'} + A_{\tau'+1}^{k-1}) 1_{\{\tau = \tau_1\}}) \\ &= \sup_{0 \leq \tau \leq \tau_1} E_0^Q (H_\tau 1_{\{\tau < \tau_1\}} + \Delta V_\tau^k 1_{\{\tau = \tau_1\}}) \end{aligned} \tag{7.207}$$

Using the ordering relations between the marginal value process and the incremental continuation value, introducing a martingale $M$ starting at zero, and again

invoking optional stopping, we see that

$$\Delta V_0^k \leq E_0^Q \max_{0\leq\tau\leq\tau_1} (H_\tau 1_{\{\tau<\tau_1\}} + E_\tau^Q \Delta V_\tau^{k-1} 1_{\{\tau=\tau_1\}} - M_\tau) \tag{7.208}$$

so that, as usual in these duality approaches, a supremum over (stochastic) stopping times is replaced by a (pathwise) maximum over deterministic times. Obviously, (7.208) can be extended (in the by-now familiar) manner of taking infimums across martingales $M$ and stopping time $\tau_1$. In fact, MH show that the duality gap is zero, with equality being attained for the optimal stopping time for the first exercise right and the martingale component of the true value function (for the level of exercise rights under consideration). Further manipulations of (7.208) lead to the desired generalization, which we simply state here:

$$\Delta V_0^k = \inf_\pi \inf_M \left\{ E_0^Q \max_{\substack{u\in\{0,\ldots,T\}\setminus \\ \{\tau_1,\ldots,\tau_{k-1}\}}} (H_u - M_u) \right\} \tag{7.209}$$

In other words, the marginal value is the double infimum of an expectation of a pathwise maximum over a restricted set of deterministic points, taken over all stopping times and all martingales (started from zero). Note that for the case of single exercise right ($k = 1$), (7.209) reduces to the Kogan-Haugh result (7.187).

## 7.7 Complex variable techniques for characteristic function applications

In this section we will consider some numerical issues associated with some of the techniques introduced in Chapter 5, specifically methods involving the use of characteristic functions. In particular we will investigate the role complex analysis plays in connecting several underlying concepts.

### 7.7.1 Change of contour/change of measure

*7.7.1.1 Importance sampling revisited*

In this subsection we discuss how certain change-of-measure techniques for facilitating the valuation of deep out-of-the-money (OTM) probabilities in affine diffusion models can be related to changes of contours in the complex plane in the quadrature formulas for calculating these probabilities. For background, let us review the discussion in Section 7.5.2, and consider the evaluation of the following probability:

$$\Pr(z > \gamma) \tag{7.210}$$

where $z$ is a unit normal and $\gamma$ is some number significantly greater than two standard deviations, say, four. As is well known, using simulation to evaluate this probability (we ignore for exposition the fact that there is a well-known algorithm for computing the cumulative distribution function for the standard normal) will give very poor results, due to the fact that it is highly unlikely that a given random sample will produce very many outcomes in the tail of the distribution, so a very large number of draws is necessary to yield accurate results of the corresponding probability. As we will see, quadrature methods suffer from the same problem. The standard textbook solution to this problem is to complete the square in the normal density function and transform the problem into one of taking an expectation of a new function under a normal distribution with nonzero mean. However, this is actually a change-of-measure result, as can be seen here:

$$\begin{aligned}\Pr(z > \gamma) &= E^P 1(z > \gamma) \\ &= E^P e^{-\varepsilon z} e^{\varepsilon z} 1(z > \gamma) = E^P e^{-\varepsilon z} \cdot E^{P_\varepsilon} e^{\varepsilon z} 1(z > \gamma)\end{aligned} \tag{7.211}$$

where $P$ denotes a measure under which $z$ is a standard normal and the new measure $P_\varepsilon$ is defined by

$$\frac{dP_\varepsilon}{dP} = \frac{e^{-\varepsilon z}}{E^P e^{-\varepsilon z}} = e^{-\varepsilon^2/2 - \varepsilon z} \tag{7.212}$$

Under $P_\varepsilon$, it is not hard to see that $z$ is normally distributed with variance 1 and mean $-\varepsilon$ . Thus, if we choose $\varepsilon = -\gamma$ , we see that if we simulate under $P_\varepsilon$, we will not be sampling the tail of the distribution, so we should expect Monte Carlo to be much more effective. This is indeed what we see in practice; see Table 7.3 for some typical results.

#### *7.7.1.2 Convergence of quadrature*

A similar issue arises in quadrature-based methods. To see this, recall the basic result (to be generalized shortly)

$$\frac{1}{2} + \frac{1}{2\pi} \int_{-\infty}^{\infty} d\phi \frac{e^{i\gamma\phi - \phi^2/2}}{i\phi} = N(\gamma) \tag{7.213}$$

**Table 7.3** Importance sampling for calculating $\Pr(z > 3)$ for $z$ a standard normal. 2,000 paths.

| Straight sim | Sim imp samp | Exact |
|---|---|---|
| 0.00050 | 0.00141 | 0.00135 |

We can see here that for values of $\gamma$ large in absolute magnitude (*i.e.*, corresponding to the tails of the distribution), this integrand will be affected by highly oscillatory terms that can hamper numerical accuracy. This is indeed what happens when this expression is used for deep OTM probabilities. A remedy to this problem is as follows.

We consider any random process for which we have suitable (either analytical or numerical) knowledge of the characteristic function: $f(\phi) = E^P e^{i\phi z}$. Then recalling the results from Section 5.2.4, we have that

$$E1(z > \gamma) = \int_{\gamma}^{\infty} dz \frac{1}{2\pi} \int_{-\infty+i\varepsilon}^{\infty+i\varepsilon} d\phi f(\phi) e^{-i\phi z} \tag{7.214}$$

Here, in the inversion of the Fourier transform, we have shifted the contour of integration to ensure that the integrand converges (*e.g.*, we take $\varepsilon < 0$ for $\gamma > 0$[82]). This allows us to reverse the order of integration in (7.214):

$$E1(z > \gamma) = \frac{1}{2\pi} \int_{-\infty+i\varepsilon}^{\infty+i\varepsilon} d\phi f(\phi) \frac{e^{-i\phi\gamma}}{i\phi} \tag{7.215}$$

Now, the approach usually encountered in the literature is to shift the contour back onto the real axis (in $\phi-$ space), taking into account the pole at the origin:

$$E1(z > \gamma) = \frac{1}{2} + \frac{1}{\pi} \int_{-\infty}^{\infty} d\phi \mathrm{Re}\left[\frac{e^{-i\phi\gamma} f(\phi)}{i\phi}\right] \tag{7.216}$$

However, there really is no compelling reason to do so. In fact, it proves judicious to keep the integration contour in the complex plane, by choosing the offset from the origin appropriately. For example, the characteristic function for a standard normal is $e^{-\phi^2/2}$. If we choose $\varepsilon = -\gamma$ , then along this path the oscillatory terms will be completely suppressed. But note that this is precisely the choice made for the change-of-measure result used to ensure adequate sampling. Note further that this is also the choice used in familiar asymptotic methods from contour integration such as stationary phase and steepest descent (more generally, referred to as saddle-point methods), namely that the exponent have a local minimum along the chosen contour:

$$-i\gamma - \phi = 0 \tag{7.217}$$

Table 7.4 shows a comparison of results.

**Table 7.4** Quadrature methods for computing $\Pr(z > 3)$ for $z$ a standard normal. Here, "adj" denotes the contour-shifted result.

| Quad | Quad adj | Exact |
|---|---|---|
| 0.001248 | 0.00135 | 0.00135 |

*7.7.1.3 Linear extensions*

In fact, this result is quite general. Consider our usual general affine diffusion model with $P$–dynamics:

$$dz_i = (A_{ij}z_j + b_i)dt + \sigma_i^k \sqrt{z_k} dw_i^k + \sigma_i^0 dw_i^0 \tag{7.218}$$

where the summation convention over repeated indices not on the left-hand side is adopted. We further assume that $dw_i^k dw_j^l = \delta_{kl}\rho_{ij}^k$ and adopt the notation $X_{ij}^k \equiv \rho_{ij}^k \sigma_i^k \sigma_j^k$. To provide a rather concrete framework, consider two particular assets 1 and 2, and the following probability:

$$\Pr(z_1(T) > \gamma_1, z_2(T) > \gamma_2) = E_t^P 1(z_1(T) > \gamma_1, z_2(T) > \gamma_2) \tag{7.219}$$

Here, $\gamma$ is a positive (in the element-by-element sense) vector. To evaluate the expression in (7.219), we need the (conditional) characteristic function $E_t^P e^{i\phi_1 z_1(T) + i\phi_2 z_2(T)}$. As we saw in Section 5.2.3, the affine form permits a solution of the following form: $e^{\alpha_0 + \alpha_j z_j}$, with the coefficients $\alpha$ satisfying the following system of ODEs:

$$\begin{aligned} \dot{\alpha}_k + A_{ik}\alpha_i + \tfrac{1}{2}X_{ij}^k \alpha_i \alpha_j = 0 \\ \dot{\alpha}_0 + b_i \alpha_i + \tfrac{1}{2}X_{ij}^0 \alpha_i \alpha_j = 0 \end{aligned} \tag{7.220}$$

The terminal conditions are $\alpha_1(T) = i\phi_1$, $\alpha_2(T) = i\phi_2$, and $\alpha_k(T) = 0$ for $k \neq 1, 2$. Then, the probability (7.219) can be written as

$$\begin{aligned} &E_t^P 1(z_1(T) > \gamma_1, z_2(T) > \gamma_2) \\ &= \frac{1}{(2\pi)^2} \int_{-\infty + i\varepsilon_1}^{\infty + i\varepsilon_1} \int_{-\infty + i\varepsilon_2}^{\infty + i\varepsilon_2} d\phi_1 d\phi_2 \frac{e^{-i\phi_1 \gamma_1}}{i\phi_1} \frac{e^{-i\phi_2 \gamma_2}}{i\phi_2} e^{\alpha_k(\phi_1, \phi_2) z_k + \alpha_0(\phi_1, \phi_2)} \end{aligned} \tag{7.221}$$

(Note that we require $\varepsilon_{1,2} < 0$ for convergence.[83]) Now, the saddle-point condition requires that, along the chosen contour, we have:

$$\begin{aligned}\partial_{\phi_1}\alpha_k(\phi_1,\phi_2)z_k + \partial_{\phi_1}\alpha_0(\phi_1,\phi_2) &= i\gamma_1\\ \partial_{\phi_2}\alpha_k(\phi_1,\phi_2)z_k + \partial_{\phi_2}\alpha_0(\phi_1,\phi_2) &= i\gamma_2\end{aligned} \tag{7.222}$$

where the notation for partial derivatives with respect to $\phi_i$ should be clear. A natural choice is to require that this minimum occur where the contour intersects the imaginary $\phi$–axis in each dimension, in which case we require $\varepsilon$ to satisfy

$$\begin{aligned}\partial_{\phi_1}\alpha_k(i\varepsilon_1,i\varepsilon_2)z_k + \partial_{\phi_1}\alpha_0(i\varepsilon_1,i\varepsilon_2) &= i\gamma_1\\ \partial_{\phi_2}\alpha_k(i\varepsilon_1,i\varepsilon_2)z_k + \partial_{\phi_2}\alpha_0(i\varepsilon_1,i\varepsilon_2) &= i\gamma_2\end{aligned} \tag{7.223}$$

We can now show that this is the same condition that will hold under a suitable change of measure that aligns the expected values of $z_{1,2}$ suitably. Consider the follow measure change:

$$\frac{dP_\varepsilon}{dP} = \frac{e^{-\varepsilon_k z_k(T)}}{E_t^P e^{-\varepsilon_k z_k(T)}} \tag{7.224}$$

Now, under $P_\varepsilon$ the condition characteristic function is given by

$$\begin{aligned}E_t^{P_\varepsilon} e^{i\phi_1 z_1(T)+i\phi_2 z_2(T)} &= \frac{E_t^P e^{(-\varepsilon_1+i\phi_1)z_1(T)+(-\varepsilon_2+i\phi_2)z_2(T)}}{E_t^P e^{-\varepsilon_1 z_1(T)-\varepsilon_2 z_2(T)}}\\ &= \exp((\alpha_k(i\varepsilon_j+\phi_j)-\alpha_k(i\varepsilon_j))z_k + \alpha_0(i\varepsilon_j+\phi_j)-\alpha_0(i\varepsilon_j))\end{aligned} \tag{7.225}$$

where the coefficients $\alpha$ satisfy the same system (7.220), except with terminal conditions $\alpha_1(T) = -\varepsilon_1 + i\phi_1$, $\alpha_2(T) = -\varepsilon_2 + i\phi_2$, and $\alpha_k(T) = 0$ for $k \neq 1,2$. From this, we can see that the requirement that the mean of $z_1$, say, under $P_\varepsilon$ be equal to $\gamma_1$ is given by[84]

$$i\gamma_1 = E_t^{P_\varepsilon} iz_1(T) = n\partial_{\phi_1}\alpha_k(i\varepsilon_j)z_k + \partial_{\phi_1}\alpha_0(i\varepsilon_j) \tag{7.226}$$

and a similar result for the expectation of $z_2$. But these are precisely the conditions stated in (7.223). In other words, the usual change-of-measure techniques that facilitate importance sampling for calculation of OTM probabilities exactly correspond to saddle-point methods that facilitate the same computation when done in terms of quadrature.

As an example, consider a standard bivariate normal, with characteristic function

$$f = e^{-\frac{1}{2}(\phi_1^2+2\rho\phi_1\phi_2+\phi_2^2)} \tag{7.227}$$

The saddle-point condition becomes

$$\begin{aligned} i\gamma_1 &= \phi_1 + \rho\phi_2 \\ i\gamma_2 &= \rho\phi_1 + \phi_2 \end{aligned} \tag{7.228}$$

for $(\phi_1\phi_2) = (i\varepsilon_1, i\varepsilon_2)$. But this is also of course the Girsanov-type result that would arise from a mean-matching condition. Results are shown in Table 7.5.

*7.7.1.4 Nonlinear extensions*

The analysis we have just considered is essentially local in nature. That is, it is appropriate for processes that are linear in some sense. Alternatively, the optimal contour (in each dimension) can be characterized as linear. In general, we can deform the contour of integration in any appropriate way (taking into account poles, branch points, *etc.*). We will concentrate here on the one-dimensional case. Thus we can write

$$E1(z > \gamma) = \int_{\gamma}^{\infty} dz \frac{1}{2\pi} \int_{c} d\phi f(\phi) e^{-i\phi z} \tag{7.229}$$

where $C$ is some contour below the real $\phi$-axis (continuing with the assumption of positive $\gamma$ ). Now, reversing the order of integration in (7.229), we get

$$E1(z > \gamma) = \frac{1}{2\pi} \int_{c} d\phi f(\phi) \frac{e^{-i\phi\gamma}}{i\phi} \tag{7.230}$$

Now, we require that along the contour of integration, the imaginary part of any exponential form is zero:

$$\mathrm{Im}\{\alpha_k(\phi_r + i\phi_i)z_k + \alpha_0(\phi_r + i\phi_i)\} = \gamma\,\phi_r \tag{7.231}$$

where $\phi_r$ represents some, say, set of quadrature points along the real $\phi-$ axis (*e.g.*, Gauss-Laguerre) and $\phi_i$ is chosen to satisfy the condition in (7.231). There are a number of ways to approach this problem. For each quadrature point, we can solve a nonlinear equation for the corresponding location in the complex $\phi-$plane. Alternatively, for a set of quadrature points, we can implement the above condition as

**Table 7.5** Quadrature results for standard bivariate normal. $\Pr(z_1 > \gamma_1, z_2 > \gamma_2)$ for $\gamma_1 = 3, \gamma_2 = 4, \rho = 0.8$; 30 grid points. Here, "exact" refers to Sheppard's method (7.127).

| Quad | Quad adj | Exact |
|---|---|---|
| −3.73614E-05 | 2.30821E-05 | 2.30593E-05 |

a minimization problem (*e.g.*, in terms of squared sums of differences). The point is this: we now consider a global or nonlinear version of conventional saddle-point methods.

As an example, consider a standard CIR process (*e.g.*, the Heston variance process):

$$dv = \kappa(\theta - v)dt + \sigma\sqrt{v}dw \tag{7.232}$$

For the parameters $\kappa = 6$, $\theta = 0.09$, $\sigma = 0.4$, $\rho = -0.6$ and initial conditions $v = 0.09$. We take time-to-maturity $\tau = 0.25$. We look for the probability that stochastic variance will be well above its long-term mean, *i.e.*, $E_t 1(v_T > 0.2)$.[85] We show a comparison of results in Table 7.6, for both (global) contour deformation and the (local) approach described in Section 7.7.1.3.

As can be seen, the contour deformation approach converges much more quickly than the conventional approach where the contour is shifted back to the real $\phi$ – axis. This reflects the fact that oscillations in the integrand are suppressed, thus facilitating the use of quadrature-based techniques. The contour of integration for 20 quadrature points is shown in Figure 7.14.

**Table 7.6** Comparison of OTM probabilities for Heston variance. Contour deformation, nonlinear vs. linear.

| QuadPts | Contour (nonlinear) | Contour (linear) |
|---|---|---|
| 10 | 0.005126 | −0.016043 |
| 20 | 0.005139 | 0.003193 |
| 30 | 0.005139 | 0.004644 |

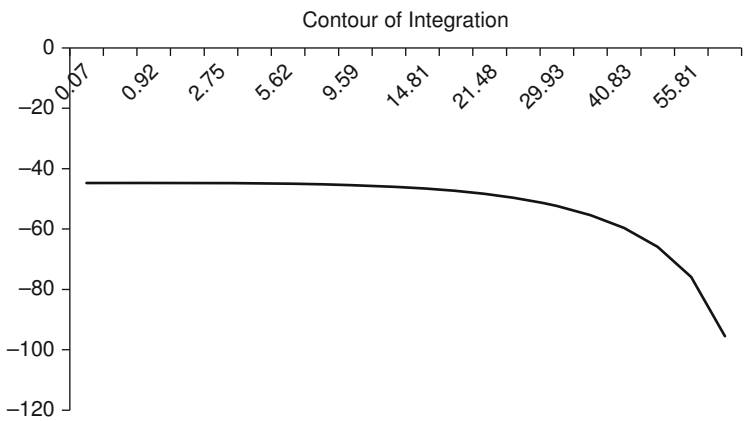



**Figure 7.14** Integration contour for quadrature

We leave as an open question the issue of how this technique can be extended to higher dimensions (*e.g.*, to calculate the joint probability of a Heston-driven price *and* integrated variance). Given the feasibility of the idea in one dimension, however, this does seem like a topic worth pursuing.

### 7.7.2 FFT and other transform methods

*7.7.2.1 Strike space vs. price space*

We have already seen how Fourier methods find application to the efficient computations of conditional expectations in Section 7.3.2 (on the Grid Model). We have also found great utility in a class of canonical processes (the class of affine jump diffusions) for which option valuation can be carried out in terms of the Fourier transform of the underlying price processes (*i.e.*, the characteristic function). In those problems, the analysis essentially takes place in (log-) price space. We now wish to consider an alternative approach, where the (log) strike is the primary variable. More accurately, we wish to consider analyzing the problem with reference to the Fourier space of the strike.[86] So, we start with a standard European option, with value in terms of log entities given by

$$V(k) = E_t^Q(e^{z_T} - e^k)^+ \tag{7.233}$$

under an appropriate measure $Q$. Now, let us formally take the Fourier transform of (7.233) wrt. $k$:

$$\tilde{V}(\phi) = E_t^Q \mathfrak{F}(e^{z_T} - e^k)^+ = E_t^Q \int_{-\infty}^{z_T} dk(e^{z_T} - e^k)e^{i\phi k} \tag{7.234}$$

We assume that the Fourier variable $\phi$ satisfies $\mathrm{Im}\{\phi\} < 0$ so that the integrals in (7.234) converge. Then we can write

$$\begin{aligned}\tilde{V}(\phi) &= E_t^Q e^{z_T} \int_{-\infty}^{z_T} dk e^{i\phi k} - E_t^Q \int_{-\infty}^{z_T} dk e^{(1+i\phi)k} \\ &= E_t^Q \frac{e^{(1+i\phi)z_T}}{i\phi} - E_t^Q \frac{e^{(1+i\phi)z_T}}{1+i\phi} = \frac{f(\phi - i; z)}{i\phi(1+i\phi)}\end{aligned} \tag{7.235}$$

where $f$ denotes the (conditional) characteristic function of $z$. Applying the Fourier inversion theorem, we thus get

$$V(k) = \frac{1}{2\pi} \int_\Gamma d\phi e^{-i\phi k} \frac{f(\phi - i; z)}{i\phi(1+i\phi)} \tag{7.236}$$

where $\Gamma$ denotes any contour in the complex $\phi$–plane lying below the real axis. As we saw in the preceding sections, the actual location of this contour should be chosen to facilitate numerical convergence and stability. (This approach should be contrasted with the approach adopted in the original paper by Carr and Madan [1999], who considered a "modulated" value function given by $v(k) \equiv e^{\alpha k} V(k)$ with $\alpha > 0$. The exponential-damping factor ensures square integrability and thus validity of the Fourier analysis. Note that $-\alpha$ effectively plays the role of the imaginary part of the Fourier variable $\phi$ in (7.236).)

#### *7.7.2.2 An adjoint relationship*

Thus, for any process for which the (conditional) characteristic function is known, (7.236) can be used to calculate the option price *as a function of (log) strike*. This result can be contrasted with results such as (7.95), which are essentially performed in (log-)price space. In fact, we have yet another duality result connecting the two approaches. Continue working in one dimension, and assume we have a product with terminal payoff $G(z_T, k)$. Then we have that

$$\begin{aligned} V(z,k) &= E_t^Q G(z_T,k) = E_t^Q \frac{1}{2\pi} \int_{\Gamma_k} d\phi_k e^{-i\phi_k k} \tilde{G}_k(z_T,\phi_k) \\ &= \frac{1}{2\pi} \int_{\Gamma_k} d\phi_k e^{-i\phi_k k} E_t^Q \tilde{G}_k(z_T,\phi_k) \end{aligned} \tag{7.237}$$

in terms of the Fourier transform $\tilde{G}_k$ of the payoff wrt. (log-) price $k$, and where $\Gamma_k$ is a suitably chosen contour of integration in the complex $\phi_k$–plane. Alternatively, we can also operate in terms of the Fourier transform of the underlying price process, *i.e.*, the characteristic function. We also have

$$\begin{aligned} V(z,k) &= \int_{-\infty}^{\infty} dz_T G(z_T,k) \Pr(z_T|z) \\ &= \int_{-\infty}^{\infty} dz_T G(z_T,k) \frac{1}{2\pi} \int_{\Gamma_z} d\phi_z e^{-i\phi_z z_T} E_t^Q e^{i\phi_z z_T} \\ &= \frac{1}{2\pi} \int_{\Gamma_z} d\phi_z E_t^Q e^{i\phi_z z_T} \int_{-\infty}^{\infty} dz_T e^{-i\phi_z z_T} G(z_T,k) = \frac{1}{2\pi} \int_{\Gamma_z} d\phi_z E_t^Q e^{i\phi_z z_T} \tilde{G}_z^*(\phi_z,k) \end{aligned} \tag{7.238}$$

where $\tilde{G}_z$ is the Fourier transform of the payoff wrt. (log-) price $z$, and $\Gamma_z$ is a suitably chosen contour of integration in the complex $\phi_z$ –plane. (Complex conjugation is denoted by the superscript *.) Combining (7.237) and (7.238) we get the following duality relation:

$$V(z,k) = \frac{1}{2\pi} \int_{\Gamma_k} d\phi_k e^{-i\phi_k k} E_t^Q \tilde{G}_k(z_T, \phi_k) = \frac{1}{2\pi} \int_{\Gamma_z} d\phi_z \tilde{G}_z^*(\phi_z, k) E_t^Q e^{i\phi_z z_T} \quad (7.239)$$

which is reminiscent of adjoint relationships between operators in the theory of ODEs.

*7.7.2.3 Numerical implementation*

The primary benefit of the formula in (7.235) comes in situations where valuations for large numbers of strikes are required (*e.g.*, management of a book of options with a wide range of moneyness). In such cases, the FFT[87] proves to be of great use. In particular, the so-called fractional FFT is especially helpful (see the excellent paper of Chourdakis [2005b], as well as Bailey and Swartztrauber [1995]). Since the fractional FFT (and the need for it) is likely less familiar than the standard FFT, we will confine the bulk of our exposition here to outlining this technique.

Integrals of the form (7.236) can be evaluated by a suitable discretization in $\phi$ –space, requiring the evaluation of summations of the form $\hat{V}(k) = \Delta_\phi \sum_j e^{-i\phi_j k} h(\phi_j)$ , where $\phi_j$ are the $\phi$ –grid points of resolution $\Delta_\phi$. Now, as is well known, summations of this form can be efficiently computed for a range of (log-strike) values $k$ if *both* this range and the $\phi$ –grid take special forms. Specifically, consider the following ensemble:

$$\begin{aligned} k_l &= \Delta_k l, l = -N/2, \ldots, N/2 - 1 \\ \phi_j &= \frac{2\pi j}{N\Delta_k}, j = -N/2, \ldots, N/2 - 1 \\ H_j &= \sum_{l=-N/2}^{N/2-1} h_l e^{2\pi ijl/N} \end{aligned} \quad (7.240)$$

Then, the components of the vector $H$ can be computed with cost $O(N \log N)$, as opposed to $O(N^2)$ for ordinary matrix multiplication, via the celebrated FFT; see Press *et al.* (2007) for greater detail.

The point that concerns us here is the fact that in the ensemble (7.240), there is a *necessary* relationship between the resolution of the (log-) strike grids and the Fourier grids, namely

$$\Delta_k \Delta_\phi = \frac{2\pi}{N} \quad (7.241)$$

(This is essentially the Heisenberg uncertainty principle.) The immediate implication of (7.241) is that there is an inverse relationship between the resolution of the (log-) strike grid and the Fourier grid. In other words, for a fine-resolution Fourier grid (necessary for accurate quadrature in the Fourier inversion in (7.236)), the corresponding (log-) strike grid will be very coarse. In particular, there will be unnecessary and inefficient valuation of options that are very deep ITM (corresponding to $-\frac{1}{2}N\Delta_k = -\frac{1}{2}\Delta_\phi^{-1}$) and very deep OTM (corresponding to $\frac{1}{2}N\Delta_k = -\frac{1}{2}\Delta_\phi^{-1}$).[88] Even if performing these evaluations (where extrinsic value is miniscule) is deemed tolerable, to get a sufficiently fine resolution of (log-) strike space, an extremely high number of grid points is required (see (7.241)), greatly adding to the computational burden. It is clearly desirable to be able to separate the discretizations in the two spaces. This is where the fractional FFT comes in.

The fractional FFT is a method for rapid computation of matrix multiplications of the form

$$H_j \sum_{l=-N/2}^{N/2-1} h_l e^{2\pi ijl\gamma} \tag{7.242}$$

for *arbitrary* $\gamma$. Thus, the particular choice $2\pi\gamma = \Delta_k\Delta_\phi$ can be freely made, *i.e.*, *without* the requirement that $\gamma = 1/N$ in the regular FFT case. Consequently, the separation of grids is attained. Of course, like anything else in life, this achievement is not free, and the cost is the need to perform three FFT calculations, each of twice the size of the original problem. However, since the size of the grids needed for a given accuracy and resolution are typically much smaller than the regular FFT case, the gains are substantially more than the pain. The trick is to exploit the fact that the product of a Fourier transform is the Fourier transform of the convolution.

Introduce the following auxiliary vectors (with appropriately interpreted negative indices):

$$y = \begin{pmatrix} \left(h_j e^{-i\pi j^2\gamma}\right)_{j=-N/2}^{N/2-1} \\ (0)_{j=N/2}^{3N/2-1} \end{pmatrix}, \quad z = \begin{pmatrix} \left(e^{i\pi j^2\gamma}\right)_{j=-N/2}^{N/2-1} \\ \left(e^{i\pi (N-j)^2\gamma}\right)_{j=N/2}^{3N/2-1} \end{pmatrix} \tag{7.243}$$

Then the fractional FFT is obtained via

$$(e^{-i\pi l^2\gamma})_{l=-N/2}^{N/2-1} \otimes \mathfrak{J}^{-1}\{\mathfrak{J}(y) \otimes \mathfrak{J}(z)\} \tag{7.244}$$

where $\mathfrak{J}$ denotes Fourier transform and $\otimes$ is the Hadamard (element-by-element) matrix product. Again, see Chourdakis (2005a) for more details.

#### *7.7.2.4 Higher dimensions*

We now consider extensions of (7.235) to higher dimensions. In particular, consider the case of a standard spread option. The valuation requires the following

expectation:

$$V(k) = E_t^Q(e^{z_2(T)} - e^{z_1(T)} - e^k)^+ \tag{7.245}$$

If we attempt to proceed as we did in (7.234) by taking the Fourier transform in (7.245) wrt. $k$, we immediately run into a problem. Because the exercise boundary is now nonlinear, the subsequent expectations (after the Fourier transformation) cannot be carried out analytically:

$$\tilde{V}(\phi) = E_t^Q\left[1(z_2(T) > z_1(T)) \int_{-\infty+i\alpha}^{\log(e^{z_2(T)}-e^{z_1(T)})+i\alpha} d\phi(e^{z_2(T)} - e^{z_1(T)} - e^k)e^{i\phi k}\right] \tag{7.246}$$

with $\alpha > 0$. It can easily be seen from (7.246) that, after the Fourier integration, the resulting expectation *cannot* be reduced to a characteristic function evaluation (as occurs in the one-dimensional case). However, these Fourier methods can be applied if we approximate the exercise boundary in a suitable manner. The idea (adopted by Dempster and Hong [2000]) is to operate in reverse, in a sense, by specifying a framework in which Fourier methods are applicable, which *indirectly* results in an approximation to the exercise boundary (as opposed to directly approximating this boundary).

We start by noting a (somewhat artificial) product for which Fourier methods can be directly applied, with bilinear payoff $(e^{z_1(T)} - e^{k_1})^+(e^{z_2(T)} - e^{k_2})^+$. Since the exercise region is linear (in the $(k_1, k_2)$ plane), the Fourier transform of the option value can easily be expressed in terms of the (joint) characteristic function of the process $(z_1, z_2)$. Note, in fact, as in the one-dimensional case, inverse transformation yields the option value across a *grid* of strike values, call them $(k_1^i, k_2^j)$. We can use this fact to construct an approximation to (7.246) that implicitly entails an approximation to the true exercise boundary. First, we again employ change-of-measure techniques to reduce the valuation problem to the form

$$E_t^{Q_z} 1(e^{z_2(T)} - e^{z_1(T)} - e^k > 0) \tag{7.247}$$

where $Q_z$ is a shorthand notation for a numeraire change, *e.g.*, $\frac{dQ_1}{dQ} = \frac{e^{z_1}}{Ee^{z_1}}$. Now, consider the following entity:

$$U(k_1, k_2) = E_t^{Q_z} 1(z_1(T) > k_1)1(z_2(T) > k_2) \tag{7.248}$$

Fourier transformation (with the Fourier variable suitably restricted to ensure convergence) yields

$$\tilde{U}(\phi_1, \phi_2) = E_t^{Q_z} \frac{e^{i\phi_1 z_1(T)}}{i\phi_1} \frac{e^{i\phi_2 z_2(T)}}{i\phi_2} = \frac{f_z(\phi_1, \phi_2)}{-\phi_1\phi_2} \tag{7.249}$$

where $f_z$ is the characteristic function of $(z_1, z_2)$ under the measure change in question. Inverting (7.249) (using an appropriate contour of integration) gives

$$E_t^{Q_z} 1(z_1(T) > k_1) 1(z_2(T) > k_2) = \frac{1}{(2\pi)^2} \int_{\Gamma_1} \int_{\Gamma_2} \frac{f_z(\phi_1, \phi_2)}{-\phi_1 \phi_2} e^{-i\phi_1 k_1 - i\phi_2 k_2} d\phi_1 d\phi_2 \tag{7.250}$$

Now, as in the one-dimensional case, multidimensional integrals such as (7.250) can be efficiently evaluated via the FFT,[89] producing values across a grid of (log) strikes. We will now put this fact to our advantage. For the specified grid of strikes, we can easily (having applied the FFT) evaluate expectations of the form

$$E_t^{Q_z} 1(k_1^i < z_1(T) < k_1^{i+1}) 1(z_2(T) > k_2^j) \tag{7.251}$$

for indices $(i, j)$ on the grid. In particular, for a given $k_1$–index $i$, we can choose the smallest $k_2$ –index $j(i)$ st. $e^{z_2} - e^{z_1} - e^k > 0$ for the region over which the integration in (7.251) occurs. Thus, the following ensemble

$$\sum_i E_t^{Q_z} 1(k_1^i < z_1(T) < k_1^{i+1}) 1(z_2(T) > k_2^{j(i)}) \tag{7.252}$$

represents a *lower-bound* valuation (because the payoff is strictly positive in the effective exercise region, hence some profitable exercises get ignored). Essentially, the exercise region is being approximated by a collection of rectangles; we refer the reader to figures 2 and 3 in Dempster and Hong (2000).

# 8 Dependency Modeling

## 8.1 Dependence and copulas

As should be quite clear to this stage, the need to understand the joint dependence between multiple stochastic entities is of critical importance in energy markets. We have of course considered numerous examples in the context of spread-option structures, where the relevant measure of dependence in Gaussian scenarios is correlation.[1] We have also considered some fairly rich classes of canonical processes that extend the standard Gaussian framework (affine jump diffusions and Lévy processes). In addition, we examined the interplay between short-term co-movements and long-term stationarity through cointegration analysis. We now examine another concept useful for modeling joint structure, namely, copulas. As will be seen, an interesting facet of copulas is the ability to construct joint dependency in terms of specified *marginal* distributions. To the extent that marginal distributions may often be extracted from market information (through, say, option prices), the flexibility offered by copula-based models can be quite enticing. Not surprisingly, there is a voluminous literature available, including book-length treatments by Nelsen (1999) and detailed survey articles in Embrechts *et al.* (2002, 2003). Our objective here is to provide an overview and highlight the most promising directions as they pertain to energy modeling.

### 8.1.1 Concepts of dependence

#### *8.1.1.1 Introduction*

A copula is simply the joint distribution function of a set of random variables with uniformly distributed marginals:

$$C(u_1,\ldots,u_n) = \Pr(U_1 \leq u_1,\ldots,U_n \leq u_n) \tag{8.1}$$

where $U_i \sim U(0,1)$.[2] (Thus, in the case of independence we would have $C = u_1 \cdots u_n$). For a general (continuous) random variable $X$ with CDF $F$, it is not hard to see that $F(X)$ is uniformly distributed.[3] Thus, (8.1) implies that

$$\begin{aligned} &C(u_1,\ldots,u_n) = \Pr(F_1(X_1) \le u_1,\ldots,F_1(X_n) \le u_n) \\ &\Pr((X_1 \le F_1^{-1}(u_1),\ldots,X_n \le F_1^{-1}(u_n)) = F(F_n^{-1}(u_1),\ldots,F_n^{-1}(u_n)) \end{aligned} \tag{8.2}$$

using continuity and monotonicity of $F_i$, and where $F$ is the *joint* distribution function of the random variables $X_i$. A relationship can clearly be seen between distribution functions of general random variables and copulas. It turns out this relationship is rather strong, as according to Sklar's theorem, *any* (multivariate) CDF can uniquely be written as a copula function:[4]

$$F(x_1,\ldots x_n) = C(F(x_1),\ldots,F(x_n)) \tag{8.3}$$

Note that this result also holds for the "reverse" distribution function, defined as $\tilde{F}(x) \equiv \Pr(X > x) = 1 - F(x)$. *E.g.*, in two dimensions we have

$$\begin{aligned} \tilde{F}(x_1,x_2) &= 1 - F_1(x_1) - F_2(x_2) + F(x_1,x_2) \\ &= \tilde{F}_1(x_1) + \tilde{F}_2(x_2) - 1 + C(1-\tilde{F}_1(x_1), 1-\tilde{F}_2(x_2)) = \tilde{C}(\tilde{F}_1(x_1),\tilde{F}_2(x_2)) \end{aligned} \tag{8.4}$$

and it is easy to show that $\tilde{C}$ satisfies the requirements for being a valid copula function.[5]

Keeping in line with Endnote 2, the so-called Fréchet-Höffding bounds put constraints on the degree of codependency embodied by a copula (and thus by Sklar's theorem, the joint dependency of any set of random variables). Specifically, we have that

$$\max\left(1-n+\sum_{i=1}^{n} u_i, 0\right) \le C(u_1,\ldots,u_n) \le \min(u_1,\ldots,u_n) \tag{8.5}$$

The upper bound corresponds to random variables with (perfect) comonotonicity, *i.e.*, they can be expressed as functions of a single random variable. In two dimensions, the lower bound corresponds to the case of countermonotonicity, or perfect negative dependence.[6] (While the upper bound is always a valid copula function, it turns out that the lower bound is only a copula in two dimensions.)

#### *8.1.1.2 Measures of dependency*

Having just mentioned two categories of dependence between random variables (co- and countermonotonicity), it is worth discussing some other measures characterizing joint structure, especially those that can be readily computed and contrasted for specific examples of copulas. An obvious, and very familiar, measure of

dependence is simply (linear) correlation. Though doubtless not necessary at this stage, we nonetheless write out the formula:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}} \tag{8.6}$$

As the name suggests, linear correlation is a measure of linear dependence between two random variables. It is not hard to see from (8.6) that correlation is, effectively, invariant under *linear* transformations. (We say "effectively" because of the different effects of increasing vs. decreasing transformations, *e.g.*, $\rho(-X,Y) = -\rho(X,Y)$). Clearly, by Cauchy-Schwartz, we have $-1 \leq \rho \leq 1$, with the bounds being attained only for the cases of perfect *linear* monotonicity (*i.e.*, when $Y = kX$ for some constant $k$) . Despite the widespread popularity and use of correlation, we will see that is only an appropriate measure of dependency within a certain class of joint distributions, namely the class of elliptical distributions.[7,8]

An alternative set of measures refers to the notion of concordancy, or the degree to which ordering is retained across pairings of the variables in question. For example, the pairs $(x_1, y_1)$ and $(x_2, y_2)$ are concordant if $(x_2 - x_1)(y_2 - y_1) > 0$ and discordant if $(x_2 - x_1)(y_2 - y_1) < 0$. A specific (and popular) example is Kendall's tau, which measures the relative difference between the degree of concordancy and discordancy between two random variables. The population version can be written as

$$\tau(X,Y) = \Pr((X_2 - X_1)(Y_2 - Y_1) > 0) - \Pr((X_2 - X_1)(Y_2 - Y_1) < 0) \tag{8.7}$$

It can be shown (see Embrechts *et al.* [2003]) that this result can be expressed in terms of the copula via

$$\tau(X,Y) = 4EC(U,V) - 1 = 4\int_0^1\int_0^1 C(u,v)dC(u,v) - 1 \tag{8.8}$$

For the case of joint normality, there is an explicit expression for Kendall tau in terms of the primary dependency characteristic, namely correlation:

$$\tau(x_i, x_j) = \frac{2}{\pi}\arcsin\rho_{ij} \tag{8.9}$$

This so-called arcsin formula is in fact generalizable to the class of elliptical distributions (see Section 8.1.2), which include standard normality as a special case; see Lindskog *et al.* (2003).

Another common concordancy measure is Spearman's rho, which is essentially the linear correlation of rank: $\rho_S(X,Y) = \rho(F(X), G(Y))$, which can be shown to satisfy

$$\rho_s(X,Y) = 12\int_0^1\int_0^1 uvdC(u,v) - 3 \tag{8.10}$$

Perfect positive/negative dependence correspond to $\rho_S = \pm 1$.

Note that both Kendall's tau and Spearman's rho are invariant under strictly increasing (general) transformations, a property that does not hold for linear correlation. This invariance property affords an interesting contrast between the two dependency measures. For a given pair of marginals, *any* value of Kendall's tau or Spearman's rho can be attained for a suitable choice of joint dependency (*i.e.*, choice of copula). This is *not* true for correlation, as can be seen from the following simple example from Embrechts *et al.* (2002). Assume $X$ and $Y$ are lognormally distributed, with $\log X \sim N(0,1)$ and $\log Y \sim N(0,\sigma^2)$. Maximal dependence, *i.e.*, comonotonicity, is characterized by $(X,Y) \stackrel{d}{=} (e^z, e^{\sigma z})$, and minimal dependence, *i.e.*, countermonotonicity, is characterized by $(X,Y) \stackrel{d}{=} (e^z, e^{-\sigma z})$, with $z \sim N(0,1)$. It is thus easy to see that the maximal and minimal attainable correlations between $X$ and $Y$ are given by

$$\rho_{\max} = \frac{e^{\sigma}-1}{\sqrt{e-1}\sqrt{e^{\sigma^2}-1}}, \rho_{\min} = \frac{e^{-\sigma}-1}{\sqrt{e-1}\sqrt{e^{\sigma^2}-1}} \tag{8.11}$$

Thus, the linear correlation can be made as small as desired by taking $\sigma$ sufficiently large, despite the perfect (positive or negative) dependence between the two variables. This phenomenon will not arise with measures of association such as Kendall's tau or Spearman's rho.

Another concept here is tail dependence. In many applications (*e.g.*, value at risk or insurance), we are not interested in joint dependence as such, but the joint dependence *conditional* on some extreme events happening together. For example, we might be concerned with the probability of one part of a portfolio suffering large losses given that a different part of the portfolio suffers a large loss. Mathematically this is characterized by the so-called coefficient of (upper) tail dependence, given by

$$\lambda_U \equiv \lim_{u\to 1-} \Pr(Y > F_Y(u) | X > F_X(u)) \tag{8.12}$$

provided the limit exists. Lower tail dependence may be similarly characterized; indeed, in financial contexts it is useful to express it as

$$\lambda_L \lim_{u\to 0+} \Pr(Y < -VaR_u(Y) | X < -VaR_u(X)) \tag{8.13}$$

where $VaR_u$ represents value at risk at the $u$–th percentile. If (say) $\lambda_U = 0$, $X$ and $Y$ are said to not exhibit (upper) tail dependence.[9] We will see when we consider

some specific examples of copulas that not all joint dependency structures possess tail dependence. This fact has obvious implications for risk management, as it can be shown that popular measures of dependency such as correlation can give a misleading picture of, say, portfolio exposure. Two useful results that follow from the definition are

$$\lambda_U = \lim_{u \to 1-} \frac{1 - 2u + C(u,u)}{1-u}, \lambda_L = \lim_{u \to 0+} \frac{C(u,u)}{u} \tag{8.14}$$

*8.1.1.3 Key points*

Having laid out the bare essentials here, we wish to emphasize a few points. First, in much the same way as general random variables in one dimension can be simulated by first generating a uniform variate and then inverting the CDF of the variable in question, so, too, can general multidimensional random variables be simulated via generation of multidimensional uniform variates related to the original variables via a suitable copula. Some specific examples will be considered shortly, with algorithms tailored to the specific form of the dependency structure in question. As a very generic approach, one can in theory always condition on the copula iteratively as follows. In $n$–dimensions, define lower dimensional CDFs via

$$C_k(u_1, \ldots, u_k) = C(u_1, \ldots, u_k, 1, \ldots, 1), k = 2, \ldots, n-1 \tag{8.15}$$

with $C_1(u_1) = u_1$, and then construct conditional CDFs via Bayes rule:

$$\begin{aligned} C_k(u_k | u_1, \ldots, u_{k-1}) &= \Pr(U_k \leq u_k | U_1 = u_1, \ldots, U_{k-1} = u_{k-1}) \\ &= \frac{\partial^{k-1} C_k}{\partial u_1 \ldots \partial u_{k-1}} \Big/ \frac{\partial^{k-1} C_{k-1}}{\partial u_1 \ldots \partial u_{k-1}} \end{aligned} \tag{8.16}$$

1. Simulate $u_1$ from $U(0,1)$.
2. Simulate $u_2$ from $C_2(u_2|u_1)$.
3. Simulate $u_n$ from $C_n(u_n|u_1, \ldots, u_{n-1})$.

Obviously, this approach assumes some degree of analytical/computational tractability of the underlying conditional expressions (or more accurately, their inverses), but as noted can always be applied in principle, and so the general problem of $n$–dimensional simulation can be reduced to a sequence of one-dimensional simulations.

A second point is that the consideration of the joint structure of a set of random variables can be separated from the *marginal* structure of the constituent variables via Sklar's theorem. This point is significant, as in many valuation problems, this marginal information can be inferred (in the proper sense) from market-supplied

information, such as option prices. As we alluded to in Chapter 7, spread-option valuations can be carried out under a rather general class of processes for which the conditional density is readily obtained. In such cases, we have some degree of flexibility in choosing the joint dependency structure while retaining the marginal structure that replicates observed market prices.

As a third point, we have already seen an example where correlation can give a misleading indication of joint dependency. Correlation is of course very widely used as a dependency measure, and one might say misused, as well. It is not hard to construct examples of joint distribution functions that are quite different, yet yield the *same* numerical value of linear correlation; see Embrechts *et al.*, 2002. (This point refers to the notion of tail dependence [or lack thereof] discussed above.) There are a number of other shortcomings of correlation as a dependency measure. We have already considered the non-invariance of correlation under rank-preserving transformations. While independent random variables are of course uncorrelated, the reverse is not true (as shown by the trivial example $Y = X^2$, with $X$ a standard normal). Thus, some amount of caution should be employed when using correlation as dependency measure. (We have already emphasized this point from a somewhat different angle when considering the difference between volatilities as value drivers under static vs. dynamic trading strategies in Chapter 3.) In general, knowledge of marginal distributions and correlations are *not* sufficient to determine the joint distribution, except in certain special cases (the class of so-called elliptical distributions to be considered below). (Nor, as we saw above, is it possible in general to attain any arbitrary value of correlation for a given set of marginals.)

#### *8.1.1.4 Relevance for energy modeling (and beyond)*

Before going further, it is necessary to situate the discussion within the context of energy modeling.[10] There is probably little need at this stage to emphasize the inherent interest in studying dependency structures for commodity processes. In particular, there is clearly value in considering generalizations of, if not alternatives to, correlation as the paragon of dependency metrics. This is not to deny that correlation is a very useful concept in many applications, only that it must be used properly, and that there may be applications where it is completely inappropriate. The simple heat-rate model in (2.42) offers a good illustration of how correlation may be operative in some regimes, but not others (depending on demand conditions and stack convexity). As well, there is some merit in being able to move away from correlation while still retaining some degree of rigor.[11]

In addition, there is the fact that all markets are, ultimately, interconnected. We mean here not simply the relation between two classes of commodities that stand in a production relationship, such as natural gas and electricity or crude oil and refined products such as gasoline and heating oil. There are also cross-commodity and cross-asset dependencies, such as between natural gas and crude oil or between commodities and equities (see Delatte and Lopez [2012] or Grégoire *et al.* [2008]).

Recent history bears this latter claim out in particular. It is well known that in the aftermath of the financial crisis of 2008, commodity volatilities were sharply down across the board. At the same time, some commodities (most notably crude) displayed a volatility term structure more reminiscent of equities than commodities (*i.e.*, relatively flat; recall Figure 2.11). These kinds of dependencies can entail not just tail relationships as such, but symmetric or antisymmetric effects (that is to say, different kinds of behaviors on the upside as opposed to the downside). To the extent that it remains preferable to model individually the dynamics of distinct markets, there needs to be means for connecting these dynamics in a joint model. Copula methods provide just such an approach.

Let us now consider different classes of copulas.

### 8.1.2 Classification

*8.1.2.1 Archimedean*

Archimedean copulas are defined by specifying a function $\varphi$ with domain on the unit interval and range on the non-negative reals, and then taking

$$C(u_1, \ldots, u_n) = \varphi^{[-1]}(\varphi(u_1) + \ldots + \varphi(u_n)) \tag{8.17}$$

where $\varphi^{[-1]}$ denotes the pseudo-inverse, given by

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & 0 \leq t \leq \varphi(0) \\ 0, & \varphi(0) \leq t \leq \infty \end{cases} \tag{8.18}$$

Some common examples include the Gumbel family:

$$\varphi(t) = (-\log t)^{\theta}, C(u, v) = \exp(-((-\log u)^{\theta} + (-\log v)^{\theta})^{1/\theta}) \tag{8.19}$$

for $0 < \theta \leq 1$, and the Clayton family:

$$\varphi(t) = (t^{-\theta} - 1)/\theta, \; C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta} \tag{8.20}$$

for $\theta \geq 0$. Note that the common feature of these members is the *single* parameter $\theta$. This aspect permits high-dimensional dependencies to be modeled with a high degree of sparseness. Archimedean copulas in general have tail dependence; *e.g.*, for the Gumbel family, a straightforward application of l'Hôpital's rule yields $\lambda_U = 2 - 2^{1/\theta}$. Similarly for the Clayton family, we have that $\lambda_L = 2^{-1/\theta}$. (Note that the case $\theta \to 0+$ corresponds to independence and $\theta \to \infty$ corresponds to perfect comonotonicity for the Clayton family, while the analogous situations for the Gumbel family are $\theta = 1$ and $\theta \to 0+$, respectively.)

*8.1.2.2 Elliptical*

Elliptical distributions[12] are a very popular category that include, as a special case, joint Gaussians. They are characterized by their characteristic functions, which as we have seen provide a very powerful computational framework. We first note the definition of a *spherical* random vector, which is distributionally invariant under orthogonal transformations:

$$X \stackrel{d}{=} OX \tag{8.21}$$

where $OO^T = O^T O = I_n$. (Unless otherwise noted, we will assume that we are operating in $n$ dimensions.) Plainly, then, the characteristic function $f(\phi) = Ee^{i\phi^T X}$ of a spherical random vector satisfies $f(O\phi) = f(\phi)$, and in fact it can be shown that there exists a function $g$ (termed the generator) such that $f(\phi) = g(\phi^T \phi)$. (Alternatively, spherical random vectors can be defined as $X \stackrel{d}{=} R \cdot U$ where $R$ is a positive random variable and $U$ is uniformly distributed on the unit hypersphere in $n-1$ dimensions, independent of $R$; note that under either convention spherical random variables have zero mean.) Elliptical random vectors extend this concept via affine maps on spherical random vectors. Namely, $Y$ is elliptical if there exists a spherical random vector $X$, a matrix $A$, and a vector $\mu$ such that

$$Y \stackrel{d}{=} AX + \mu \tag{8.22}$$

From the definition, it is easy to see that the characteristic function of $Y$ satisfies

$$f(\phi) = Ee^{i\phi^T Y} = e^{i\phi^T \mu} Ee^{i\phi^T AX} = e^{i\varphi^T \mu} g(\phi^T AA^T \phi) = e^{i\phi^T \mu}(\phi^T \Sigma \phi) \tag{8.23}$$

where $\Sigma = AA^T$.[13] (Note that in general $X$ could be taken to be $k-$dimensional, in which case $A$ is $n \times k$, *i.e.*, of reduced rank.)

Elliptical distributions are thus characterized by the triplet $(\mu, \Sigma, g)$[14]. (Alternatively, this characterization can serve as the definition of ellipticity.) Clearly, multidimensional Gaussian random variables are elliptical, with $g(x) = \exp(-x/2)$ and $\mu$ and $\Sigma$ having the interpretation of the mean and covariance, respectively. (Other examples are usefully documented in Hamada and Valdez [2004] and Landsman and Valdez [2003].) In general, $\mu$ will always correspond to the mean of the distribution (when the mean exists), but obviously $g$ and $\Sigma$ are only unique up to a constant scaling, which can always conventionally be chosen so that $\Sigma$ corresponds to the covariance matrix (again, for variables with finite second moments). Note that only in the Gaussian case does zero correlation between variables imply independence. Elliptical distributions have important applications in portfolio analysis, for obvious reasons: they offer a generalization of standard mean-variance analysis (*i.e.*, in terms of scale and location).

General elliptically distributed variables retain a number of useful features from the Gaussian case, chiefly linearity and conditionality. It is not hard to see that if

$Y \in \mathbb{R}^n$ is elliptical with triplet $(\mu, \Sigma, g)$, then $b + BY$ is also elliptical with triplet $(b + B\mu, B\Sigma B^T, g)$. If $Y$ is partitioned as $\begin{pmatrix} Y_1 & Y_2 \end{pmatrix}^T$ with $Y_1 \in \mathbb{R}^p$, $Y_2 \in \mathbb{R}^q$ and $p + q = n$, then clearly $Y_1$ and $Y_2$ are elliptical.[15] $Y_2|Y_1$ is also elliptical with triplet

$$(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(Y_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}, g') \tag{8.24}$$

where $g'$ is a different generator and we use the following partitions: $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. This result is of course very well known for the Gaussian case; the more general discussion can be found in Fang *et al.* (1987). It is actually useful to sketch out the general derivation using characteristic functions, as it also serves to illustrate some subtle points regarding elliptical distributions. Using Bayes's theorem and the definition of ellipticity, we have that[16]

$$\begin{aligned} E(e^{i\phi_2^T Y_2}|Y_1) &= \int dY_2 e^{i\phi_2^T Y_2} \Pr(Y_2|Y_1) \\ &= \frac{\frac{1}{(2\pi)^n} = \int dY_2 e^{i\phi_2^T Y_2} \int d\tilde{\phi} e^{-i\tilde{\phi}_1^T Y_1 - i\tilde{\phi}_2^T Y_2} g(\tilde{\phi}^T \Sigma \tilde{\phi})}{\frac{1}{(2\pi)^p} \int d\phi_1 e^{-i\phi_1^T Y_1} g(\phi_1^T \Sigma_{11} \phi_1)} \\ &= \frac{\int d\phi_1 e^{-i\phi_1^T Y_1} g(\phi_1^T \Sigma_{11}\phi_1 + \phi_2^T \Sigma_{21}\phi_1 + \phi_1^T \Sigma_{12}\phi_2 + \phi_2^T \Sigma_{22}\phi_2)}{\int d\phi_1 e^{-i\phi_1^T Y_1} g(\phi_1^T \Sigma_{11}\phi_1)} \end{aligned} \tag{8.25}$$

Upon introducing the Cholesky factorization $C_1 C_1^T = \Sigma_{11}$ and the substitution $\xi_1 = C_1^{-T}\phi_1$, (8.25) can be written as

$$\begin{aligned} &E(e^{i\phi_2^T Y_2|Y_1}) \\ &= \frac{\int d\xi_1 e^{-i\xi_1^T C_1^{-1} Y_1} g(\xi_1^T\xi_1 + \phi_2^T \Sigma_{21} C_1^{-T}\xi_1 + \xi_1^T C_1^{-1}\Sigma_{12}\phi_2 + \phi_2^T \Sigma_{22}\phi_2)}{\int d\xi_1 e^{-i\xi_1^T C_1^{-1} Y_1} g(\xi_1^T\xi_1)} \\ &= e^{-i\phi_2^T \Sigma_{21}\Sigma_{11}^{-1} Y_1} \frac{\int d\xi_1 e^{-i\xi_1^T C_1^{-1} Y_1} g(\xi_1^T\xi_1 + \phi_2^T(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})\phi_2)}{\int d\xi_1 e^{-i\xi_1^T C_1^{-1} Y_1} g(\xi_1^T\xi_1)} \end{aligned} \tag{8.26}$$

Now, we note the following about (8.26). First, the Fourier variable $\phi_2$ appears *only* in the following two forms: $e^{-i\phi_2^T \Sigma_{21}\Sigma_{11}^{-1} Y_1}$ (as a multiplicative factor) and as a function of $\phi_2^T(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})\phi_2$ (also multiplicatively). This confirms the claims made about the conditional means and variances. Second, the conditional characteristic generator (given by the ratio of integrals in (8.26)) is in general different from the unconditional generator, and more importantly depends on the

conditional variable $Y_1$. In truth, it is more accurate to speak of a family of conditional generators, all dependent on the variable upon which conditioning takes place. (Obviously, in the case of joint normality, the characteristic generator factors nicely, thus greatly simplifying (8.26) and recovering the standard results for conditional normality.) We speculate that this behavior is related to the inconsistency of elliptical marginal distributions mentioned in Endnote 14.

In general, elliptical distributions do not always have tail dependence. It proves useful to derive an alternative formula for (upper) tail dependence to illustrate this point. Since $\Pr(V > v|U = u) = 1 - \frac{\partial}{\partial u}C(u, v)$, it follows that

$$\begin{aligned}\lambda_U &= \lim_{u\to 1-}\left(2 - \frac{d}{du}C(u,u)\right)\\ &= \lim_{u\to 1-}\left(2 - \frac{\partial}{\partial s}C(s,t)\bigg|_{s=t=u} - \frac{\partial}{\partial t}C(s,t)\bigg|_{s=t=u}\right)\\ &= \lim_{u\to 1-}(\Pr(V > u|U = u) + \Pr(U > u|V = u)) = 2\lim_{u\to 1-}\Pr(V > u|U = u)\end{aligned} \tag{8.27}$$

where the last equation in (8.27) only holds true for symmetric copulas (*i.e.*, those for which $C(u, v) = C(v, u)$). Applying this result to a Gaussian copula and using standard results for conditional normals, we get

$$\begin{aligned}\lambda_U &= 2\lim_{u\to 1-}\Pr(V > v|U = u) = 2\lim_{u\to\infty}\Pr(Y > x|X = x)\\ &= 2\lim_{u\to\infty}\left(1 - N\left(\frac{x - \rho x}{\sqrt{1-\rho^2}}\right)\right) = 0\end{aligned} \tag{8.28}$$

where $\rho$ is the correlation parameter of the Gaussian copula and $X$ and $Y$ are defined as follows: $X = N^{-1}(U)$, $Y = N^{-1}(V)$. Equation (8.28) shows that the Gaussian copula is (asymptotically) tail independent.[17] Although this property represents a potential drawback to using (popular) Gaussian models for many risk-management applications, it is *not* true in general that elliptical distributions lack tail dependence. For example, the Student t-copula has tail dependence, even when the correlation parameter is zero (see Schmidt [2007]).

#### *8.1.2.3 Generalized elliptical*

We also note here the so-called generalized elliptical distributions (*e.g.*, Frahm and Jaekel [2007, 2008]). These are defined very similarly to regular elliptical distributions, namely in law we have that the random vector $X \in \mathbb{R}^d$ satisfies

$$X = \mu + R\Lambda U^{(k)} \tag{8.29}$$

where $U^{(k)}$ is uniformly distributed on the unit hypersphere $S^{k-1}$, $\Lambda \in \mathbb{R}^{d\times k}$, $\mu \in \mathbb{R}^d$, and $R$ is a random variable. In contrast to the usual elliptical case, $R$ need not be positive or independent of $U$. The ramifications of this fact are that any spherical invariance is lost and $\mu$ need not correspond to the expected value of $X$, and $\Lambda$ need not be associated with the Cholesky factorization of the covariance matrix of $X$. This class allows for the modeling of tail dependence and radial asymmetry. Generalized elliptical variables find application in random matrix theory, which we will discuss in Section 8.2.1.

#### *8.1.2.4 Empirical*

The empirical copula is simply based on the empirical distribution of some sample, modulo a rank transformation. In other words, suppose we have some (vector) sample $(x_1^k, \ldots, x_d^k)$ for $k = 1, \ldots, N$. Then we can define empirical CDFs via

$$\tilde{F}_i(x) = \frac{1}{N} \sum_k 1(x_i^k \leq x) \tag{8.30}$$

The empirical copula is then defined as

$$\tilde{C}(u_1, \ldots, u_d) = \frac{1}{N} \sum_k 1(\tilde{F}_1(x_1^k) \leq u_1, \ldots, \tilde{F}_d(x_d^k) \leq u_1) \tag{8.31}$$

Clearly, $N \cdot \tilde{F}_i(x_i^j)$ is the rank of the point $x_i^j$ within the $i^{\text{th}}$ dimension of the sample.

#### *8.1.2.5 Product (generalized)*

We have already seen a rather trivial example of a copula corresponding to independent variables, namely the product copula $C_\perp(u_1, \ldots, u_d) = u_1 \cdots u_d$. We consider here some extensions of this structure for constructing more general copulas, with the particular application in mind of capturing correlation skew. Introduce a $k$–vector $C$ of copulas (with range $[0,1]^d$), and a $k \times d$ matrix $G$ of strictly increasing functions from $[0,1]$ to $[0,1]$. The elements of $G$ satisfy $\prod_{i=1}^{k} G_{ij}(v) = v$, $\forall j$. Then, the following function

$$\tilde{C}(u_1, \ldots, u_d) = \prod_{i=1}^{k} C_i(C_{i1}(u_1), \ldots, G_{id}(u_d)) \tag{8.32}$$

can be shown to be a copula (see Liebscher [2008] and Lucic [2012]).[18] A sample parameterized example would be $G_{ij}(u) = u^{\theta_{ij}} \cdot$ with $\cdot \sum_{i=1}^{k} \theta_{ij} = 1, \forall j$.

What does such a model gain us? In many applications, there is an asymmetry in how states are to be jointly valued. For example, consider the case of a basket option,

for simplicity on two variables. The payoff takes the form $(w_1 S_1 + w_2 S_2 - K)^+$. (To further fix matters, one may think of $S_{1,2}$ as traded entities whose marginals may be "implied" from market data, say, option prices.) It may be the case that we want a pricing functional that weights joint downward movements differently than joint upward movements. Now, let $C$ denote a 2 vector of Gaussian copulas with different correlation parameters. (It will prove useful to conduct the analysis via transformations to uniforms.) That is,

$$C_i(u_1, u_2) = N_2(N^{-1}(u_1), N^{-1}(u_2); \rho_i) \tag{8.33}$$

where $N$ denotes standard cumulative normal distribution functions (of the appropriate dimension). Take the following choice for $G$:

$$G = \begin{pmatrix} u^{\theta_1} & u^{\theta_2} \\ u^{1-\theta_1} & u^{1-\theta_2} \end{pmatrix} \tag{8.34}$$

and so we write (8.32) as

$$\tilde{C}(u_1, u_2) = N_2(N^{-1}(u^{\theta_1}), N^{-1}(u^{\theta_2}); \rho_1) \cdot N_2(N^{-1}(u^{1-\theta_1}), N^{-1}(u^{1-\theta_2}); \rho_2) \tag{8.35}$$

It can be seen from (8.35) that, depending on the relative magnitudes of the parameters $\theta_{1,2}$, the behavior of the copula $\tilde{C}$ will have different behavior when $u_{1,2}$ are near 1 (corresponding to joint upward movement in prices) and when they are near 0 (corresponding to joint downward movement in prices). For example, if $\theta_{1,2}$ are close to 0 (say, 0.1), then $\tilde{C}$ behaves like $N_2(N^{-1}(u^{\theta_1}), N^{-1}(u^{\theta_2}); \rho_1)$ when $u_{1,2}$ are both small, and like $N_2(N^{-1}(u^{1-\theta_1}), N^{-1}(u^{1-\theta_2}); \rho_2)$ when $u_{1,2}$ are both near 1. Thus, different correlation behaviors can be captured in different regimes (*i.e.*, skew) with models such as (8.32).

#### *8.1.2.6 Vine*

As we have seen, the primary strength of copulas is the ability to separately model marginal and joint distributions. However, in the models considered thus far, the dependency structure is in some sense fixed across the individual RVs in question. For example, a Gaussian or multivariate $t$ copula imposes the same category of dependency across pairs of RVs (*e.g.*, correlation). In some applications, it may be desirable to have different categories of dependency within the overall ensemble of RVs. For example, some pairs (but not others) may exhibit nonsymmetric dependency, whole other pairs may possess heavy tail dependency. So-called vine copulas provide a means of incorporating such heterogeneous dependency structures.

The basic idea is well described by pair copula construction (PCC). Let us illustrate in three dimensions, for variables denoted by $(x_1, x_2, x_3)$. By Bayes law, the underlying density can be written as[19]

$$f_{1,2,3}(x_1, x_2, x_3) = f_{3|1,2}(x_3|x_1, x_2) f_{2|1}(x_2|x_1) f_1(x_1) \tag{8.36}$$

Note that from Bayes law and Sklar's theorem[20] we have

$$\begin{aligned} f_{2|1}(x_2|x_1) &= \frac{f_{1,2}(x_1, x_2)}{f_1(x_1)} = \frac{\frac{\partial^2}{\partial x_1 \partial x_2} C_{1,2}(F_1(x_1), F_2(x_2))}{f_1(x_1)} \\ &= c_{1,2}(F_1(x_1), F_2(x_2)) f_2(x_2) \end{aligned} \tag{8.37}$$

where $c_{1,2}(u_1, u_2) \equiv \frac{\partial^2}{\partial u_1 \partial u_2} C_{1,2}(u_1, u_2)$ is the copula density associated with the copula $C_{1,2}$. Applying the result in (8.37) repeatedly to (8.36), we see that

$$\begin{aligned} &f_{1,2,3}(x_1, x_2, x_3) \\ &\quad = c_{1,3|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \\ &\quad \times c_{2,3}(F_2(x_2), F_3(x_3)) c_{1,2}(F_1(x_1), F_2(x_2)) \\ &\quad \times f_1(x_1) f_2(x_2) f_3(x_3) \end{aligned} \tag{8.38}$$

In addition, we note the result for conditioning on a univariate variable $v$:[21]

$$\begin{aligned} &f(x|v) = c_{x,v}(F_x(x), F_v(v)) f_x(x) \Rightarrow \\ &F(x|v) = \int_{-\infty}^{x} c_{x,v}(F_x(u), F_v(v)) f_x(u) du \\ &\quad = \int_{-\infty}^{x} \frac{\partial^2}{\partial F_x(u) \partial F_v(v)} C_{x,v}(F_x(u), F_v(v)) f_x(u) du = \frac{\partial}{\partial F_v(v)} C_{x,v}(F_x(x), F_v(v)) \end{aligned} \tag{8.39}$$

Observe what (8.38) and (8.39) allow us to do. Making the simplifying assumption that any *parameterization* (but not arguments) of the constituent pair copulas are independent of the conditioning variables, we have great flexibility in constructing multivariate dependency structures using *only* the dependency structures across *pairs* of the RVs in question (as well as the individual marginals, of course). We can thus create models where some pairs of variables have an Archimedean dependence, while others have Gaussian dependence, and still others have Lévy dependence.

In general, these constructions are not unique, and as the number of dimensions increases, the number of possible combinations of pairings increases. There are several different variants for addressing this problem (R-vines, D-vines, C-vines, *etc.*), and we refer the reader to such sources as Czado (2010).

*8.1.2.7 Lévy*

An important extension of copula concepts concerns the class of Lévy processes, which we studied in Chapter 5. For review,[22] these may be thought of as encompassing Brownian motion with drift, plus jumps broadly understood (*i.e.*, standard compound Poisson processes as well as so-called infinite activity jump processes). This very rich class can be extended to include (at least some) stochastic volatility models via stochastic time changes. It is probably useful to recall the cornerstone result, the Lévy-Khintchine formula: $Ee^{i\phi^T x_t} = e^{t\psi(\phi)}$, where $x$ is a $d$–dimensional Lévy process and $\psi$ is called the characteristic exponent of the process and given by

$$\psi(\phi) = i\phi^T\alpha - \frac{1}{2}\phi^T A\phi + \int_{\mathbb{R}^d/\{0\}} (e^{i\phi^T x} - 1 - i\phi^T x 1_{|x|<1})\nu(dx) \tag{8.40}$$

The entity of central importance here is $\nu$, the so-called Lévy measure of the process. It has the interpretation that the expected number of jumps within the time interval $[0, t)$ with amplitudes inside some (Borel) set $B$ is given by

$$E\{\#\text{ jumps } \in B \text{ up to time } t\} = t\int_B \nu(dx) \tag{8.41}$$

Marginal measures may be obtained as follows. Consider component $x_1$. Then we have

$$\begin{aligned} Ee^{i\phi_1 x_1(t)} &= \exp\left(t\left(i\phi_1\alpha_1 - \frac{1}{2}A_{11}\phi_1^2 + \int_{\mathbb{R}^d/\{0\}} (e^{i\phi_1 x} - 1 - i\phi_1 x 1_{|x|<1})\nu\, dx\right)\right) \\ &= \exp\left(t\left(\begin{array}{c} i\phi_1\left(\alpha_1 + \int_{\mathbb{R}^d/\{0\}} x(1_{x_1<1} - 1_{|x|<1})\nu(dx)\right) - \frac{1}{2}A_{11}\phi_1^2 + \\ \int_{\mathbb{R}/\{0\}} dx_1(e^{i\phi_1 x} - 1 - i\phi_1 x 1_{x_1<1}) \int_{R^{d-1}/\{0\}} \nu(x)dx_2\ldots dx_d \end{array}\right)\right) \end{aligned} \tag{8.42}$$

from which we can infer that $\nu_1(x_1) = \int_{\mathbb{R}^{d-1}/\{0\}} \nu(x)dx_2\cdots dx_d$. Observe that independence of the components of $x$ requires that $\nu(x) = \sum_k \left[\nu_k(x_k)\prod_{j\neq k}\delta(x_j)\right]$. In two dimensions, perfect codependence can be represented by $\nu(x_1, x_2) = \nu_1(x_1)\delta(x_2 - x_1)$.

Note that the Lévy measure need not be integrable over the entire real space $\mathbb{R}^d$. However, it is required that $\min(1, |x|^2)\nu(x)$ be integrable near 0 (hence the truncation term in the integrand in (8.40)). Thus, in general it is not possible to separate the arrival (or intensity) of jumps from their amplitude. Only in the special (familiar) case of the compound Poisson process is this possible, where the normalization of the Lévy measure over $\mathbb{R}^d$ allows one to conceive of the probability of *a* jump occurring, distinct from the probability of a particular *sized* jump occurring (because the normalized measure can serve as a valid probability measure of jump amplitudes and the normalization can be interpreted as the jump intensity). The more general framework (of non-integrable Lévy measures) gives rise to the notion of so-called infinite activity jump processes, which have a nonvanishing expected number of jumps of arbitrarily small size over any finite time interval.[23]

One can clearly see now the interpretation of the various terms of the characteristic exponent in (8.40). The first two terms plainly correspond to linear drift and (Gaussian) diffusion, respectively. The term in the integral comprises two effects: a standard Poisson process representing "large" jumps, and a countably infinite number of (Poisson) jumps of increasingly large intensity and decreasingly small amplitudes. This latter point can be seen by writing the relevant terms as

$$\begin{aligned}\int_{\mathbb{R}^d/\{0\}} (e^{i\phi^T x} - 1 - i\phi^T x 1_{|x|<1})\nu(dx) &= \int_{|x|\geq 1} (e^{i\phi^T x} - 1)\nu(dx) \\ &+ \int_{0<|x|\leq 1} (e^{i\phi^T x} - 1 - i\phi^T x)\nu(dx) \\ &= \int_{|x|\geq 1} (e^{i\phi^T x} - 1)\nu(dx) + \sum_n \int_{\frac{1}{2^{n+1}}\leq|x|\leq\frac{1}{2^n}} (e^{i\phi^T x} - 1 - i\phi^T x)\nu(dx) \qquad (8.43)\end{aligned}$$

The first term in the last equation in (8.43) clearly represents the contribution of a standard Poisson process, over jumps of magnitude greater than 1, with intensity given by $\lambda_0 \equiv \int_{|x|\geq 1} \nu(dx)$ and the distribution of jump amplitudes given by $\lambda_0^{-1}\nu(dx)1_{|x|\geq 1}$. The second term corresponds to an infinite sum of Poisson jumps with linear drift, over decreasingly small size and increasingly large arrival rates. (The intensity of the $n$th jump is $\nu(\frac{1}{2^{n+1}} \leq |x| \leq \frac{1}{2^n})$, which is divergent due to the non-integrability of the Lévy measure, and the distribution of jump amplitudes is confined to infinitesimally small intervals.)[24]

Now, a multidimensional Lévy process obviously entails a certain kind of dependency structure (namely, the matrix $A$ in (8.40), which captures the covariance structure of the continuous, diffusive components and the Lévy measure $\nu$, which drives the structure of the discontinuous, jump components). A natural question

then concerns how restrictive[25] this default (so to speak) structure is for modeling joint Lévy processes. We will now begin to investigate this question.

*8.1.2.8 Precursor to dynamics*

With the introduction of Lévy *processes*, we have implicitly introduced the notion of dynamics to the dependency problem, which had hitherto been absent. Recall one of the chief motivations for the copula concept: the ability to specify individual *marginal* behavior separately from joint. However, it does not follow that individual *dynamics* can be specified independently from collective dynamics.[26] In other words, dynamic behavior can be thought of as a continuum of marginal distributions, and it is certainly not clear how the copula concept carries over in such cases, at least not without introducing some notion of dynamics in the copula itself. At a minimum, the standard definition of a copula provides little guidance on the matter.[27]

Time-dependent behavior presents some challenges for applying copulas, and we will see in the context of Lévy processes that certain copulas (called, not surprisingly, Lévy copulas) can be constructed as a *particular* means of dealing with these challenges. Of course, Lévy processes constitute a specific structural assumption, and therefore sacrifice some generality. It is therefore worth briefly discussing the character of dependency in rather more general processes before turning to the particulars of the Lévy class.

### 8.1.3 Dependency: continuous vs. discontinuous processes

*8.1.3.1 General diffusive dynamics*

Generically, a continuous (diffusive) process can be written as

$$dx = \mu(x)dt + \sigma(x)dw_t \tag{8.44}$$

for some vector-valued process $x$. It is important to understand that the dynamics in (8.44) are not well defined as they stand, but are really a shorthand notation for an integral representation

$$x_T = x + \int_t^T \mu(x_s)ds + \int_t^T \sigma(x_s)dw_s \tag{8.45}$$

which *can* be well defined. The obvious challenge lies in specifying what, exactly, is meant by the stochastic differential $dw_t$ (and the associated integration with respect to it). Of course, the approach that dominates the literature (virtually without challenge) is the renowned Itô calculus, in which the appropriate limiting representation of the integrals in (8.45) entails a dominant balance (so to speak) between

deterministic first-order and stochastic second-order terms. Needless to say, Gaussian dynamics (completely defined by their underlying covariance structure) are the natural way of viewing the system in (8.44). (It goes without saying that this issue is of great interest from the perspective of valuation via [dynamic] portfolio formation.)

It is far beyond the scope of this book to consider the possibility of non–Itô-constructed dynamics. The only point we wish to make concerns the compelling reasons to adopt correlation as a dependency measure for continuous processes. This follows from the need to give the dynamics in (8.44) (the natural way to represent stochastic change) a rigorous foundation via (8.45). While, formally, one can always repeatedly draw stochastic differentials from any (marginal) distribution they like, then combine these differentials via their preferred copula, it is very, very far from obvious that this formal procedure has any kind of mathematical meaning (as does the case of, say, affine diffusions).

#### *8.1.3.2 Jump processes*

Continuing with these themes, defining jump dynamics also requires a coherent foundation in an integral representation. As in the continuous case, a particular approach is dominant, namely the compound Poisson process (or more generally a Poisson point process). In this framework, the only kinds of dependencies that can be incorporated between jump drivers are either dependence or independence. A typical example is the canonical class of affine jump diffusions from Section 5.2.3, *e.g.*, (5.76). Here the Itô isometry (familiar from the diffusive case) reads

$$dq_i dq_j = \delta_{ij} dq_j \tag{8.46}$$

Admittedly, this restriction is somewhat restrictive. It is of course compensated (pardon the pun) by the great flexibility that jump modeling provides in capturing certain structural effects of interest (see Sections 5.2.1 and 5.2.3). Still, it remains of interest to see whether the dependencies implicit in the construction of standard diffusive and Poisson dynamics can be made more general.

As already noted, the question of dependency between (dynamic) processes is very general, and does not necessarily concern Lévy processes as such; rather, the latter are a special case of the former, and any notion of dependency that is crafted to Lévy processes comes at the cost that any structural assumption brings. Having said this, the application of copula concepts to joint Lévy processes does serve to nicely illustrate the underlying challenges, and introduces a rich set of analytical tools, as well. We will therefore give the topic some attention now.

### 8.1.4 Consistency: static vs. dynamic

Extending the copula concept to Lévy processes presents a number of challenges that actually serve to clarify the essence of these small-jump effects (as well as facilitating the simulation of such processes, it turns out). Let us recall what the primary objective is. We want a framework for building up joint dependencies while retaining a given marginal structure. However, a few difficulties are encountered. At the risk of belaboring the obvious, Lévy processes are processes, and hence any appropriate copula structure would have to be time dependent, and it is not entirely clear how to incorporate such dynamics in a way that retains the underlying Lévy marginal structure. It is actually worthwhile to consider this point in more detail, as it illustrates themes we have emphasized throughout, namely the idea of volatility as a measure of information accumulation over particular time horizons.

#### *8.1.4.1 Stable processes*

We employ an example used by Tankov and Cont (2003). First we introduce the class of so-called $\alpha$-stable Lévy processes,[28] which have the property that $ax_1 + bx_2 \stackrel{d}{=} cx + d$ with $a^\alpha + b^\alpha = c^\alpha$ for any constants $a, b, c$, and $d$ and some constant $\alpha$ satisfying $0 < \alpha \leq 2$. Here $x_{1,2}$ are independent copies of the underlying Lévy process. An obvious example is a multidimensional zero-mean normal, with $\alpha = 2$. By considering the characteristic function, it can be shown that the Lévy measure for a stable pure jump process in $\mathbb{R}^n$ has the form $r^{-\gamma}\, g(d\mathbf{\Omega})$ in generalized spherical coordinates (here $\mathbf{\Omega}$ is short hand for the angle variables/solid angle element), in which case $\alpha = \gamma - n$. Further consideration of the characteristic function leads to the result

$$Ee^{i\phi^T x_{\beta t}} = e^{t\psi(\beta^{1/\alpha}\phi)} \tag{8.47}$$

from which we conclude the following scaling law:

$$x_{\beta t} \stackrel{d}{=} \beta^{1/\alpha} x_t \tag{8.48}$$

This is clearly true for Gaussian processes, with $\alpha = 2$.

An example with $\alpha = 1$ is the well-known Cauchy process $Z_t$, with (in two dimensions) Lévy measure $\nu(x, y) = (x^2 + y^2)^{-3/2}$ and terminal (time $t$) density $\frac{t}{2\pi}((x^2 + y^2)^2 + t^2)^{-3/2}$. The corresponding copula can be explicitly written out, but its precise form is not important here (see Tankov and Cont [2003] for the actual result). What concerns us here is that this Copula, denoted by $C_Z$, is *not* time dependent, but is also *not* the independence copula: $C_Z(u, v) \neq C^\perp(u, v) \equiv uv$. Now introduce a standard Brownian motion $W_t$ (governed by the independence copula and independent of $Z$) and consider the process $X_t = Z_t + W_t$. Note that $Z$ is 1-stable and $W$ is 2-stable. From the scaling law in (8.48), we see that $X_t/t \stackrel{d}{=} Z_1 + W_{1/t}$ and $X_t/t^{1/2} \stackrel{d}{=} Z_{t^{1/2}} + W_1$. Hence, upon invoking the suitable limiting theorems

(again, see Tankov and Cont [2003] for the details), we see that $X_t/t \xrightarrow{d} Z_1$ as $t \to \infty$ and $X_t/t^{1/2} \xrightarrow{d} W_1$ as $t \to 0+$. Consequently, the large and small time behavior of $X$ is very different, and in particular its copula *is* time dependent,[29] despite the fact that its constituent components are independent with time-independent copulas. This example illustrates two points. One, as claimed, extension of the copula concept for modeling joint dependency for Lévy processes is not straightforward, and there are some complicating issues that require careful attention. Second, we see here the role that time scales play in the analysis of even a simple toy problem. The time scales over which the two processes are distributionally invariant are different, and this fact has implications for how their joint structure behaves, with obvious ramifications for modeling that structure. At a minimum, it would appear that the copula approach is not ideally suited for studying multivariate Lévy processes. But, let us continue the fight.

#### *8.1.4.2 Lévy measures*

More generally, another complication is the fact that the natural characterization of Lévy processes is through the Lévy measure, and *not* through the density or distribution function, which of course is the natural basis (so to speak) of standard copula models. It turns out that many of the usual copula results can be adapted if this alternative viewpoint (*i.e.*, measures as opposed to densities) is adopted, as shown in the work of Tankov (2003) and Tankov and Cont (2003). We will confine attention here to the case where the underlying processes have only positive jumps.[30] The relevant entity here is the so-called (upper) tail integral, defined by

$$U(x) \equiv \nu([x,\infty]) = \int_x^\infty \nu(dx) \tag{8.49}$$

in one dimension, with obvious extensions to higher dimensions. Akin to the relationship between joint distribution functions and marginal distribution function, we have that $U(0,\ldots,x_k,\ldots 0) = U_k(x_k)$, where $U_k$ is the tail measure of the $k^{\text{th}}$ component. (Conventionally, $U(0)$ is defined to be $\infty$ to avoid having to notationally distinguish between infinite and finite activity processes.) We also define the generalized inverse of $U$ to be

$$U^{-1}(y) \equiv \inf\{x > 0 : U(x) \leq y\} \tag{8.50}$$

Another useful result is the series representation of pure jump processes with only positive-valued jumps. These processes are termed subordinators and their characteristic function takes the form[31]

$$Ee^{i\phi x_t} = \exp\left(\int_{\mathbb{R}^d/\{0\}} \left(e^{i\phi^T x} - 1\right) \nu(dx)\right) \tag{8.51}$$

We claim that, in law, the process $x_t$ is equivalent to the following entity:

$$\sum_{i=1}^{\infty} U^{-1}(\Gamma_i) 1(V_i \leq t) \tag{8.52}$$

where $\Gamma_i$ are the arrival times of a standard (unit) Poisson process (so that $\Delta\Gamma_i$ are independent with $\Delta\Gamma_i$ distributed as $e^{-\Delta\Gamma_i}$), and $V_i$ are standard uniform deviates independent of $\Gamma_i$. This equivalence follows from verifying that the characteristic functions agree. First let $n_\theta$ denote the largest $n$ s.t. $\Gamma_n \leq \theta$ for some (arbitrarily large) number $\theta$ and let $X_\theta = \sum_{i=1}^{n_\theta} U^{-1}(\Gamma_i) 1(V_i \leq t)$. Then, by conditioning on $n_\theta$ and using well-known results concerning Poisson arrival times,[32] we find that (letting $u$ and $v$ denote independent uniform random variables)

$$\begin{aligned} Ee^{i\phi X_\theta} &= EE_{n_\theta} e^{i\phi X_\theta} = \sum_n \frac{e^{-\theta}\left(\theta E e^{i\phi U^{-1}(\theta_u) 1(v \leq t)}\right)^n}{n!} \\ &= \exp\left(\theta\left(t\theta^{-1}\int_{U^{-1}(\theta)}^{\infty} e^{i\phi x}\nu(dx) + 1 - t - 1\right)\right) \\ &= \exp\left(t\int_{U^{-1}(\theta)}^{\infty} (e^{i\theta x} - 1)\nu(dx)\right) \end{aligned} \tag{8.53}$$

from which the required result follows by taking the limit $\theta \to \infty$.[33] In fact, this approach is commonly used for simulating Lévy processes (see Asmussen and Glynn [2007]), since it obviously is not very difficult to generate the underlying sequence of exponential and uniform deviates. (This argument employs an approximation that effectively truncates the number of jumps. It is interesting to note that the result can also be derived by considering a small amplitude limit of the truncated Lévy measure, by essentially arguing in reverse: the truncated measure can be normalized to a valid probability distribution, which when applied to the ordered jumps of the process produce a set of ordered uniforms that are equal in law to the arrival times of a standard Poisson process. See El-Bachir [2008]. There is thus a duality between the size of jump amplitudes and the number of jump arrivals.)

*8.1.4.3 Sklar's theorem extended*

We are now in a position to state Tankov's extension of Sklar's theorem for Lévy copulas. First we define a positive Lévy copula. This is a function $F : [0,\infty]^d \to [0,\infty]$ that is (1) increasing (so that $dF$ is a positive measure), (2) satisfies $F(u) = 0$ if at least one component of $u$ is zero, and (3) has uniform marginals ($F_i(z) \equiv F(\infty,\ldots,z_i,\ldots,\infty) = z$). Note that we are essentially operating in terms of the "reverse" distribution function, so to speak: $\Pr(X \geq x) = 1 - \Pr(X \leq x)$. (Compare with Endnote 5, and see (8.4).) Let $x_t$ be a Lévy process in $\mathbb{R}^d$ having only positive jumps in every component, with joint tail integral $U$ and marginal tail integral $U_i$. Then there exists a positive Lévy copula $F$ s.t. $U(x) = F(U_1(x_1),\ldots,U_d(x_d))$. (The converse is also true.) This result not only serves as an analogue to the case of a regular copula, but also allows the time dynamics to be concentrated in the marginal (where it naturally arises due to the underlying Lévy structure), and not the joint structure. Note that the usual representations of, say, independence and perfect co-monotonicity have to be modified. From the discussion following (8.42), the independence Lévy copula becomes

$$F_{\perp}(x_1,\ldots x_n) = \sum_{i=1}^{n} x_i \prod_{j\neq 1} 1(x_j = \infty) \tag{8.54}$$

while in two dimensions, perfect codependence becomes

$$\begin{aligned} U(x_1,x_2) &= \int_{x_1}^{\infty}\int_{x_2}^{\infty} \nu_1(\xi_1)\delta(\xi_2-\xi_1)d\xi_1 d\xi_2 = \int_{x_1}^{\infty} d\xi_1 \nu_1(\xi_1)H(\xi_1-\xi_2) \\ &= U_1(\min(x_1,x_2)) = \min(U_1(x_1),U_1(x_2)) \end{aligned} \tag{8.55}$$

from which we write $F_{\|}(x_1,x_2) = \min(x_1,x_2)$.

An example is the extension of the Archimedean class of copulas. This includes the so-called Clayton family:

$$F_\theta(u,v) = (u^{-\theta} + v^{-\theta})^{-1/\theta} \tag{8.56}$$

for $\theta > 0$. (Contrast with (8.20), in particular the ramifications of the different ranges and domains.) Another example is the generalization of the result in (8.52). The result involves conditional Lévy copulas for generating dependent Poisson arrival times and is reminiscent of the result in (8.16). To demonstrate the intuition, we simply outline the result in two dimensions, for the specific case of the Clayon

Lévy copula. We introduce the conditional distribution function and its inverse from (8.56):

$$\begin{aligned} F_\theta(v|u) &= \frac{\partial}{\partial u} F_\theta(u,v) = \left(1 + \left(\frac{u}{v}\right)^\theta\right)^{-1-1/\theta}, \\ F^{-1}(y|x) &= x(y^{-\frac{\theta}{1+\theta}} - 1)^{-1/\theta} \end{aligned} \tag{8.57}$$

We then have the following representation for a two-dimensional subordinator:

$$\begin{aligned} X_t &= \sum_{i=1}^{\infty} U^{-1}(\Gamma_i) 1(V_i \le t) \\ Y_t &= \sum_{i=1}^{\infty} U^{-1}(F^{-1}(W_i|\Gamma_i)) 1(V_i \le t) \end{aligned} \tag{8.58}$$

where $W_i$ is another sequence of uniforms, independent of $V_i$ (and of course $\Gamma_i$). For more details, see Tankov (2003).

*8.1.4.4 Application: spark spreads*

Finally, we note some applications of Lévy copulas to spark spread modeling in Benth and Kettler (2010) and Meyer-Brandis and Morgan (2014). These models are actually inspired by the original work of Barndorff-Nielsen and Shephard (2001), who considered Ornstein-Uhlenbeck processes driven by Lévy innovations, and also have similarities to the approach taken by Hikspoors and Jaimungal (2007). The basic model takes the form $P^a(t) = \Lambda^a(t)(Y_1^a(t) + Y_2^a(t))$, where the superscript $a$ denotes either electricity or gas (the spark spread is of course given by $P^e - HR \cdot P^g$ for some heat rate). The factor $\Lambda$ represents a seasonality factor, while $Y^1$ and $Y^2$ capture effects of autocorrelation and spikes, respectively, through the follow dynamics:

$$\begin{aligned} dY_1^a &= \kappa_1^a(\theta^a - Y_1^a)dt + \sigma^a dw^a \\ dY_2^a &= \kappa_2^a Y_2^a dt + dL^a \end{aligned} \tag{8.59}$$

where $L^a$ is a subordinator (*i.e.*, only positive-valued jumps are modeled). A joint structure can be imposed on the Lévy terms through a suitable Lévy copula, say Clayton. The Brownian terms can be correlated, but are independent of the Lévy components. Not surprisingly, given the central role the characteristic exponent plays in defining Lévy processes, option-pricing results can be obtained via methods previously discussed (in Chapter 5). For example (dropping the superscripts for convenience), $Y_2(T) = e^{-\kappa_2(T-t)}Y_2 + \int_t^T e^{\kappa_2 s} dL_s$ so using the i.i.d. property of Lévy processes, we find that

$$E_t e^{i\phi Y_2(T)} = \exp(i\phi e^{-\kappa_2(T-t)} Y_2) E_t \exp\left(\int_t^T e^{-\kappa_2(T-s)} dL_s\right)$$
$$= \exp\left(i\phi e^{-\kappa_2(T-t)} Y_2 + \int_t^T \psi_{L_2}(\phi e^{-\kappa_2(T-s)}) ds\right) \tag{8.60}$$

where $\psi_{L_2}$ is the characteristic exponent of $L_2$. Meyer-Brandis and Morgan (2014) derive expressions for the spread option formula, using by-now familiar methods involving characteristic functions.

### 8.1.5 Wishart processes

While we have endeavored to examine concepts of dependency that are much broader than the familiar class of (linear) correlation, it remains the case that correlation (and more generally, covariance) is a very useful concept, when properly employed. In the canonical affine processes studied in Chapter 5, instantaneous correlation/covariance between the underlying Brownian drivers[34] manifests itself indirectly in some global (so to speak) dependency structure that is in general difficult to characterize systematically. With the Lévy copulas of the preceding subsection, a global dependency structure could be imposed upon a special category of pure jump processes. Here, we consider a third alternative, namely *directly* modeling the dynamics of (stochastic) correlation/covariance. This leads us to investigate *matrix* affine jump diffusions, a special case of which include the (somewhat) well-known class of Wishart processes.

#### *8.1.5.1 Affine representation*

We start with a model from Leippold and Trojani (2010). Denoting by $S_n^+$ the cone[35] of symmetric, positive semi-definite $n \times n$ matrices, the dynamics of a matrix $\Sigma \subset S_n^+$ are specified as

$$d\Sigma = (\Omega\Omega^T + M\Sigma + \Sigma M^T)dt + \sqrt{\Sigma} \cdot dW \cdot Q + Q^T \cdot dW^T \cdot \sqrt{\Sigma} + dJ \tag{8.61}$$

with $\Omega$, $M$, and $Q$ $n \times n$ real matrices, $W$ a $n \times n$ matrix of standard Brownian motions, and $J$ a pure jump process in $S_n^+$. The matrix square root for symmetric matrices is defined in terms of the familiar eigenvalue-eigenvector factorization by

$$\Sigma = V^T \Lambda V \Rightarrow \sqrt{\Sigma} = V^T \sqrt{\Lambda} V \tag{8.62}$$

with $V$ a matrix of eigenvectors of $\Sigma$ (arranged by column) and $\Lambda$ the corresponding diagonal matrix of eigenvalues.[36] (Of course the square root of a diagonal matrix is simply given by $(\sqrt{\Lambda})_{ij} \equiv \delta_{ij}\sqrt{\Lambda_{ii}}$.) As a technical point, we impose

the additional restriction that $\Omega\Omega^T \gg (n-1)Q^TQ$ to ensure that $\sum$ is positive semi-definite. With these restrictions, and the assumption that the jump intensity have the affine form $\lambda_J(\sum) = \lambda_0 + \mathrm{Tr}(\lambda_1 \Sigma)$ for $\lambda_0 \geq 0$ and $\lambda_1 \in S_n^+$, the process in (8.61) provides coherent dynamics for modeling (positive-definite) stochastic covariances/correlations. (See Leippold and Trojani [2010] and the references therein.)

The special form $\Omega\Omega^T = \beta Q^TQ$ (for non-negative $\beta$) and no jumps leads to the so-called Wishart process, with the stronger requirement that $\beta > n+1$ ensuring that $\Sigma$ is positive definite a.s. We call attention to this special case because of the connection to the Wishart distribution in sampling distributions discussed in Section 6.2.2. However, there is obviously merit in studying the more general case in (8.61), so we will take a broader perspective here. (On Wishart processes as such, see Bru [1991].) We illustrate with a generalization of our old friend, the Heston stochastic volatility model (see da Fonseca *et al.*, 2007, 2008). We start with a (vector) price process with dynamics given by

$$\frac{dS_i}{S_i} = \mu_i dt + \Sigma_{ij}^{1/2} dw_j \tag{8.63}$$

where the matrix process $\Sigma$ follows (8.61) with $J = 0$ (*i.e.*, no jumps). The vector $w$ in (8.63) is a vector of standard Brownian motions. We will shortly specify the relation between the Brownian drivers of the return processes (8.63) and the covariance processes (8.61). In terms of log-prices $z$ we have that

$$\begin{aligned} dz_i = d\log S_i &= \mu_i dt + \Sigma_{ij}^{1/2} dw_j - \frac{1}{2}\Sigma_{ij}^{1/2} dw_j \Sigma_{ij'}^{1/2} dw_{j'} \\ &= \left(\mu_i - \frac{1}{2}\Sigma_{ii}\right) dt + \Sigma_{ij}^{1/2} dw_j \end{aligned} \tag{8.64}$$

For convenience we write (8.61) (*sans* jumps) in index notation for the Wishart case:

$$d\Sigma_{ij} = (\beta Q_{ki}Q_{kj} + M_{ik}\Sigma_{kj} + \Sigma_{ik}M_{jk})dt + \Sigma_{ik}^{1/2} dW_{kl} Q_{lj} + Q_{ki} dW_{lk} \Sigma_{lj}^{1/2c} \tag{8.65}$$

*8.1.5.2 Matrix Riccati structure*

Owing to the *individual* affine form of (8.64) and (8.65), we have hope that we can apply the methods developed in Chapter 5 to derive a system of ODEs that determine the characteristic function of the (log-) price process, and consequently the terminal distribution of the process. This requires, as already mentioned, that we specify the relation (correlation) between the constituent Brownian drivers. To retain the desirable affine structure this requires that we assume (See Da Fonesca *et al.*, 2007)

$$dw_j dW_{kl} = \delta_{jk}\rho_{kl} dt \tag{8.66}$$

We already saw in Section 6.3 the important role played by the Wishart distribution in the question of sampling distributions.[37] However, since the presence of stochastic processes that are inherently matrix entities introduces some important complications to the basic affine problems studied in Chapter 5, we devote a bit more space to this issue here. So, letting $f = E_t e^{i\phi_k z_k(T)}$ denote the (conditional) characteristic function of the log-price $z$, we see that $f$ satisfies the following PDE (employing the summation convention):

$$f_t + (\mu_{i-} \tfrac{1}{2}\Sigma_{ii}) f_{z_i} + (\beta Q_{ki} Q_{kj} + M_{ik}\Sigma_{kj} + \Sigma_{ik} M_{jk}) f_{\Sigma_{ij}} + \tfrac{1}{2}\Sigma_{ij} f_{z_i z_j} + 2\Sigma_{ii'} Q_{lj} Q_{lj'} f_{\Sigma_{ij}\Sigma_{i'j'}} + (\Sigma_{ii'}\rho_l Q_{lj} + \Sigma_{i'j}\rho_k Q_{ki}) f_{\Sigma_{ij} z_{i'}} = 0 \tag{8.67}$$

where we have used the independence within $W$ and $w$, the cross-structural relation in (8.66), and exploited the underlying symmetry of $\Sigma$. By looking for a solution to (8.67) of the familiar form $f = \exp(A_{ij}\Sigma_{ij} + B_i z_i + C)$, we find that $B = i\phi$ and that $A$ and $C$ satisfy the following *matrix* Riccati ODE:

$$\begin{aligned} \dot{A} &= A(M + iQ^T\rho\phi^T) + (M^T + i\phi\rho^T Q)A + 2AQ^T QA + \gamma \\ \dot{C} &= i\phi^T\mu + \beta \mathrm{Tr}(AQ^T Q) \end{aligned} \tag{8.68}$$

where $\gamma \equiv -c\frac{1}{2}(\phi\phi^T + i\mathrm{diag}(\phi))$. In (8.68) we have made the substitution $t \to T - t$ so the initial conditions become $A(0) = C(0) = 0$.

Matrix Riccati ODEs, like their scalar counterparts, are reasonably amenable to analysis. For example, a solution for $A$ can be written in the following form: $A = A_2 A_1^{-1}$, where $A_{1,2}$ satisfy the following *linear* system:

$$\begin{pmatrix} \dot{A}_1 \\ \dot{A}_2 \end{pmatrix} = \begin{pmatrix} -M - iQ^T\rho\phi^T & 2Q^T Q \\ \gamma & M^T + i\phi\rho^T Q \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \tag{8.69}$$

as can be verified using $\dot{A} = \dot{A}_2 A_1^{-1} - A_2 A_1^{-1}\dot{A}_1 A_1^{-1}$. For more on matrix Riccati ODEs, see Benner and Mena (2004).

## 8.2 Signal and noise in portfolio construction

As we have emphasized repeatedly, valuation entails the formation of portfolios via hedging and proper accounting for unhedged risk. Dynamic aspects of this problem were examined in Chapter 3. In this section we will consider aspects of the static portfolio construction problem that are greatly dependent on estimation, or more accurately, sensitive to estimation noise. To provide further context, a central concern here will be sample size in comparison to the number of assets in question. It is

not uncommon in energy markets to encounter situations where the sample size is of the same order of magnitude as the number of underlyings. For example, in load valuation problems, we are often concerned with how expected price and load co-move over some time horizon. Often, such contracts apply to a seasonal term, such as summer months. To the extent that liquid futures markets convey useful information about realized prices (and to the extent that we can form projections of load based on other, more stationary drivers such as weather), we can investigate historically how realizations diverged from expectations, as comprised as a set tailored to the term in question. In many cases, the available data pertinent to this situation is comparable in size to the number of constituent factors (*e.g.*, pricing a summer load deal requires knowledge of 6 drivers over a specific term, and there may be less than 10 years of [seasonal] futures data). It thus becomes challenging to be able to distinguish structure from noise in these co-movements, as the dimensionality of the problem effectively increases.

### 8.2.1 Random matrices

#### *8.2.1.1 Markowitz reviewed*

Let us first review very briefly the textbook Markowitz portfolio optimization problem. A generic formulation is the selection of portfolio weights (long and short positions are allowed) $w$ across $N$ assets, with returns $\mu$ and covariance matrix $\Sigma$, such that a given rate of (expected) return $\mu_0$ is attained at minimal portfolio variance:

$$\begin{aligned} &\min\ w^T \Sigma w \\ &\text{st } w^T \mu = \mu_0 \end{aligned} \tag{8.70}$$

Standard application of Lagrange multipliers yields optimal portfolio weights

$$w^* = \mu_0 \frac{\Sigma^{-1}\mu}{\mu^T \Sigma^{-1} \mu} \tag{8.71}$$

As already noted, portfolio optimization theory is standard material, and this very simple framework is more than adequate for our purposes here. Let us now look at alternative mathematical representations of (8.71) that will help tease out the econometric issues that are of interest to us.

#### *8.2.1.2 Eigenportfolios*

Since a covariance matrix is necessarily symmetric, it permits an eigenvalue-eigenvector factorization of the following form:

$$\Sigma V = V\Lambda \Rightarrow \Sigma = V\Lambda V^T \tag{8.72}$$

where $\Lambda$ is a diagonal matrix of the eigenvalues $\lambda_i$ of $\sum$ and $V$ is an orthogonal matrix whose columns are the eigenvectors of $\Sigma$ (column $i$ corresponds to the $i^{\text{th}}$ diagonal element of $\Lambda$). Now, each of these eigenvectors can be thought of as a portfolio itself, which we will term *eigenportfolios* (for obvious reasons)[38]. The (expected) returns of the eigenportfolios are given by $g = V^T \mu$. Note that any portfolio $w$ can be expressed in terms of eigenportfolios $u$ via $w = Vu$; the expected return is independent of the representation since $w^T \mu = u^T g$. (As we know, however, representations that are equivalent mathematically may not necessarily be equivalent econometrically, and we will soon see how an eigenportfolio representation can be superior in identifying noise components in the portfolio optimization problem.)

We see from (8.71) that the optimal portfolio weights can be expressed in terms of eigenportfolios as

$$w^* \infty V \Lambda^{-1} g \tag{8.73}$$

The vector $\Lambda^{-1} g$ has elements $g_i/\lambda_i$. Thus, the optimal weights can be decomposed into eigenportfolios, with more weight given to eigenportfolios corresponding to small eigenvalues (or more accurately, small in comparison to the associated eigenreturn). Note that, from $V^T \Sigma V = \Lambda$, we see that the eigenportfolios can be ranked in terms of riskiness according to the size of the eigenvalues of the covariance matrix. We will understand better the significance of the covariance eigenvalues in general when we look at principal component analysis in Section 8.2.2. The point we want to emphasize here is that the optimal portfolio weights will be highly sensitive to any small eigenvalues (as a general rule, the more highly correlated the assets are, the more small eigenvalues there will be). This is not too surprising: low-risk eigenportfolios will tend to be favored in a portfolio variance minimization problem.

Why is the question of sensitivity important here? In practice, we of course never know the true value of the covariance matrix, and can only use an estimated value in any portfolio optimization.[39] For example, a common estimator for the covariance is the sample covariance:

$$\hat{\Sigma} = \frac{1}{T} \sum_t x_t x_t^T \tag{8.74}$$

for some sample of a vector of returns $x_t$.[40] In matrix form, the expression in (8.74) can be written as $\hat{\Sigma} = \frac{1}{T} X X^T$, where $X$ is a $N \times T$ matrix of returns (each column corresponds to a return in the sample, each row is a sample of returns for a particular asset). The behavior of ensembles such as (8.74) obviously has relevance for the question of sampling distributions. The point here is that the estimator in (8.74) obviously depends on the realized sample, which is of course random. (Trivially, a different realized sample produces a different estimate of the covariance matrix.) We have called attention throughout this chapter to the challenges presented by small sample sizes. The issue of interest here is related, but slightly different. Note that there are two dimensions (so to speak) to the problem: sample size, *and* the

number of assets. If the number of assets is small relative to the sample size (say, four vs. 1,000), then one would expect the sample covariance to be a reasonable estimator. However, if the number of assets is comparable to the sample size (*e.g.*, 500 samples of 100 assets), then the answer is no longer clear. In other words, we are interested in the limiting cases $T \gg 1$, $N \gg 1$, $q \equiv N/T = O(1)$.

The role of eigenvalues is again important here. Note that, for $N > T$ (*i.e.*, more assets than samples), the matrix $XX^T$ has $N - T$ eigenvalues equal to zero.[41] This follows simply from the fact that the null-space of $X^T$ has dimensionality $N - T$ in this situation. Now, these zero eigenvalues (of the sample covariance) are spurious, as they do not correspond to any corresponding singular property of the population covariance. Indeed, as the sample size increases for the assets in question (that is, $T \to \infty$ for $N$ fixed), all of the eigenvalues of the sample covariance are nonzero and have real informational content. These considerations lead to the following question in the intermediate case of interest: how significant (in the common language sense) are small eigenvalues in a given sample (referring to the number of assets in relation to sample size)? In other words, for a particular sample covariance, are the small eigenvalues meaningful, or do they simply represent noise? This is a very important question in the context of portfolio optimization since, as we have seen, the small eigenvalue/low risk eigenportfolios are given a disproportionately large share of the weight. It is imperative that we avoid optimizing to noise. In other words, do low-risk eigenportfolios really exist, or are they just a manifestation of insufficient data?

To understand the effects of covariance estimator sensitivity on portfolio optimization, it is useful to examine some results from random matrix theory.

#### *8.2.1.3 Eigenvalue distribution*

Since we anticipate that some of the sample covariance matrix eigenvalues have real informational content (while some do not), we can approach the problem as follows. Can we reconstruct the estimated covariance matrix using the large ("real") eigenvalues, while filtering or otherwise cleaning up the small ("spurious") eigenvalues? Our objective is to have a systematic means of doing so.[42] Results from random matrix theory that concern the spectrum of entities of $XX^T$ provide some insights. As usual, we will mainly focus on the basics; the interested reader can consult sources such as Bouchaud and Potters (2009) or Laloux *et al.* (1999).

The central ideas can be conveyed with the simplest possible example, namely the spectrum of the empirical correlation matrix of i.i.d. assets (with unit variance). In this case, the density of the eigenvalues is given by the celebrated Marcenko-Pastur (MP) law:

$$p(\lambda) = \frac{1}{2\pi\lambda q}\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})} \tag{8.75}$$

for $q < 1$ and where $\lambda_{\max/\min} = (1 \pm \sqrt{q})^2$. (For $q > 1$ the density has a point mass [Dirac delta function] of weight $1 - 1/q$ at zero; for $q = 1$ the familiar Wigner semicircle law can be recovered.)

The result in (8.75) indicates that the spectrum of the correlation matrix of a large sample of a large number of i.i.d. random variables is bounded. We should therefore take the existence of eigenvalues of an empirical correlation matrix outside this range as an indication that the sample in question is *not* i.i.d., and that those eigenvalues may indeed have informational content. It is not uncommon for a large dimensional stochastic system to be well described by a much smaller system of primary drivers, along with a larger system that can be characterized as noise. A well-known example is the natural gas (Henry Hub) forward curve. Individual contracts on the curve tend to move up and down in concert, with neighboring months (especially within a season) being highly correlated. We will see more on this concept in Section 8.2.2 on principal component analysis, when it will be shown that it is precisely the largest eigenvalues that correspond to these more primary drivers, with the associated eigenvectors indicating the direction (properly understood) of those drivers. (In the natural gas example, the primary direction would correspond to an equal weighting of all months under consideration, *e.g.*, the winter contracts.)

We thus expect those (larger) eigenvalues outside the MP band to be significant (in the plain language sense of the word) while those (smaller) ones inside the band to be representative of noise, and filtered out (so to speak) from subsequent analysis. How can this be done? A straightforward approach is to simply not distinguish between these noise components. That is, they are essentially blended while retaining the total trace of the empirical correlation matrix.[43] Consider the spectral decomposition of a (symmetric) matrix in terms of its eigenvalues/eigenvectors:

$$V = \sum_{i=1}^{N} \lambda_i v_i v_i^T \tag{8.76}$$

with the eigenvalues sorted in decreasing order. Suppose we retain the $K$ largest eigenvalues. Then we can define a cleaned correlation matrix via

$$\tilde{V} = \sum_{i=1}^{k} \lambda_i v_i v_i^T + a I_N \tag{8.77}$$

with $a = \frac{1}{N} \sum_{i=k+1}^{N} \lambda_i$ and $I_N$ the $N \times N$ identity matrix. Technically, $\tilde{V}$ is not a correlation matrix, since its diagonal elements are not all ones, so a suitable normalization as in Rebonato and Jäckel [1999] should be applied. An alternative cleaning would be to simply replace the diagonal elements of $\sum_{i=1}^{K} \lambda_i v_i v_i^T$.

*8.2.1.4 Spectral estimators and robust covariance estimation*

Let us return to the elliptically distributed random variables from Section 8.1.2. (as well as their generalizations). Recall that these were defined through their characteristic function and contained as a special case (joint) Gaussians. Specifically, if $y$ is elliptically distributed with zero mean, then $Ee^{i\phi^T y} = g(\phi^T \Sigma \phi)$ and, if the covariance matrix exists, it can be associated (up to a scaling) with the so-called dispersion matrix $\Sigma$. Apart from the special case of normality, however, it will not in general *be* the covariance matrix, and in particular the usual ML covariance estimator (that is, $\frac{1}{N}\langle y_i y_i^T \rangle$) will not correspond to the dispersion matrix. The correct MLE result is obtained as follows. Recall from Endnote 14 that the density of an (zero-mean) elliptical variable can be written as $\frac{1}{\sqrt{\det \Sigma}} h(y^T \Sigma^{-1} y)$. Applying some matrix derivative results, we find the ML estimator is given by the fixed point equation

$$\hat{\Sigma} = \frac{2}{N} \left\langle L'(y^T \hat{\Sigma}^{-1} y) y y^T \right\rangle \tag{8.78}$$

where $L(x) \equiv -\log h(x)$. For the Gaussian case, $h \propto e^{-x/2}$ and the usual MLE result is easily recovered; note also the scale invariance in (8.78).

Departures from the special Gaussian case give rise to some interesting phenomena. We recall again (from Section 8.1.2) the fact that, for elliptically distributed variables, lack of correlation is *not* the same as independence (the reverse is obviously always true). Consequently, results such as the MP law that consider the spectrum of the correlation matrix with i.i.d. run the risk of being misapplied, such as in eigenvalue cleaning. A good example of this problem can be found in Frahm and Jaekel (2008).

*8.2.1.5 Shrinkage*

So-called shrinkage methods shift the empirical correlation matrix closer to the identity:

$$\Sigma_\alpha = \alpha \hat{\Sigma} + (1 - \alpha) I \tag{8.79}$$

for $0 \leq \alpha \leq 1$. Thus, the new eigenvalues are given by $\lambda_\alpha = 1 + \alpha(\hat{\lambda} - 1)$. Shrinkage estimators are essentially Bayesian techniques that appropriately adjust the "prior" weighting of assets in accordance to the their relative returns, *i.e.*, minimal diversification; see (8.71). As such, the largest eigenvalues (that is, those with the most informational content) are generally least affected by this "updating." In truth, any appropriate matrix that could serve the role as a prior estimate could be employed in (8.79), such as a matrix with ones along the diagonal and all off-diagonal elements equal to the average empirical correlations (akin to the cleaning method in (8.77)). The choice of the adjustment parameter $\alpha$ is rather problem dependent. In line with the Bayesian underpinnings, for problems where the signal-to-noise ratio is believed to be large/small, $\alpha$ should be chosen close to 1/0. Alternatively, if a high

(low) degree of diversification is desired, $\alpha$ should correspondingly be small (large). For a nonlinear extension of shrinkage, see Ledoit and Wolf (2014).

### 8.2.2 Principal components and related concepts

Another useful application of the eigenvalue-eigenvector structural decomposition is the well-known Principal Components Analysis (PCA). To understand the main idea, consider a random vector $x$ with zero mean and covariance matrix $\Sigma$. Now, introduce a linear transformation of $x$ via $y = Ax$. We would first like to choose the matrix $A$ to render the transformed variable independent (or at least de-correlated). Equation (8.72) provides the necessary result: if $A = V^T$ (where $V$ is the matrix of [orthogonal] eigenvectors of the covariance matrix), then $Eyy^T = V^T Exx^T V = V^T \Sigma V = \Lambda$. Note that the inverse transformation is trivial: $x = Vy$. Thus, we have a decomposition of the variable $x$ in terms of the eigenvectors of the covariance matrix, with coefficients given by the (uncorrelated) components of the random vector $y$.

There is yet another aspect of this expansion. In many applications, a very large percentage of the covariance matrix eigenvalues are much smaller than the largest eigenvalues (*e.g.*, the top two or three). For example, for the popular correlation form $\rho^{|i-j|}$, with $\rho = 0.95$ we have the following eigen decomposition:

$$\begin{pmatrix} 1 & 0.95 & 0.90 & 0.86 \\ & 1 & 0.95 & 0.90 \\ & & 1 & 0.95 \\ & & & 1 \end{pmatrix} \Rightarrow$$

$$\Lambda = \text{dig}\begin{pmatrix} 3.76 \\ 0.16 \\ 0.05 \\ 0.03 \end{pmatrix}, V = \begin{pmatrix} 0.49 & -0.65 & -0.51 & 0.27 \\ 0.51 & -0.27 & 0.49 & -0.65 \\ 0.51 & 0.27 & 0.49 & 0.65 \\ 0.49 & 0.65 & -0.51 & -0.27 \end{pmatrix} \tag{8.80}$$

It can be seen that one eigenvalue is much larger than the rest. The variance of the transformed entity $y_1$ is (in this case) 94% of the total variance (across all the components).[44] The corresponding eigenvector thus represents the basis of a subspace capturing the bulk of the variability of the original variable $x$, which here has an effective (so to speak) dimension of one.[45] We thus anticipate that in many applications, the following approximation can be utilized:

$$x_i \approx V_{i1} y_1 \tag{8.81}$$

with $y_1$ having variance $\lambda_1$ (the largest eigenvalue of the covariance matrix). So, a common application of PCA is to employ these primary drivers/factors in place of the constituent variables of the full system. Instead of having to model (say) all

the variables, it is often sufficient to consider just a few, representing significant dimension reduction.

There is a well-known interpretation to the first three components of a PCA analysis, especially when these components comprise the great bulk of total variance. The first component is of course the so-called market mode or market portfolio, with more-or-less equal weights assigned to all (normalized) assets. The next component consists of antisymmetric positions across the assets, and thus various spread positions. The third component consists of symmetric positions across assets, and so represents aggregations of pairs of assets.[46] For a given portfolio with normalized weights denoted by $w_i$, the portfolio variance can be written as

$$w^T \Sigma w = w^T V \Lambda V^T w = u^T \Lambda u = \lambda_i u_i^2 \tag{8.82}$$

where $u \equiv V^T w$ are the portfolio weights in terms of eigenportfolios. Since $u^T u = w^T w = 1$, we see that the portfolio variance can be decomposed into a weighted average of the variance of the various factors. In particular, the more aligned the portfolio is with the market mode, the higher the portfolio variance. Consequently, PCA can provide a means of identifying portfolios that are market-neutral, *i.e.*, portfolios whose variance is attributable to asset spreads rather than overall market moves as such.[47]

# Notes

## 1 Synopsis of Selected Energy Markets and Structures

1. More recent expositions include Wolyniec (2015), Swindle (2014), and Geman (2009).
2. We assume that the reader has a basic familiarity and understanding of such basic instruments as futures and options.
3. We will define the precise manner in which we use terms such as "stable" in Chapter 2.
4. Supply shocks such as outages can further force high-cost units into service.
5. For crude oil. For the most part, natural gas time series date back to the mid-1990s, and electricity data typically begins in the early 2000s.
6. Despite obvious biases, McLean and Elkind (2004) has a useful overview of this transition.
7. And going even further, the exposure may be related not to the Northeast as such, but to Boston as opposed to New York.
8. As is well known, futures and forwards are not the same, although they are numerically equal as prices when discount rates are independent of the underlying asset. Futures are marked-to-market daily (futures are costless to enter into and entail an accrued cash flow), whereas forwards are not. (Mathematically, futures are martingales with the money market unit of account as numeraire, while forwards are martingales with the zero-coupon bond as numeraire.) Futures are thus appropriate for (dynamic) hedging strategies. We will generally ignore the distinction throughout. For more details see Björk (2009).
9. We are referring of course to the distinction between periods of high demand (on-peak, typically weekdays during business hours) and low demand (off-peak, typically weekends and nighttime).
10. In addition, there is always the possibility of operational failure (outages), which may amount to simply a derate of expected revenue (when the presence or absence of the unit in the generation stack does not materially affect power prices) or may substantially alter the hedging/replication strategy we employ (*e.g.*, how many option positions we put on against the plant), and thus how we value the plant.

11. In the terminology of econometrics, the issue here is one of stationarity vs. non-stationarity, or the extent to which expectations of the future can be formed based on unconditional as opposed to conditional information. These concepts will be covered in Chapter 2.
12. U.S. power prices are denominated in units of $/MWh (megawatt hours) and natural gas prices in units of $/MMBtu. Heat rates are conversion factors that reflect the efficiency of a unit in converting fuel inputs into power outputs.
13. As well, there could be fixed volumes of fuel required for start-up.
14. And if possible, finding upper bounds as close as possible to the lower bound.
15. The convention in (1.2) is in terms of delivery volume, which requires the (per-unit) receipt price to be grossed up.
16. In truth, there can be multiple delivery and/or receipt points, pipe segmentation, *etc.* These remain generalized spread structures, however.
17. We leave aside the question of whether hub variability, which for Henry Hub is reflected in market instruments through option prices, is related to basis variability.
18. Mathematically, basis looks (at least far enough from maturity) more like a (discontinuous) pure jump process as opposed to a (continuous) diffusive process.
19. We mention here that some storage facilities permit injection from/withdrawal to multiple physical locations.
20. And summed over each hour in the term, of course. We ignore this feature here as it is not material to the exposition.
21. The reason for this is not too surprising. Power prices and load (demand) are both related economic entities, hence the information flows that drive power should be expected to drive load, as well. (We will see in Chapter 2 the connection between information flow and variance accumulation.) Thus, loads, even system loads that have a seasonal pattern (similar to temperature), should be expected to be fundamentally distinct from such more-or-less periodic processes such as weather (even when the latter is an important driver of the former).
22. Load is an interesting structure because although it has (or may have) a vega, as a bilinear product it has a delta but no gamma.
23. We are merely illustrating a point here. As we will see, variances, correlations, *etc.* may not be relevant for a particular valuation problem (although they may well be).
24. For obvious reasons, similar points apply to projecting (monthly) temperature.
25. In truth, even loads dominated by seasonal effects display such features; *e.g.*, within their overall seasonal structure, system loads have had a discernable downward shift (demand destruction) after the crisis in 2008.
26. In truth, as we shall see, hedging does not so much create residual risk as it entails a *transformation* of risk.

## 2 Data Analysis and Statistical Issues

1. Often it is better to craft the problem and subsequent analysis in terms of *ratios*, but this will merely clutter the notation so we will employ an additive representation here.
2. The disturbance is a *population* entity. That is, it is a property of the DGP as such. In contrast, the more widely used term residual to be discussed later (and which we have already used in its common language sense) is a *sample* property. That is, it is a property of a particular realization of some DGP, or more accurately of some estimator operating on a realization of that DGP. The distinction between population and sample is extremely important, but we will tend not to overemphasize the subtle (but nonetheless real) difference as it applies to disturbance vs. residual.
3. Plainly, for long exposure we would want to reduce the price we bid, and for a short exposure we would want to increase the price we ask (in which case we would be more concerned with, say, the 75th percentile of the residual distribution).
4. For example, depending on the particular product, the entity of interest (or more accurately its statistical properties) may be the ratio $y_T/x_T$ (for heat-rate products) or $y_T - x_T$ (for basis products). Thus, the structure in question may dictate that a particular *function* of the portfolio components be analyzed econometrically, as opposed to analyzing the components as such.
5. Another practical example would be load modeling, with a single year's worth of hourly data. With $365 \cdot 24 = 8760$ points, this sample may appear to be quite sizeable, and at the *hourly* level, it may well be. However, it nonetheless represents only one *year* of data, so an econometric analysis that fails to adequately account for time scales (or more accurately, the conditional information that is operative at different time scales) may simply involve optimizing to the particulars of that year (*i.e.*, fitting to noise). Depending on the relative importance of information flows at different time scales (a question that is of course deal dependent), a particular sample may be huge or tiny.
6. As we shall see, even this distinction is somewhat empty, as there are in fact degrees of detail that can be encompassed continuously under the umbrella of variance scaling laws. Over the appropriate time horizon, however, the distinction does indeed make sense.
7. This is actually not entirely true because, as we shall see in the next section, the variance scaling law of temperature *does* imply some propagation of information across more than one month's time, and this effect can have great impact on valuing products which are highly sensitive to weather (such as load-serving contracts). However, from an illustrative purpose, such effects are not critically important. True enough that a hot month presently will likely be followed by a

hotter-than-normal month, but the relevant question is *how much* hotter, and there is little doubt that the effect dissipates fairly quickly with time.

8. We actually alluded to this concept in the introductory example.
9. Since spot commodities as such do not trade (physical delivery or possession must be made, implying the necessary infrastructure or architecture for doing so), there is no economic or financial reason that commodity (spot) prices must be martingales, even under a pricing measure.
10. More generally (and accurately), the important concept is ergodicity, which essentially means that population entities (such as expectations) can be well proxied by sample entities (such as time averages), when the sample is sufficiently large. It is a question of inferring information about a population from pathwise information. For a *long enough* sample, stationarity should be the dominant effect for ergodic series. We will have much more to say on this topic of asymptotic diagnostics.
11. We will discuss shortly what we mean here by "degree."
12. Or more generally, any deal with volumetric risk. Volume as such does not trade, hence its covariation with price must be related to some other market instrument (usually options or other volatility-dependent structures [such as tolls]; typically expected volume is accounted for via futures positions).
13. Put differently, it is hard to subdivide any realization of this series into independent pieces unless the sample is extremely large: current information persists for a long time.
14. By which we mean: the estimator will behave very differently from its associated theory. Consequently, this (asymptotic) theory can give a very misleading picture of estimator stability (that is, in any given sample there is a very definite possibility that the estimator is simply optimizing to noise).
15. The chief assumption, apart from the normality of the deviations, is non-stochasticity of the regressors $x$. More generally, the regressors can be stochastic but independent of the deviations, in which case the properties of the estimator must be thought of as conditional (on the regressors). Another common generalization is auto-correlation between deviations. (The critical feature that must be retained is independence of regressors and disturbance, lest one confront so-called identification issues.) Even very slight weakening of the classical OLS assumptions renders many statistical properties of the resulting estimators only asymptotically valid, *i.e.*, valid in the limit of large sample size. Consult Hamilton (1994) for a thorough discussion.
16. The manner in which the estimated variance is obtained will be clear momentarily when the notion of unbiasedness is discussed.
17. A similar expression could be crafted for $\hat{\beta}$ but it contributes nothing to the point we wish to make.
18. For the standard assumptions of OLS, (2.8) implies normality of the estimator, with (2.9) and (2.10) providing the mean and variance, respectively. However,

this result is not useful for diagnostics about the estimator, as the variance $\sigma^2$ is in general not known and so could not be employed in any diagnostics. Rather, *estimates* of the residual variance must be used (see (2.7)), hence the precise diagnostic form will in general be non-normal (but in the case of standard OLS, *will* be standard, *e.g.*, involving the $t$-distribution [hence the term $t$-statistic]). However, many asymptotic results *are* normal, so results such as (2.10) do have some utility, if only pedagogical. Note that as the number of (non-stochastic) regressors $x$ increases, the variance in (2.10) shrinks like the reciprocal of sample size.

19. To understand the idea, one can ask how likely it would be for a putative unit normal to manifest itself as 5 standard deviations, say, above or below its mean. For the classic OLS model the exact distributional properties of the estimator can be obtained, but in general there is only recourse to *asymptotic* results for the diagnostics (*i.e.*, the distribution of the estimators for very large sample size). We will discuss some of the perils and pitfalls of relying on asymptotic results later, but thinking in terms of normally distributed estimators will serve to convey the necessary intuition.
20. We do not intend to dwell on philosophical matters to any great extent, but technically speaking non-rejection or failure to reject is not the same thing as acceptance.
21. We will not go into various notions of convergence for random variables, such as almost sure convergence, convergence in distribution, convergence in probability, *etc.*, which play a role in econometric analysis. See Hamilton (1994) for a proper discussion.
22. These results can be thought of as analogues to (2.9) and (2.10). For normally distributed deviations, OLS produces the same estimators as MLE for the regression coefficients. A subtle point that we shall not elaborate upon here is that OLS estimator for the deviation variance is *not* equivalent to that from MLE, although for large enough samples the distinction is not of great consequence; see Hamilton (1994).
23. More accurately, conditional ML as the estimator is conditional on the initial value of the time series; *i.e.*, the (unconditional) distribution of the initial value is ignored.
24. Although the random noise $\varepsilon_t$ is independent of the prior $x_{t-1}$ (and unconditionally zero mean), this is a *pathwise* property. The relevant expectation for analyzing (2.19) is taken *across* paths. In addition, the joint dependence of the estimator in (2.19) on the components of the vector $\varepsilon$ is inherently nonlinear. Hence, we cannot appeal to iterated expectations to establish unbiasedness, as in the case of OLS.
25. So-called because the polynomial lag operator representation of the time series (2.17) has a root on the unit circle. *I.e.*, with $Lx_n = x_{n-1}$ the process in (2.17) can be written as $\Phi(L)x_n \equiv (1 - \phi L)x_n = \varepsilon_n$, so that stationarity ($\phi < 1$), say, is characterized by a root outside the unit circle.

26. These are just statements of the classical Central Limit Theorem (CLT), to be discussed in Section 6.5.1.
27. It exhibits standard square root of sample size convergence. More generally, we anticipate that variance estimators, by better conforming to the requirements of the CLT (see the previous endnote), will tend to exhibit (distributional) behavior that is less model-dependent than mean reversion estimators. It is in this sense we can characterize variance estimators as being (comparatively) more robust than mean reversion estimators.
28. We trust that it will be clear from the context when the variable $T$ denotes matrix transpose and when it denotes terminal time/number of data.
29. We will later discuss ramifications of dropping this assumption, *e.g.*, the distinction between algebraic and geometric multiplicity of an eigenvalue, Jordan normal forms, *etc.*
30. We are appealing to the fact that, for stationary processes, sample properties converge to population properties (the so-called ergodic principle).
31. Slightly nonstandard as it satisfies $dwdw^T = \Omega dt$.
32. The analogue in higher dimensions of the process in (2.23) that is technically stationary but econometrically indistinguishable from non-stationary part is a process whose response matrix (so to speak) has (some) eigenvalues nearly on/just within the unit circle. It would be a useful exercise for the reader to derive the analogue of (2.24) to deduce the comparative robustness of the covariance estimator.
33. I am grateful to Krzysztof Wolyniec for emphasizing the importance of this topic.
34. We will switch between discrete- and continuous-time formulations as convenient. See the Appendix to Chapter 6 for a brief discussion of the connections (and disconnections) between the two.
35. A somewhat unrelated point, but as Keynes supposedly said, the market can remain irrational longer than an investor can remain liquid.
36. We ignore throughout the case of so-called explosively growing processes, with roots of the characteristic polynomial *inside* the unit circle.
37. In the technical econometric sense, exogeneity (in contrast to endogeneity) refers to variables that determine some equilibrium relationships, but are not themselves subject to that equilibrium.
38. We cannot deny: there is a Bayesian flavor to this discussion. The point being, in any analysis, resort must be made to *some* kind of prior information.
39. Again, forwards are of interest to us because they are the primary trading/hedging instruments available in energy markets.
40. The phenomena described here are distinct from estimation biases as traditionally understood; in the context of mean reversion see, *e.g.*, Parsons (2008) or Yu (2009).
41. For convenience, we assume that the process starts from zero.

42. We have shown here how the sample variance (itself a random variable) corresponds to a population entity of the sample on which the estimator operates.
43. And of course, the weaker the mean reversion rate, the less relevant the distortionary effect is.
44. It can be seen that, even on a monthly basis, there can be extreme events, such as the spikes due to severe supply disruptions associated with Hurricane Katrina in the fall of 2005, as well as the crisis-associated run-up in prices (and subsequent) collapse in the summer of 2008.
45. Owing to the non-stationarity of (monthly) natural gas, we have not attempted to deseasonalize this time series.
46. Although natural gas displays clear commodity-like behavior throughout the overall sample, even it shows a delineation precrisis and postcrisis, specifically a general decline in volatility (attributable both to demand destruction and structural market changes [shale, *etc.*]).
47. Recall that we treat futures as equivalent to futures in our exposition.
48. Under a *pricing* measure, forward/futures prices are of course martingales. The question here concerns the nature of prices under the *physical* measure. The claim of efficient markets theory is that *risk-adjusted* prices are martingales, which of course, absent a theory of risk adjustment, is untestable. Nonetheless, we can simply ask in plain-language terms if (liquid) futures markets display any readily exploitable opportunities, reminiscent of mean reversion. We will see that generally speaking they do not. We cannot here discuss the technicalities associated with efficient market theory; see EW for a discussion of the inefficiency of energy futures markets for long-dated contracts (say, with more than three years to maturity).
49. A period roughly corresponding to the end of fighting in Libya in early 2011 and the start of the collapse in crude prices in summer 2014.
50. Recall Figure 2.11 as an indication of the recent financialization of forward crude. Fundamentals desks eagerly await weekly releases of inventory levels by various reporting agencies, but it is generally fallacious to believe that in liquid futures markets one can systematically trade on the basis of publicly available information.
51. In power markets summer typically comprises June through September, whereas in (U.S.) gas markets winter typically spans November through March, so seasonality in the two markets do not coincide.
52. There is little doubt that man-made, non-stationary effects such as urbanization and general economic growth have had an effect on temperature, as can be discerned from longer term time series (say, back to the 1960s) for locations such as Las Vegas or Phoenix (or even Dallas). (We take no position on the issue of AGW.) It is a common practice in weather derivative markets to look at discrepancies between ten- and twenty-year temperature averages.

53. We will ignore here pure supply-side effects such as unit outages forcing inefficient (*i.e.*, high marginal cost) units into service (giving rise to spikes).
54. It is worth noting a feature of forward heat rates. We will see in Section 5.2.5 a model (due to Schwartz) of a mean-reverting spot model with a non-stationary (stochastic) mean. The resulting forward dynamics give rise to the term structure shown in Figure 5.2. (This figure actually shows the volatility scaling law for the spot process, but in this model the spot scaling law is equivalent to the forward volatility term structure.) The increase and leveling off of volatility is in fact seen in actual markets, as we will see in Chapter 3 (see Figure 3.5). This behavior is due to (non-stationary) capital structure effects, namely changes in the generation stack. We briefly note here that such capital effects are also operational in regards to spreads, which reflect consumption/production across time.

## 3 Valuation, Portfolios, and Optimization

1. For example, an injection/withdrawal schedule for a storage facility, or a dispatch schedule for a power plant.
2. The classic Black-Scholes paradigm, which we will analyze shortly, is a prime example of the role different measures play in hedging and valuation.
3. For simplicity we will assume that the temporal dependence is only on expiry.
4. We ignore here the complication that futures typically settle against the *average* price over some period, usually a particular month, *e.g.*, November 2014, or a particular season, *e.g.*, summer 2015.
5. So in general the set of available instruments can consist (depending on the particular market) of both observables such as prices, and unobservables such as volatility (as implied by options prices). Keep in mind that option prices provide projections of realized *cumulative* volatility, *i.e.*, integrated volatility over some term. Thus, stochastic volatility (really, variance) models can only be useful to the extent that they provide a ready connection between implied and realized volatility (which is generally unobservable).
6. Note that this does not mean that prices equal expected values under the everyday, real-world probability measure (usually referred to as the physical measure). However, in liquid futures markets there is abundant evidence that current prices are in fact good projections of future values (in other words, over most time horizons of interest, there is little evidence of [systematic] bias in these markets).
7. Think of a storage deal, spanning a single injection-withdrawal season, say from April 2015 to March 2016.
8. For simplicity we ignore transaction costs here.

9. Notice the ordering of the arguments is not arbitrary; it is meant to suggest (and we will clarify in due course) that the value driver and its projection are the primary variables, and the associated value function/hedges/actions are in some sense "adapted" (in the plain language, not technical probabilistic, sense) to these drivers.
10. It is well-known that delta-hedging an option creates exposure to realized volatility; we will greatly elaborate on this point later.
11. Here is a very simple example. It is often not possible to rebalance forward positions intra-month (exceptions being markets where balance-of-the-month ["balmo"] contracts trade). Thus, it is irrelevant how volatile or "spiky" (spot) prices get within a month, if we can only hedge those prices with static contracts for that month. In general the volatility that can be collected through such portfolios is much less than the usual volatility estimated from the standard deviation of daily returns.
12. To say nothing of the typically great computational challenges presented by optimizing this operational flexibility.
13. By physical measure we simply mean the probability measure under which the time series of prices is actually observed. We will discuss different probability measures as they pertain to valuation and hedging in this volume, but we will assume that the reader already has a good grasp of these concepts. Björk (2009) is an outstanding reference on these topics, especially as they pertain to finance.
14. Shreve (2004a, b) is a good reference for the specifics of stochastic calculus and stochastic differential equations, with applications to finance in mind. Again, we assume sufficient familiarity on the reader's part here.
15. A put option would be handled no differently, except the payoff would be $(K - S)^+$.
16. Formally, any arbitrage opportunity refers to a portfolio whose price is $\leq 0$ and whose terminal payoff is $> 0$ almost surely. This can easily be seen to be operationally equivalent to the law of one price.
17. We trust that it will be clear from the context when a subscript refers to a time index and when it refers to a partial derivative.
18. We will discuss both of these points in subsequent sections. The first point can be seen from inspecting the payoff function graphically and invoking Jensen's inequality. The second point is related to the notion of a numeraire, and can be understood by seeing that if the units in which the prices of the underlying and strike are doubled (say), then from an economic perspective, the price of the option should simply double.
19. As a further preview, we will argue against the standard approach that mimics (*e.g.*, in the case of stochastic volatility models) the derivation of BS by introducing fictitious instruments (such as options) and establishing consistency relationships across the (augmented) set of assets.

20. Here, "projection" does not refer to its technical, mathematical meaning but the plain language sense of an (risk-adjusted) estimate (or guess).
21. It should be clear that this framework can be trivially adjusted for the case of selling a structure.
22. Note that there is no *explicit* reference to prices (or other traded entities) in (3.14). This is not an oversight. In this portfolio, the dependence on prices is *indirect*, through the value drivers (which are certainly a characteristic of the underlying price processes) and their projections. The notation is meant to emphasize that for a structured product, the bet is on a value driver, and that any price bet should be excluded from the portfolio in question (because it is generally more efficient to bet on prices *directly* via traded instruments such as futures).
23. For a general diffusive process satisfying $dS = \mu\, dt + \sigma\, dw$, from Ito we have that a function $V(S,t)$ satisfies $dV = (V_t + \mu V_S + \frac{1}{2}\sigma^2 V_{SS})dt + \sigma V_S dw$, so if $V$ is a martingale under this measure it can be readily seen that the following partial differential equation (PDE) is satisfied: $V_t + \mu V_S + \frac{1}{2}\sigma^2 V_{SS} = 0$. The Feynman-Kac formula establishes this connection between PDEs and martingales.
24. It is well-known that for GBM this is the only measure that is equivalent in some sense (to be made precise later) to the original data-generating process.
25. Technically, the precise result (obscured due to our suppression of interest rates) is that *discounted* prices/payoffs are martingales under the risk-neutral measure.
26. As we saw, the issue ultimately comes down to a question of the time scales over which information accumulates. *E.g.*, knowing that the current month is hotter than normal or that a large generation unit has experienced an outage affects our short-term projections much more than our long-term projections.
27. We ignore effects here like outages that give rise to jumps and spikes.
28. As we saw in Chapter 2, the issue is not so much the presence or absence of mean reversion as such, but rather the *time scales* over which such effects operate. For our purposes here we will treat the phenomenon as binary.
29. The (instantaneous) power-gas correlation $\rho'$ can be extracted from $\rho'\sigma_g\sigma_p = \rho\sigma_g\sigma_h + \sigma_g^2$.
30. Obviously, when bidding on a structure, we can always attribute zero value to extrinsic (optionality minus intrinsic), but it is just as obvious that few counterparties will part with this value for nothing.
31. Of course, cases where dynamic hedges can only be put on closer to maturity can be readily handled in this framework.
32. Notice the implicit assumption we make here: the value driver $\sigma_\alpha$ is a function *only* of moneyness and time-to-maturity (both at inception). In other words, we assume that the value driver does not depend on prices as such. This is actually not always a valid assumption; periods of large structural change (such as

the introduction of the Rockies Express [REX] pipeline in 2007) can engender such a dependence. An important consideration here concerns sample selection, and the ability to project historical information about the value driver in question (in the econometric language employed in Chapter 2, the issue comes down to stationarity vs. non-stationarity). This objective (projection from an actual sample) has consequences for the question of when (and whether) it is better to analyze cash flows (as was done in the analysis around (3.31)) as opposed to value drivers. See also the following endnote.

33. As an estimator, the idea behind this approach is the following. In terms of realized and projected value drivers (vectors $\sigma$ and $\hat{\sigma}$ resp.), write the portfolio as $\Pi(\sigma) = \Pi(\hat{\sigma}) + \Pi_{\hat{\sigma}}^{T}(\sigma - \hat{\sigma}) + \cdots$. If the realized value driver satisfies $\Pi(\sigma) = 0$, then to leading order the portfolio (constructed in terms of the projected value driver) is driven by the difference between realized and projected value drivers. In particular, under mild assumptions regarding dependence between value driver and price, if $\hat{\sigma} = E_0\sigma$, then $E_0\left[\Pi(\hat{\sigma})\right] = 0$ (again, to leading order). (We must stress that these expectations are wrt. the *physical* ["real world"] measure; *pricing* [martingale] measures, as frequently [and unthinkingly] employed in the industry are irrelevant here.) We are being somewhat sloppy here, as the value driver in question is clearly a pathwise entity, and so *does* depend on prices. However, for suitably chosen value drivers (indeed, this is precisely one of the defining criteria of a good value driver, as is definiteness of sign of the components of the gradient $\Pi_{\hat{\sigma}}$), the dependence will be weak enough such that the ensemble averages used in estimation (*i.e.*, as sample analogues of population properties) have meaning. (To understand the issue, the reader should ask why it makes sense to average, say, the last ten years of August temperatures in Dallas, while it does not make sense to average the last ten years of GDP.) We will return to these themes in the subsequent chapters.
34. Note that an algorithm similar to (3.32) can be crafted in terms of Gaussian (Bachelier) options, which would be appropriate for natural gas basis/transport options (recall the underlying market structure from Section 1.2.2).
35. This is not entirely true, as it is sometimes possible to dynamically hedge intramonth via balmo contracts. We do not show the results here, but this hedging strategy typically collects a volatility between the return volatility and static volatility shown in Figure 3.6.
36. By extrinsic value we mean the difference between total (option) value and intrinsic value. We should stress that although "intrinsic" commonly means evaluation of terminal payoff based on current market prices, this conception is basically operationally meaningless. Intrinsic can only mean that value that can be locked in *right now*, which of course, assuming liquid (forward) markets, does indeed amount to the same evaluation. It should simply be stressed

that this valuation entails a definite hedging strategy (*e.g.*, full position in underlying for positive intrinsic, no position at all for zero intrinsic). Note then, that if you cannot even get off a static hedge for some product (say, illiquid gas transport), it makes no sense to speak of "intrinsic value."

37. The second derivative in (3.36) is to be understood in the sense of a generalized function or distribution.
38. In (3.37), $< x >_t$ is the quadratic variation process for $x$, and of course is simply $t$ for Brownian motion.
39. The expectation of the delta function term produces a conditional density, which is of lognormal form, and the resulting integral can be analytically evaluated.
40. Note that this counter-hedge (based on the BS delta) effectively synthesizes a put position, which of course has the same extrinsic as the corresponding call (via put-call parity).
41. In truth, the vast majority of structured deals around storage are rentals/leases of various terms (usually one to three years, occasionally as long as five [although such long-term deals are very rare in the postcrisis world of dried-up liquidity and collapsing volatilities]).
42. As well, there are typically fuel costs on injection (*i.e.*, you must inject an extra volume due to physical losses) and fixed charges on either injection or withdrawal. These features do not substantially affect our points here.
43. We will spell out later the form of optimal intrinsic, as well as a near-optimal static replication strategy in terms of spread options, for a very general case of non-salt domes units, with fuel and commodity charges, and ratchets (flow rates dependent on the inventory of the facility). The problem can broadly be represented as either a dynamic programming problem, or approximately as a linear program, conditional on no operational constraints being violated regardless of option exercise. See the Appendix to Chapter 4.
44. Of course, after gas has been injected, it is still possible to reverse the corresponding forward sale and sell spot gas from storage. We will consider such valuations later, but for now this aspect is tangential to our main point here.
45. It is a lower bound partly because, as we have noted, we ignoring spot-forward optionality that arises after gases have been injected. However, in general there are additional representations of monthly option value in terms of max/min options and similar such structures. We will describe these later. In truth, for the salt dome example starting with zero initial inventory, this spread option value is in fact the best monthly representation.
46. There are traded spread option products in energy markets, of course. However, note the particular structure here: the option on (say), June injection and November withdrawal is exercised at the end of May, at which point the withdrawal leg (November) has not expired. This pre-expiry feature has

ramifications on the underlying ratio volatility structure (via the well-known Samuelson effect), which we will discuss later.

47. We will note here that the optimal allocation of spread options (referred to in Endnote 44) can be determined analytically (it is simply the pair-wise collection of injection-withdrawal months along the subdiagonal, *e.g.*, Apr–May, May–Jun, *etc.*) and gives a value of 3.56.
48. Even here, though, we should note how rapidly moving market events have rendered even the best expositions incomplete. Since EW's publication in 2003 there have been major structural changes to the U.S. energy market, including the Rockies Express (REX) pipeline, the transition of ERCOT from a zonal to a nodal market, and of course the ramifications of the shale revolution.
49. Options struck against the international benchmark for crude oil, Brent, also trade.
50. These settle against average spot price for a given month, or effectively futures prices at expiry.
51. These are struck/exercised every day within a given month, if in-the-money (*i.e.*, if the spot price is above the strike).
52. Precrisis, floating strike (cash) options traded in some gas markets, *e.g.*, SoCal.
53. The only value driver in this table we have not already explicitly referred to is convexity for load deals. This is simply the relation between realized volume (volumetric risk is the central feature of such deals) and realized price volatility.
54. Gamma is commonly, and indeed usefully, viewed as a measure of sensitivity of delta, the underlying hedge, and thus plays an important role in assessing dynamic trading costs and the feasibility of conducting such hedging strategies. However, this concern is separate from the one we have here.
55. This assumes valuation under a measure for which probability of future prices is scale invariant; see Alexander and Nogueira (2006). Thus, these results would not hold true for a measure under which prices were log-mean-reverting, say. But this of course highlights the point that these various greeks are intimately related to *dynamic* rather than static hedging strategies, and little significance should be attached to them outside of such a context.
56. This assumes, of course, that the variability associated with projecting correlation does not vitiate the utility of locking in (so to speak) the leg volatilities through vega hedging. In general, correlation estimates can be quite noisy, especially in comparison to volatility estimates.
57. Vega is always positive for a single asset option, but as we will see, the leg vegas for a spread option have indeterminate sign, depending on the leg volatilities *and* correlation. The ratio vega (for a Margrabe-type option) is of course always positive.
58. The cost is always positive, and typically (although in principle not necessarily) the same for either long or short positions.

59. Generally speaking, Leland's formula overstates hedge costs, as can be seen from simulation studies. Given that the structures of interest typically exist within a larger portfolio, it is important to take advantage of aggregation as much as possible in reducing these (hedging) costs.
60. At expiry, $V = (S-K)^+$ so $\Delta = H(S-K)$ and $\Delta_S = \delta(S-K)$.
61. Many of the concepts in this section originated in joint work with Krzysztof Wolyniec.
62. Or, when vega hedges are introduced, a bet on realized correlation.
63. And usually does, in energy markets.
64. A full discussion of these technical concepts can be found in any standard text, *e.g.* Björk (2009). For our purposes here a filtration is a (nested) sequence of collections of events (themselves sets of outcomes) describing the manner in which information about a stochastic process is revealed through time (basically, the dynamics of the process), and a process is said to be adapted to a filtration if it is measureable wrt. the filtration (a fancy way of simply saying that information provided by the filtration at a particular point in time is sufficient to determine the value of the process at that time).
65. Essentially this means the value of the process is known at the current time; *e.g.*, in a financial context you would know what hedge to put on right now.
66. Recall from Section 3.1.6 that a bond term is implicitly included in the replication strategy derived from (3.61), to maintain self-financing. The bond term is not explicitly present in the (portfolio) dynamics because of our prevailing assumption of zero interest rates.
67. A far more technical, but still useful, exposition can be found in Davis (2005).
68. We will in fact be interested in weakening the standard assumption further by considering non-Gaussian price and/or state dynamics. This characteristic is of course not equivalent to market incompleteness as such. There are special instances of market completeness in non-Gaussian settings, but these are very much the exception and not the norm. However, even Gaussian models, *e.g.*, the mean-reverting model with stochastic mean in (5.124), can exhibit incompleteness. The tools we will develop here can in fact be applied to a wide range of processes, so it will be useful to conceive the problem around prototypical incomplete market models such as stochastic volatility. (Under a pricing measure, linear [*i.e.*, Gaussian] models are in fact complete across structures with payoffs that depend only on the subset of tradeables, as opposed to nonlinear [*i.e.*, non-Gaussian] models that are not.)
69. Note that static hedges can easily be incorporated in the framework of (3.62) by taking $\Delta_s = \Delta_t$ for all $s \geq t$.
70. Although the primary focus in energy markets is on spread structures, we are going to start with the analysis of the standard option to better relate our results to the standard literature. This simplification will not affect the conclusions in any substantial way.

71. Technically, these results refer to discounted tradeables; as usual we will neglect effects due to discounting throughout this discussion.
72. By which we mean without reference to the (subjective) preferences of any agents.
73. Note that in general the reverse is also true, in the context of *derivative* securities: for a given EMM, there is not a unique physical measure that gives rise to it, even when (underlying) prices themselves are martingales under both physical and pricing measures. This highlights the basic pointlessness of the common practice of "calibrating" price models to match market option quotes (which, even when they exist, are often illiquid anyway), at least when such models are divorced from actual portfolios consisting of the calibrating instruments. Even in this case, however, the point of the portfolio is to efficiently create some desired exposure, *not* to guarantee consistency with market prices (*e.g.*, you should not bake your view on prices into a delta-hedged option portfolio: not because of some inconsistency but rather because there are more efficient ways of betting on price [futures], and the option portfolio is meant to extract exposure to realized volatility, not price).
74. A common, but mistaken, interpretation of the MMM is that it is the EMM that leaves the DGP under the original, physical measure as unchanged as possible. It is actually the *residuals* from a particular hedging strategy (namely, a [dynamic, orthogonal] projection onto the space of tradeables) that are minimally affected.
75. For example, there may be parametric dependence through projected quadratic variation and the use of a BS pricing/hedging functional. We will see just such an example shortly.
76. We will soon see examples of how this can be done, but for now we appeal to the example of BS pricing/hedging, with exposure being realized vs. projected volatility.
77. It is worth pointing out the contrast between static and dynamic hedging strategies, which as we have stressed is extremely important in commodity markets. Consider the case of rolling intrinsic. Since here we specify the hedging program (always hold intrinsic positions), we proceed in the reverse manner, namely, we must find a value component (of the value function) corresponding these particular hedge dynamics. Note that if we only use price information in the value function, we could only satisfy a condition such as (3.67) if we take the value component to be intrinsic value, and this is not realistic as it is highly implausible that some counterparty would transact extrinsic value for nothing. Hence we anticipate that the corresponding value component *must* use non-price information, or more accurately *non-observable* price-related information. Since we know from Section 3.1.4 that rolling intrinsic creates exposure to realized local time, which of course depends on

the physical dynamics of the price process (under the physical measure, obviously, which is where the residual exposure is reckoned), we plainly must know something about these dynamics. It is here we can see the problem with rolling intrinsic, at least in comparison to other valuation/hedging strategies (that is, value functions) such as BS: the informational requirements are quite high. In financial markets we must know something about the price drift. In commodity markets we must know something about the mean reversion rate *and* level. In both cases these entities are notoriously hard to estimate (see Chapters 2 and 6). In fact, we can see here that in general the informational requirements will be *higher* in commodity markets, due to these kinds of effects, which are ultimately traceable to the particular kinds of time scales that are operational in those markets (again, see Chapter 2). We can start to see here the informational efficiency of programs such as BS valuation/hedging.

78. Admittedly, under the canonical class of affine jump diffusions that we will consider in Chapter 5, independence is preserved under the two measures. But this fact is a consequence of the specialized assumptions about the underlying price dynamics, and *not* a general conclusion.
79. Note that while a pricing measure may of course be identical to the physical measure, it is meaningless to operate from an *assumption* that they are equal. This amounts to putting the cart before the horse. (One is reminded of the joke about the economist on a deserted island with a can of beans: "Assume we have a can opener.")
80. For an ATM option, the BS value is approximately linear in volatility, but not for OTM or ITM options. Again, this point is not itself relevant unless we beg the question by assuming that price and volatility are independent under the pricing measure.
81. Models such as (3.131) where the mean and covariance have an affine form (*i.e.*, constant plus linear in the state variables) are very popular due to their great tractability, and will be considered in great detail in Chapter 5. (See Section 5.2.5 for more details on Heston in particular.)
82. We present here the spot process. Technically, we should be considering the corresponding forward process, as these are the actual instruments available for hedging in energy markets (spot commodities do not trade as such, as physical possession must be taken at some point). However, using the techniques developed in the prior sections, we know that the affine framework readily allows transition between spot and forward formulations. Since the extra complications (*e.g.*, time-dependence) introduced by forward modeling is not relevant to the points we wish to make, we will instead employ spot models.
83. Examples in energy markets would include NYMEX natural gas and PJM West Hub electricity (on-peak).

84. Note that for jump processes, we must introduce a *continuum* of auxiliary options to be able to derive the martingale pricing equation.
85. With appropriate vega adjustment, of course.
86. We should call attention to somewhat similar work by Poulsen *et al.* (2009). The use of so-called plug-in estimators has some affinity with our approach; see Gandy and Luitgard (2013).
87. In the econometrics terminology employed in Chapter 2, in this case variance becomes less stationary/more non-stationary.
88. Although we do not present the results here, we note that the optimal hedging parameter for the mean reversion level is fairly flat across rates and nearly equal to the physical level; it asymptotes to the optimal local value as the reversion rate increases.
89. In fact, much of the material presented here will receive a more focused discussion in Section 7.6.
90. Analogous decisions must be made when the unit is currently up, *i.e.*, to stay on or shut down.
91. We will see an example in terms of portfolio optimization shortly.
92. Note that strictly speaking the discrete-time form of (3.101) renders the exercise policy Bermudan, not American, in nature (*i.e.*, exercise can only take place at specific times prior to expiry).
93. We trust it will be clear from the context when a subscript $t$ represents a partial derivative wrt. time, and when it refers to a driver/state at a particular time.
94. The dependence is implicit; the point at which the arguments in the max function are equal represents the decision boundary separating exercise regions from hold regions in time/driver space; see Kwok (1998) for a discussion in the context of standard American options.
95. Forgive us this academic dalliance. Simply think of utility as a proxy for wealth.
96. The calculations are reminiscent of the ones employed for affine jump diffusions. We note that multidimensional extensions are possible, although we omit the details here.
97. Each multiplier represents the incremental value of a specific constraint, *i.e.*, the sensitivity of optimal value to replacing the 0 in a constraint in (3.107) by $\varepsilon$. That is, they are shadow prices.
98. The original problem (3.107) is typically referred to as the *primal* problem.
99. This does not conflict with the usual connection between prices of (discounted) tradeables and martingales, once the reinvested payoff at exercise is accounted for. Put differently, the optimally stopped value process *is* a martingale.
100. We can mention here that many numerical approaches to these kinds of problems, such as simulation, produce lower bounds on value, hence it is quite useful to be able to find upper bounds on value.
101. See also Section 4.1. Examples from tolling will be considered in Section 4.2.

102. It is possible to craft financially settled deals (so-called virtual storage) or even physical deals (so-called park and loans) where inventory can go negative.
103. The spread options represent injection and withdrawal decision between appropriate temporal blocks (typically, contract months, but we leave it quite general here). More generally, the portfolio consists of max and min options across different withdrawal and injection periods. While (by put-call parity) these appear at first blush to be simply identical to spread options and forwards, they are in fact not redundant when fuel losses are taken into account. See the Appendix to Chapter 4 for the actual algorithm.
104. In practice the constituent options typically must be replicated via delta-hedging, creating a complex set of exposures across decision months, effectively a term surface of quadratic variations.
105. It is a trivial, but nonetheless useful, exercise to verify this statement for general affine jump diffusions and their associated forward relations and dynamics using results such as (5.76), (5.149), and (5.152).
106. As a side note, observe the necessary role played by mean reversion for the existence of extrinsic value. Intuitively it can be seen from (3.121) or (3.122) that, apart from deterministic or seasonal effects, the value function is only dependent on state (inventory). Note also, in the absence of a volatility term structure (*i.e.*, no mean reversion), it is clear that the spread options in the basket formulation have only intrinsic value. As a general rule, it can be shown (empirically) that as the resolution of the (forward) hedging instruments increases, the "monthly" lower bound approaches the "daily" true value; see Section 4.1. This illustrates an important theme we will revisit in Section 4.2 when we consider tolling: so-called spot and forward valuation methodologies are *not* alternatives to one another but are in fact closely related through particular market structures.
107. For more on control-based approaches to storage valuation, see Ahn *et al.* (2002), Thompson *et al.* (2009), and Ludkovski and Carmona (2010).
108. As we will see, the (purely) diffusive joint dynamics will not be altered under a change of measure.
109. In particular, we examined the consequences of eschewing the standard approach of introducing auxiliary (fictitious, really) assets by means of which consistency relations can be established across valuations.
110. In other words, we only use information that has just been revealed, and not some longer history.
111. As well as estimates of the volatility of variance and correlation with price, both of which must be filtered as well as even standard techniques such as a sample covariance estimator cannot be applied. In addition we must use as-yet-unspecified risk-adjusted parameters for the drift of the variance in the valuation functional.
112. Indeed, it may possess jumps, as well.

## 4 Selected Case Studies

1. I would like to acknowledge Krzysztof Wolyniec for motivating this investigation.
2. We have discussed here the valuation of storage from a monthly perspective (however "monthly" may be conceived operationally) and its convergence to daily value with *finer* levels of market resolution. We have not said anything about the important daily component of value for a *given* monthly resolution. In other words, after a set of spread options has been allocated at inception, once a particular month is entered into, one may deviate from committed monthly flows (either injection or withdrawal) on a daily basis (*e.g.*, if prices spike on some day, one may decide to release more gas from storage than the initial monthly schedule requires, *etc.*). There is clearly additional (option) value to this aspect of storage, and the obvious question is: how much *more* value relative to the monthly value can be captured through suitable hedging/trading strategies? We cannot address this question in detail here, but see Endnote 13 below for a sketch of possible approaches.
3. The reader should not hesitate to refer to the relevant parts of Chapter 7 for greater clarification.
4. In general there will be a distinction between power prices for weekdays and weekends, as well as between onpeak and offpeak periods within a given day. For convenience we ignore this important distinction here; it should be clear that these effects can be incorporated in a straightforward manner.
5. We will assume throughout that the start charges are already appropriately normalized.
6. Note that we implicitly assume that, for the time period in question, the number of switches (changes in operational state) is effectively unlimited (*e.g.*, there can be as many as one per day).
7. Some recent work offering computational benefits (by avoiding nested Monte Carlo for the martingale construction) is Schoenmakers *et al.* (2012).
8. Many of the results to be discussed here stem from joint work with Krzysztof Wolyniec.
9. Although we focus here on tolling, the methodology presented is in fact a quite general approach. In light of the representations of storage valuation that we have already considered, it should be evident that other deal types can be readily incorporated in the duality framework.
10. In Section 7.4.1 we will discuss some of the issues involved in evaluating the $Q$-expectations of these kinds of payoffs. As a practical matter of valuation, we note the following:

    These expectations will obviously depend upon the joint distribution of power and gas. Valuation in terms of a replicating hedging strategy will in

turn depend on the availability of liquid instruments for relative pricing. In some markets, there will be traded leg options, and occasionally ATM spark-spread (heat-rate) options. In other markets we may only be able to hedge price exposure, but there may be the possibility to dynamically hedge with balance-of-month contracts.

The above payoff structure is conditional on expectations/projections at the start of a month. In general we will be faced with a forward-starting valuation, *e.g.*, next summer valued today. Thus there is a contribution to value arising from forward variability that can also be captured by a suitable hedging strategy. Just as in the cash/intra-month case, the relevant forward-value drivers depend on the underlying market structure.

The upshot of these two points is that correlation may not be the relevant value driver in all, or even most, cases. To begin with, on a cash basis there may not be sufficiently liquid option markets on which to base a projection of leg volatility, in which case separate nonmarket (*i.e.*, historical) projections of volatilities and correlations is likely non-robust and probably pointless. Second, the aggregate (or blended) correlation that may accrue on a daily basis (that is, arising from the separate forward- and cash-hedging regimes) may have little relevance for deals with any kind of stringent physical constraints (*i.e.*, anything besides vanilla spark spread options).

11. The peaks and troughs in Figure 4.3 reflect the different distribution of on-peak and off-peak hours within the deal months; recall from Table 4.3 that there is no seasonal structure in the price curves.
12. We present simulation-based results for the lower bound here, but in truth the quadrature techniques that we discussed in Section 7.4 can also be brought to bear.
13. We note here that the duality methods outlined here for tolling can also be applied to assess the lower-bound storage valuation from Section 4.1. The main idea was presented in Section 3.3.2 and involves using the constituent spread options as the basis of the martingale proxy. In practice, the shadow price (wrt. inventory; see [3.125]) that arises naturally out of the spread-option linear program in (4.18) can be employed for daily management of injection/withdrawal decisions in light of prevailing spot prices. In fact, prompt futures prices are typically a good proxy for shadow prices (at least when the facility is not nearly empty or nearly full), so incremental daily storage value can be viewed as a kind of spread option between spot and prompt.

## 5 Analytical Techniques

1. A longer overview (but still very much on the short and sweet side) can be found in an appendix in Björk (2009). See also Shreve (2004a, b).

2. Events are sets of outcomes, and the associated probability measure is defined over the set of events and countable unions and complements thereof, *i.e.*, the sigma algebra of events.
3. Over some time horizon $[t, T]$.
4. We will state general results later, but the relevant result in one dimension is $E\exp(i\phi z) = \exp(i\phi\mu - \frac{1}{2}\phi^2\sigma^2)$ for $z \sim N(\mu, \sigma^2)$.
5. By this we simply mean that continuous processes remain continuous processes under a measure change; a continuous process cannot be transformed into a jump process via measure change (although in general a process's canonical form can change, *e.g.*, a Gaussian process may become non-Gaussian, *etc.*).
6. Thus, in contrast with the diffusive case, where the structure of the stochastic driver is not altered by the change of measure (the covariance remains the same), for jump processes there *is* a structural change: the jump intensity is different. It is not hard to show (via characteristic function methods) that, also in contrast with the Brownian case, the drift of a jump process does not change *numerically*. However, it changes *relative* to the (new) jump driver, and it is precisely the (unconstrained) freedom of selecting the jump amplitude of the RN process that gives rise to the nonuniqueness of the martingale measure.
7. As we shall see subsequently, characteristic functions provide an extremely nice framework for establishing the precise form of the induced dynamics under a measure change, for a wide range of processes.
8. Examples would include spark spread options between power and gas (in which case $\alpha$ would represent a heat rate) and natural gas transport options (in which case $\alpha$ would represent fuel losses along the pipe).
9. Note that, as remarked previously, the RN derivative is in general a process itself.
10. For more on the use of numeraires in option pricing, see the aforementioned papers by Benninga *et al.* (2002) and Schroder (1999).
11. For general continuous processes we have a result of the form $df = (f_t + \zeta f)dt + f_z dw$, from which the martingale condition on $f$ requires that the drift term vanish.
12. Note that all terms involving $\zeta$ vanish, reflecting the status of $e^{\zeta(T)}$ as a $P$-martingale.
13. Of course this result could have been obtained directly from (5.37) by expanding out the results for the characteristic function under $Q$ and collecting terms appropriately in $\phi$, but, apart from being a tad burdensome (so we see again the great analytical utility offered by measure change approaches), the approach laid out here will prove greatly useful in more general, affine problems where the manipulations would be far less tractable. It is thus worth introducing the tactic here.
14. It is worth noting the contribution of Geman and Eydeland (1995) to the problem. Carr and Schröder (2004) employ analytic continuation to extend the

original results of Geman and Yor (1993), another example of the great utility afforded by complex analysis (a theme we emphasize throughout here).

15. In particular, Curran (1992) provides very accurate lower bounds for the Asian option price.
16. Note, using (5.50), that the requirement that the RN derivative be a $Q$–martingale is satisfied.
17. Conditioned on the current time $t$.
18. Although Girsanov is of course essentially a characteristic function result, in practice it often proves rather awkward to employ directly (that is, without explicit reference to its underlying construct).
19. More precisely, Itô calculus.
20. A semi-martingale is a stochastic process that can be decomposed into a local ("ordinary") martingale and an adapted, finite variation process (by which we informally mean a process whose value is known at the current time and which only has a countable set of discontinuities). Among other desirable properties, they are the largest class of processes for which stochastic integration can be reasonably defined, quadratic variation always exists, and reducing the available information set does not affect semi-martingale status. As already noted in the text, these properties are crucial for constructing portfolio dynamics necessary for valuation problems.
21. We thus ignore, for the most part, entities such as fractional Brownian motion (fBM), which is similar in many ways to ordinary Brownian motion, except that the assumption of independent increments is dropped. For an overview of fBM, see Nualart (2006). Broadly speaking, such effects represent long-range dependencies. We also (largely) neglect effects at the other end of the spectrum, namely on very short time scales (*e.g.*, high frequency). Here, due to market microstructure effects, estimates of quadratic variation (say) do not scale proportionately (so that, *e.g.*, entities based on minute-by-minute returns do not converge to scaled versions based on daily returns). We very briefly discuss high frequency issues in the Appendix to Chapter 6.
22. We will confine attention here to one-dimensional processes, although it should be fairly obvious how to extend the discussion to higher dimensions. This latter issue will be considered in Chapter 8, on the subject of joint dependency structures.
23. This definition is related to the property of right continuity with left limits (*càdlàg* in French), which is obviously suitable for financial applications (as the customary assumption is that one puts on a hedge based on current [price] information).
24. As usual, we provide no formal proofs here, as there is a copious literature on Lévy processes that the reader can draw upon for technical details. We can recommend Kyprianou (2006) and Papapantoleon (2008) as good sources of information.

25. We assume unit time interval for convenience.
26. In truth, the nature of the divergence does need to be restricted by requiring that $\min(1, |x|^2)\nu(x)$ be integrable near 0, as should be clear from the integrals in (5.59).
27. Roughly speaking, again, "activity" is an inherently two-dimensional entity; to repeat, only in the case of integrable Lévy measures can jumps be decomposed into categorically distinct (Poisson) entities such as "intensity" and "amplitude."
28. We note in passing that this process is built up from time-changed Brownian motion; see the next subsection on stochastic volatility models as they relate to time-changed processes.
29. We do not propose to critically evaluate this position here. What we wish to note is that, as a practical matter, the data set that is typically available in energy markets (the aforementioned studies were based on equity markets) simply do not support the resolution necessary to conclusively answer the question of whether the data-generating process is an infinite activity process. Of far greater importance is the fact that the hedging strategies that are available in these markets (and as will be shown, these strategies are central to the question of valuation) will take place over time horizons that effectively render this question (of jump activity) somewhat moot. (This is certainly the case with the rather mundane case of compound Poisson processes.) We should point out here that, even in the continuous process case (*e.g.*, joint normality), that not all representations are created equal. A simple example is the case of correlation vs. ratio volatility. These are obviously isomorphic (so to speak), but as a practical matter it is generally preferable to estimate volatilities as opposed to correlations. *A fortiori*, while a pure jump representation may be theoretically (or at least aesthetically) superior to a diffusive representation, we would generally expect the latter to be more robust than the former in terms of the important operational task of estimation. It is for these reasons that we will, for the most part, eschew consideration of pure jump processes in this book.
30. Of course, the menagerie of GARCH models prevalent in the literature should also be mentioned. We will briefly discuss GARCH models in the next chapter, but it is worth noting here that these models are models of directly observable residuals (*e.g.*, of some *other* model of returns, say), as opposed to the kind of (continuous-time) unobservable stochastic volatility models we are considering here. There is no necessary correspondence between the two kinds of models (*e.g.*, in the limit of infinitesimal time steps), despite popular impressions to the contrary.
31. In the United States there are weekly announcements of various fundamental statistics such as crude oil and natural gas inventories, which of course may come in above or below expectations and thus impact (short-term) price movements. However, inasmuch as the arrival of this information is publicly known

and the drivers in question cannot realistically vary by extreme amounts week over week (recall the discussion of variance scaling laws in Chapter 2), these events are better characterized as jumps of limited extent. In informationally efficient (*i.e.*, liquid) futures markets, there is little reason to think that such events are systematically exploitable.

32. Clearly, $T$ must be positive valued and increasing.
33. Obviously, the assumption of independence between the process being time-changed and the time change itself is somewhat restrictive. Carr and Wu (2004) claim to circumvent this restriction and retain the tractable structure of (5.60) by applying a complex-valued measure change via the optional stopping theorem. However, as we understand the argument, this approach largely amounts to a notational/computational device. Wu (2008) seems to confirm this suspicion. In truth, the types of stochastic volatility models of interest are more conveniently analyzed in terms of the canonical affine processes we will discuss in the next subsection, so we will not pursue this particular issue further.
34. A useful exercise for the reader is to use this result to verify the standard result that a partitioned normal variable is also conditionally normal, with conditional mean and covariance given by $\mu_{x_2|x_1} = \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1}(x_1 - \mu_1)$ and $\Sigma_{x_2|x_1} = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}$ where $\mu$ and $\Sigma$ are the unconditional means and the notation for the partitioned means and covariances should be clear from the context.
35. As will be clear in the next subsection, the ensemble (5.65) retains the affine structure of the model (5.63) (essentially, instantaneous drifts and variances are one degree polynomials).
36. The expression in (5.66) finds use in simulations of the Heston model; see Broadie and Kaya (2006).
37. Alternatively, we can say that it is possible to condition price on integrated (cumulative) variance. This is a standard approach for pricing options under (independent) stochastic volatility: the option price is *e.g.*, a Black-Scholes expression integrated over the (occasionally known) distribution of stochastic (integrated) volatility.
38. Note that when we speak of an "unconditional" expectation here, we are being a bit sloppy, as there is always conditioning, but only on information at inception (*i.e.*, we will take "conditional" here to mean conditional across the entirety of a path).
39. Note that we can view the dynamics of $x = \Omega_T - \Omega$, conditional on the above information set, as the following Lévy process (with an abuse of notation): $dx = \rho(\Omega'_T - \Omega')dt + \rho_s\sqrt{V_T - V}\,dw$. The second term is of course a time-changed (standard) Brownian motion. The first term can be thought of as a stochastic drift, with mean zero. It is thus akin to a regression residual.

40. The result in (5.75) follows from the independence of increments of $L$ and the proportionality of the log-characteristic function of Lévy processes to the time horizon in question.
41. We refer the reader to Tankov and Cont (2003) for the technical details for rigorously defining stochastic dynamics for general jump diffusions.
42. The difficulties associated with problems exhibiting very different time scales is in fact a very well-studied topic in the numerical analysis of ODEs, where such behavior is called *stiffness*. We call attention to this issue here, but will not discuss it in any depth, as the subject matter is well covered in either standard texts (*e.g.*, Press *et al.* [2007]) or with explicit reference to affine jump diffusions (Huang and Yu [2007]). Suffice to say that the issue entails a question of stability vs. efficiency of any particular scheme for integrating the underlying system. As we have emphasized throughout this volume, the presence of time scales of differing orders of magnitude are a hallmark feature of commodity markets.
43. Or more accurately, the current values of the state variables.
44. We are being a bit sloppy in (5.104), as in general $\alpha$ will have a multidimensional dependence on $\phi$, making $a$ a matrix-valued entity. However, it should be clear to the reader how to proceed, so we will not dwell on the details here.
45. *I.e.*, let $C = \frac{2}{\sigma^2}\dot{w}/w$.
46. In addition, since $v$ represents a (positive) variance, the process parameters must satisfy certain conditions to ensure that the process never goes negative, specifically the so-called Feller condition: $2\kappa\theta \geq \sigma^2$. We note that we have not experienced any numerical problems with the formulation in (5.123), even in cases where the Feller condition is violated (the scenario studied by Lord and Kahl [2008]). The issue in this case is that the amplitude of the function $D$ in (5.121) does not decay fast enough to suppress oscillations arising from phase shifts due to the branch cut singularities. These phase shifts are themselves a numerical phenomenon dependent on the particular (analytical) representation (the effect does not arise from a direct integration of the governing equations (5.120)), giving a good example that mathematical equivalence does not necessarily imply operational equivalence.
47. As is well known, a realistic feature of electricity markets is spikes, which of course can be modeled by jump processes. We will consider such an example later.
48. This representation is, strictly speaking, false, as actually existing futures/forwards contracts (we will ignore throughout the subtle but important difference between the two) settle not against terminal spot as such, but rather the *average* spot price over some specified time block, usually a particular calendar month, *e.g.*, July 2014. (For sufficiently long times to maturities [obviously a market-dependent criterion], traded futures contracts may only settle on lower resolutions such as seasonal or quarterly blocks, *e.g.*, Q1 of 2018.) However,

this fact does not detract from the usefulness of (5.147) in elucidating several important points, namely how certain features of physical spot markets are manifested in financial futures markets, and how value drivers are impacted by different hedging strategies with futures (*e.g.*, static vs. dynamic).

49. *I.e.*, injection takes place at $T_1$ along with commitment (also undertaken at $T_1$) to withdraw at $T_2$.
50. Or more accurately, they do not *necessarily* vanish. In general the parameters under the two measures will be distinct (except of course for the covariance matrices).
51. The magnitude of the difference between the two mean reversion rates will of course be a reflection of the preferences of market participants. The point is, in liquid forward markets, the precise nature of these preferences is irrelevant and the forward price is simply a given, at least from the perspective of valuing structured products.
52. In truth, fuel prices themselves exhibit some mean reverting, *over a long enough time horizon*. In other words, (5.162) is a representation of *relative* time scales over which the effects of interest are operational. This point was given much attention in Section 2.2.
53. Probably better termed a stationary relationship, in the terminology of econometrics; see Chapter 2 for more details.
54. For simplicity we ignore jumps here, but it should be clear that they can be readily incorporated.
55. As a forward price, the expectation in (5.165) must be wrt. some pricing measure. However, the techniques involved in the subsequent analysis do not particularly depend on any kind of martingale structure, so we simply omit any explicit reference to a probability measure.
56. Of course, the full solution to (5.165) involves the entity $\gamma_0$, which also satisfies an ODE that is not hard to derive. However, in general this entity does not enter into the forward dynamics in affine models, so we neglect it here. (Its evolution is dependent on the coefficients of the state variables, but not *vice versa*.)
57. That is to say, a block diagonal matrix with block elements of the form $\begin{pmatrix} \lambda & 1 & 0 \\ & \ddots & 1 \\ 0 & 0 & \lambda \end{pmatrix}$ where $\lambda$ is an eigenvalue and the dimension of the block corresponds to the multiplicity of the eigenvalue.
58. And also, apart from the jump terms, (5.154).
59. The idea that information accumulation associated with prices may vary on a much longer time scale than the information accumulation associated with some fundamental relationship(s) between those prices is of course simply a more abstract formulation of the well-known econometric concept of cointegration. The role of exogenous variables considered here can be seen to fit in

the (also well-known econometrically) category of Granger causation. There will be more on all of this in Chapter 6.

60. Examples would include exchanging price exposure for volatility exposure in a delta-hedged leg option, or volatility exposure for correlation exposure in a vega-hedged spread option.
61. The reader may want to revisit Section 5.1.
62. Recall the applications in Section 5.1.
63. An interesting commodity application of models such as (5.186) is to augment the system with a similar log-load process and the integrated variance $dV = vdt$. A pathwise relationship can then be seen between power-load covariance and realized power variance, illustrating the claims made about vega-hedging load deals in Section 1.2.4.
64. The drift of tradeables under the pricing measure is of course zero.
65. For a volatility-weighted RN process with coefficients $\alpha$ (akin to (5.171)), the relationship between means under physical and pricing measures is given by $\mu^Q = \mu^P + X\alpha$ in terms of the process covariance $X$. Now, for another $P$-martingale with volatility coefficients $\beta$, the condition of orthogonality with tradeables requires that $X\beta = 0$ for *only* those rows corresponding to tradeables, and the requirement that this process remain a martingale under $Q$ requires that $\alpha^T X\beta = 0$. These conditions imply that the RN coefficients can be partitioned as $\alpha = \begin{pmatrix} \alpha_S^T & 0^T \end{pmatrix}^T$, where the nonzero components correspond to tradeables. Similarly partitioning $X$, the equation imposing zero drift of tradeables under $Q$ can be solved, from which the expression for the $Q$-drift of the non-traded entities can be shown to be equivalent to (5.190) and (5.191).
66. For the extension of these results to processes with jumps, as well as some discussion of the discrete-time case, see Mahoney (2015a). Jumps actually introduce some very nontrivial complications to the problem. The equivalence between MMM and the entropy measure no longer holds, and in general the MMM under jump processes entails a *signed* measure. The equivalence with local variance minimization no longer holds, either. In fact, it is a useful exercise to extend the analysis of 5.2.9.4 to, say, Merton's model (the resulting expression for the local variance-minimizing hedge can be cross-checked with Tankov and Cont [2003]). These results will show that, unlike in the diffusive case, the optimal hedge ratios are dependent on the value function itself. In other words, it is not clear how pathwise properties that hold true in the diffusive case carry over to the discontinuous case.
67. Properly discounted, of course.
68. Using the shorthand notation $\mu(-dx) = \mu(-x)dx$.
69. The "truncation" function (for lack of a better term) $h$ retains the salient feature of being confined about the origin.

70. This result was first published by Eberlein and Papapantoleon (2005a). In fact, there are similar symmetries across other structures that can be crafted as either fixed or floating payoffs, such as lookback options.
71. We are referring to options with arithmetically averaged payoffs; as in the Brownian case, geometrically averaged payoffs clearly present little problem for such models.
72. Note that, as in the Lévy case, this is a rather different question from asking whether the *option* valuation in question is feasible after the measure change. This nontrivial difficulty does not detract from the obvious use and power of symmetry relationships, however.
73. A natural extension of the standard Merton jump diffusion in (5.112) to include mean reversion directly runs into two problems in practice. To mimic actual data, both the mean-reversion rate (necessary to equilibrate jumps) and diffusive volatility (necessary to reflect randomness during "normal" periods) must be unreasonably high. Hence the appeal of model such as (5.220), where the two effects (jumps vs. diffusions) are in some sense separated.
74. It is easy, but not terribly illuminating, to write out the ODE solved by $\alpha$.

## 6 Econometric Concepts

1. As we will see, when there is a single cointegrating relationship, OLS can consistently be used to extract estimates of it, although of course the relevant diagnostics will not be standard.
2. Since the row rank of a matrix equals its column rank, we assume that $A$ has been suitably arranged so that $A_1$ in (6.1) is nonsingular.
3. We are being somewhat abusive of notation in referring to "cointegrated processes" in Figure 6.1. We can think of that scenario (involving a random walk and white noise) as entailing a linear stochastic relationship between *two* non-stationary variables, with the multiplicative coefficient equal to zero and additive stationary noise. Technically, the trivial linear relationship $y = 0 \cdot x + \varepsilon$ is excluded by the formal definition of cointegration. Nonetheless, the sloppiness of notation does not detract from the overall message, which is that conventional econometric techniques can yield extremely misleading results when applied to non-stationary time series except under very special (and fortuitous) conditions. Roughly speaking, there has to be some kind of underlying stationary relationship in order to apply these methods to non-stationary processes.
4. For convenience we omit a constant term.
5. See Hamilton (1994) for a discussion of what OLS produces in the presence of multiple cointegrating relationships. We will discuss a more general approach to the problem in Section 6.1.3.

6. The joint dynamics in (6.14) are linear and hence Gaussian in nature. It would be a useful exercise for the reader to derive the characteristic function of the process (in the continuous-time limit) using the methods of Chapter 5. It can be seen that the relevant system of ODEs involves a matrix with a zero eigenvalue, which manifests itself in a process variance that grows linearly with large time horizon, a tell-tale sign of non-stationarity. (A second eigenvalue is the negative [in this case] number $b_2 - \gamma\, b_1$, indicating a variance scaling reminiscent of mean reversion over some smaller time horizon.) We will examine this idea in more detail in Section 6.1.4.
7. This result also follows from the continuous-time limit of (6.13), from which it can be seen that the (log-)heat rate is a standard mean-reverting process (note that there is an implicit time step factor baked into the coefficients of the discrete-time model, and the $\kappa^2$ drops out in the limit).
8. It is not hard to see from (6.15) that for a general cointegrating coefficient $\gamma$ as in (6.14), the long-term correlation will be $\mathrm{sgn}(\gamma)$.
9. This can be seen by applying the same kind of transformation to (6.19) as in (6.12).
10. Let us stress that we do not seriously believe that such spread-trading opportunities are very prevalent in liquid futures markets, or that if they are, that they are painlessly exploitable (recall Keynes's dictum regarding market irrationality and investor liquidity). We only wish to illustrate how the necessary objective is to identity drivers whose variance scaling grows much less rapidly than that of the constituent legs.
11. Not to be confused with value at risk!
12. This is simply a nonlinear eigenvalue problem, which is reducible to a standard eigenvalue problem; see Press *et al.* (2007).
13. We ignore the cases of explosive growth, where some roots lie strictly within the unit circle. We further ignore, except in passing, cases where the roots on the unit circle are complex, which is a hallmark feature of seasonal non-stationarity. In other words, we will assume that any roots not outside the unit circle are equal to 1.
14. In case it is not clear from the format, the first matrix is block lower triangular starting with the (3, 3) block, with all nonzero blocks being the negative identity. In the second matrix, starting in the (3, 2) block, the blocks are shifted pairs of $+/-$ identity matrices (surrounded by zero matrices). The objective is to recraft the dynamics to involve only lagged differences, plus a term in the previous level. Conceptually the level at any other prior time could be used instead, but using the prior time is conventional.
15. These results fall under the umbrella of the celebrated Granger Representation Theorem; see Hamilton (1994).
16. Which are also the eigenvalues since $R$ is diagonal.

17. In truth we should use the Schur decomposition, but this will be done in the sequel.
18. We are somewhat glossing over the issue of multiple eigenvalues, but this is not important at this stage.
19. More accurately $I$ is the filtration generated by this process. With this shorthand notation understood, we have, *e.g.*, that $I_t = z_t \cup I_{t-1}$, which will be used subsequently in (6.59).
20. See, for example, Javaheri *et al.* (2003), Johannnes and Polson (2003), Doucet and Johansen (2008), or Fulop (2011). These techniques are typically applied to the recursion in terms of joint densities in (6.60), rather than the filtering (marginal) densities in (6.59). The central challenge concerns evaluation of the (normalization) integral in the denominator in (6.60), specifically both its high-dimensional nature and the fact that the joint density from the prior step, $\Pr(x_n|z_{0:n})$, generally will not be available in a form amenable to numerical quadrature. However, note that, since the normalization integral can be written as $\int dx_n \Pr(x_n|z_{0:n}) \int dx_{n+1} \Pr(z_{n+1}|x_{n+1}) \Pr(x_{n+1}|x_n)$, it takes the form of an expectation, and we anticipate that simulation can be applied to evaluate it. (Simulation as a computational tool will be discussed in Section 7.5.) In particular, since the density $\Pr(x_n|z_{0:n})$ can be evaluated point-wise but not (usually) directly sampled, importance sampling, that is evaluation of expectations via $E^p f = E^q \frac{p}{q} f$ (where the superscript refers to the density with respect to which the expectation is taking place) can be applied, with a judicious choice of auxiliary density for which direct sampling *is* possible. (Importance sampling will be discussed in Section 7.5.2.) Recursive methods known as Sequential Monte Carlo are employed in particle filtering to carry out this routine.
21. We say "deceptively" because the typical textbook-style exposition (see, *e.g.*, Welch and Bishop [2006]) entails an overemphasis on computational issues (algorithmic flow charts, *etc.*) and a proliferation of unnecessary jargon (*a priori* and *a posteriori* estimates/updates, *etc.*). The main point is under-emphasized, if not missed altogether.
22. We are considering linear models for the moment, which of course exclude most stochastic volatility models. We will later relax the assumption of linearity.
23. The situation is reminiscent of the phenomenon of multicollinearity in OLS, where there is (near) linear dependence between a subset of regressors. The estimator can retain predictive power *as a whole*, but the diagnostics associated with *individual* parameters can become extremely unstable. (Recall that OLS entails an inversion of a matrix that becomes singular [or nearly so] in this case.) This discussion serves to distinguish engineering applications of filtering from financial applications: financial observables (prices) are typically untainted by any signal noise.
24. As can the conditional means, but we omit the details.

25. *E.g.*, from $C_{11}C_{11}^T = HQ_1H^T$ and $C_{11}C_{21}^T = HQ_1F^TH^T$ we can extract an expression for $C_{21}C_{21}^T$ that is used in the calculation of $C_{22}C_{22}^T$, *etc.*
26. We note again that (as in the linear case) noise can also be incorporated in the measurement relation, as well.
27. If $x \sim N(0,1)$, then $Ex^2 = 1$, but linearization about the mean of $x$ would suggest $Ex^2 = 0$!
28. Nor is there a need for any Jacobians.
29. The analogy to quadrature techniques should be clear to many readers; see Section 7.4.
30. There are in fact some subtle implementation details that do not really concern us here, such as augmentation of the space state in the case of nonlinear noise and guidance on the judicious choice of scaling constants in the construction of the sigma points; see Wan and van der Merwe [2001] or Javaheri et al. [2003].
31. It must be stressed that this filtration is distinct from the natural filtration of the underlying (partially observed) process.
32. The system in (6.89) and (6.93) may be thought of as analogues of the (continuous-time) Kalman-Bucy equations, which typically includes measurement noise and does not make reference to the observation filtration.
33. Wealth is modeled via a power utility function. The academic nature of this assumption does not, however, detract from the usefulness of the subsequent results.
34. Acronyms for (resp.) autoregressive conditional heteroskedasticity and generalized ARCH. There is a bewildering number of offshoots of these models, as documented in Bollerslev (2008), such as the amusingly named PARCH.
35. The estimation is conditional on some set of initial data points, as is the typical case with autoregressions. Note that a simple test for heteroskedasticity (in the ARCH case) is to run a standard regression (say, OLS with $F$-test diagnostics) on the (lagged, squared) residuals from an ordinary autoregression in (6.96).
36. Alternative techniques such as quasi-maximum likelihood estimation (QMLE) can be employed when the innovations are non-Gaussian, but estimators based on Gaussian disturbances nonetheless produce useful results. See Section 6.5.
37. Heston and Nandi allow a general number of lags in the conditional variance, which for convenience we eschew.
38. EW should be consulted on these points.
39. To review, for integer degrees of freedom a chi-squared variable is distributed as a sum (over the degrees of freedom) of independent, squared standard normals.
40. In (6.115) we have used results for the matrix logarithm, which is defined (akin to the matrix exponential) via formal Taylor series expansion; *e.g.*, $\log(I + A) = A - \frac{1}{2}A^2 + \frac{1}{3}A^3 - \cdots$ for matrices with a suitably defined norm satisfying $\|A\| < 1$. We have also used the result that for diagonal matrices, $\log(V\Lambda V^{-1}) = V \cdot \log \Lambda \cdot V^{-1}$ with the logarithm of a diagonal matrix defined in the obvious way. Finally we have used the well-known relation between the matrix eigenvalues

and the matrix determinant. The end result is the well-known formula relating the trace of a matrix logarithm to the logarithm of the determinant.

41. In general the ensemble in (6.120) would only be taken over those elements of the process that correspond to log prices, and not entities such as stochastic volatility. For ease of exposition we will suppress such explicit notation. Needless to say we will ignore any issues concerning the filtering of unobserved state variables.
42. Compare with (6.113). For more on linear-quadratic jump diffusions, see Cheng and Scaillet (2007).
43. Obviously, the inverse in (6.133) is to be viewed as a notational device.
44. Note that this formalism can be extended to models where some of the unobservables are in fact structural and categorically distinguished from the random drivers/noise, as in stochastic volatility models (*e.g.*, Heston). With such models, the estimator in (6.132) must entail some kind of filtering of the unobserved state variables, as discussed in Section 6.2, and the extraction of unobservables (now broadly understood to mean not just Brownian drivers [say] but also stochastic variance terms) will be far more involved than suggested by (6.133). (*I.e.*, projected state variables are distinguished from proper residuals.) Since these complications will only obscure the main points we wish to make here, we will not consider them in any greater detail.
45. Certainly, pure computational challenges (*e.g.*, optimizing likelihood functions, computing high-dimensional integrals in filtering, *etc.*) can also be significant, but this topic will be the focus of Chapter 7.
46. By which we mean, the principal question asked concerns hypotheses about model parameters *given* a model, invariably in some operationally unhelpful limit of very large samples. By contrast, we argue for conditioning on the *actual* (finite-sized) sample being analyzed, for which the concern is with robustness and stability of the model and estimator being employed.
47. As distinct from resampling without replacement, *e.g.*, the so-called jackknife.
48. This question is intimately related to the question of how close the eigenvalues of $\Phi$ are to the unit circle. As always, the issue concerns the connection of population to sample.
49. *E.g.*, the resampled indices might be $\{3,1,8,3,\dots\}$, in which case the second resampled residual is the first residual from the original set.
50. I would like to thank Krzysztof Wolyniec for impressing upon me the significance of this paper.
51. We also note its close cousin, the Law of Large Numbers (LLN). Both results refer to a (asymptotic) relation between sample mean and population mean, but the CLT is stronger in that it specifies the distributional form of that relation.
52. In truth, the estimator $\hat{\sigma}$ of the standard deviation of the noise term is slightly different; for MLE it is the sum of the squared (realized) residuals divided by the number of observations, for OLS it is the sum-squared residuals divided by the

number of observations less two. This obviously does not affect the consistency of the MLE estimator (in the statistical sense that, as the number of observations increases, the bias of the estimator decreases).

53. We do not mean here the sense in which cross-sectional is often used in econometrics, *e.g.*, across futures contracts of a given time to maturity.
54. This problem is not dissimilar to the situation encountered when applying simulation techniques to the valuation of American-style options, as we will see in Chapter 7.
55. Or more accurately, the first-order optimality conditions based on (6.145).
56. If $X_t$ is a martingale, then $X_0, X_1 - X_0, X_2 - X_1, \ldots$ is a martingale difference sequence; see Hamilton (1994).
57. It would be a good workout for the reader to apply the kind of argument used in (6.149), along with the results associated with the Wishart process (*e.g.*, (6.116)), to re-derive the asymptotics for the standard VAR process in (2.36).
58. When the auxiliary model equals the true model, integration by parts and (6.142) shows that $J = H$; see also (6.150).
59. For example, the lag coefficients of an AR(1) process with non-Gaussian disturbance term can be consistently estimated by (counterfactually) assuming the disturbance is Gaussian, although consistent estimation of the disturbance depends on its actual nature.
60. $Z$ is simply the set of realizations $\{z_t\}$.
61. In the typical application, the simulation would in fact take the form of a very long time series, with the ensemble average in (6.167) taking place along the path. In such cases, there is not really an auxiliary density as such, rather the constituent terms in (6.167) are *conditional* auxiliary densities. Recall the discussion associated with (6.149).
62. There are a host of simulation-based techniques whose essence is clear enough and we will not go into them here. These include Simulated Method of Moments and Simulated Maximum Likelihood. The idea of course is that the necessary expectations are computed via simulation.
63. Note that, to leading order in (small) $\Delta t$, (6.170) conforms to a (Euler) discretization of (6.137).
64. Although we have a natural interpretation of the density in (6.171) as an approximation to the true density in (6.169), recall from Section 6.5.2 that in general QMLE selects that approximation that minimizes the KL divergence between the two densities, or effectively the relative entropy.
65. There actually remains some bias in the mean reversion-rate estimator.
66. Indeed, this issue led to some misleading results in Zhou (2001).
67. Attempting to identify possible trading opportunities requires knowing *both* mean reversion rate and level. Seeing that a price is above or below its long-term mean is useful only if there is some idea as to how long it will take the price to revert (and therefore being able to quantify how significant the deviation

is). It goes without saying that being wrong about this point can have serious monetary consequences. You can be right about your bet long term, but this is of little matter if you are crushed short term waiting for the bet to play out. (This is not to say that the econometrics give much confidence in being able to identify the mean itself, either.)

68. *I.e.*, knowing the true underlying parameters. It is not enough to *know* that a process is driven by jumps, stochastic volatility, mean reversion, *etc.* One must obtain actual, numerical values for the parameters underlying such processes, *if* such models are to be used for valuation purposes.
69. Bates (2006) also considers characteristic function-based estimation of affine processes with latent variables, using a conditional result reminiscent of the results in Section 5.2.2 as a filtering device. Chacko and Viceira (2003) also apply matching methods discussed here, but by simply integrating the conditional characteristic function across the latent space (weighted by the unconditional latent density, of course). Although this is viewed as a Markovian partial information characteristic function, it cannot really be, as stochastic volatility processes (*e.g.*, Heston) are not Markovian and the density conditional only on price (say) depends on the *entire* past history of prices. See Section 6.2.
70. Yu (2014) also notes a third option of increasing both sample size and resolution.
71. It should be clear from this expression why the population mean of a process is rather difficult to estimate from a sample average: only two data points are used in the estimator.
72. We trust there will be no confusion between using $N$ as both the number of variables and the normal CDF.
73. The terms involving exponentials simply comprise a (convergent) geometric series, the contribution from which vanishes in the limit when divided by the sample size.
74. We will casually switch between treating the time parameter as a subscript or as a (proper) function argument.
75. The subsequent analysis will of course take $\theta = 0$.
76. For simplicity we will assume different mean-reversion rates.
77. Across the time points $\tau_k$.
78. We leave aside the question of whether technology is merely revealing an underlying market microstructure, or in fact creating it.
79. *Continue à droite, limite à gauche* (pardon our French.) In English, right continuous with left limits. This conception is important for modeling jumps in financial applications, as it captures the notion of unanticipated shocks but continuity *after* such shocks; see Tankov and Cont (2003). In fact, in many applications $A$ is actually previsible/predictable/left continuous, which means $A_t$ is $\mathfrak{F}_{t-}$-measureable (heuristically, its value is known an instant ahead).
80. To the filtered probability space on which $X$ is defined.

81. The distinction between a local martingale and a martingale is a rather subtle technical distinction that need not greatly concern us here. Suffice to say the relevance concerns the construction of Itô integrals that are central to so much modeling. (The set of local martingales is closed under the operation of stochastic integration.) All martingales are local martingales, but not *vice versa*; roughly speaking, local martingales typically possess some kind of anomaly (to engage in some physics speak) that amounts to a singularity on the domain of definition. A good example in the context of mathematical finance (the CEV process) can be found in Carr *et al.* (2007). Further intuition may be gained by considering the difference between futures and forwards.
82. Note that QV is in general a process.
83. RV converges in probability to QV in the limit of infinitesimal resolution.
84. A somewhat related approach can be found in Aït-Sahalia (2004).
85. The first expression in (6.224) is fairly easy to establish. The second expression requires a bit more work, in particular a judicious application of the triangular rule for inequalities involving absolute values of sums and differences.
86. Recall from Section 3.2 the important role of QV in the valuation of options in incomplete markets. In particular, for the Heston stochastic volatility model we saw that valuation with a pricing functional *not* derived from an EMM provided a robust framework for extracting QV *in conjunction with a specific hedging strategy*. We will say here that we observe broadly similar results for valuing options under jump diffusions. It is worth noting that in the familiar Merton jump diffusion formula (6.224), the QV from (6.224) is only vaguely apparent, even in expected value. This is not too surprising, as the conventional EMM result for Merton is based on replication via a *continuum* of hedging instruments, which amounts to assuming the solution to the problem at hand. We see again the critical need for identifying entities such as QV which are not only robust to estimation but have a *direct* connection to valuation in terms of a specific hedging regime.
87. In other words, do not look for jumps as such, but rather where their effects matter the most, *e.g.*, close to maturity.

## 7 Numerical Methods

1. The log-asset ratio is a difference of normal variables, whose covariance structure has a well-known quadratic form. In the context of spark spread options, this volatility is commonly referred to as a heat-rate volatility.
2. Recall again from Euler's formula that, since the option value is homogeneous of degree one, $V = S_1 V_{S_1} + S_2 V_{S_2}$.

3. An interesting fact is that, although the vega of the spread option with respect to the ratio volatility is of course positive, the individual *leg* vegas are of indeterminate sign. This follows simply from the chain rule, *e.g.*, $\frac{\partial \sigma}{\partial \sigma_1} = \frac{\sigma_1 - \rho\sigma_2}{\sigma}$. Note that what determines the sign is the regression coefficient between the (normalized) returns ($\frac{\rho\sigma_2}{\sigma_1}$).
4. Obviously these two models are not consistent with one another. As always, the primary criterion for deciding between models comes down to the question of: what trades, and how do tradeables relate to the payoff in question?
5. This situation also applies to other fuel markets, and a few electricity markets.
6. Indeed, standard valuation via GBM may give misleading, if not spurious results, as it may simply reflect the variability of the common backbone, and not the variability of the spread as such. See Figure 1.3.
7. In truth, Kirk's original idea was to write the dynamics of the aggregate strike as $\frac{d(S_1+K)}{S_1+K} = \frac{S_1}{S_1+K}\sigma_1 dw_1$, from which the association in (7.14) naturally follows. See Li *et al.* (2008) for a similar (equivalent lognormal) approach.
8. The combination $\rho\sigma_1 - \sigma_2$ also appears in the valuation of the digital, second-order term in (7.11).
9. A similar idea is employed in the trigonometric approximation of Carmona and Durrleman (2003).
10. So, *e.g.*, in terms of inception prices the means are given by $\mu_i = \log S_i - \frac{1}{2}\sigma_i^2$.
11. Note that when $K = 0$, the expression for $d$ becomes linear in $z_1$ and the resulting integrals can be integrated exactly via the well-known result $\int_{-\infty}^{\infty} dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} N(az + b) = N(\frac{b}{\sqrt{1+a^2}})$. The reader can verify that the Margrabe formula is recovered in this case.
12. There are in fact two discretizations, one for each normal CDF appearing in the different integrands in (7.26).
13. Typically for quadrature over infinite intervals, the set of grid points is truncated at some suitably large finite value, where "suitable" refers to the range for which the omitted contributions are numerically (and probabilistically) negligible. In truth, the asymptotic behavior of the integrand may be effectively employed to enhance accuracy, as we will see in the next subsection.
14. We will see the general result for this claim in the subsection on quadrature methods.
15. As will be seen, elliptical distributions are defined by their characteristic functions, which have the form $e^{i\phi^T \mu}\Psi(\phi^T \Sigma \phi)$ for some vector $\mu$ and matrix $\Sigma$. These obviously encompass the Gaussian case.
16. The results in this section arose from joint work with Alexander Eydeland.
17. In truth we are considering the expectation in (7.44) with respect to an arbitrary measure, *i.e.*, prices need not be martingales. Although this means that the resulting valuation cannot be an option price in the sense of representation of a replication strategy, expectations such as that in (7.44) prove useful (and

meaningful) in control problems to be studied later, so it is worth introducing such problems now.

18. By embedding the underlying Toeplitz matrix in a larger (twice as big and padded with zeros) circulant matrix (a special kind of Toeplitz matrix where each column is a cyclic permutation [periodic wraparound] of the first column), the problem can be crafted as a convolution and, as is well known, the Fourier transform of a convolution is the product of the Fourier transform of the two entities being convolved. Toeplitz matrix multiplication thus requires two Fourier transforms and one inverse transform.
19. We note in passing here another very fast transform method, the so-called fast Gauss transform, applied to option pricing in Broadie and Yamamoto (2003).
20. This general form encompasses both mean reversion and GBM, obviously. Pearson's method can likewise be viewed as solving a PDE (in two spatial dimensions).
21. We use the fact that $\varphi(x-\sigma) = \varphi(x)e^{x\sigma-\sigma^2/2}$. This is essentially the same trick that allows terms to cancel out in a straightforward calculation of deltas (via differentiation) in the BS model.
22. *I.e.*, take $x_T = \mu_x + \sigma_x\xi$, $y_T = \mu_y + \sigma_y\zeta$ and use the assumptions in (7.61).
23. It is not hard to verify that (7.66) is essentially a two-dimensional convolution, so that its Fourier transform is the product of the Fourier transforms of the convolved entities.
24. In (7.61) this requirement would be satisfied if $a_{xy} = a_{yx} = 0$.
25. Essentially, the grid is a tile pattern of parallelograms. Note that the time-$T$ grid in (7.63) will be unchanged.
26. In truth, as happens in the one-dimensional case, the multiplication in (7.66) must be embedded in a larger (twice the size in both dimensions, with zero padding), circulant-type matrix multiplication for which the 2D FFT can be directly applied. See Eydeland and Mahoney (2003) for details.
27. More accurately, if $\tau$ is a stopping time with respect to the filtration $\mathfrak{J}_t$ of some random process, then the event $\{\tau \le t\} \in \mathfrak{J}_t$. Equivalently, the random variable $1\{\tau \le t\}$ is $\mathfrak{J}_t$–measureable.
28. As is well known, in the usual GBM/BS framework, an American call option has no early exercise optionality unless there are dividends. (The put option always has early exercise value.) Since we will typically be considering valuation and hedging strategies with futures (which are driftless under the pricing measure), this condition will be satisfied (so to speak) as long as interest rates are nonzero. Although the discounting issue will not really be relevant when we consider physical control problems for non-martingale spot prices, it is in general important and so we will reintroduce explicit discounting in the exposition here.
29. We are here appealing, in a very heuristic manner, to the modern definition of conditional expectation.

30. Here $p$ and $q$ are the probabilities of up/down moves, respectively (so $p+q=1$) and $u$ and $d$ are the size of the up/down moves, respectively (under the pricing measure). The binomial model is essentially a finite difference approximation to the underlying valuation PDE.
31. Note that, even under the pricing measure, spot prices are *not* martingales. This is not problematic in any sense, as spot assets are (typically) *not* traded on a forward, financial basis. That is to say, transactions in terms of spot must be settled *physically*, *i.e.*, actual possession of the asset must be assumed by one of the transacting parties.
32. The typical convention is to render per-unit cash flows relative to *delivered* volumes, or equivalently relative to state changes.
33. Typically, some kind of penalty is applied to ensure that constraints are satisfied. For example, if the terminal condition requires an empty tank, then we would take $V_T(S;Q_j) = -M$ (where $M$ is very large), except for $Q_j = 0$ where $V_T = 0$ for all $S$.
34. For simplicity we here consider only time-homogenous problems, which are functions only of time-to-maturity $\tau = T - t$.
35. Recall the discussion of load-serving products from Section 1.2.4. There, we saw that a load deal (statically) hedged at expected load creates an exposure to relative power-load covariance (pathwise), and that in many deals (primarily industrial) this entity has a relationship to realized price variation. We thus saw the feasibility of further hedging such deals with volatility-derived instruments. A delta-hedged option is one possibility, a variance swap is another. Hence the interest in being able to evaluate such structures here.
36. Obviously, this grid will only occupy the first quadrant (*i.e.*, only positive values for both variables).
37. Meaning at the present time $n$ of the iteration in (7.100).
38. Boundary terms do indeed present a challenge in more than one dimension. In this case, it is difficult to appeal to asymptotic behavior as there is no single "preferred" or "natural" direction to invoke in the asymptotics. *E.g.*, it makes little sense to speak of the behavior of a spread option for large values of the long leg without specifying the order of magnitude of the short leg. In this case, there is probably little more that can be done beyond truncating the discretized asset space at very large values.
39. The actual quadrature points can be generated using algorithms found in any standard text, *e.g.*, Press *et al.* (2007).
40. A useful exercise for the reader would be to employ these methods to gauge the (continued surprising) effectiveness of Kirk-type approximations for spread products with fixed strike components.
41. We considered tolling deals in great detail in Section 4.2, and this slightly out-of-context example provides a good example of quadrature-based techniques for rather complex problems.

42. Incremental heat rate is derived from $HR_{\min} C_{\min} + HR_{inc}(C_{\max} - C_{\min}) = HR_{\max} C_{\max}$.
43. Of course, the range $\left[0, \sin^{-1} \rho\right]$ must be (linearly) mapped to the range $[-1, 1]$ appropriate for this quadrature scheme.
44. An extension of Newton's method. See Acklam (2002).
45. Note that most documented algorithms for Gauss-Hermite quadrature (*e.g.*, Press *et al.* [2007]) need to be suitably modified to take into account the factors of $1/2$ in the exponent and $\frac{1}{\sqrt{2\pi}}$ multiplying the exponential.
46. A simple algorithm is given in Press *et al.* (2007). It can easily be shown that the transformation $L_{ij} \to L_{n+1-i,n+1-j}$ produces a factorization in terms of an *upper* triangular matrix.
47. The Jacobi rotation algorithm is a standard (and robust) algorithm for computing the eigenvalues and eigenvectors of a symmetric matrix; again consult Press *et al.* (2007) for the relevant details. Recall that the eigenvectors of distinct eigenvalues of a real, symmetric matrix are orthogonal (and in fact form a complete basis set, in which case can always be expressed as an orthonormal set via Gramm-Schmidt).
48. The so-called triangular factorization; see Hamilton (1994). The diagonal elements of $D$ are the same as the diagonal elements of $L$ in the Cholesky factorization.
49. It turns out that the form (7.134) is well suited to obtaining various greeks/option sensitivities via the so-called likelihood ratio method, which will receive a fuller exposition in the next section on simulation methods. As we will see, for payoffs $V(x)$ under some $n$–dimensional Gaussian process $x$, the deltas can be obtained from the expected value of $V\Sigma^{-1}x$, where $\Sigma$ is the process covariance matrix.
50. *E.g.*, for a 6–dimensonal problem with 10 quadrature points in each dimension (often quite sufficient for one-dimensional problems), 1 million total points are required.
51. We are speaking here in the context of commonly encountered problems in energy markets; *e.g.*, tolling problems with on-peak and off-peak components and fuel switching, gas transport problems with multiple receipt and delivery points, *etc.* It is safe to say Gaussian quadrature will not be applicable to 30-year mortgage-backed security valuation.
52. In general, standard Gaussian schemes such as (7.111) and (7.112) are *not* embedded, so that as higher accuracy is sought, a completely new set of quadrature weights/points (and hence a completely new set of evaluations of the function being integrated) are required. It is possible to extend Gaussian quadrature to retain a nested structure (*e.g.*, Gauss-Kronrod-Patterson or Clenshaw-Curtis schemes), but this is a somewhat advanced topic in numerical analysis that we

cannot present in any kind of depth here. Again, Press *et al.* (2007) serves as a good starting point for pursuing these topics further.

53. In other words, if you have a function well approximated by level *l*–type functions (so that schemes such as (7.138) should be expected to perform well), then going to a higher level of resolution will not gain much in way of accuracy. We will exploit this point shortly for facilitating higher-dimensional schemes.
54. Examples would be Gaussian schemes that are exact for polynomials of a given degree that remain exact (for that degree) as the number of quadrature points is increased. Another example would be functions that are piecewise linear over intervals of size *h*, and thus can be exactly integrated via the trapezoidal rule; such functions are still exactly integrated by this scheme when the number of quadrature points is doubled (so that the step size is halved).
55. In fact, one-dimensional embedded schemes also suffer from this problem.
56. The old adage "garbage in, garbage out" comes to mind here.
57. Note the claim here. We are *not* saying that (historical) hourly prices are irrelevant to good valuation of such products. Rather, we are saying that hourly prices should not be *directly* modeled (at least in such markets where hourly prices do not trade). It is the *relationship* of hourly prices to those entities that clear against traded products (*e.g.*, monthly prices vs. futures) that should be modeled. (At any rate, there is seldom sufficient data in energy markets to permit robust estimation of hourly models.)
58. Or more absurdly, supply-demand conditions for time horizons well beyond the maturity of any (traded) forward curve.
59. The Central Limit Theorem (CLT) states that the distribution of the sample average of $N$ IID variables is asymptotically normal about the population mean, with variance decreasing such as $1/N$. (More accurately, the entity $\frac{1}{N^{1/2}} \sum_i \frac{x_i - \mu}{\sigma}$ is asymptotically distributed as a standard normal, where $\mu$ and $\sigma$ are respectively the population mean and standard deviation.) The basic intuition can be gleaned from looking at the characteristic function of the sample average (indeed, the standard proof follows the same path). A classic interview question applies the CLT by asking how many coin flips out of, say, 100, would need to be heads (or tails) for the observer to conclude that the coin was not fair.
60. We should mention here a popular alternative to linear congruential generators, the so-called Mersenne twister, which is a matrix linear recursion over a binary field with very long period (of the form $2^n - 1$, a Mersenne prime [hence the name]).
61. Here $\lambda$ would be the jump intensity. As is well known, the inter-arrival times of the jumps of a Poisson process are independent and exponentially distributed.
62. Another useful trick for simulating points uniformly distributed over an *n*-dimensional hypersphere is to simulate *n* independent unit normals, and then normalize these by the $L^2$ norm, as can be seen from transforming the CDF

via generalized spherical coordinates. Recall that such random variables (not surprisingly, termed spherical random variables) played a central role in the study of the important topic of joint dependency, considered in Chapter 8 in the context of copulas.

63. In truth, the real requirement is that RVs can be readily drawn from $G$, whether via the inverse operating on a uniform RV or some other tractable method.
64. It is worth noting that the basic idea of acceptance/rejection also underlies the use of Markov Chain Monte Carlo (*e.g.*, the so-called Gibbs sampler or the Metropolis-Hastings algorithm) and particle filters used to implement the general filtering algorithm discussed in Section 6.2.
65. We also see here the claim made previously about the $O(1/\sqrt{N})$ computational efficiency of simulation.
66. This approach will only be effective for right-tailed probabilities, *i.e.*, for $\alpha > 0$, since otherwise the additional exponential factor will contribute divergently growing terms that will hamper convergence. The left-tailed probabilities (*i.e.*, the case $\alpha < 0$) are easily handled, however, by considering $\Pr(z > \alpha) = 1 - \Pr(z < \alpha)$. This is the same situation that arises in option pricing, where it is advisable (either numerically or econometrically) to always consider OTM options; *e.g.*, value OTM puts rather than ITM calls.
67. This fact also holds true for eigenvalue/eigenvector calculations, which often proves useful in dimension reduction techniques such as principal components analysis (PCA), which we will consider in Chapter 8.
68. We will explain shortly the distinction between forward and cash vegas. The alert reader will no doubt anticipate that the distinction relates to the structure of the underlying market, specifically the market for options (volatility), *e.g.*, whether the extent of traded options is monthly, daily, *etc.*
69. This was previously pointed out in Section 3.1.6.
70. This argument is also used to deduce certain properties of the information matrix used in diagnostics for maximum likelihood estimation, as we noted in Chapter 6.
71. For ease of exposition we assume the region of integration in question is $(0,1)^d$ in $\mathbb{R}^d$.
72. Pseudo-clustering is a more appropriate characterization, although in any finite sample the ramifications are quite real.
73. *I.e.*, for for bits (0 or 1) $b_1$ and $b_2$, $b_1$ XOR $b_2 = 1$ iff $b_1$ and $b_2$ are different.
74. *I.e.*, the primitive polynomial and resulting direction numbers.
75. We revert to our usual custom of ignoring discounting effects, although we should stress that, in the absence of mean-reverting effects (or more generally volatility scaling laws; see Chapter 2), many control problems have *no* operational/early exercise premium. Recall the well-known example of a non-dividend paying stock under GBM.

76. These regressors are, necessarily, meant to reflect *pathwise* relationships. As such, they cannot really be used in conjunction with the quasi-Monte Carlo methods of Subsection 7.5.3, which are by nature a means of (non-stochastically) filling out regions in hyperspace, and thus are best suited for problems which can be crafted as (high-dimensional) quadratures.
77. These are certainly the most well-known (and popular) expositions. However, the basic idea appears to have been used as early as Carriere (1996). For an early application in energy markets, see Ghiuvea *et al.* (2001).
78. It also bears some similarity to tree-based regression from machine learning; see Hastie *et al.* (2009).
79. *E.g.*, a spark spread option. Recall, in the context of tolling, the concrete example in Section 4.2.
80. By exploiting the nonincreasing nature of $A$, observe that the duality result in (7.187) can actually be crafted in terms of *martingale* value proxies, not supermartingales. This fact will be used when we apply duality to control problems.
81. *I.e.*, it is the incremental value (of losing an exercise right). Note that the difference operator here applies to the exercise rights, not the usual case in econometrics where it applies to the time index (as in Chapters 2 or 6). Trivially (using telescoping sums), the value function for a given level of exercise rights can be obtained from all marginal value functions of lower order (so to speak).
82. Note that the case $\gamma < 0$ is suitably handled by using the identity $\Pr(z > \gamma) = 1 - \Pr(z < \gamma)$ and considering a contour *above* the real axis (*i.e.*, $\varepsilon > 0$). This approach amounts to shifting the contour of integration and accounting for the crossing of a pole.
83. Recall that $\gamma$ is positive. In the case where one (or both) of the components of $\gamma$ is negative, we can (as in the one-dimensional case) employ identities such as $\Pr(x > a, y > b) = \Pr(x > a) - \Pr(x > a, y < b)$ and shift contours above (instead of below) the real axis to facilitate the desired calculation.
84. From the usual result that partial derivatives of the characteristic evaluated at $\varphi = 0$ yield appropriate moments of the underlying distribution.
85. In truth, we are usually more interested in *integrated* variance, which requires introducing the process $dV = vdt$ . However, it is simpler to consider instantaneous variance here for the approach we wish to illustrate.
86. This problem was first studied by Carr and Madan (1999); a useful overview can be found in Borak *et al.* (2005).
87. Already encountered in Section 7.3.2.
88. These are the so-called Nyquist frequencies arising from the implicit, imposed periodicity of the value function (truncated in the frequency regime).
89. In $d$ dimensions and resolution size $N$ (in each dimension) the operation cost is $O(N^d \log N)$ vs. $O(N^{2d})$ for ordinary matrix multiplication.

## 8 Dependency Modeling

1. Allowing for the important caveat about market structure, namely the existence of liquid option markets in determining whether it is correlation, or a ratio volatility, that is the relevant measure.
2. Note that an arbitrary function $C$ will not be a valid copula function unless it satisfies certain consistency and other technical conditions, such as $C(1,\dots,u_k,\dots,1) = u_k$ and $C(u_1,\dots,0,\dots,u_n) = 0$; see Nelsen (1999). These are mainly common-sense requirements that allow interpretation as a distribution function, *e.g.*, in two dimensions we must have $C(x_2,y_2) - C(x_1,y_2) - C(x_2,y_1) + C(x_1,y_1) \geq 0$, which amounts to the requirement that the probability mass of the box $[x_1,x_1) \times \left[y_1,y_2\right)$ be non-negative.
3. This follows from $\Pr(F(X) < x) = \Pr(X < F^{-1}(x)) = x$.
4. The uniqueness result only holds for continuous random variables; for discrete random variables, the CDF can still be written in terms of a copula function, but the result is nonunique (more specifically, it is only unique on the Cartesian product of the ranges of the marginal distributions).
5. *E.g.*, if $\tilde{C}(q,r) = q + r - 1 + C(1-q,1-r)$ for some copula $C$, then $\tilde{C}(q,0) = 0$ and $\tilde{C}(q,1) = q$. It should be clear why the convention in terms of the standard distribution function is chosen, although this alternative framework will prove suitable when we consider applications to Lévy processes.
6. In such a case random variables $u$ and $v$ can be written as $F_1^{-1}(u)$ and $F_2^{-1}(1-u)$ for some uniform variate $u$ and some CDFs $F_i$.
7. Even here, there may be cases where the second moments do not exist.
8. We have already seen, from the perspective of valuation, that correlation may not be the appropriate entity of interest, even when the underlying physical (joint) distribution is indeed characterized by correlation.
9. More technically, they are asymptotically independent in the upper tail.
10. There is as always the risk of getting distracted by particular, mathematical details. Obviously, like any other tool, copulas are useful in certain situations but not in others, and one should be careful not to overstate their applicability. As marginal information is necessarily contained in any joint dependency structure, one can reasonably ask if it is not better in some cases to directly model this joint behavior instead of breaking a problem into separate analyses of the marginals and the copula. For a spirited debate on the issue, see Mikosch (2005) and Genest and Rémillard (2006).
11. We are here thinking in opposition to strange creatures such as Dynamic Conditional Correlation (DCC), an escapee from the GARCH bestiary. (See Caporin and McAleer [2013] for a critical discussion.) Such models are essentially *ad*

*hoc* attempts to formally model the manifestations of information accumulation over different time scales, at the cost of severing the explicit link to the mechanisms giving rise to those time scales. As a result, DCC dependency structure varies with the choice of (individual) marginal distributions. See again Section 6.2.3.

12. We should confess to a bit of subterfuge here, as elliptical *distributions* are a completely separate entity from copulas, as such. An elliptical *copula* is simply a copula derived from the CDF of an elliptical distribution. An elliptical distribution has a very specific set of marginals (see Endnote 14), whereas an RV with an elliptical copula can of course have arbitrary marginals. However, elliptical distributions represent a sufficiently important class of joint dependency, and are commonly presented in conjunction with copulas, so we feel justified in including them in our discussion here. The reader should keep in mind these subtleties, however.
13. The matrix $\Sigma$ can be thought of as characterizing the linear dependencies of the distribution, while the radial variable $R$ (in the base spherical representation) generates nonlinear dependencies. It is this separation that creates (in the non-Gaussian) a nonequivalence between independence and noncorrelation.
14. Yet another characterization of elliptical distributions is through the resulting density, which can be seen to have the form $\frac{1}{\sqrt{\det \Sigma}} c_n h_n(\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$ for some function $h$ (also termed a generator) and normalization constant $c$, both dimension dependent. In fact, it is far more common to operate in terms of the density generator rather than the characteristic function generator. Note that in general the marginals will not necessarily correspond to the one-dimensional versions of the families characterizing the joint structure. For example, the marginal of a $N(0, I_n)$ normal is $N(0, 1)$, but the marginals of a multivariate logistic variable are not themselves logistic. This phenomenon is known as inconsistency and can induce rather odd effects when considering loss distributions of portfolios.
15. Recall Endnote 14: the marginals are elliptical, but they do not necessarily belong to the same class or family that characterizes the joint distribution, except in certain special cases (*e.g.*, normal).
16. For convenience we set the unconditional means to zero.
17. Obviously the result only holds for correlations strictly less than one.
18. The proof basically follows from the definition of a copula and the product requirement on $G$ (note that for the case under consideration, $G_{ij}(0) = 0$ and $G_{ij}(1) = 1$).
19. Density functions will be denoted by lowercase ($f$), distribution functions by uppercase ($F$).
20. See Patton (2006) for a discussion of conditional copulas, in particular an extension of Sklar's theorem.

21. Multivariate generalizations are possible; see Czado (2010).
22. To facilitate the exposition, we will unavoidably encounter some overlap with material previously presented in Chapter 5.
23. Since this kind of behavior clearly mimics diffusive (continuous) behavior, some researchers (*e.g.*, Carr *et al.* 2002) have gone so far as to advocate modeling in terms of pure jump processes. Whatever one may think of this approach, Lévy processes clearly offer great flexibility in modeling a wide range of processes.
24. Note, then, that the choice of truncation interval (*e.g.*, $|x| \leq 1$ in (8.43)) is essentially arbitrary (conventional is probably a better term), so long as it excludes the origin. (Note that, since $1_{|x|\leq 1} = 1_{|x|\leq \varepsilon} + 1_{\varepsilon \leq |x| \leq 1}$ for small $\varepsilon$, and the integrand corresponding to the infinite activity jumps is [to leading order] a quadratic form in $\phi$, there is thus some freedom in how the structure of a given Lévy process can be allocated [so to speak] between diffusive and pure [infinite activity] jump components.) This fact can be employed to show that the Lévy property is preserved under linearity for *independent* Lévy processes.
25. In the sense of precluding more general dependency structures.
26. Obviously, the presupposition is that we care about dynamics as such, and not simply terminal distributions.
27. For a (formal) recipe in the context of GARCH modeling, see Patton (2006).
28. For more on stable processes, see Borak *et al.* (2005).
29. The small time copula is the Brownian component's independence copula, and the large time copula is the Cauchy process's copula, which are of course different.
30. The more general case is considered in Kallsen and Tankov (2006). This restriction is not as stringent as it may seem, as it has obvious relevance for modeling stochastic volatility with Lévy processes; see Barndorff-Nielsen and Shephard (2001).
31. Note that the non-integrability of the Lévy measure is weaker than in the standard case: only $\min(1, |x|)\nu(x)$ need be integrable.
32. Namely, that, conditional on the number of events within a given time interval, the arrival times are distributed as a ranked set of uniforms. Given the complete symmetry inherent in the summation in (8.53), the terms may be treated as independent uniforms.
33. This result can be extended to more general pure jump processes; see Rosinski (2001).
34. Along with either perfect dependence or complete independence between the Poisson jump components.
35. Essentially a vector space closed under multiplication by a positive scalar.
36. The result in (8.62) can be generalized for nonsymmetric matrices without multiple eigenvalues by using the inverse matrix of eigvenvectors (note that the

matrix square root does not always exist). We will also use the notation $\Sigma^{1/2}$ to denote matrix square root.

37. Specifically, the distribution of various entities [such as variance] under the typical case in practice of small sample sizes.
38. The term *factors* may perhaps be more familiar to some readers.
39. Nor do we know the expected returns, which are notoriously hard to estimate robustly. We do not even know (usually) whether joint normality is a good model for any particular situation. We will ignore these concerns here.
40. We will ignore the contributions of means, both sample and population, in this discussion.
41. Recall the discussion in Section 6.1.3 on the eigenvalues of matrices of less than full rank.
42. This kind of problem arises in other applications involving correlation matrices. For example, oftentimes the Cholesky factorization of a correlation matrix is required (*e.g.*, in simulations; see Chapter 7), the algorithm for which requires the underlying matrix be positive definite, which is equivalent to requiring that the matrix has all positive eigenvalues. In such cases where the (estimated) correlation matrix is "slightly" negative definite (meaning the most negative eigenvalue is small in absolute value), it is possible to reconstruct a valid (*i.e.*, positive definite) correlation matrix by imposing a ceiling on the negative eigenvalues (say, replacing them by a small, but positive number) and then using the eigenvalue-eigenvector factorization in (8.72), suitably rescaling the eigenvectors so the resulting product has ones along the diagonal. See Rebonato and Jäckel (1999) or Schöttle and Werner (2004).
43. This is akin to the sum of the eigenvalues in a principal components analysis representing the total variance of the driving factors.
44. Note that the corresponding eigenvector has nearly equal elements. This is a typical pattern in the case where there is high pair-wise correlation among the original entities. There is an interpretation in terms of the eigenportfolios discussed in Section 8.2.1: the highest risk eigenportfolio, with nearly equal weights across assets, can be thought of as the so-called market portfolio. The lower risk eigenportfolios can be seen to consist of various spread positions across the assets.
45. We encountered the notion of effective dimensionality when we considered approaches for increasing the efficiency of simulation in Chapter 7.
46. In interest rate work one often sees the terms level, slope, and curvature for these components, respectively.
47. PCA represents, essentially, a change in basis. We should therefore mention recent work on a similar concept, so-called balanced baskets. See Bailey and López de Prado (2012).

# Bibliography


[1] Acklam, Peter John, 2002, "A Small Paper on Halley's Method," available at http://home.online.no/~pjacklam.

[2] Adhikari, Sondipon, 2007, "Matrix Variate Distributions for Probabilistic Structural Dynamics," *AIAA Journal*, 45, 7.

[3] Ahn, Hyungsok, Danilova, Albina, and Swindle, Glen, 2002, "Storing Arb," *Wilmott Magazine* available at http://www.wilmott.com/pdfs/020923_storing_arb.pdf.

[4] Aït-Sahalia, Yacine, 2004, "Disentangling Diffusion from Jumps," *Journal of Financial Economics*, 74.

[5] Aït-Sahalia, Yacine, and Jacod, Jean, 2014, *High-Frequency Financial Econometrics*, Princeton: Princeton University Press.

[6] Aït-Sahalia, Yacine, and Kimmel, Robert L., 2010, "Estimating Affine Multifactor Term Structure Models Using Closed-Form Likelihood Expansions," *Journal of Financial Economics*, 98.

[7] Albrecher, H., 2004, "The Valuation of Asian Options for Market Models of Exponential Levy Type," available at http://www.hec.unil.ch/halbrecher/files/AsianLevyExp.pdf.

[8] Albrecher, H., Dhaene, J., Goovaerts, M., and Schoutens, W., 2005, "Static Hedging of Asian Options under Levy Models," *Journal of Derivatives*, 12, 3.

[9] Albrecher, H., Mayer, P., and Schoutens, W., 2008, "General Lower Bounds for Arithmetic Asian Option Prices," *Applied Mathematical Finance*, 15, 2.

[10] Albrecher, Hansjörg, Mayer, Phillip, Schoutens, Wim, and Tistaert, Jurgen, 2006, "The Little Heston Trap," available at http://www.schoutens.be/HestonTrap.pdf.

[11] Alexander, Carol, 2001, *Market Models: A Guide to Financial Data Analysis*, Chichester: John Wiley and Sons.

[12] Alexander, Carol, and Lazar, Emese, 2004, "The Continuous Limit of a GARCH Process," ISMA Centre Discussion Papers in Finance 2004–10.

[13] Alexander, Carol, and Nogueira, Leonardo M., 2006, "Hedging Options with Scale-Invariant Models," ICMA Centre Discussion Papers in Finance DP2006-03.

[14] Amos, D. E., 1969, "On Computation of the Bivariate Normal Distribution," *Math. Comp.*, 23.

[15] Andersen, Leif, 2006, "Efficient Simulation of the Heston Stochastic Volatility Model," available at http://www.javaquant.net/papers/LeifAndersenHeston.pdf.

[16] Andersen, Leif, and Broadie, Mark, 2004, "A Primal-Dual Simulation Algorithm for Pricing Multi-Dimensional American Options," *Management Science*, 50, 9.

[17] Andersen, Torben G., Chung, Hyung-Jin, and Sørensen, Bent E., 1999, "Efficient Method of Moments Estimation of a Stochastic Volatility Model: A Monte Carlo Study," *Journal of Econometrics*, 91.

[18] Asmussen, Søren, and Glynn, Peter W., 2007, *Stochastic Simulation: Algorithms and Analysis*, New York: Springer.
[19] Avellaneda, Marco, and Lee, Jeong-Hyun, 2010, "Statistical Arbitrage in the U.S. Equities Market," *Quantitative Finance*, 10, 7.
[20] Bailey, David H., and Swartztrauber, Paul N., 1995, "The Fractional Fourier Transform and Applications," available at http://www.davidhbailey.com/dhbpapers/fracfft.pdf.
[21] Bailey, David H., and López de Prado, Marcos M., 2012, "Balanced Baskets: A New Approach to Trading and Hedging Risks," *The Journal of Investment Strategies*, 1, 4.
[22] Barndorff-Nielsen, Ole E., and Shephard, Neil, 2001, "Non-Gaussian Ornstein-Uhlenbeck-Based Models and Some of Their Uses in Financial Economics," *Royal Statistical Society*, 63, 2.
[23] Barndorff-Nielsen, Ole E., and Shephard, Neil, 2002, "Estimating Quadratic Variation Using Realized Variance," *Journal of Applied Econometrics*, 17.
[24] Barndorff-Nielsen, Ole E., and Shephard, Neil, 2003, "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society*, 64, 2.
[25] Barndorff-Nielsen, Ole E., and Shephard, Neil, 2004a, "Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation," available at http://www.people.fas.harvard.edu/~shephard/papers/split.pdf.
[26] Barndorff-Nielsen, Ole E., and Shephard, Neil, 2004b, Power and Bipower Variation with Stochastic Volatility and Jumps (with Discussion)," *Journal of Financial Econometrics*, 2.
[27] Barndorff-Nielsen, Ole E., and Shephard, Neil, 2012, "Basics of Lévy Processes," available at www.nuff.ox.ac.uk/users/shephard/introlevy120608.pdf.
[28] Bates, David S., 2006, "Maximum Likelihood Estimation of Latent Affine Processes," *The Review of Financial Studies*, 19, 3.
[29] Baxter, Martin, and Rennie, Andrew, 1996, *Financial Calculus*, Cambridge: Cambridge University Press.
[30] Bellman, Richard, 1957, *Dynamic Programming*, Princeton: Princeton University Press.
[31] Benninga, Simon, Björk, Tomas, and Wiener, Zvi, 2002, "On the Use of Numeraires in Option Pricing," *The Journal of Derivatives*, 10, 2.
[32] Benner, Peter, and Mena, Hermann, 2004, "BDF Methods for Large-Scale Differential Riccati Equations," in B. De Moor *et al.* (eds.), *Proceedings 16th International Symposium on Mathematical Theory of Network and Systems*, Leuven.
[33] Benth, Fred Espen, and Kettler, Paul C., 2010, "Dynamic Copula Models for the Spark Spread," *Quantitative Finance*, 11, 3.
[34] Billingsley, Patrick, 1961, "The Lindeberg-Lévy Theorem for Martingales," *Proc. Amer. Math. Soc.*, 12.
[35] Bjerksund, Petter, and Stensland, Gunnar, 2006, "Closed Form Spread Option Valuation," available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1145206.
[36] Bjerksund, Petter, Stensland, Gunnar, and Vagstad, Frank, 2008, "Gas Storage Valuation: Price Modeling v. Optimization Methods," available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1288024.
[37] Björk, Tomas, 2009, *Arbitrage Theory in Continuous Time, 3rd Ed.*, Oxford: Oxford University Press.

[38] Black, F., and Scholes, M., 1973, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81.
[39] Boguslavsky, Michael, and Boguslavskaya, Elena, 2004, "Arbitrage Under Power," *Risk* June 2004.
[40] Bollerslev, Tim, 2008, "Glossary to ARCH (GARCH)," CREATES Research Paper 2008-49.
[41] Boogert, Alexander, and de Jong, Cyriel, 2008, "Gas Storage Valuation Using a Monte Carlo Method," *The Journal of Derivatives*, Spring 2008, 81–98.
[42] Borak, Szymon, Detlefsen Kai, Härdle, Wolfgang, 2005, "FFT-Based Option Pricing," SFB 649 Discussion Paper 2005-011.
[43] Borak, Szymon, Härdle, Wolfgang and Weron, Rafal, 2005, "FFT-Based Option Pricing," SFB 649 Discussion Paper 2005-008.
[44] Bouchaud, J. P., and Potters, M., 2000, *Theory of Financial Risks: From Statistical Physics to Risk Management*, Cambridge: Cambridge University Press.
[45] Bouchaud, J. P., and Potters, M., 2009, "Financial Applications of Random Matrix Theory: A Short Review," available at http://arxiv.org/abs/0910.1205v1.
[46] Broadie, M., and Kaya, O., 2006, "Exact Simulation of Stochastic Volatility and Other Affine Jumpt Diffusion Processes," *Operations Research*, 54, 2.
[47] Broadie, M., and Cao, M., 2008, "Improved Lower and Upper Bound Algorithms for Pricing American Options by Simulation," *Quantitative Finance*, 8, 8.
[48] Broadie, M., and Yamamoto, Y., 2003, "Application of the Fast Gauss Transform to Option Pricing," *Management Science*, 49, 8.
[49] Brooks, Chris, 2002, *Introductory Econometrics for Finance*, Cambridge University Press.
[50] Bru, M. F., 1991, "Wishart Processes," *Journal of Theoretical Probability*, 4.
[51] Caflisch, R., 1998, "Monte Carlo and Quasi-Monte Carlo Methods," *Acta Numerica*, 1–49.
[52] Caflisch, R., Morokoff, W., and Owen, A., 1999, "Valuation of Mortgage-Backed Securities Using Brownian Bridges to Reduce Effective Dimension," in B. Dupire (ed.), *Monte Carlo Simulation in Finance*, Risk Publications.
[53] Cai, Ning, and Kou, S. G., 2011, "Pricing Asian Options Under a Hyper-Exponential Jump Diffusion Model," available at http://www.columbia.edu/∼sk75/asianOR.pdf.
[54] Caporin, Massimiliano, and McAleer, Michael, 2013, "Ten Things You Should Know About the Dynamic Conditional Correlation Representation," available at http://eprints.ucm.es/21803/1/1320.pdf.
[55] Capriotti, Luca, 2008, "Reducing the Variance of Likelihood Ratio Greeks in Monte Carlo," available at http://www.luca-capriotti.net/pdfs/Finance/LSIS_lrm.pdf.
[56] Carmona, René, and Coulon, Michael, 2012, "A Survey of Commodity Markets and Structural Models for Electricity Prices," available at http://orfe.princeton.edu/rtg/ fmsummer/sites/orfe.princeton.edu.rtg.fmsummer/files/CarmonaCoulon_Survey_June2012_v2.pdf.
[57] Carmona, René, and Durrleman, Valdo, 2003, "Pricing and Hedging Spread Options," *SIAM Review*, 45, 4.
[58] Carmona, René, and Durrleman, Valdo, 2005, "Generalizing the Black-Scholes Formula to Multivariate Contingent Claims," *Journal of Computational Finance*, 9.

[59] Carmona, René, and Ludkovski, Michael, 2003, "Spot Convenience Yield Models for the Energy Markets," available at http://orfe.princeton.edu/∼rcarmona/download/fe/convenienceyield.pdf.
[60] Carmona, René, and Ludkovski, Michael, 2008, "Pricing Asset Scheduling Flexibility Using Optimal Switching," *Applied Mathematical Finance*, 15, 6.
[61] Carr, Peter, 2002, *FAQ's in Option Pricing Theory*, available at http://www.math.nyu.edu/research/carrp/papers/pdf/faq2.pdf.
[62] Carr, Peter, Cherny, Alexander, and Urusov, Mikhail, 2007, "On the Martingale Property of Time-Homogeneous Diffusions," available at http://homepage.alice.de/murusov/papers/07ccu-mart.pdf.
[63] Carr, Peter, Geman, Hélyette, Madan, Dilip, and Yor, Marc, 2002, "The Fine Structure of Asset Returns: An Empirical Investigation," *Journal of Business*, 75, 2.
[64] Carr, Peter, Geman, Hélyette, Madan, Dilip, and Yor, Marc, 2003, "Stochastic Volatility for Lévy Processes," *Mathematical Finance*, 13, 3.
[65] Carr, Peter, and Jarrow, Robert, 1990, "The Stop-Loss Start-Gain Strategy and Option Valuation," *Review of Financial Studies*, 3, 3.
[66] Carr, Peter, and Madan, Dilip, 1999, "Option Pricing and the Fast Fourier Transform," *Journal of Computational Finance*, 2, 4.
[67] Carr, Peter, and Schrder, M., 2004, "Bessel Processes, the Integral of Geometric Brownian Motion, and Asian Options," *SIAM Theory of Probability and Its Applications*, 48, 3.
[68] Carr, Peter, and Wu, Liuren, 2004, "Time-Changed Lévy Processes and Option Pricing," *Journal of Financial Economics*, 17, 1.
[69] Carrier, George F., Krook, Max, and Pearson, Carl E., 1966, *Functions of a Complex Variable: Theory and Technique*, New York: McGraw-Hill.
[70] Carriere, J. F., 1996, "Valuation of the Early-Exercise Price for Options Using Simulations and Nonparametric Regression," *Insurance: Mathematics and Economics*, 19.
[71] Chacko, George, and Viceira, Luis M., 2003, "Spectral GMM Estimation of Continuous-Time Processes," *Journal of Econometrics*, 116.
[72] Cheng, Peng, and Scaillet, Olivier, 2007, "Linear-Quadratic Jump-Diffusion Modeling," *Mathematical Finance*, 17, 4.
[73] Chi, Hongmei, Beerli, Peter, Evans, Deidre W., and Mascagni, Michael, 2005, "On the Scrambled Sobol' Sequence," in V. S. Sunderam *et al.* (eds.), *Lecture Notes in Computer Science 3516*, Berlin: Springer.
[74] Chourdakis, Kyriakos, 2005a, "Option Pricing Using the Fractional FFT," available at http://faculty.baruch.cuny.edu/lwu/890/chourdakisfrft_jcf2005.pdf.
[75] Chourdakis, Kyriakos, 2005b, "Switching Lévy Models in Continuous Time: Finite Distributions and Option Pricing," available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=838924.
[76] Cochrane, John H., 1997, "Time Series for Macroeconomics and Finance," available at http://faculty.chicagobooth.edu/john.cochrane/research/papers/time_series_book.pdf.
[77] Cools, Ronald, 2002, "Advances in Multidimensional Integration," *Journal of Computational and Applied Mathematics*, 149.

[78] Crosby, John, Le Saux, Nolwenn, and Mijatović, Aleksandar, 2010, "Approximating Lévy Processes with a View to Option Pricing," *International Journal of Theoretical and Applied Finance*, 13.
[79] Curran, Michael, 1992, "Beyond Average Intelligence," *Risk*, 5, 10.
[80] Czado, C., 2010, "Pair Copula Constructions of Multivariate Copulas," in F. Durante *et al.* (eds.), *Workshop on Copula Theory and Applications*, New York: Springer.
[81] da Fonseca, José, Grasselli, Martino, and Tebaldi, Claudio, 2007, "Option Pricing When Correlations are Stochastic: An Analytical Framework," *Review of Derivatives Research*, 10, 2.
[82] da Fonseca, José, Grasselli, Martino, and Tebaldi, Claudio, 2008, "A Multifactor Volatility Heston Model," *Quantitative Finance*, 8, 6.
[83] Dahlquist, Germund, and Bjrck, Åke, 1974, *Numerical Methods*, Englewood Cliffs: Prentice-Hall.
[84] Davis, Mark H. A., 1998, "Option Pricing in Incomplete Markets," in M. A. H. Dempster and S. R. Pliska (eds.), *Mathematics of Derivative Securities*, Cambridge: Cambridge University Press.
[85] Davis, Mark H. A., 2001, "Mathematics of Financial Markets," in B. Engquist and W. Schmid (eds.), *Mathematics Unlimited: 2001 and Beyond*, Berlin: Springer-Verlag.
[86] Davis, Mark H. A., 2004, "Valuation, Hedging, and Investment in Incomplete Financial Markets," in J. M. Hill and R. Moore (eds.), *Applied Mathematics Entering the 21st Century*, Philadelphia: Society for Industrial and Applied Mathematics.
[87] Davis, Mark H. A., 2005, "Martingale Representation and All That," in E. H. Abed (ed.), *Advances in Control, Communication Networks, and Transportation Systems: In Honor of Pravin Varaiya*, New York: Birkhauser.
[88] Davydov, D., and Linetsky, V., 2003, "Pricing Options on Scalar Diffusions: An Eigenfunction Expansion Approach," *Operations Research*, 51.
[89] Deelstra, Griselda, and Petkovic, Alexandre, 2010, "How They Can Jump Together: Multivariate Lévy Processes and Option Pricing," available at http://homepages.ulb.ac.be/~grdeelst/DP.pdf.
[90] Delatte, Anne-Laure, and Lopez, Claude, 2012, "Commodity and Equity Markets: Some Stylized Facts from a Copula Approach," available at http://mpra.ub.uni-muenchen.de/39860/.
[91] Dempster, M. A. H., Eswaran, A., and Richards, D. G., 2000, "Wavelet Methods in PDE Valuation of Financial Derivatives," Judge Institute of Management Working Paper 31.
[92] Dempster, M. A. H., and Hong, S. S. G., 2000, "Spread Option Valuation and the Fast Fourier Transform," Judge Institute of Management Working Paper 26.
[93] Dempster, M. A. H., Medova, Elena, ad Tang, Ke, 2008, "Long Term Spread Option Valuation and Hedging," *Journal of Banking and Finance*, 32.
[94] Deng, Shijie, 1998, "Stochastic Models of Energy Commodity Prices and Their Applications: Mean-Reversion with Spikes and Jumps," available at http://www.ucei.berkeley.edu/PDF/pwp073.pdf.
[95] De Wiart, B. Carton, and Dempster, M. A. H., 2011, "Wavelet Optimized Valuation of Financial Derivatives," *International Journal of Theoretical and Applied Finance*, 14, 7.

[96] Doucet, Arnaud, and Johansen, Adam, 2008, "A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later," available at https://www.seas.harvard.edu/courses/cs281/papers/doucet-johansen.pdf.
[97] Duan, Jin-Chuan, 1995, "The GARCH Option Pricing Models," *Mathematical Finance*, 5, 1.
[98] Duffie, Darrell, Pan, Jun, and Singleton, Ken, 2000, "Transform Analysis and Asset Pricing for Affine Jump Diffusions, *Econometrica*, 68.
[99] Eberlein, Ernst, and Papapantoleon, Antonis, 2005a, "Equivalence of Floating and Fixed Strike Asian and Lookback Options," available at http://www.stochastik.uni-freiburg.de/∼eberlein/papers/equivalence.pdf.
[100] Eberlein, Ernst, and Papapantoleon, Antonis, 2005b, "Symmetries and Pricing of Exotic Options in Levy Models," available at http://www.stochastik.uni-freiburg.de/∼eberlein/papers/survey-paper.pdf.
[101] El-Bachir, Naoufel, 2008, "Dependent Jump Processes with Coupled Lévy Measures," ICMA Centre Discussion Papers in Finance DP2008-3.
[102] Embrechts, Paul, Lindskog, Filip, and MacNeil, Alexander, 2003, "Modeling Dependence with Copulas and Applications to Risk Management," in S. Rachevn(ed.), *Handbook of Heavy Tailed Distributions in Finance*, Amsterdam: Elsevier.
[103] Embrechts, Paul, McNeil, Alexander, and Straumann, Daniel, 2002, "Correlation and Dependence in Risk Management: Properties and Pitfalls," in M. A. H. Dempster (ed.), *Risk Management: Value at Risk and Beyond*, Cambridge: Cambridge University Press.
[104] Ericsson, Neil R., and MacKinnon, James G., 1999, "Distributions of Error Correction Tests for Cointegration," Board of Governors of the Federal Reserve System International Finance Discussion Papers 655.
[105] Etheridge, Alison, 2002, *A Course in Financial Calculus*, Cambridge: Cambridge University Press.
[106] Eydeland, Alexander, 1994, "A Fast Algorithm for Computing Integrals in Functions Spaces: Financial Applications," *Computational Economics*, 7.
[107] Eydeland, Alexander, 1996, "A Spectral Algorithm for Pricing Interest Rate Options," *Computational Economics*, 9.
[108] Eydeland, Alexander, and Mahoney, Dan, 2002, "An Efficient and Accurate Computational Technique for Dynamic Programming with Markov Processes," in Ehud I. Ronn (ed.), *Real Options and Energy Management*, London: Risk Books.
[109] Eydeland, Alexander, and Mahoney, Dan, 2003, "A Fast Convolution Method for Option Pricing," Mirant Technical Report.
[110] Eydeland, Alexander, and Wolyniec, Krzysztof, 2003, *Energy and Power Risk Management*, Hoboken: John Wiley and Sons.
[111] Fang, K.-T., Kotz, S., and Ng, K.-W., 1987, *Symmetric Multivariate and Related Distributions*, London: Chapman & Hall.
[112] Filipović, Damir, and Mayerhofer, Eberhard, 2009, "Affine Diffusion Processes: Theory and Applications," available at http://arxiv.org/pdf/0901.4003.
[113] Föllmer, Hans, and Schweizer, Martin, 2010, "The Minimal Martingale Measure," in R. Cont (ed.), *Encyclopedia of Quantitative Finance*, Wiley.
[114] Frahm, Gabriel, and Jaekel, Uwe, 2007, "Tyler's M-Estimator, Random Matrix Theory, and Generalized Elliptical Distributions with Applications to Finance," available at http://www.econstor.eu/bitstream/10419/26740/1/527784575.PDF.

[115] Frahm, Gabriel, and Jaekel, Uwe, 2008, "Random Matrix Theory and Robust Covariance Matrix Estimation for Financial Data," available at http://arxiv.org/pdf/physics/0503007.pdf.
[116] Frees, Edward W., and Valdez, Emiliano A., 1998, "Understanding Relationships Using Copulas," *North American Actuarial Journal*, 3, 1.
[117] Fulop, Andras, 2011, "Filtering Methods," in Jin-Chuan Duan, James E. Gentle, and Wolfgang Haerdle (eds.), *Handbook of Computational Finance*, Berlin: Springer-Verlag.
[118] Gallant, A. Ronald, and Tauchen, George, 2010, "EMM: A Program for Efficient Method of Moments Estimation, Version 2.6, User's Guide," available at http://aronaldg.org/courses/compecon/.
[119] Gandy, Axel, and Veraart, Luitgard A. M., 2013, "The Effect of Estimation in High-Dimensional Portfolios," *Mathematical Finance*, 23, 3.
[120] Geman, Hélyette, 2002, "Pure Jump Levy Processes for Asset Price Modeling," *Journal of Banking and Finance*, July.
[121] Geman, Hélyette, and Yor, Marc, 1993, "Bessel Processes, Asian Options, and Perpetuities," *Mathematical Finance*, 3.
[122] Geman, Hélyette, and Eydeland, Alexander, 1995, "Domino Effect," *Risk*, 8.
[123] Geman, Hélyette (ed.), 2009, *Risk Management in Commodity Markets: From Shipping to Agriculturals and Energy*, Chichester: Wiley.
[124] Genest, Christian, and Rémillard, Bruno, 2006, "Discussion of "Copulas: Tales and Facts," by Thomas Mikosch," available at http://brunoremillard.com/Papers/mikosch-response.pdf.
[125] Genest, Christian, and Favre, Anne-Catherine, 2007, "Everything You Always Wanted to Know About Copula Modeling but Were Afraid to Ask," available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.160.8266.
[126] Ghiuvea, Cristian S., Lehoczky, John P., and Seppi, Duane, 2001, "Pricing of Generalized American Options with Applications to Real and Financial Energy Derivatives," Working Paper, Carnegie Mellon University.
[127] Gibson, R., and Schwartz, E. S., 1990, "Stochastic Convenience Yield and the Pricing of Oil Contingent Claims," *Journal of Finance*, 45.
[128] Glasserman, Paul, 2004, *Monte Carlo Methods in Financial Engineering*, New York: Springer-Verlag.
[129] Glasserman, Paul, and Yu, Bin, 2004, "Simulation for American Options: Regression Now or Regression Later?," in Harald Niederreiter (ed.), *Monte Carlo and Quasi-Monte Carlo Methods 2002*, Berlin: Springer.
[130] Gouriéroux, Christian, and Sufana, Razvan, 2003, "Wishart Quadratic Term Structure Models," available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=757307.
[131] Grégoire, Vincent, Genest, Christian, and Gendron, Michel, 2008, "Using Copulas to Model Price Dependence in Energy Markets," *Energy Risk*, March 2008.
[132] Grzywacz, Piotr, and Wolyniec, Krzysztof, 2011, "Mutli-Scale Volatility in Commodity Markets," *Energy Risk*, August 2011.
[133] Gyurkó, L. G., Hambly, B. M., and Witte, J. H., 2011, "Monte Carlo Methods via a Dual Approach for Some Discrete Time Stochastic Control Problems," http://people.maths.ox.ac.uk/hambly/PDF/Papers/dualmc.pdf.

[134] Gurrieri, Sebastien, 2011, "An Analysis of Sobol Sequence and the Brownian Bridge," available at http://ssrn.com/abstract=1951886.
[135] Hamada, Mahmoud, and Valdez, Emiliano A., 2004, "CAPM and Option Pricing with Elliptical Distributions," Quantitative Finance Research Centre Paper 120.
[136] Hamilton, James D., 1994, *Time Series Analysis*, Princeton: Princeton University Press.
[137] Hansen, Lars Peter, 1982, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 4.
[138] Hansen, Lars Peter, 2007, "Generalized Methods of Moments Estimation," *Palgrave Dictionary of Economics*, June 2007.
[139] Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed.*, New York: Springer.
[140] Haugh, Martin B., and Kogan, Leonid, 2004, "Pricing American Options: A Duality Approach," *Operations Research*, 52, 2.
[141] Haugh, Martin B., and Kogan, Leonid, 2008, "Duality Theory and Approximate Dynamic Programming for Pricing American Options and Portfolio Optimization," in J. R. Birge and V. Linetsy (eds.), *Handbooks in OR & MS, Vol. 15*, Amsterdam: North Holland.
[142] Heiss, Florian, and Winschel, Viktor, 2006, "Estimation with Numerical Integration on Sparse Grids," Munich Discussion Paper No. 2006-15.
[143] Henderson, Vicky, and Wojakowski, Rafal, 2002, "On the Equivalence of Floating and Fixed Strike Asian Options," *Journal of Applied Probability*, 39.
[144] Heston, Steven L., 1993, "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options," *The Review of Financial Studies*, 6, 2.
[145] Heston, Steven L., and Nandi, Saikat, 2000, "A Closed-Form GARCH Option Valuation Model," *The Review of Financial Studies*, 13, 3.
[146] Hikspoors, Samuel, and Jaimungal, Sebastian, 2007, "Energy Spot Price Models and Spread Options Pricing," *International Journal of Theoretical and Applied Finance*, 10, 7.
[147] Hinch, E. J., 1991, *Perturbation Methods*, Cambridge: Cambridge University Press.
[148] Holtz, Markus, 2011, *Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance*, Berlin: Springer.
[149] Hull, John, 2005, *Options, Futures, and other Derivatives, 6th Ed.*, Upper Saddle River: Prentice Hall.
[150] Hurd, T. R., and Zhou, Zhuowei, 2010, "A Fourier Transform Method for Spread Option Pricing," *SIAM J. Financial Math.*, 1, 1.
[151] Huang, Shirley J., and Yu, Jun, 2007, "On Stiffness in Affine Pricing Models," *Journal of Computational Finance*, 10, 3.
[152] Jaillet, Patrick, Ronn, Ehud I., and Tompaidis, Stathis, 2004, "Valuation of Commodity-Based Swing Options," *Management Science*, 50, 7.
[153] Javaheri, Alireza, Lautier, Delphine, and Galli, Alain, 2003, "Filtering in Finance," *Wilmott Magazine*.
[154] Jiang, George J., and Knight, John L., 2002, "Estimation of Continuous Time Processes via the Empirical Characteristic Function," *Journal of Business and Economic Statistics*, 20, 2.

[155] Joe, S., and Kuo, F. Y., 2003, "Remark on Algorithm 659: Implementing Sobol's Quasirandom Sequence Generator," *ACM Transactions on Mathematical Software*, 29.
[156] Joe, S., and Kuo, F. Y., 2008a, "Constructing Sobol Sequences with Better Two-Dimensional Projections," *SIAM Journal on Scientific Computing*, 30.
[157] Joe, S., and Kuo, F. Y., 2008b, "Notes on Generating Sobol Sequences," available at http://web.maths.unsw.edu.au/~fkuo/sobol/index.html.
[158] Johannes, Michael, and Polson, Nicholas, 2003, "MCMC Methods for Continuous-Time Financial Econometrics," available at http://home.uchicago.edu/~lhansen/JP_handbook.pdf.
[159] Johansen, S., 1988, "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12.
[160] Johansen, S., 1991, "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models". *Econometrica*, 59.
[161] Joshi, M. S., 2003, *The Concepts and Practice of Mathematical Finance*, Cambridge: Cambridge University Press.
[162] Kahl, C., and Jäckel, P., 2005, "Not-So Complex Logarithms in the Heston Model," *Wilmott*, September 2005.
[163] Kall, Peter, and Wallace, Stein W., 1994, *Stochastic Programming, 2nd Ed.*, Chichester: Wiley.
[164] Kallsen, J., and Tankov, P., 2006, "Characterization of Dependence of Multidimensional Lévy Processes Using Lévy Copulas," *Journal of Multivariate Analysis*, 97.
[165] Kalman, R. E., 1960, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering* 82, 1.
[166] Karatzas, Ioannis, and Shreve, Steven E., 1991, *Brownian Motion and Stochastic Calculus*, New York: Springer-Verlag.
[167] Kettler, Paul, 2006, "Lévy Copula-Driven Financial Processes," available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.170.3968.
[168] Kirk, E., 1996, "Correlation in Energy Markets," in V. Kaminski (ed.), *Managing Energy Price Risk*, pp. 71–78, London: Risk Publications.
[169] Kohler Michael, 2010, "A Review on Regression-Based Monte Carlo Methods for Pricing American Options," in L. Devroye *et al.* (eds.), *Recent Developments in Applied Probability and Statistics*, Berlin: Springer.
[170] Kwok, Y. K., 1998, *Mathematical Models of Financial Derivatives*, Berlin: Springer.
[171] Kyprianou, Andreas, 2006, *Introductory Lectures on Fluctuations of Levy Processes with Applications*, Berlin: Springer-Verlag.
[172] Laloux, Laurent, Cizeau, Pierre, Bouchaud, Jean-Philippe, and Potters, Marc, 1999, "Noise Dressing of Financial Correlation Matrices," *Physical Review Letters*, 83, 7.
[173] Landsman, Zinoviy M., and Valdez, Emiliano A., 2003, "Tail Conditional Expectations for Elliptical Distributions," *North American Actuarial Journal*, 7, 4.
[174] Ledoit, Olivier, and Wolf, Michael, 2014, "Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks," available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2383361.
[175] Lewis, Alan, 2001, "A Simple Option Formula for General Jump-Diffusion and Other Exponential Levy Processes," available at http://www.optioncity.net/pubs/ExpLevy.pdf.

[176] Leippold, Markus, and Trojani, Fabio, 2010, "Asset Pricing with Matrix Jump Diffusions," available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1274482.
[177] Li, Minqiang, Deng, Shijie, and Zhou, Jieyun, 2008, "Multi-Asset Spread Option Pricing and Hedging," available at http://mpra.ub.uni-muenchen.de/8259/.
[178] Liebscher, Eckhard, 2008, "Construction of Asymmetric Multivariate Copulas," *Journal of Multivariate Analysis*, 99.
[179] Lindskog, Filip, McNeil, Alexander, and Schmock, Uwe, 2003, "Kendall's Tau for Elliptical Distributions," in Georg Bol *et al.* (eds.), *Credit Risk: Measurement, Evaluation and Management*, Heidelberg: Physica-Verlag.
[180] López de Prado, Marcos M., and Leinweber, David, 2012, "Advances in Cointegration and Subset Correlation Hedging Methods," *The Journal of Investment Strategies*, 1, 2.
[181] Longstaff, F., and Schwartz, E., 2001, "Valuing American Options by Simulation: A Least Squares Approach," *Review of Financial Studies*, 14.
[182] Lord, R., Fang, F., Bervoets, F., and Oosterlee, C. W., 2008, "A Fast and Accurate FFT-Based Method for Pricing Early-Exercise Options Under Lévy Processes," *SIAM Journal on Scientific Computing*, 30, 4.
[183] Lord, R., and Kahl, C., 2007, "Optimal Fourier Inversion in Semi-Analytical Option Pricing," *Journal of Computational Finance*, 10, 4.
[184] Lord, R., and Kahl, C., 2008, "Complex Logarithms in Heston-Like Models," available at http://www2.math.uni-wuppertal.de/~kahl/publications/complexlogarithmsheston.pdf.
[185] Longstaff, Francis A., and Schwartz, Eduardo S., 2001, "Valuing American Options by Simulation: A Simple Least-Squares Approach," *Review of Financial Studies*, 14, 1.
[186] Lucic, Vladimir, 2012, "Correlation Skew via Product Copula," available at http://www.cass.city.ac.uk/__data/assets/pdf_file/0006/154923/Correlation-Skew-via-Product-Copula.pdf.
[187] Ludkovski, Michael, 2008, "Financial Hedging of Operational Flexibility," *International Journal of Theoretical and Applied Finance*, 11, 8.
[188] Ludkovski, Michael, and Carmona, René, 2010, "Valuation of Energy Storage: An Optimal Switching Approach," *Quantitative Finance*, 10, 4.
[189] MacKinnon, James G., Haug, Alfred A., and Michelis, Leo, 1999, "Numerical Distribution Functions of Likelihood Ratio Tests for Cointegration," *Journal of Applied Econometrics*, 14, 5.
[190] MacKinnon, James G., 2006, "Bootstrap Methods in Econometrics," *The Economic Record*, 82.
[191] Madan, Dilip, Carr, Peter, and Chang, Eric, 1998, "The Variance Gamma Process and Option Pricing," *European Financial Review*, 2, 1.
[192] Mahoney, Dan, 2015a, "Minimal Martingale Results for Jump Processes," working paper.
[193] Mahoney, Dan, 2015b, "Cointegration and Variance Scaling Laws," working paper.
[194] Mahoney, Dan, and Wolyniec, Krzysztof, 2012, "Valuation of Spread Commodity Structures in Cointegrated Futures Markets," *Energy Risk*, February 2012.
[195] Mallory, Mindy L., and Lence, Sergio H., 2012, "Testing for Cointegration in the Presence of Moving Average Errors" *Journal of Time Series Econometrics*, 4, 2.

[196] Margrabe, William, 1978, "The Value of an Option to Exchange One Asset for Another," *Journal of Finance*, 33, 1.
[197] McAleer, Michael, and Medeiros, Marcelo C., 2008, "Realized Volatility: A Review," *Econometric Reviews*, 27.
[198] McLean, Bethany, and Elkind, Peter, 2004, *The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron*, New York: Portfolio Trade.
[199] Meinshausen, N., and Hambly, B. M., 2004, "Monte Carlo Methods for the Valuation of Multiple Exercise Options," *Mathematical Finance*, 14.
[200] Merton, Robert C., 1990, *Continuous-Time Finance*, Malden: Blackwell.
[201] Meyer-Brandis, Thilo, and Morgan, Michael, 2014, "A Dynamic Lévy Copula Model for the Spark Spread," in Fred Espen Benth (ed.), *Quantitative Energy Finance*, New York: Springer.
[202] Mikosch, T., 2005, "Copulas: Tales and Facts," available at www.math.ku.dk/∼mikosch/Preprint/Copula/s.pdf.
[203] Moler, Cleve, and van Loan, Charles, 2003, "Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later," *SIAM Review*, 45, 1.
[204] Monroe, I., 1978, "Processes That Can Be Embedded in Brownian Motion," *Annals of Probability*, 6.
[205] Nelsen, Roger B., 1999, *An Introduction to Copulas*, New York: Springer-Verlag.
[206] Newey, Whitney K., and Steigerwald, Douglas G., 1997, "Asymptotic Bias for Quasi-Maximum Likelihood Estimators in Conditional Heteroskedasticity Models," *Econometrica*, 65, 3.
[207] Niederreiter, H., 1988, "Low-Discrepancy and Low-Dispersion Sequences," *Journal of Number Theory*, 30.
[208] Nielsen, Lars B., 2001, "Pricing Asian Options," Masters Thesis, Aarhus University (Denmark).
[209] Nocedal, Jorge, and Wright, Stephen J., 2006, *Numerical Optimization, 2nd Ed.*, New York: Springer.
[210] Nualart, David, 2006, "Fractional Brownian Motion: Stochastic Calculus and Applications," available at http://www.icm2006.org/proceedings/vol3.html.
[211] Papageorgiou, A., 2002, "The Brownian Bridge Does Not Offer a Consistent Advantage in Quasi-Monte Carlo Integration," *Journal of Complexity*, 18, 1.
[212] Papapantoleon, Antonis, 2008, "An Introduction to Levy Processes with Applications in Finance," available at http://page.math.tu-berlin.de/∼papapan/papers/introduction.pdf.
[213] Park, Stephen K., and Miller, Keith W., 1988, "Random Number Generators: Good Ones are Hard to Find," *Communications of the ACM*, 31, 10.
[214] Parsons, Cliff, 2008, "Explaining Bias in Mean-Reversion Speed Estimates for Energy Prices," *Energy Risk*, July 2008.
[215] Patton, A. J., 2006, "Modeling Asymmetric Exchange Rate Dependence" *International Economic Review*, 47.
[216] Pearson, Neil, D., 1995, "An Efficient Approach for Pricing Spread Options," *Journal of Derivatives*, 3.
[217] Pesaran, M. H., Shin, Y., and Smith, R. J., 2000, "Structural Analysis of Vector Error Correction Models with Exogenous I(1) Variables," *Journal of Econometrics*, 97, 2.
[218] Pfaffel, Oliver, "Wishart Processes," 2012, available at http://arxiv.org/pdf/1201.3256v1.pdf.

[219] Pham, Huyên, 2010, "Lectures on Stochastic Control and Applications in Finance," available at https://sites.google.com/site/phamxuanhuyen/.
[220] Potters, M., Bouchaud, J. P., and Laloux, L., 2005, "Financial Applications of Random Matrix Theory: Old Laces and New Pieces," available at http://arxiv.org/abs/physics/0507111v1.
[221] Poulsen, Rolf, Schenk-Hoppé, Klaus Reiner, and Ewald, Christian-Oliver, 2009, "Risk Minimization in Stochastic Volatility Models: Model Risk and Empirical Performance," available at http://www.math.ku.dk/∼rolf/Klaus/pse.pdf.
[222] Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., 2007, *Numerical Recipes 3rd Edition*, Cambridge: Cambridge University Press.
[223] Rebonato, Riccardo, and Jäckel, Peter, 1999, "The Most General Methodology to Create a Valid Correlation Matrix for Risk Management and Option Pricing Purposes," available at http://www.quarchome.org/correlationmatrix.pdf.
[224] Rebonato, Riccardo, 2004, *Volatility and Correlation, 2nd Ed.*, Chichester: Wiley.
[225] Rockinger, Michael, and Semenova, Maria, 2005, "Estimation of Jump-Diffusion Processes via Empirical Characteristic Functions," FAME Research Paper no. 150.
[226] Rogers, L. C. G., 2002, "Monte Carlo Valuation of American Options," *Mathematical Finance*, 12, 3.
[227] Rogers, L. C. G., 2003, "Duality in Constrained Optimal Investment and Consumption Problems: A Synthesis, in *Paris-Princeton Lectures on Mathematical Finance 2002*, Berlin: Springer.
[228] Rogers, L. C. G., 2007, "Pathwise Stochastic Optimal Control," *SIAM Journal on Control and Optimization*, 46.
[229] Rosinksi, Jan, 2001, "Series Representations of Lévy Processes from the Perspective of Point Processes," in O. E. Barndorff-Nielsen (ed.), *Lévy Processes – Theory and Applications*, Boston: Birkhauser.
[230] Schmidt, T., 2007, "Coping with Copulas," in J. Rank (ed.), *Copulas: From Theory to Applications in Finance*, London: Risk Books.
[231] Schoenmakers, John, Zhang, Jianing, Huang, Junbo, 2012, "Optimal Dual Martingales, Their Analysis and Application to New Algorithms for Bermudan Products, available at http://www.wias-berlin.de/people/schoenma/SchoenZhangHuangAcc.pdf.
[232] Schöttle, K., and Werner, R., 2004, "Improving the 'Most General Methodology to Create a Valid Correlation Matrix'," available at http://www.risklab.de/Dokumente/Aufsaetze/Schoettle,Werner[04]-ImprovingTheMostGeneralMethodologyToCreateAValidCorrelationMatrix.pdf.
[233] Schroder, Mark, 1999, "Changes of Numeraire for Pricing Futures, Forwards, and Options," *The Review of Financial Studies*, 12, 5.
[234] Shreve, Steven, E., 2004a, *Stochastic Calculus for Finance I: The Binomial Asset Pricing Model*, New York: Springer-Verlag.
[235] Shreve, Steven, E., 2004b, *Stochastic Calculus for Finance II: Continuous-Time Models*, New York: Springer-Verlag.
[236] Skipper, Max, and Buchen, Peter, 2003, "The Quintessential Option Pricing Formula," available at http://www.maths.usyd.edu.au/u/pubs/publist/preprints/2003/skipper-22.html.
[237] Schwartz, Eduardo S.,1997, "The Stochastic Behavior of Commodity Prices: Implications for Valuation and Hedging," *Journal of Finance*, 52, 3.

[238] Schwartz, Eduardo S., and Smith, James E., 2000, "Short-Term Variations and Long-Term Dynamics in Commodity Prices," *Management Science*, 46, 7.
[239] Singleton, Kenneth J., 2001, "Estimation of Affine Asset Pricing Models Using the Empirical Characteristic Function" *Journal of Econometrics*, 102.
[240] Smith, Tony, 2008, "Indirect Inference," in *The New Palgrave Dictionary of Economics, 2nd Ed.*, London: Palgrave MacMillan.
[241] Smolyak, S., 1963, "Quadrature and Interpolation Formulas for Tensor Products of Certain Classes of Functions," *Dokl. Akad. Nauk SSSR*, 4.
[242] Stentoft, Lars, 2012, "Value Function Approximation or Stopping Time Approximation: A Comparison of Two Recent Numerical Methods for American Option Pricing Using Simulation and Regression," available at http://ssrn.com/abstract=1315306.
[243] Swindle, Glen, 2014, *Valuation and Risk Management in Energy Markets*, New York: Cambridge University Press.
[244] Tankov, Peter, 2003, "Dependence Structure of Multivariate Lévy Processes with Applications in Risk Management," available at http://www.cmap.polytechnique.fr/preprint/comment.php?showdetails=1%5C&paper=502.
[245] Tankov, Peter, 2007, "Lévy Processes in Finance and Risk Management," *Wilmott Magazine*, Sep-Oct 2007.
[246] Tankov, Peter, and Cont, Rama, 2003, *Financial Modeling with Jump Processes*, Boca Raton: Chapman & Hall/CRC Financial Mathematics Series.
[247] Thompson, Matt, Davison, Matt, and Rasmussen, Henning, 2009, "Natural Gas Storage Valuation and Optimization: A Real Options Application," *Naval Research Logistics*, 56, 3.
[248] Trolle, Anders B., and Schwartz, Eduardo S., 2009, "Unspanned Stochastic Volatility and the Pricing of Commodity Derivatives," *The Review of Financial Studies*, 22, 11.
[249] Tsitsiklis, J., and van Roy, B., 2001, "Regression Methods for Pricing Complex American-Style Options," *IEEE Transactions on Neural Networks*, 12, 4.
[250] Venkatramanan, Aanand, and Alexander, Carol, 2011, "Closed Form Approximations for Spread Options," *Applied Mathematical Finance*, 18, 5.
[251] Villar, Jose A., and Joutz, Frederick L., 2006, "The Relationship Between Crude Oil and Natural Gas Prices," Energy Information Administration, Office of Oil and Natural Gas, October 2006.
[252] Wan, Eric A., and van der Merwe, Rudolph, 2001, "The Unscented Kalman Filter," in Simon Haykin (ed.), *Kalman Filtering and Neural Networks*, New York: John Wiley and Sons.
[253] Watson, M. W., 1994, "Vector Autoregressions and Cointegration," in R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics, Volume IV*, Amsterdam: North-Holland..
[254] Wasilkowski, G., and Woźniakowski, H., 1995, "Explicit Cost Bounds of Algorithms for Multivariate Tensor Product Problems," *J. Complexity*, 11.
[255] Welch, Greg, and Bishop, Gary, 2006, "An Introduction to the Kalman Filter," available at http://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf.
[256] West, Graeme, 2004, "Better Approximations to Cumulative Normal Functions," available at https://lyle.smu.edu/~aleskovs/emis/sqc2/accuratecumnorm.pdf.
[257] Wilmott, Paul, 2000, *Paul Wilmott on Quantitative Finance*, Chichester: Wiley.

[258] Wilmott, Paul, Howison, Sam, and DeWynne, Jeff, 1997, *The Mathematics of Financial Derivatives*, Cambridge: Cambridge University Press.
[259] Wolyniec, Krzysztof, 2015, *Quantitative Methods in Commodity Markets*, New York: John Wiley and Sons.
[260] Wu, Liuren, 2008, "Modeling Financial Security Returns Using Lévy Processes," in J. Birge and V. Linetsky (eds.), *Handbooks in Operations Research and Management Science: Financial Engineering, 15*, Elsevier.
[261] Yu, Jun, 2009, "Bias in the Estimation of the Mean Reversion Parameter in Continuous Time Models," SMU Economics & Statistics Working Paper No. 16-2009.
[262] Yu, Jun, 2014, "Econometric Analysis of Continuous Time Models: A Survey of Peter Phillips' Work and Some New Results," *Econometric Theory*, 30, 4.
[263] Zhou, Hao, 2001, "Finite Sample Properties of EMM, GMM, QMLE, and MLE for a Square-Root Interest Rate Diffusion Model," available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.8099.
[264] Zivot, Eric, and Wang, Jiahui, 2006, *Modeling Financial Time Series with S-PLUS®*, *2nd Ed.*, New York: Springer-Verlag.

# Index