Otto J.W.F. Kardaun

Classical Methods of Statistics

Springer

Classical Methods of Statistics

Otto J. W. F. Kardaun

Classical Methods of Statistics

With Applications in Fusion-Oriented Plasma Physics

Springer

Dr. Otto J. W. F. Kardaun
MPI für Plasmaphysik (IPP)
Boltzmannstr. 2
85748 Garching, Germany
E-mail: ojk@ipp.mpg.de

About the author

Gymnasium-beta in Zwolle, the Netherlands, 1972. Ranked seventh in the national mathematics olympiad 1971. The author studied physics, mathematics, and philosophy (BS degrees) at Groningen University, 1972–1976, minoring in chemistry. He received his 'doctoraal' (MS) degree in theoretical physics (with a research period in plasma physics at FOM Utrecht) in 1979, and his PhD in mathematical statistics at Groningen University in 1986. Guest research periods in England (JET), Japan (JSPS, JAERI) and the USA (Courant Institute). He is research scientist at the Max Planck Institute for plasma physics in Garching, co-founding member of an internationally collaborating working group on fusion, and (co-)author of multifarious papers, some of them often cited.

Library of Congress Control Number: 2005927933

ISBN-10 3-540-21115-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-21115-0 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springeronline.com

Printed in Germany

Typesetting by the author
Data conversion: Frank Herweg, Leutershausen
Cover production: Erich Kirchner, Heidelberg

Printed on acid-free paper 55/3141/mh 5 4 3 2 1 0

両親と先生方へ

Γονεῦσι και ’Επίστημοις Προγεγένημενοις

PREFACE

This monograph presents a selection of fundamental ideas and procedures in probability theory and statistics and has been written to provide a working background for physicists involved in the interpretation of scientific data.

As a framework for empirical science, statistics is applicable in many domains of investigation, and in a broad range of sciences, varying from natural sciences (including technological applications) and life sciences (including medical applications) to economics, environmental science and social sciences. While they often explain some generally usable statistical methods and theory as well, many of the applications in the text are oriented towards physics. This is apparent from the greater part of the examples and analogies employed, and sometimes also from the kind of mathematics selected.

Elements of probability theory are presented in so far as they are needed for an appropriate understanding of our primary subject, statistical inference, with a special view to the analysis of data sets obtained from plasma-physical experiments. Occasionally the author could not resist the temptation to present (in small print) mathematical or probabilistic material not of immediate interest for an intermediate understanding of our primary subject.

The mathematical level of the text is such that it should be accessible for advanced undergraduate and graduate students in physics and mathematics, and hence also in astronomy, computer science, statistics, etc., with a thorough but modest acquaintance of mathematics. At incidental locations (especially in the first chapter) some parts of mathematics are used with which the reader may be less familiar. In most cases, these concern intrinsic remarks the understanding of which is not necessary to follow the rest of the text, and which therefore can be skipped at first reading. Since statistics is an inductive science, even if probability theory is used as its deductive underpinning, the basic difficulty in grasping (new) statistical concepts does not primarily lie in their mathematical aspects.

In the first two chapters, the text is interspersed with a number of exercises (some with an asterix to indicate an extension or an increased level of difficulty) to help the reader consolidate his understanding of the subject and to stimulate further critical reflection. A few of the exercises in Chap. 1 presuppose some elementary knowledge of basic algebraic structures, usually taught in first or second year university courses.

The selection of topics reflects, on one hand, the author's education at Groningen University and, on the other hand, the practical needs expressed, during numerous discussions, by research workers involved in the analysis of statistical data, in plasma physics as well as in medicine.

In order to give the reader an idea what not to expect in this book, we briefly mention a number of topics, related to statistical science, which are not treated here. In the chapter on probability theory, neither stochastic processes nor communication theory is included. In the statistical chapters, the

emphasis is on motivation and conveyance of the basic practical concepts without paying attention to the decision-theoretical framework, and without considering the interesting ramifications of statistical optimality theory. Not included either are more sophisticated topics (errors-in-variables models, time series analysis, sequential analysis and design of experiments, wavelet analysis, optimisation and neural networks) nor statistical simulation techniques such as Monte Carlo and bootstrap methods. Also, parameter-free methods, multiple decision procedures and censored data analysis have been omitted. The main emphasis of this monograph is indeed on classical statistical methods. For foundational issues about Bayesian probability theory and distributional inference, the reader is referred to [358,359].

Acknowledgements. The author has benefited much from and is most grateful for the varied discussions and often intensive collaboration with a number of colleagues. This has led to the different parts of the book being prepared first, in preliminary form, as internal reports, while the author worked under helmsmanship of K. Lackner in the 'Abteilung Tokamakphysik' at the Max Planck Institute for Plasma Physics (IPP) in Garching near Munich. Especially mentioned should be the collaboration and interaction with K. Itoh, S.-I. Itoh, J. Kardaun, A. Kus, P. McCarthy and K. Riedel, as well as the education received from the H–mode Database Working Group [122,360], a cooperative group between several plasma physical laboratories involved in plasma performance prediction of future devices, stimulated by the International Toroidal Experimental Reactor (ITER) project [671] and during the engineering design period effectively guided by G. Cordey and V. Mukhovatov. Also a thought-provoking interest in Bayesian statistics was, always vividly, expressed by several colleagues from the former plasma–surface physics division of IPP, afterward at the Center for Interdisciplinary Plasma Science, under guidance of V. Dose.

Chapters 1 and 2 in fact originated from exercise classes in a statistics course for students in Pharmacy at Groningen University in 1980, when the author worked under supervision of his PhD advisor W. Schaafsma. Somewhat later, it grew from introductory courses in statistics, for staff, post-docs and PhD students majoring in physics, held by the author at several occasions, being organised as a guided staff seminar, during the nineties at IPP in Garching (for the tokamak physics, surface physics, and experimental divisions). Once it was also held at a summer school of plasma physics in Santander, Spain.

The first draft of Chap. 3 on applied regression analysis was initiated by A. Kus, after he had been invited to hold a lecture during such a course at IPP. Several iterative improvements of this chapter were made subsequently and also many practical aspects for preparing this text in LaTeX were resolved in congenial collaboration with him during the period that he worked at the Wendelstein 7–AS stellarator in Garching. The material of this chapter can be viewed as an extension in several directions of the more concisely written

description of the general linear regression model in Sect. 2.4. It contains, therefore, at some places a rehearsal of important statistical arguments in the context of linear regression. Primarily for didactic reasons, this redundancy has been intentionally retained. During the revision phase of this book, notably Chaps. 1, 2, 6 and 7 have been expanded compared to their former versions in IPP report [354].

Statistical profile analysis is described in Chap. 4. The basic investigations were performed during the early nineties through intensive iterative discussion with P. McCarthy and K. Riedel, which was stimulated further by an invitation of the latter to visit the Courant Institute in New York, and resulted in [343] and IPP report [357]. The author is grateful for having much learned from the interaction with these two scientists. A practical companion of this chapter, containing a detailed analysis of experimental temperature and density profiles from the Thomson scattering YAG–laser diagnostic for Ohmic discharges at the ASDEX tokamak, has appeared in [343] and [451]. The subject of this chapter also benefited from firsthand interaction with K. Lackner.

Discriminant analysis, another topic [585] inherited from what has been termed the Groningen school of statistics in [358], is described in Chap. 5. The subject was introduced in Garching and notably stimulated by a collaboration starting with a visit of K. Itoh and S.-I. Itoh to the ASDEX tokamak division led by G. von Gierke and F. Wagner. A presentation was held at the third H–mode Workshop at the JET tokamak in Culham near Oxford. The investigations were further pursued during a working visit by the author in the group of H. Maeda and Y. Miura at the JFT–2M tokamak in Naka, Japan, and the ensuing cooperation with S.-I. Itoh (Kyushu University) and K. Itoh (National Institute for Fusion Science), which a few years later was extended by their invitation to visit the Research Institute for Applied Mechanics (RIAM) at Kyushu University. Also the practical support and the previous experience in medical discriminant analysis from J. Kardaun (Statistics Netherlands) definitely contributed to shape this chapter, which is mainly based on NIFS report [351]. The decision was made to retain in the practical part of Chap. 5 the original analysis as a case study, after having scrutinised the text for adequacy after an elapse of several years, and while making only a limited number of necessary adaptations, such as an update of the references. In particular, the temptation has been resisted to present the elaboration of a more recent, extended dataset, which may have increased direct relevance from a practical point of view, but presumably with only a limited additional value for the general reader.

A concise overview of various statistical software products has been included in Chap. 6, which is adapted to more or less the current state-of-the art. Most of the practical data sets in Chap. 7 were prepared by several colleagues from the ASDEX Upgrade Team (in particular, H.-U. Fahrbach, A. Herrmann, M. Maraschek, P. McCarthy, R. Neu, G. Pautasso, F. Ryter,

J. Stober). The data they have assembled are the result of years of successful operation of the ASDEX Upgrade tokamak, a cooperative undertaking, imprinted by experimental leadership of O. Gruber, W. Köppendörfer and M. Kaufmann. The author provided a motivating plasma-physical introduction to these data sets and formulated the exercises. One data set, corresponding to case ID in Chap. 7, contains data from several tokamaks (ASDEX, ASDEX Upgrade, DIII-D, JET, JT-60U and TFTR), and is actually a subset of a public release of the H-mode database maintained at EFDA by K. Thomsen. For the assemblage of these invaluable data, deep recognition should be given to the various individuals and tokamak teams, which is only concisely expressed in the introduction of this exercise.

The author is particularly indebted to W. Schaafsma for his critical reading and numerous comments on the first two chapters of the book, and to F. Engelmann for valuable comments on parts of Chaps. 4, 5 and 7. The author also acknowledges discussion with G. Becker. In addition, a number of editorial improvements were made through attentive reading of the text by A. Geier (from IPP) and W. Kardaun (from Maastricht) as well as by the copy-editor from Springer–Verlag. The Russian list of keywords was inspired by the language course of Mrs. Г. Арбейтер-Zehrfeld at MPI Garching and discussed with G. Pereverzev. Around 1997, several students of S.-I. Itoh assisted with composing the Japanese keyword list, which was latterly also discussed with Y. Nishimura and H. Urano. Motivated by the language course of Mrs. Tojima Herdtle in Garching, the component meanings and Unicode numbers were found with help of the Kodansha Kanji Learner's dictionary. This list was type-set on a MacIntosh computer using MS–WORD under OS X. All of the remaining text was composed using LaTeX under UNIX on IBM and SUN computers.

Finally, the author wishes to acknowledge K. Lackner for proving continual support and a genuine research environment during many years, as well as the hospitality in the tokamak physics group of S. Günter during the period that the former was accredited at the European Fusion Development Agency (EFDA) in Garching. This enabled the author to finalise the present book. It is hoped that, in spite of the occasional diversions and the many revisions that have been made (which may have added precision, albeit possibly at the cost of some directness and simplicity), the enthusiastic vein arising from (small) discoveries and from the gratifying self-education incurred during the write-up of the text will still be transparent to the reader. The author obviously feels intellectual responsibility for any remaining obscurity or error of representation, even while a legal disclaimer is mandatory, and therefore wishes to encourage readers to bring any of such deficiencies to his attention.

Garching, April 2005 Otto J.W.F. Kardaun

Contents

1 Elements of Probability Theory

..., nos, qui sequimur probabilia nec ultra id, quod veri simile occurrit, progredi possumus, et refellere sine pertinacia et refelli sine iracundia parati sumus[†]

M.T. Cicero, 106–43 BC

1.1 Introduction

Probability statements are often used in a colloquial sense, for example in expressions such as 'the probability of winning a lotto is small', 'the probability of dying from a heart disease is 40%', 'it is quite probable that it will rain tomorrow', and 'the probability of getting a 5 when you throw a die is one out of six'. In statistics and probability theory, a foundation has been given for such kind of statements. While the concept of probability has been made more precise, it was embedded into empirical experience and put into a logical framework by axiomatising the rules for reasoning with probabilities. A statistical foundation of reasoning with probabilities lies in the concept of sampling at random from a finite population (an 'urn' in probabilistic literature). Here probabilities appear as relative frequencies in a reference population, see [426, 545]. Games of chance provided a slightly different motivation for the emergence of probability [99]. We consider in this chapter one historical and two modern interpretations of the concept of probability and two axiomatic systems for their calculus.

(I) According to *Laplace* [420], who survived the turmoil of the French revolution twice in a high-ranking position [640, 644], the probability of a certain event is the ratio between the number of 'favourable outcomes' (with respect to the event considered) and the number of all possible, 'equally likely' outcomes. In this interpretation, the probability that a newly born child will be a boy is exactly 1/2. This approach is appealing because of its simplicity. However, it is not without its difficulties. A delicate point is often how to determine whether 'elementary' events exist which can be regarded as 'equally likely'. In the case that the basic events cannot be considered as equally likely, a definition of probability is not provided. Note that, strictly

[†] ..., we who pursue probable states of affairs and not [anything] beyond that which occurs similarly to the truth, are able to progress and are disposed to refute without obstinacy as well as to be refuted without anger (Tusculanae disputationes, Liber II, Prooemium)

speaking, the M/F birth rate differs from 1:1, see, e.g., [17] and Example 6 in Chap. 2 of [696]. Moreover, one would also like to consider the probability of a boy compared with the probability of, for instance, a still-born child or of a twin.

In addition, in the elementary definition of Laplace, room is left for a number of paradoxes constructed along the line that the space of possible outcomes is changed (reduced or extended) and partitioned in various ways, such that 'logically' the same favourable event is one out of n events in one situation and one out of m events in another ($m \neq n$). An interesting problem illustrating this is: Suppose you meet an acquaintance, who has two children. The child who happens to be with the acquaintance is a girl. What is the probability that the other child is a boy, $1/2$, $2/3$ or ...? Of a similar nature is the quiz–master's paradox, sometimes called the Monty Hall dilemma [391], which resembles a 'cooperative game with imperfect spying': There are n closed doors, behind one of them there is a prize. After the participating player has made a choice, pointing at one door, the quiz–master (knowing both the choice of the player and the location of the prize) opens $n-2$ doors without a prize, while leaving a second door, not pointed at by the player, closed. What is the probability that the prize is behind the door first chosen? In fact, such type of questions lie at the root of classical probability theory, see [7, 57, 391], and are also investigated in possibility theory, where uncertainty is expressed by sub- and super-additive rather than by additive measures, see [358, 609].

The objectivistic and subjectivistic interpretations of probability, which we will discuss now, can be viewed as two different extensions of Laplace's interpretation.

(II^a) A *frequentist* (or: objectivistic) *interpretation.* A probability is defined as (the limit of) the relative frequency of a certain type of events in an increasingly large number of experiments, conducted under similar ('macroscopically identical') situations. The above sentence about the die, for example, is interpreted as: When we throw a (fair) die a sufficiently large number of times (i.e., if we repeat the experiment under the same macroscopic conditions), then the relative fraction of the number of times a 5 will show up, converges to $1/6$. A similar, but somewhat different, approach is to consider the relative frequency based on an 'a-select' drawing from a finite reference population, such as a deck of cards. While physicists are more accustomed to think in terms of repeatable experiments (with 'random perturbations'), statisticians tend to prefer the second empirical model. Furthermore, somewhat delicately, the principle of randomness has to be imposed, which means, for a population, that it is either not composed of inhomogeneous sub-populations, or at least that the drawing mechanism does not favour any sub-population. For a series of experiments the principle requires that various rules in selecting sub-sequences lead to the same limit, i.e. in a fair coin tossing game, no legitimate strategy should exist to 'beat the opponent'. A population or a

series of experiments satisfying the principle of randomness is called a collective. This concept of probability is in line with Von Mises and Reichenbach, see [545,709,710], who in the logical positivistic tradition of the Wiener Kreis tried to minimise the role of metaphysics in scientific discourse. Naturally, such probabilities are a mathematical idealisation to describe (and predict) the behaviour of a certain class of physical phenomena. In some situations, this idealisation describes more accurately the class of observed, and in the future observable, phenomena than in others. Anyhow, since attention is restricted to statements that are – in principle – accessible to intersubjective, empirical observations, the frequency interpretation is also called the objectivistic interpretation of probability.

(II^b) A *personal* (or: subjectivistic) *interpretation.* A probability is interpreted as a numerical representation of the 'strength of belief' a person attaches to a certain proposition. The strength of personal belief can be (and in many cases is indeed) adjusted in the light of the available data from pertinent experiments. This process ('démarche de pensée') is also called 'the transition from prior to posterior probabilities (with respect to the data from the experiment under consideration)'. The origin of this interpretation is usually attributed to *Thomas Bayes* whose fundamental paper [39] was published after his death around the middle of the eighteenth century. His ideas were propagated and popularised through Laplace's early re-derivation of Bayes' theorem [419, 644] and by Poisson [517], who distinguished between 'chance' (with a frequentist meaning) and 'probabilité' (personal probability), an issue also discussed in [198, 201, 511] and [236, 642]. Statistical analysis employing such an interpretation of probability is sometimes called Bayesian statistical analysis.

The distinction between (II^a) and (II^b) draws the main line and has been made here for didactic convenience. In the second case, some further differentiation can be made between the purely subjectivistic sub-school, in more recent times associated with the names of Savage [582] and de Finetti [138], and the sub-school of formal Bayesians such as Jeffreys [323], Jaynes [321], and, in spite of some anti-Bayesian criticism [200], in a sense also Fisher [205], see [359]. In the first case, De Finetti's famous representation theorem is sometimes used with the intention to justify the approach by 'deriving' the existence of a prior distribution, while the somewhat awkward problem of specifying prior probabilities is addressed by eliciting them from the client on the basis of his betting behaviour. The formal Bayesians use other means to choose a proper or improper prior distribution. In this case, it is maintained that if every rational man behaves 'coherently', i.e. processes his probabilities according to the axiomatic rules of probability (compatible with the Kolmogorov axioms), and if the prior distributions are chosen according to conventions based on invariance properties (Jeffreys) or a maximum entropy principle (Jaynes), rather than to a personal strength of belief, then the subjective element in the Bayesian approach can be replaced by an intersub-

jective one, at least with regard to the exchange of opinion among 'rational scientists'. For further material and discussion on such type of issues, the reader is referred to [6, 321, 323, 358, 573, 722].

Following Kolmogorov, *frequentists* use the word probability in a more restricted set of circumstances than do *subjectivists*. That is why they insist on using some other expression, such as 'degree of conviction' or 'strength of belief' to describe a degree of (personal) confidence with respect to propositions that does not have a relative frequency interpretation, see [358]. By extending the domain of discourse, at least the formal Bayesians are more normative with respect to a coherent use of probability statements (in a non-frequentist context) than are the frequentists. The two major schools in probability theory lead to different approaches in inferential statistics, which lead to different results notably for small samples. For large samples, the difference tends to be small from a practical point of view, see Question 12 in [358]. There have been several approaches (all of them 'revisionist' in stainless subjectivistic eyes) to provide a viable synthesis of the conflicting schools by providing an extension of classical objectivistic statistics to certain problem areas, without reverting to a subjective interpretation of the concept of probability itself. For instance, the theory of *empirical Bayes* estimation [554], and the concepts of *fiducial*, [203] *structural* [209] and *predictive* [224] inference. The latter expresses ignorance as a uniform distribution on observable quantities, rather than on (unknown) parameters, which happens to be close to Bayes' original scholium [39], see Stigler [642]. More recently, the theory of *distributional inference* has been developed by Schaafsma et al., see [6, 347, 359, 400, 573, 583, 693], where, employing foundational work in statistical decision theory by Wald and many others, see e.g. [191], epistemic probabilities are derived using relevant classes of ('proper') loss-functions. The purpose of 'distributional inference' is to construct *epistemic probability* distributions (instead of just point estimates and confidence intervals) for unknown parameters as well as future observations, on the basis of data, (objectivistic) probability models and ('proper') loss functions. Epistemic probability distributions are sometimes also called credence distributions [349] or confidence distributions [165,358], since, even if the probabilities used are defined objectivistically, the modelling itself is based on, often more or less disguised, subjective elements. Stated otherwise, statistical inferences are *stami-factions*, i.e., statistical mixtures of facts (data) and fruitful fictions (models and loss functions), where the distinction between the two is less clear than their appearance may suggest at first sight, since data usually contain a considerable amount of fiction, and models (may) contain a considerable amount of factual information. For various further considerations about the statistical relationship between data and models, the reader is referred to [136, 358, 520].

Probabilists constructed, similarly to what logicians did with parts of colloquial reasoning, axiomatic frameworks defining the rules to which precise

statements about probability should adhere. One such axiomatic system has been devised in the early thirties by *Kolmogorov*. It attaches, at least in principle since in practice numerical specifications are often avoided, probabilities to 'events', identified as sets, and is particularly suited for the frequentist interpretation of probability. Conditional probabilities are easily defined in terms of absolute (i.e., unconditional) probabilities, as long as the conditioning event has a positive (nonzero) probability. For continuous distributions, the theory of conditional probabilities can be cast, through conditional expectations of indicator functions, in a mathematical measure-theoretic framework with some appealing generality, see [102, 167, 682]. Another, slightly more general, axiomatic system has been developed by *Popper*, see [520]. It considers assigning conditional probabilities as an extension of assigning truth values to the Boolean algebra of propositions. Absolute probabilities are defined in terms of conditional probabilities, and the formal structure of the system is oriented towards a semantic interpretation by subjective probabilities. In Sect. 1.2 we discuss the Kolmogorov axioms and, as an example of an axiomatic system starting with conditional probabilities, in Sect. 1.3 the axioms by Popper. For other work on axiomatic and fundamental aspects of probability theory we refer to [130, 193, 235, 407, 721].

The text in this chapter constitutes a mixture of elementary concepts and theory, intertwined with remarks which provide more details or elaborate on a certain aspect. Most of these remarks may be skipped at first reading without serious loss of continuity. It will be of convenience for the reader to have some practice with operations in basic set theory, usually taught at high school or undergraduate university level. It is to be realised, however, that set theory itself is by no means a superficial area, and has implications at the roots of the foundations of mathematics, see for instance [77], and Chap. 14 of [162]. Although it is not a prerequisite for following the flow of the main text, it is surmised that some modest acquaintance with elementary, undergraduate algebraic concepts may help the reader to apprehend more fully some of the remarks and exercises in this chapter, which are typeset in a smaller font. The algebraic discipline itself, which has been developed as an abstraction of the (compositional) properties of numbers, matrices, sets, function spaces, and other mathematical categories, constitutes together with topology and the theory of ordering relations, one of the main structuring ingredients of modern mathematics [19, 60, 73, 162, 416, 697]. For an historical overview of mathematics the reader is referred to [9, 75, 147, 364, 385] and for a modern history of classical electrodynamics to [134]. Corporate history spanning periods over binary multiples of 75 years and which eventually, in a more or less indirect fashion, has influenced also the present book, can be found in [241, 252, 705].

1.2 Probability Measures and Their Generalisation

Definition 1.1. *A probability space is defined by a triple* $(\Omega, \mathfrak{B}(\Omega), P)$*, where* Ω *is a set,* $\mathfrak{B}(\Omega)$ *is a* σ*–algebra of subsets of* Ω*, and* P *is a probability measure.*

Interpretation. The elements of $\mathfrak{B}(\Omega)$, denoted by $A, B, \ldots$ are interpreted as *events*. For example, $\Omega = \{1, 2, \ldots, 6\}$, $A = \{4, 5, 6\}$, $B = \{2\}$. B is called an elementary event or outcome, and A is a compound event. The usual set operations are interpreted accordingly: $A \cap B$ denotes the (compound) event that event A as well as event B takes place, $A \cup B$ means that event A or event B takes place (we consistently adhere to the logical use of non-exclusive 'or'), A^c that the event A does not occur, $A \subset B$ that B has to occur if A occurs, $B \backslash A = B \cap A^c$ that B takes place, but A does not occur. The empty set, $\varnothing = A \cap A^c$ is used to indicate a 'logically impossible' event, which cannot happen.

Definition 1.1 (cont.) A family $\mathfrak{A}$ of subsets of Ω is called a σ–algebra if it has following (axiomatic) properties:

(1) $\varnothing \in \mathfrak{A}$,

(2) $A \in \mathfrak{A}$ implies $A^c \in \mathfrak{A}$,

(3) $A_1, A_2, \ldots \in \mathfrak{A}$ implies $\bigcup_{i=1}^{\infty} A_i \in \mathfrak{A}$ for any denumerable collection of subsets A_i.

Remark. From (1) and (2), it follows immediately that any σ–algebra contains, as an element, the set Ω.

Definition 1.1 (cont.) A *probability measure* P is a set-function $\mathfrak{B}(\Omega) \to [0, 1] \subset \mathbb{R}$ satisfying:

(1) $\mathrm{P}(\varnothing) = 0$ and $\mathrm{P}(\Omega) = 1$ ('positivity and normalisation');

(2) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$ if, and only if, $A \cap B = \varnothing$ ('finite additivity'); for infinite sets Ω we require even

(2^a) $\mathrm{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$ if $A_i \cap A_j = \varnothing$ for all i and j with $i \neq j$ ('$\sigma-$additivity').

This definition marks a natural end-point of a long development. Some of its contextual background is described in the following seven remarks.

Remarks.

1. These (axiomatic) properties of a probability measure are known as the Kolmogorov axioms [389]. The condition in (2^a) is phrased: the events $A_1, A_2, A_3, \ldots$ are mutually exclusive. In set theory, an algebra (σ–algebra) is a family of subsets which is closed under taking complements and finite (denumerable) unions, and hence also under finite (denumerable) intersections of its elements. The properties (2)

and (2^a) are called finite and denumerable (or: σ–) additivity, respectively. From a mathematical viewpoint, a σ–algebra has quite a complete structure, and can be constructed starting with simpler families of sets such as *rings* (which are closed under set differences and finite set unions) and *semi-rings*. The latter are defined by being closed under finite intersections, while the difference between two subsets (i.e., between two elements of the semi-ring) is not necessarily an element of the semi-ring, but can always be written as the disjunct union of finitely many elements of the semi-ring. (A semi-ring always contains the empty set, and a ring containing the set Ω as an element is easily shown to be an algebra, see Exercise 1.) An example of a semi-ring is the set of all intervals on the real line. For a concise introduction to algebraic set theory, the reader is referred to [263] and [167].

2. A probability measure can be *generalised* to a positive measure, i.e., a countably additive set function $\mu\colon \mathfrak{B}(\Omega) \to [0, \infty]$, to a signed measure $\nu\colon \mathfrak{B}(\Omega) \to (-\infty, \infty)$, which is the difference between two finite, positive measures, and even to a signed measure $\nu\colon \mathfrak{B}(\Omega) \to [-\infty, \infty]$ which is the difference between two positive measures, provided of course that at least one of these two measures is finite, i.e., satisfies $\mu(A) < \infty$ for all $A \in \mathfrak{B}(\Omega)$, see, e.g., Chap. 7 in [167]. A translation invariant, positive measure is a model to assign volumes to sets, a positive measure represents a mass distribution, and a signed measure represents a distribution of electric charge.

3. The collection of all subsets of a finite or denumerable set Ω (e.g., $\Omega = \mathbb{N}$ or $\mathbb{Z}$) forms a σ–algebra on which probability measures can be defined. In an 'ideal situation', the same would hold for all subsets of $\Omega = \mathbb{R}^p$.
   In reality, the state of affairs is more involved, since one can construct rather complicated sets. In practical mathematical analysis, one considers in $\mathbb{R}^p$ the σ–algebra generated (through taking at most countably many intersections, unions, and differences) by all open sets, which is called the σ–algebra of the Borel measurable sets, and the more extended σ–algebra of all Lebesgue measurable sets. By default, all open sets (and also all closed, and hence all compact sets) in $\mathbb{R}^p$ are Lebesgue measurable. The axiom of choice has to be invoked to find subsets in $\mathbb{R}^p$ which are not Lebesgue measurable. There exist many textbooks on measure and integration theory, in which special attention is paid to the relationship between measures and topology. We mention in particular [74, 167, 263, 390, 568], which have been used for Remarks 5 and 6. The less versed reader may return to these two remarks after having skipped them at first reading. Measures on fractal sets, which are not Borel measurable and have Lebesgue measure zero, are briefly discussed in Remark 7. Fractal sets enjoy considerable recent attention, the reason of which is not only to be sought in the

fact that they can conveniently be drafted using graphical computer software.

4. A non-degenerate probability is assigned to an event when it still has to take place, and also, according to a subjectivistic interpretation, when it has already taken place, but we were not yet able 'to look at it'. Consider throwing one fair die in a poker play. Before the die is thrown, the probability that a 6 turns up is 1/6. After the die is thrown, and player A looked at it, but player B did not, the probability for player A that a 6 has turned up is 0 or 1, while, according to a subjectivistic interpretation, for player B it is still a sensible statement to say that it is (close to) 1/6. A logician refrains in such a situation from making probability statements. He will say that the event did or did not occur, and that the true state of affairs is known to A, but not known to B.
   We mention an instructive example from [7], which bears some resemblance to the quiz–master paradox [391]. One of two players ('A' and 'B') has thrown a die. After having looked at the result, player A is instructed to tell player B (in accordance with the 'true state of nature') one of three elements of partial information: (a) the outcome is 1 or 5, (b) the outcome is even or (c) the outcome is a triple. (In case a six is thrown, player A has a choice between providing information (b) or (c). His 'strategy' is unknown to player B.) Using the Laplacian principle of 'insufficient reason', player B may be tempted to assign 1/3 to the three possibilities in case (b) and 1/2 to the two possibilities in case (c). This choice is far from optimal from a decision-theoretic point of view, however. In a Bayesian approach, it is assumed that player A provides in case a six is thrown, information (b) with probability ρ and information (c) with probability $1-\rho$. Then, by a conditioning argument (Bayes' rule, see below), it is derived that, if information (b) is given, the probabilities for obtaining 2, 4, and 6, are $\frac{1}{2+\rho}, \frac{1}{2+\rho}$, and $\frac{\rho}{2+\rho}$, respectively. The class of all Bayes' rules $0 < \rho < 1$ has some attractive properties. However, the practical question is *which* ρ is 'effectively' utilised by player A? The reader is referred to [7] for a clearly described comparison between various approaches in this situation, both in the framework of 'game theory' and of 'distributional inference'.
5. The theoretically inclined reader may ask himself why a probability measure is defined on merely a $\sigma-$algebra, and not, for instance, on all subsets of $\Omega = \mathbb{R}^p$. This question, lying at the root of measure theory and real analysis, is discussed in this remark. Let $I_p = [\underline{0}, \underline{1}]^p$ be the unit cube in $\mathbb{R}^p$, $p \geq 1$. By a 'paradoxical' no-go theorem of Banach and Tarski (1924), see [27,716], it is not possible to construct a countably additive, translation invariant measure with a fixed normalisation, say $\mathrm{P}(I_p) = 1$, defined for all subsets of I_p. Formulated otherwise, by breaking the unit cube into (complicated) subsets and translating these, it is possible to assemble from them a number

of unit cubes. On the other hand, sub-additive ('outer') measures, for which the inequality $(2^{a'})$ $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathrm{P}(A_i)$ holds instead of the equality in (2^a), can be defined on *all* subsets of I_p. To avoid the type of irregularities following from the Banach–Tarski paradox, one is apt to consider (1) the 'Borel measurable sets' $\mathfrak{B}$, which are those subsets that can be formed from all open (or, equivalently, from all closed) sets by making, at most countably many, standard set operations, i.e., complements, intersections and unions, see Chap. 1 of [568], and (2) the 'Lebesgue measurable sets', denoted by $\mathfrak{L}$ (notation courteously introduced by Radon, see [167]), which are constructed as follows. On the (set-theoretic) semi-ring $\mathfrak{R}$ of the semi-closed rectangles in $\mathbb{R}^p$, the p-dimensional Lebesgue pre-measure λ_p is defined which satisfies $\lambda_p((\boldsymbol{a}, \boldsymbol{b}]) = \prod_{i=1}^{p}(b_i - a_i)$, for any $\boldsymbol{a}$ and $\boldsymbol{b}$, where $\boldsymbol{a}$ denotes the 'left–lower' point and $\boldsymbol{b}$ the 'right–upper' point of the rectangle. This pre-measure is extended to an outer (or upper) measure on all subsets $\mathfrak{P}(\mathbb{R}^p)$ by defining

$$\lambda_p^*(A) = \inf\Big\{ \sum_{i=1}^{\infty} \lambda_p(R_i), A \subset \bigcup_{i=1}^{\infty} R_i, R_i \in \mathfrak{R} \Big\} \,. \tag{1.1}$$

Subsequently, following the approach by Carathéodory (1914), the Lebesgue measurable sets are defined as those subsets A of $\mathbb{R}^p$ for which $\lambda_p^*(D) = \lambda_p^*(D \cap A) + \lambda_p^*(D \cap A^c)$ for all $D \in \mathfrak{P}(\mathbb{R}^p)$, i.e., (A, A^c) splits every set S in two additive parts. Similarly to outer measures, one can introduce inner (or lower) measures, which are super-additive. While in general $\lambda_{p,*}(A) \leq \lambda_p^*(A)$, for Lebesgue measurable sets equality holds: $\lambda_{p,*}(A) = \lambda_p^*(A) \equiv \lambda_p(A)$. In some sense, inner and outer measures have complementary (dual) properties. Another general relation, which we shall return to later in this chapter, is $\lambda_{p,*}(A) + \lambda_p^*(A^c) = \lambda_p(\Omega)$. It can be shown that $\mathfrak{B} \subset \mathfrak{L} \subset \mathfrak{P}(\mathbb{R}^p)$, where the inclusions are strict. The set $\mathfrak{B}$ has the cardinality of the set of real numbers, $\aleph_1$, whereas $\mathfrak{L}$ has the cardinality of all subsets of the real numbers, $2^{\aleph_1}$.

(i) An example of a set which is not Borel measurable, but has Lebesgue measure zero is the p-dimensional Cantor discontinuum, which is a non-denumerable, compact set constructed by deleting middle parts of a convex polytope in a self-similar, recursive fashion such that k^j parts are deleted at step j for some fixed integer k, $j = 1, 2, 3, \ldots$, and the deleted parts form a family of disjunct open sets, the content ('measure') of which constitutes a geometric sequence whose sum equals the content of the original figure. Hence, a Cantor set itself is closed, compact, non-denumerable and has Lebesgue measure zero, but is not Borel measurable. It can be shown that every Lebesgue measurable set can be approximated by a Borel measurable set, except for a residual set of *Lebesgue* measure zero, called 'negligible' in [74]. From the difference in cardinality one can infer that there are many sets with Lebesgue measure zero, a topic discussed further in Remark 7.

(ii) Sets which are not Lebesgue measurable can be constructed by invoking the axiom of choice as was first observed by Vitali (1905), see [322]: Consider on $[0, 1]$ with addition modulo 1, the equivalence relation, $x \sim y$ if $x - y$ is rational. This splits $[0, 1]$ in a (non-denumerable) family of disjunct sets. Any set V which chooses one representative from this family,

is not Lebesgue measurable, since $[0,1]$ can be split up into a denumerable family of subsets, each of which is a translation of V. Hence, many subsets in $\mathbb{R}^p$ are not Lebesgue measurable. This is not the full story, however. As described for instance in [167], a result obtained in [628] states that a theory of real numbers can be made in which every subset in $\mathbb{R}^p$ is Lebesgue measurable, through replacing (loosely speaking) the axiom of choice, introduced by Zermelo (1904), see [475,749], and, similar to Euclid's postulate in plane geometry, being independent [114] of the other axioms in the Zermelo–Fränkel set-theory, by a weaker form of the axiom of choice. This (alternative) theory of real numbers is consistent under some relatively weak meta-mathematical assumptions [628].

6. In this remark we describe some of the underlying mathematical structure. After introducing the (abstract) notion of a measurable function, we look at the relationship between measures and linear functionals, and describe the road of generalisation of the 'standard', uniform (Lebesgue) measure on $\mathbb{R}^p$ to more general situations. Once measurable sets are introduced, measurable functions are defined in analogy with continuous functions between topological spaces. A function $f : \big(\Omega, \mathfrak{B}(\Omega)\big) \to \big(\mathfrak{X}, \mathfrak{B}(\mathfrak{X})\big)$, where $\mathfrak{B}(\Omega)$ and $\mathfrak{B}(\mathfrak{X})$ are σ–algebras, is called measurable if $f^{-1}(A) \in \mathfrak{B}(\Omega)$ for all $A \in \mathfrak{B}(\mathfrak{X})$. Since the inverse function mapping respects basic set operations, it suffices to verify this condition for any generating set of the σ–algebra $\mathfrak{B}(\mathfrak{X})$, for instance, $\{(a,\infty); a \in \mathbb{R}\}$ if $\big(\mathfrak{X}, \mathfrak{B}(\mathfrak{X})\big) = \big(\mathbb{R}, \mathfrak{B}(\mathbb{R})\big)$ where $\mathfrak{B}(\mathbb{R})$ is the family of the Borel sets. Unless it is clear from the context, one has to mention the two σ–algebras with respect to which a function is measurable. A measure generalises the notion of a primitive function $F(x) - F(a) = \int_a^x f(u)du$, $\mu_f(E) = \int_E f(u)du$ being defined for every measurable set E. For any fixed E, $\mu_E(f) = \int_E f(u)du$ is a linear functional, which associates a real number to any element f in a vector space of functions, such that $\mu_E(f+g) = \mu_E(f)+\mu_E(g)$ and $\mu_E(cf) = c\mu_E(f)$ for all $c \in \mathbb{R}$. This property is exploited in [74] to define measures on $\mathbb{R}^p$ as linear functionals which are continuous in a suitable topology. They are first considered on $\mathcal{K}(\Omega, \mathbb{R})$, the real-valued continuous functions on Ω with compact support. Subsequently, the class of functions is extended, and at the same time the corresponding class of linear functionals is reduced, such that, at the end of this process, measures, or equivalently integrals, are defined on an appropriate class of measurable functions. In this approach, a set is called measurable if its indicator function χ_E (which is one for all elements in E and zero otherwise) is a measurable function and a set is called integrable (i.e., has finite measure) if it is the intersection of a measurable set with a compact set, or, equivalently, if its indicator function has a finite measure.
Some rather analogous formal steps are used in detail while defining measures either as additive functions on σ–algebras of subsets or as linear functionals on function spaces. In the latter case, the connection with distribution theory and functional analysis is rather tight, see [568,601], permitting a natural introduction of topological concepts such as the point-wise ('weak*–') convergence of distributions, called 'convergence in the vague topology' in measure theory, see also [602]: $\mu_n \to \mu$ if and only if $\mu_n(f) \to \mu(f)$ for all

$f \in \mathcal{K}(\Omega, \mathbb{R})$. Given these two ways to define a measure, one can understand that, 'par abus de notation', μ_E is usually identified with μ_f and written as μ, even if a positive measure μ can also be considered as a mapping $\mathfrak{B}(\Omega) \times \mathscr{F}^+(\Omega) \to \mathbb{R}$, where $\mathfrak{B}(\Omega)$ is a class of 'μ'-measurable subsets of Ω and $\mathscr{F}^+$ is a class of positive, 'μ'-measurable functions on Ω.
The Lebesgue measure on $\mathbb{R}^p$ is naturally generalised to a Lebesgue–Stieltjes measure. On the real line, one considers $\lambda_F((a,b]) = F(b) - F(a)$ for any (right-continuous) nondecreasing function F defined on $\mathbb{R}$. For $F(x) = x$, the, translation-invariant, Lebesgue measure is recovered. In more dimensions, finite differences are taken, and the detailed formulation, albeit conceptually similar, is slightly more complicated, see e.g. Chap. 2 of [167]. On $\mathbb{R}^p$, a Lebesgue–Stieltjes measure, by being defined on a larger σ–algebra, generalises a Borel measure, which is a countably additive (signed) measure on $\mathbb{R}^p$ supplied with the Borel sets. (A Borel measure need not be translation invariant.) Sets which are Lebesgue–Stieltjes measurable are also Lebesgue measurable, even though the Lebesgue–Stieltjes measure of a set is usually different from the Lebesgue measure of that set. In line with Borel, we shall refer to Lebesgue measurable sets on $\mathbb{R}^p$ as being 'non-pathological'.
In some situations, it is worthwhile to consider measures on spaces that are more general than $\mathbb{R}^p$. The Lebesgue–Stieltjes measure on $\mathbb{R}^p$ can be generalised to a Radon measure on a Polish space, which is a separable topological space that can be metrised by a complete metric, see [74, 115, 167]. Some special cases of a Polish space are $\mathbb{R}^p$, a separable Banach space, a separable, compact Hausdorff space, and subsets of $[0,1]^{\mathbb{N}}$ which can be written as a denumerable intersection of open sets, see, e.g., [711]. On a locally compact Hausdorff space Ω (a Polish space being a special case), a (general) Borel measure is by definition a measure which is finite for any compact set, and a Radon measure is, by definition, a Borel measure which is 'regular', which means 'inner compact regular', i.e., $\mu(A) = \sup_{K \subset A, \text{compact}} \mu(K)$, as well as 'outer open regular', i.e., $\mu(A) = \inf_{O \supset A, \text{open}} \mu(O)$, for all measurable A.
Special Radon measures are those which are *moderate*, see Chap. 9 of [74], which means that the space Ω can be covered by denumerably many open sets of finite measure. On a Polish space every Borel measure is a moderate Radon measure, according to a theorem by Ulam (1939), see Chap. 8 of [167]. One virtue of considering such types of generalisation is that the set of Radon measures itself provided with the 'vague topology' is a Polish space, such that probability measures are conveniently defined on it and their properties exploited in the theory of 'random measures', see [93, 338].

7. The Lebesgue measure is too coarse for measuring the content of interesting fractal sets in $\mathbb{R}^p$, which generically have Lebesgue measure zero. For this purpose, the concept of a family of α–Hausdorff measures has been developed, by Hausdorff [271], based on [98], which coincides with the Lebesgue measure for Borel measurable ('regular') sets. Consider for $\Omega = \mathbb{R}^p$ (or Ω any Polish space, see Remark 6)

$$\mathcal{H}^\alpha_\delta(A) = \inf \Big\{ \sum_{i=1}^{\infty} d_i^\alpha; \quad A \subset \cup_{i=1}^{\infty} A_i, \quad d_i = \operatorname{diam}(A_i) < \delta \Big\} , \tag{1.2}$$

where $A_1, A_2, \ldots$ is a covering of the set A in $\mathbb{R}^p$ and

$$\mathrm{diam}(A_i) = \sup_{x,y} \{d(x,y) : x, y \in A_i\} \ , \tag{1.3}$$

where d is a metric measuring the distance between points in Ω. The Hausdorff measure of the set A is

$$\mathcal{H}^\alpha(A) = \lim_{\delta \to 0} \mathcal{H}^\alpha_\delta(A) \ . \tag{1.4}$$

Intuitively, it measures the 'p-dimensional length' of the set A if α is tailored to the (Hausdorff) dimension of the set, which is defined as the value of α for which the Hausdorff measure switches from infinity to zero. For instance, the Hausdorff dimension of the Cantor set (see Remark 5) on [0,1] equals $\log 2 / \log 3$.

A measure with properties complementary to those of the Hausdorff measure is the packing measure [688]. A packing of the set A in $\mathbb{R}^p$ (or, of any Polish space), is defined as a family of *disjunct*, open balls $B(x_i, r_i)$, such that $x_i \in A$ for all $i = 1, 2, \ldots$. Consider now

$$\overline{\mathcal{T}}^\alpha_\delta(A) = \sup \Big\{ \sum_{i=1}^{\infty} d_i^\alpha; \ B(x_i, r_i) \text{ is a packing of } A \Big\} \ , \tag{1.5}$$

where $d_i = \mathrm{diam}(B(x_i, r_i)) < \delta$ for $i = 1, 2, \ldots$. In this case, the set function

$$\overline{\mathcal{T}}^\alpha(A) = \lim_{\delta \to 0} \mathcal{T}^\alpha_\delta(A) \tag{1.6}$$

is not necessarily an (outer) measure. However, it can be shown, see [688], that

$$\mathcal{T}^\alpha(A) = \inf_{A \subset \cup_{i=1}^{\infty} A_i} \overline{\mathcal{T}}^\alpha(A_i) \tag{1.7}$$

is a measure. As for the Hausdorff dimension, the packing dimension is defined as the value of α where the packing measure switches from infinity to zero. The packing dimension may be finite in those cases that the Hausdorff dimension is zero, and, similarly, the Hausdorff dimension can be finite when the packing dimension is infinite. In plasma physics, fractal sets occur among others as Poincaré maps of the magnetic field in so-called ergodic regions of the plasma. The reader is referred to [163, 448, 559, 689] for further information on fractal measures.

Exercise 1.1. Derive from the axioms in Def. 1.1 and the usual set operations that
a) $P(A^c) = 1 - P(A)$,
b) $P(A) + P(B \backslash A) = P(A \cup B)$,
c) $P(A \cap B) + P(B \backslash A) = P(B)$,
d) $P(A \cup B) + P(A \cap B) = P(A) + P(B)$,
e) $A \subset B$ implies $P(A) \leq P(B)$ (monotonicity).
Draw Venn diagrams illustrating these properties.

Remark. Exercise 1.1 generalises the statement of axiom (2) in Def. 1.1. The following exercise utilises some basic definitions of set-theoretic and algebraic structures, most of them mentioned in Remark 1 above, and which can be found in almost any textbook on undergraduate algebra.

Exercise 1.2. The symmetric difference $A\Delta B$ is defined as $(A \cup B)\backslash(A \cap B)$.
a) Show that a ring is closed under the operations $\cup$ and Δ as well as under $\cap$ and Δ.
b) Show the reverse: If a family of subsets $\mathfrak{A}$ is closed under either $\cup$ and Δ or under $\cap$ and Δ, then $\mathfrak{A}$ is a ring.
c) Derive the (perhaps somewhat surprising) result that if a family of subsets is closed under $\cap$ and asymmetric differences, $\backslash$, it is not necessarily a ring.
d) Show that neither the family of open nor the family of closed intervals on $\mathbb{R}$ form a semi-ring, but that both the families of left semi-closed and of right semi-closed intervals each form a semi-ring.
e) If one identifies $\cap$ with multiplication and Δ with addition, then a ring of subsets is algebraically a *ring*, i.e., a group with respect to addition associative with respect to multiplication, and the multiplication is distributive with respect to the addition. (Remark. It has the special properties that every element is idempotent ($A \cap A = A$) and its own inverse with respect to addition ($A\Delta A = \varnothing$). Such a ring is sometimes called a Boolean ring, see Sect. 1.3.)
f) Show that if a ring $\mathfrak{A}$ contains Ω, it is a commutative algebra with unity over the trivial field $\{0, 1\}$ (with respect to standard multiplication and addition). Show also that in that case the set of functions

$$\left\{ f : \Omega \to \mathbb{R}; f(x) = \sum_i c_i \chi_{A_i}(x), A_i \in \mathfrak{A} \right\}$$

constitutes an algebra over the field $\mathbb{R}$.
g) Show that if a ring $\mathfrak{A}$ contains Ω then it is algebraically not a group with respect to multiplication; in fact, Ω being the unit element, no element A besides Ω itself possesses a (multiplicative) inverse, and hence $\mathfrak{A}$ is algebraically not a field. (Sometimes, the term σ–field is used instead of σ–algebra, but this usage does not have a strict algebraic connotation.)

Definition 1.2. *The events A and B are called* independent *if $P(A \cap B) = P(A)P(B)$.*

Definition 1.3. *For each $A, B \in \mathfrak{B}(\Omega)$ with $P(B) \neq 0$,*

$$P(A|B) = P(A \cap B)/P(B) \tag{1.8}$$

is interpreted as the probability that event A will take place, given that (it is certain that) event B will take place and is called the conditional probability *of A given B.*

Remark. If A and B are independent, then $\mathrm{P}(A|B) = \mathrm{P}(A)$. If B implies A, i.e. $B \subset A$, then $\mathrm{P}(A|B) = 1$.

Exercise 1.3. Explain the fallacy in the following argument: The probability that a virus will affect your disk when you connect (at random) to an internet site is say, 1/1000. The probability that two viruses will be placed (independently!) on your disk is 1/1000000. Hence, be sure to have a virus on your disk before connecting, in order to increase the safety of your system.

Exercise 1.4. We throw a die twice. What is the sample space Ω? What is the probability (a) to get twice a 6, (b) to get at least one 6, (c) to get precisely one 6, (d) to get no 6?

Exercise 1.5. (a) Using Exercise 1.1, derive that if $A \subset B$ then $\mathrm{P}(A) \leq \mathrm{P}(B)$. (b) Verify by drawing Venn diagrams the following (dual) properties for $k = 3$:

$$\mathrm{P}(A_1 \cup \ldots \cup A_k) = \sum_i \mathrm{P}(A_i) - \sum_{i<j} \mathrm{P}(A_i \cap A_j) + \ldots (-1)^{k+1} \mathrm{P}(A_1 \cap \ldots \cap A_k) \tag{1.9}$$

and

$$\mathrm{P}(A_1 \cap \ldots \cap A_k) = \sum_i \mathrm{P}(A_i) - \sum_{i<j} \mathrm{P}(A_i \cup A_j) + \ldots (-1)^{k+1} \mathrm{P}(A_1 \cup \ldots \cup A_k)\ . \tag{1.10}$$

The additivity property of measures can be generalised to *super-additivity* and *sub-additivity*. The corresponding set functions have been called inner and outer measures, as well as lower and upper probability measures, respectively. Interpreted as belief measures and plausibility measures, see [386, 609], they have found application in fuzzy set theory.

Definition 1.4. *A* belief measure μ_{be} *satisfies*

(1) $\mu_{be}(\varnothing) = 0$, $\mu_{be}(\Omega) = 1$*;*

(2) *if* $A \subset B$ *then* $\mu_{be}(A) \leq \mu_{be}(B)$*;*

(2^a) (*in case* Ω *is an infinite set:*) *for any nested sequence* $A_1 \subset A_2 \subset \cdots \subset A_\infty$, $\lim_{k\to\infty} \mu_{be}(A_k) = \mu_{be}(A_\infty)$*;*

(3) $\mu_{be}(A \cup B) \geq \mu_{be}(A) + \mu_{be}(B) - \mu_{be}(A \cap B)$.

A measure that satisfies axioms (1) to (2^a) in the above definition is called a *fuzzy measure*. A *plausibility measure* μ_{pl} satisfies, by definition, these axioms with the inequality in axiom (3) reversed. Evidently, probability measures, which satisfy axiom (3) with an equality sign, are a special case of belief measures, as well as of plausibility measures. Another special case of a belief measure (under suitable regularity conditions) is a *necessity measure*, which satisfies, in addition to the axioms in Def. 1.4,

$$\mu_{ne}(A \cap B) = \min\bigl(\mu_{ne}(A), \mu_{ne}(B)\bigr)\ . \tag{1.11}$$

The corresponding specialisation of a plausibility measure is a *possibility measure* which satisfies

$$\mu_{po}(A \cup B) = \max\bigl(\mu_{po}(A), \mu_{po}(B)\bigr)\ . \tag{1.12}$$

Exercise 1.6. Show that

(1) $\mu_{be}(A) + \mu_{be}(\overline{A}) \leq 1$,

(2) $\mu(A) = 1 - \mu_{be}(\overline{A})$ defines a plausibility measure,
(3) $\mu(A) = 1 - \mu_{po}(\overline{A})$ defines a necessity measure, and
(4) $\mu_{ne}(A \cup B) \geq \max\big(\mu_{ne}(A), \mu_{ne}(B)\big)$.

Definition 1.5. *A* Sugeno measure ν_y *is a fuzzy measure satisfying*

$$\nu_y(A \cup B) = \nu_y(A) + \nu_y(B) + y\nu_y(A)\nu_y(B) \tag{1.13}$$

for all A *and* B *with* $A \cap B = \varnothing$.

Exercise 1.7. Show that for $y = 0$, $y > 0$, $y < 0$, a Sugeno measure ν_y is a probability, belief and plausibility measure, respectively ($y \in \mathbb{R}$).

Remarks.

1. A motivation for using belief measures instead of probability measures to represent ignorance has been given in [609], see also reflection 14 on the cryptic issues for discussion formulated by D.R. Cox in [358], where Sugeno measures were suggested as a possible one-parametric framework for expressing (epistemic) confidence distributions [165], i.e., as a model for confidence measures.
2. The terms necessity measure and possibility measure suggest an analogy with rules from *modal logic*, see, e.g., [64, 104, 220, 294, 542], and *deontic logic*, see [26]. In fact, a number of modal logical systems exists, all of which are an extension of classical two-valued propositional logic, and in which two additional (complementary) logical operators are used, $\Box a$ and $\Diamond a$, which can be interpreted as 'the proposition a is necessarily true', and 'the proposition a is possibly true', respectively. The two corresponding deontic interpretations are 'action a is obligatory', and 'action a is allowed'. In modal logic with anankastic (i.e., necessarian) interpretation, the following properties hold, either as a definition, an axiom or a theorem (tautology):

$$\Box a = \sim \Diamond \sim a\ , \quad \Diamond a = \sim \Box \sim a\ , \quad \Box a \to a, \quad a \to \Diamond a\ . \tag{1.14}$$

Reversing the implication in either of the two last expressions in (1.14) leads to an expression of which the truth value is not always 1, i.e., which is not a theorem. In deontic logic, only the weaker form $O(a) \to P(a)$ holds, with the interpretation 'obligation implies permission', whereas in *provability logic* (see [220, 703]) $\Box a \to a$ can only be proved if a itself is a provable assertion, i.e., $\Box(\Box a \to a) \to \Box a$ ('Löb's axiom') is a theorem.[1] A valuation assigns to each proposition b a truth value $V(b)$. In *two-valued logic*, where the law of excluded middle holds ($a \vee \sim a$ is a tautology and $a \wedge \sim a$ is a contradiction), $V(b)$ only assumes the values 0 and 1.

[1] Here, $\Box a$ means 'a can be proven' and $\Diamond a$ means 'a cannot be disproved'. Neither $\Box a \to \Diamond a$ nor $\Box a \to a$ is a theorem in provability logic, since a formal axiomatic system is not necessarily consistent and must not be 'sound'.

Semantically, $V(\Box a) = 1$ can be interpreted as 'a is true in all possible worlds', $V(\Diamond a) = 1$ as 'a is true in (at least) one possible world' and $V(a) = 1$ as 'a is true in this (or a specific) possible world'. In a deontic context, the interpretations 'true in all possible worlds' is replaced by '(socially) desirable in all circumstances'. For more semantic interpretation of deontic logic in various situations, the reader is referred to [26, 294].

It is tempting to consider $\mu_{ne}(A) + \mu_{po}(A^c) = 1$, $\mu_{ne}(A) + \mu_{ne}(A^c) \leq 1$, and $\mu_{po}(A) + \mu_{po}(A^c) \geq 1$, as analogies of the three modal-logical statements $\Box a \vee \Diamond \sim a$ is a tautology, $\Box a \vee \Box \sim a$ is not a tautology, $\Diamond a \vee \Diamond \sim a$ is a tautology (and even: there exists a stronger statement than $\Diamond a \vee \Diamond \sim a$, which is a tautology), respectively.

Exercise 1.8. While using the standard rules of classical propositional logic, derive the above three modal logical statements from the properties of (1.14) and interpret them semantically. If $a \rightarrow b$ is a theorem, then (1) $\Box a \rightarrow \Box b$ and (2) $\Diamond a \rightarrow \Diamond b$ are also theorems. Derive (2) from (1), and compare this with axiom (2) of Def. 1.4.

1.3 Conditional Probablity Structures and Bayes' Theorem

Definition 1.6. *A conditional probability structure is a doublet* $(\mathcal{S},\ \mathcal{P}_c)$*, where* $\mathcal{S}$ *is a 'universe of discourse', having the structure of a Boolean algebra, and* $\mathcal{P}_c$ *a set of conditional probabilities, i.e., of binary functions* $\mathcal{S} \times \mathcal{S} \rightarrow [0, 1] \subset I\!R$*, that satisfy some axioms, to be discussed below.*

Remark. For the axiomatic properties of a Boolean algebra, see Definition 1.6.

Interpretation. The elements of $\mathcal{S}$, denoted by $a, b, c, \ldots$, are interpreted as propositions. For instance, the element a can denote the proposition 'Caesar, J., was not larger than 1.60 m', or 'Maxwell's laws of electromagnetism are correct'. The proposition 'a and b' is denoted by ab, and the negation of the proposition a is written as $\overline{a}$.

Definition 1.6 (cont.) Conditional probabilities assigned to propositions satisfy the following axioms:

(A1) There exist $a, b, c, d \in \mathcal{S}$, such that $p(a|b) \neq p(c|d)$ (Non-degeneracy);

(A2) $p(a|c) = p(b|c)$ for all $c \in \mathcal{S}$ implies $p(d|a) = p(d|b)$ for all $d \in \mathcal{S}$ (Probabilistic Equivalence);

(A3) $p(a|a) = p(b|b)$ for all $a, b \in \mathcal{S}$ (Reflexivity);

(B1) $p(a \wedge b|c) \leq p(a|c)$ for all $a, b, c \in \mathcal{S}$ (Monotonicity);

(B2) $p(a \wedge b|c) = p(a|b \wedge c)p(b|c)$ for all $a, b, c \in \mathcal{S}$ (Product Rule);

(C) $p(a|b) + p(\overline{a}|b) = 1$ unless for every $c \in \mathcal{S}$, $p(c|b) = p(b|b)$ (Sum Rule).

Remark. These axioms, derived in [520], are intended for general propositions. They have been derived from a concrete context, however, which in its simplest form is as follows. For a finite set Ω, let $\mathcal{S}$ consist of all statements of the form $\omega \in A$, where $A \subset \Omega$. In that case, propositions a are identified with subsets A, and $p(a|b)$ corresponds to

$$\mathrm{P}(A|B) = \frac{\#\{\omega; \omega \in A \cap B\}}{\#\{\omega; \omega \in B\}} \tag{1.15}$$

and one can verify, by simple computations with rational numbers, that Axioms $(A1)$ to (C) are then satisfied. (As usual, the symbol # is used to denote the number of elements in a set.)

For simplicity, we assumed here that $\mathcal{S}$ constitutes a Boolean algebra, see [2, 59, 68, 731]. In [520], only the closure of the set $\mathcal{S}$ under product and negation of its elements is postulated, and Popper derives, using the above axioms, that if one defines $a \sim b$ by $p(a|c) = p(b|c)$ for all $c \in \mathcal{S}$, then the set of equivalence classes $\mathcal{S}/\sim$ satisfies the axioms of a Boolean algebra.

Note that the Popper axioms do not presume explicitly that conditional probabilities assume values between 0 and 1. However, one can derive that for all $a, b \in \mathcal{S}$, $p(a|b) \geq 0$, while $p(a|a)$ is a constant which may be set to 1. Hence, the exception condition in axiom (C) can be rephrased as 'unless $p(c|b) = 1$ for all $c \in \mathcal{S}$', which can also be read as 'unless $p(b) = 0$', where the absolute probability $p(b)$ is defined by the following two rules: (1) $p(b) = 1$ means $p(b|c) \geq p(c|b)$ for all $c \in \mathcal{S}$, and (2) $p(b) = p(b|a)$ for some (hence for any) $a \in \mathcal{S}$ with $p(a) = 1$.

It is noted that in Popper's system conditional probabilities are also defined when the conditioning proposition has probability zero. This is an important special case, since, according to his concepts regarding the philosophy of science, inductive (universal) statements can, strictly speaking, never be verified, since only a finite number of supporting examples are possible. (They can be falsified, however, by providing appropriate counter-examples.) Hence, the assertion that some general physical theory (such as Newton's laws of gravitation) is true must have ('subjectivistic') probability zero. Nevertheless, one would like to make statements of the form: 'the probability that such and such observations are made, under the assumption that Newton's theory is correct'.

Exercise 1.9. Try to derive from the definition of absolute probabilities and the above axioms that $p(b) = 0$ is equivalent to $p(b|c) \leq p(c|b)$ for all $c \in \mathcal{S}$. Is it equivalent to $p(c|b) = 1$ for all $c \in \mathcal{S}$? Derive $p(b) = 0 \Leftrightarrow p(\bar{b}) = 1$ and also $p(b) = 0 \Leftrightarrow p(a|b) + p(\bar{a}|b) = 2$ for all $a \in \mathcal{S}$.

Remark. $p(a|b)$ has a clear analogy with a valuation of the logical statement $b \to a$ where the range of possible truth values are extended from $\{0, 1\}$ to $[0, 1]$. For instance, '$p(a|b) = 1$ for all $a \Leftrightarrow p(b) = 0$' corresponds to the

logical rule 'ex falso sequitur quodlibet'. Semantically, this rule hampers the direct application of probability to scientific statements of the type above, since in the philosophy of Popper such general laws have absolute (logical) probability zero. Therefore, in [520], he considers measures of corroboration that are not probabilities. The difficulty is related to the fact that there are not enough numbers 'infinitely close' to zero and one, respectively. A possible way of repair might be to consider $p(\cdot|\cdot) : \mathcal{S} \times \mathcal{S} \to [0,1]^*$, where $[0,1]^* \subset \mathbb{R}^*$ is the unit interval in *non-standard analysis*.[2] The elaboration of this is left as an exercise for the reader. A somewhat different approach is to express subjective probabilities by sub- and super-additive measures, rather than by additive ones, see Question 14 in [358] for a discussion.

To give conveniently a (somewhat redundant) set of axiomatic properties of a Boolean algebra, a second binary operation is introduced, defined by $a \vee c = \overline{\overline{a} \wedge \overline{c}}$, and interpreted as '$a$ or c'.

Definition 1.7. *A Boolean algebra is a set $\mathcal{S}$ with two binary operations, $\wedge$ and $\vee$, one unary operation, $\sim$, and an equivalence relation, $=$, such that for all elements a,b,c... $\in S$,*

(1) $a \wedge b = b \wedge a$, $a \vee b = b \vee a$ (*Commutativity*)
(2) $a \wedge (b \wedge c) = (a \wedge b) \wedge c$, $a \vee (b \vee c) = (a \vee b) \vee c$ (*Associativity*)
(3) $a \wedge (a \vee b) = a$, $a \vee (a \wedge b) = a$ (*Absorptivity*)
(4) *there exists a 0 (zero) and 1 (unity) in S, such that* $a \wedge 0 = 0$, $a \vee 0 = a$, $a \wedge 1 = a$, $a \vee 1 = 1$
(5) $a \wedge \overline{a} = 0$, $a \vee \overline{a} = 1$ (*Complementarity*)
(6) $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$, $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$ (*Distributivity*)

Classical set theory, with operations $\cap$, $\cup$ and $\setminus$, as well as (two-valued) propositional logic, with operations $\wedge$ (and), $\vee$ (or) and $\sim$ (not), have the structure of a Boolean algebra. (In the last case, 0 corresponds to a contradiction, and 1 to a tautology.)

There is a tight relation with the theory of ordered structures. Using again set theoretic notation, we define $A \subset B$ as $A = A \cap B$ (or, equivalently, as $B = A \cup B$). This introduces a *partial ordering* on $\mathfrak{B}(\Omega)$.[3] Any partially

[2] In non-standard analysis, the real numbers are extended by considering the set of Cauchy sequences, with a suitable equivalence relation, such that, in the spirit of Leibniz, infinitely small and infinitely large numbers are conveniently defined as well as is (most of) the usual algebraic, topological, and ordering structure of the real numbers. Hence, the real numbers form a proper subspace of the non-standard numbers, see, e.g., [8, 146, 415, 557].

[3] Recall that a partial ordering is a binary relation that is reflexive (i.e., $A \subset A$), transitive ($A \subset B$ and $B \subset C$ implies $A \subset C$) and antisymmetric ($A \subset B$ and $B \subset A$ implies $A = B$). Since there exist elements for which neither $A \subset B$ nor $B \subset A$, the ordering is not total.

ordered set for which each pair of elements has a least upper bound (sup) as well as a greatest lower bound (inf) is called a *lattice*. Evidently, $\sup(A, B)$ may sometimes, but does not always, coincide with $\max(A, B)$.

Examples.

1. The set of natural numbers $(\mathbb{N}, \prec)$ with $a \prec b$ defined by 'a is a divisor of b', constitutes a lattice with $\inf(a, b)$ being the greatest common divisor (gcd) of a and b, and $\sup(a, b)$ the least common multiple (lcm) of a and b.
2. By the basic properties of a σ–algebra, one can directly derive that if $\mathfrak{F}_1$ and $\mathfrak{F}_2$ are σ–algebras, then their intersection $\mathfrak{F}_1 \bigcap \mathfrak{F}_2 = \{A; A \in \mathfrak{F}_1 \text{ and } A \in \mathfrak{F}_2\}$ is also a σ–algebra, but, in general, $\mathfrak{F}_1 \bigcup \mathfrak{F}_2$ is not a σ–algebra. Any (at most denumerable) collection of σ–algebras of a set Ω can be extended to form a lattice, by applying repeatedly $\inf(\mathfrak{F}_1, \mathfrak{F}_2) = \mathfrak{F}_1 \bigcap \mathfrak{F}_2$ and $\sup(\mathfrak{F}_1, \mathfrak{F}_2)$, defined as the σ–algebra *generated by* $\mathfrak{F}_1 \bigcap \mathfrak{F}_2$. With respect to partial order from the inclusion relation for σ–algebras, $\inf(\mathfrak{F}_1, \mathfrak{F}_2)$ is the largest σ–algebra contained in both $\mathfrak{F}_1$ and $\mathfrak{F}_2$ ('greatest lower bound'), and $\sup(\mathfrak{F}_1, \mathfrak{F}_2)$ is the smallest σ–algebra containing both $\mathfrak{F}_1$ and $\mathfrak{F}_2$ ('least upper bound').

The principal connection between lattices and Boolean algebras is as follows: Starting with a lattice $(\Omega, \subset)$, one defines $A \cap B = \inf(A, B)$, and $A \cup B = \sup(A, B)$. A Boolean algebra is a complementary, distributive lattice with zero and unit elements that are the global minimum and global maximum, respectively, i.e., a lattice satisfying axioms (4) to (6) of Def. 1.7.

As is well known, the Boolean algebra of classical set theory and binary logic constitutes the basis of microprogramming in digital electronics [666, 731]. During the last decades an interesting extension of Boolean set theory has been developed. Here one considers *fuzzy sets* that are characterised by (continuous) 'belongingness functions' $\mu : \Omega \to L$, where $L = [0, 1]$ (or some other lattice).

By defining

$$\mu_1(\omega) \wedge \mu_2(\omega) = \max(0, \mu_1(\omega) + \mu_2(\omega) - 1) \text{ (conjunction)},$$
$$\mu_1(\omega) \vee \mu_2(\omega) = \min\big(1, \mu_1(\omega) + \mu_2(\omega)\big) \text{ (disjunction), and}$$
$$\mu_1(\omega) \Rightarrow \mu_2(\omega) = \min\big(1, 1 - \mu_1(\omega) + \mu_2(\omega)\big) \text{ (implication)},$$

one has introduced on $[0, 1]$ the properties of a *multi-valued logic* [67, 240, 550], introduced by Łukasiewicz around 1930. Just a few properties of multi-valued logic are described in the following two remarks.

Remarks.

1. With the negation defined as $\neg\mu_1(\omega) = 1 - \mu_1(\omega)$, one gets $\big(\mu_1(\omega) \Rightarrow \mu_2(\omega)\big) \equiv \big(\neg\mu_1(\omega) \vee \mu_2(\omega)\big)$. Furthermore, by defining $\big(\mu_1(\omega) \Leftrightarrow \mu_2(\omega)\big) \equiv \big(\mu_1(\omega) \Rightarrow \mu_2(\omega)\big) \wedge \big(\mu_2(\omega) \Rightarrow \mu_1(\omega)\big)$, one has the property $\big(\mu_1(\omega) \Leftrightarrow \mu_2(\omega)\big) = 1$ for $\mu_1(\omega) = \mu_2(\omega)$, which is phrased as 'the (multi-valued)

truth value of the equivalence relation is one on the diagonal', sc. of the square obtained by plotting $\mu_2(\omega)$ against $\mu_1(\omega)$.

2. The first two operators we have defined above are sometimes called *strong* conjunction, and *weak* disjunction, respectively. (To distinguish between strong and weak, we apply subscripts s and w.) A weak conjunction is $\mu_1(\omega) \wedge_w \mu_2(\omega) = \min\big(\mu_1(\omega), \mu_2(\omega)\big)$, and the corresponding strong disjunction (by the Morgan's Law) is $\mu_1(\omega) \vee_s \mu_2(\omega) = 1 - \min\big(1 - \mu_1(\omega), 1 - \mu_2(\omega)\big) = \max\big(\mu_1(\omega), \mu_2(\omega)\big)$. These lead to a strong implication $\mu_1(\omega) \Rightarrow_s \mu_2(\omega) \equiv \max\big(-\mu_1(\omega), \mu_2(\omega)\big)$, and to two forms of equivalence, $\mu_1 \Leftrightarrow_{sw} \mu_2 \equiv (\mu_1 \Rightarrow_s \mu_2) \wedge_w (\mu_2 \Rightarrow_s \mu_1)$ and $\mu_1 \Leftrightarrow_{ss} \mu_2 \equiv (\mu_1 \Rightarrow_s \mu_2) \wedge_s (\mu_2 \Rightarrow_s \mu_1)$, respectively. Neither of the two forms of equivalence has truth value one 'on the diagonal'. Interestingly, the weak implication leads with both the (default) strong and the weak conjunction to the same type of equivalence relation.

For further mathematical and philosophical background on non-classical logical systems, the reader is referred to [111, 542, 638] and the entries in [748].

Since a considerable part of mathematics can be based on set theory, the introduction of fuzzy sets leads quite naturally to fuzzy numbers, fuzzy relations, fuzzy calculus, fuzzy measures, etc., which are generalisations of their 'crisp' counterparts (and, hence, have lost some of their properties). A large number of articles have been written in the search of the most natural (classes of) generalisations and their properties.

Fuzzy-set based mathematics is used in rather precise, though often more logical than empirical, reasoning about uncertainty and ambiguous concepts and has found technical applications, among others in artificial intelligence and automatic control of consumer electronics. Some interesting textbooks on fuzzy set theory and its applications are [67, 386, 723], see also [158, 340].

In practice, it seems sensible, instead of 'fuzzifying' too much, to apply fuzzy concepts selectively where needed. For instance, one can apply fuzzy measures on crisp sets using standard (non-fuzzy) calculus.

A useful identity is *Bayes' Theorem*, first formulated in [39, 419], see also [641, 644], which states that

$$p(a|b)p(b) = p(b|a)p(a) \; . \tag{1.16}$$

In Popper's axiomatic system discussed above, the identity directly follows from the product rule in Def. 1.6. In set-theoretical notation, the theorem is often formulated as follows. Let $A_1, \ldots, A_k$ be a partition of the outcome space Ω, i.e., $\Omega = A_1 \cup \cdots \cup A_k$ with $A_i \cap A_j = \varnothing$ $(i, j = 1, \ldots, k)$, and let B be any other event, i.e., measurable subset of Ω. Then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \; , \tag{1.17}$$

with

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i) \; . \tag{1.18}$$

Interpretation. Suppose the events $A_1, \ldots, A_k$ denote an exhaustive and exclusive set of 'diseases' for a living or mechanical system (including 'normal state'), and the event B denotes a 'syndrome', i.e., a set of symptoms. In many cases, some empirical information about $\mathrm{P}(B|A_i)$ and $\mathrm{P}(A_i)$ can be obtained. Bayes' theorem can then be used to invert these probabilities to estimate $\mathrm{P}(A_i|B)$, a quantity of primary diagnostic interest. $\mathrm{P}(A_i)$ is called the *prior probability* of event A_i, and $\mathrm{P}(A_i|B)$ is called the *posterior probability* of event A given event (or: information) B. In objectivistic, 'empirical' Bayesian probability theory, $\mathrm{P}(A_i)$ is interpreted as the 'prevalence', i.e., the frequency of occurrence, of disease A_i in some, suitable, reference population. This prevalence can be estimated, at least in principle, by drawing a sample. In the subjectivistic interpretation of Bayesian probability theory, $\mathrm{P}(A_i)$ is considered as the (personal) prior strength of belief attached to the proposition that the system under investigation is affected by the disease A_i. Naturally, objectivists and either rationalist or personalist Bayesians hold different meta-statistical viewpoints, not on the mathematical correctness of Bayes' theorem, but on the empirical situations the theorem can be fruitfully applied. In the first case, attention is restricted to the properties of a physical system in repetitive situations, and in the second case to our (declared) knowledge thereof or, even more generally, of uncertain states of affairs. For a discussion on similar such issues, we refer the reader to [29, 138, 204, 323, 358, 407, 545, 582]. An application of Bayes' theorem is given in the following exercise.

Exercise 1.10. Suppose a series of plasma discharges of a certain 'standard type' are being produced, e.g., deuterium plasmas heated by deuterium neutral beams at a certain plasma current, magnetic field and density, with standardised wall conditioning. Because not all experimental parameters are completely known, even in this homogeneous class of discharges, some events still occur (seemingly) in a haphazard fashion, but we have been in a position to record what happened in a number of such similar discharges. The plasma can suffer from a disruption ($D = +$) or stay disruption-free ($D = -$). A monitor (i.e., warning) signal exists, which is in this case the ratio between two spectroscopically measured impurity signals, for instance the ratio of carbon to oxygen concentration, [C]/[O]. For simplicity, we consider this ratio to be discretised into a small number of levels $m = 1, \ldots, k$. From previous experience, a disruption occurred, on average for this type of discharges, in 1/2 of all cases.
Estimate $\mathrm{P}(D = +|M = m)$, i.e., the posterior probability of a disruption, given that the level of monitor signal is m, on the basis of the (additional) information given in Table 1.1, based on a small but representative subset of discharges. Hint: Use the formula $\mathrm{P}(D = +|M = m) =$

$$\frac{\mathrm{P}(M = m|D = +)\mathrm{P}(D = +)}{\mathrm{P}(M = m|D = +)\mathrm{P}(D = +) + \mathrm{P}(M = m|D = -)\mathrm{P}(D = -)} \,. \qquad (1.19)$$

What is $\mathrm{P}(D = +|M = m)$ if the 'prior probability' $\mathrm{P}(D = +) = 1/2$ is replaced by $\mathrm{P}(D = +) = 1/3$, as is suggested by the small subset of the data, displayed in Table 1.1? Check your answer by an argument not based on Bayes' theorem.

Table 1.1. Disruption frequencies (see Exercise 1.10).

D|m	0	1	2	3	total
+	4	8	12	8	32
−	8	2	4	2	16

For didactic purposes, fictitious numbers have been used.

1.4 Random Variables and Probability Distributions

In this section we restrict attention to real-valued random variables. Many (but not all) of the properties are extended to vector-valued random variables, or even to random variables that map the probability space Ω into more general function spaces, see Sect. 1.2, Remark 6, and [504]. Obviously, the particular properties of these more general spaces have to be taken into account.

In this section we concentrate the attention (mainly) to real-valued random variables. A number of aspects can be extended to vector-valued random variables.

Definition 1.8. *A random variable is a (measurable) function $X : \Omega \to \mathbb{R}$, where $(\Omega, \mathfrak{B}(\Omega), P)$ is a general probability space.*

Such a function induces from the given probability measure P on Ω a new probability measure P_X on $\mathbb{R}$: $P_X(A) = P\big(X^{-1}(A)\big)$, for each interval $A \subset \mathbb{R}$, and whence for each (measurable) subset in $\mathbb{R}$. Mathematically, $P(X^{-1}(A))$ is defined as $P\{\omega|X(\omega) \in A\}$, which is frequently abbreviated as $P\{X \in A\}$. P_X (and if the context is clear, 'par abus de langage', even P) is also called the probability distribution associated with the random variable X.

Example. $\Omega = \{1, 2, 3, 4, 5, 6\}^n$ is the set of outcomes when throwing n dice (subsequently or in one throw if they are identified by numbers $1, \ldots, n$), or when throwing the same die n times. X_1 is the number of times a 3 turned up, X_2 is the total number of dots that turned up, $X_3 = X_2/n$ is the average number of dots that turned up. X_1 defines a probability measure on $\{0, 1, 2, \ldots, n\}$, which can also be viewed as a probability measure on $\mathbb{R}$. As we will see later in this chapter, this probability measure can be

easily characterised, and is called the binomial distribution. The probability distributions induced by X_2 and X_3 are more complicated. The important point is here to see the difference and the connection between $(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$ and $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mathrm{P}_X)$.

Definition 1.9. *The (right-continuous) distribution function F of a random variable X is the function $F : \mathbb{R} \to [0,1]$ defined by $F(x) = P\{X \leq x\}, x \in \mathbb{R}$.*

Remarks.

1. A function F is right-continuous in a point x_0 if no jumps in F for x to the right of x_0 are allowed. The familiar $\epsilon - \delta$ definition is: F is right-continuous in x_0 if for every $\epsilon > 0$, there exists $\delta > 0$ such that $0 \leq x - x_0 < \delta$ implies that $|F(x) - F(x_0)| < \epsilon$. (A function F is right-continuous if it is right-continuous for every point $x_0 \in \mathbb{R}$.)
2. Similarly, left-continuous distribution functions are introduced by defining $\widetilde{F}(x) = \mathrm{P}\{X < x\}$. For describing probability measures one needs only to consider one of these two classes of distribution functions. (Which one is a matter of convention.)

Theorem 1.1.

(1) A (right-continuous) distribution function has the following properties: $F(-\infty) = 0, F(\infty) = 1$, F is monotonic (hence, non-decreasing).

(2) (Borel–Lebesgue) Each right-continuous function F with the properties (1) defines a probability measure by $P\{a < X \leq b\} = F(b) - F(a)$.

Proof. (1) follows directly from the axioms of a probability measure P. The most difficult part of (2) is to derive σ–additivity, on the semi-ring of *left-open, right closed* intervals, which is done using the Heine–Borel property of a closed interval (which is a compact set, such that every covering by open sets can be reduced to a finite covering by open sets), see e.g. [167]. The resulting pre-measure on this semi-ring is extended to a measure on a σ–algebra by the procedure described in Remark 5 following Def. 1 in Sect. 1.2.

Remark. For monotonically non-decreasing (non-increasing) functions F: $\mathbb{R} \to \mathbb{R}$, right-continuity is equivalent to upper (lower) semi-continuity. In our case, for any right-continuous distribution function F, the set $F^{-1}(-\infty, a) = F^{-1}[0, a)$ is open for any $a \in [0,1]$, but $F^{-1}(a, \infty) = F^{-1}(a, 1]$ is not always open, hence F is upper semi-continuous.[4]

[4] Upper semi-continuity means that the function may not jump upwards. For a real-valued function $f : \mathbb{R} \to \mathbb{R}$, it can be defined in several ways, see Chap. 3 of [76]: (1) the inverse image of any upper-bounded semi-infinite open interval is open, (2) for any x_0: $f(x_0) = \limsup_{x \to x_0} f(x)$, (3) $f(x) = \inf_{g > f} g(x)$ where g runs over all continuous functions (everywhere) larger than f, and (4) for any x_0: for every $\epsilon > 0$ there exists a $\delta > 0$ such that $|x - x_0| < \delta$ implies $f(x) < f(x_0) + \epsilon$. The

Although measure theory à la Bourbaki [74] has been conceived to comprise both cases in one framework, in practice, it is useful to distinguish between discrete and continuous probability distributions.

Definition 1.10. *A probability distribution is discrete if X can assume at most countably many values, i.e. if its distribution function F is a step function with at most countably many jumps.*

In this case $P\{X = x_i\} = F(x_i^+) - F(x_i^-)$ is just the jump of F at x_i, $i = 1, 2, \ldots$. We frequently abbreviate $P\{X = x_i\}$ by p_i.

Definition 1.11. *A probability distribution F is called continuous if F is 'absolutely continuous', i.e., practically speaking, if F is continuous and, possibly except for (at most) countably many isolated points, possesses a continuous derivative $F' = f$ (with respect to its argument x).*

Example. The negative exponential distribution with

$$F(x) = \begin{cases} 1 - e^{-x/\lambda} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{1.20}$$

and

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & \text{for } x > 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{1.21}$$

is continuous on $\mathbb{R}$.

Definition 1.12. *For a continuous (discrete) distribution one defines the expectation value:*

$$\mu = E(X) = \int_\Omega X(\omega) dP(\omega) = \int_{-\infty}^{\infty} x f(x)\, dx\ , \qquad \mu = E(X) = \sum_{i=1}^{\infty} p_i x_i\ , \tag{1.22}$$

the kth moment (around the origin):

$$\mu_k' = E(X^k) = \int_{-\infty}^{\infty} x^k f(x)\, dx\ , \qquad \mu_k' = E(X^k) = \sum_{i=1}^{\infty} p_i x_i^k\ , \tag{1.23}$$

the kth central moment (i.e. around the mean):

$$\mu_k = E(X-\mu)^k = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\, dx\ , \qquad \mu_k = E(X-\mu)^k = \sum_{i=1}^{\infty} p_i (x_i-\mu)^k\ , \tag{1.24}$$

with the important special case, the variance *(the central second moment)*

$$\sigma^2 = var(X) = E(X-\mu)^2 = \mu_2\ . \tag{1.25}$$

indicator function of the closed interval $[0, 1]$, while being upper semi-continuous, is right semi-continuous in 0 and left semi-continuous in 1. It is noted that a different concept of semi-continuity is used e.g. in [661] for multi-valued functions.

Exercise 1.11. Derive: $\mathrm{var}\,(X) = \mathrm{E}(X^2) - \mu^2$.

Definition 1.13. *For a continuous distribution, an α-quantile is defined as $F^{-1}(\alpha)$, where α is some fraction between 0 and 1. A 50% quantile, $F^{-1}(\frac{1}{2})$, is usually called median. For a discrete distribution, $F(x)$ is a step function and a difficulty arises. One way to define a (unique) inverse, see e.g. [607], is $F^{-1}(\alpha) = \inf_x\{F(x) \geq \alpha\}$. This definition has the feature that the median of two observations, $\{x_1, x_2\}$ is equal to the smaller of the two observations. Another possibility is to define $F^{-1}(\alpha) = \{x; F(x) = \alpha\}$, see [77]. In that case, for n observations, where the step-function F jumps at $x_1 < x_2 < \cdots < x_n$, $F^{-1}(\alpha)$ is not defined, unless $\alpha = \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$. For statistical applications, an improved definition is $F^{-1}(\alpha)$ is the interval $\{F_\lambda^{-1}(\alpha); \lambda \in (0,1)\}$, where*

$$F_\lambda^{-1}(\alpha) = \lambda \inf_x\{F(x) \geq \alpha\} + (1-\lambda)\sup_x\{F(x) \leq \alpha\}\ . \tag{1.26}$$

For $\alpha = 1/4,\ 1/2,\ 3/4$, the quantity $F^{-1}(\alpha)$ is called (lower, middle, upper) quartile, whereas for $\alpha = j/10$ the term jth decile is used ($j = 1, 2, \ldots, 9$).

Remark. Obviously, for any absolutely continuous F, $F^{-1}(\alpha) = x \Leftrightarrow x = F(\alpha)$.

Exercise 1.12. Verify the following monotonicity properties: (a) $F(x) > t \Rightarrow x \geq F^{-1}(t)$ and (b) $x > F^{-1}(t) \Rightarrow F(x) > t$. Note that $F(x) \in [0,1]$, but $F^{-1}(t)$ can be an (open) interval. (The possibility of an equality in the right-hand side of (a) is related to the convention that right-continuous distribution functions are considered.)

Remarks.

1. The standard deviation, $\sigma = \sqrt{\sigma^2}$, is a number characterising how much a distribution is 'spread out' or dispersed. Another such characteristic, more suitable for broad-tailed distributions, for instance the Cauchy distribution $f(x) = \frac{1}{\pi(1+x^2)}$ which has infinite variance, is half the interquartile distance: $\frac{1}{2}\big(F^{-1}(3/4) - F^{-1}(1/4)\big)$.
2. In a unified notation one can write for each continuous function $g : \mathbb{R} \to \mathbb{R}$:
$$\mathrm{E}g(X) = \int_{-\infty}^{\infty} g(x)\, dF(x)\ , \tag{1.27}$$
where F is an arbitrary distribution function. For random variables which map $(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$ into a more general measure space $(\mathfrak{X}, \mathfrak{B}(X))$, a distribution function may not be well defined, but we can still write
$$\mathrm{E}g(X) = \int_\Omega g\big(X(\omega)\big)\, d\mathrm{P}_X(\omega)\ , \tag{1.28}$$

which can be viewed as the general definition. For many practical problems with real-valued random variables it is sufficient to consider only the (fairly large) class of distribution functions that can be written as $F = F_{ac} + F_{step}$, where F_{ac} is absolutely continuous and F_{step} is a step function with (at most) countably many jumps $p_i = F(x_i^+) - F(x_i^-), i = 1, 2, \ldots$. It is noted that a more general decomposition is $F = F_{ac} + F_{step} + F_{sing}$, where F_{sing} is continuous, but not absolutely continuous. A distribution function which is zero on the complement of a Cantor set, and otherwise linearly increasing from 0 to 1 on the interval $[0, 1]$, called a Cantor 'staircase' in [390], is a classical example of a continuous function which is not absolutely continuous. The above notation for the expectation value, using distribution function F or more generally the probability measure P_X, is preferred by mathematicians. Physicists are more used to thinking in terms of

$$\mathrm{E}g(X) = \int_{-\infty}^{\infty} g(x) f(x)\, dx \;, \tag{1.29}$$

where the generalised probability density f stands for a distribution of order zero, i.e., the derivative of a probability measure or a linear functional on an appropriate function space. For a large class of practical problems, $f(x) = f_{ac}(x) + \sum_{i=1}^{\infty} p_i \delta(x - x_i)$, where f_{ac} stands for the derivative of the absolutely continuous part of the distribution function, and $\delta(x - x_i)$ is a delta function concentrated in x_i.[5] We refer the reader to [167, 601] for an exposition of the theories (Stieltjes integration and Schwartz distribution theory, respectively) that justify these notations and their algorithmic manipulation.

A note on terminology. In physical literature f_{ac} is sometimes called a distribution function (for instance: a *Maxwellian distribution function*), and in mathematical texts on 'distribution theory', the generalised density f is called a *distribution* (*of order zero*). In this text, we will adhere to the standard probabilistic terminology by calling F the distribution function, its derivative f the (generalised) density and using 'the distribution of a random variable' as a general term denoting either its probability measure P_X, or, equivalently, its distribution function F or, equivalently, its generalised density f or even some other function that characterises the distribution, see Sect. 1.6. As a brief notation for this combined concept we employ $\mathfrak{L}(X)$, called the *probability law* of the random variable X, see for instance Chap. 4 of Loève [432].

The first few central moments of a distribution determine several characteristics of its shape.

[5] The derivative of the Cantor staircase can be considered as *a formal distribution* which integrates to 1 on the interval $[0, 1]$. Note, however, that the classical Lebesgue integral of the Cantor staircase derivative, considered as a function, is zero. Such niceties are outside the just mentioned class of practical problems.

Definition 1.14. *The (moment-based) skewness of a distribution is given by*

$$\gamma_1 = \mu_3/(\mu_2)^{3/2} \,. \tag{1.30}$$

and the excess of its kurtosis by

$$\gamma_2 = \mu_4/(\mu_2)^2 - 3 \,. \tag{1.31}$$

Remarks.

1. Skewness and excess can be written as ratios of cumulants, see Exercise 1.24.
2. Traditionally, the notation $\beta_1 = \gamma_1^2$ and $\beta_2 = \gamma_2 + 3$ is used, where $\beta_2 = \mu_4/(\mu_2)^2$ is called the kurtosis.
3. The skewness measures a deviation from symmetry. The skewness is positive if the distribution has a long tail to the right. Distributions with kurtosis 3 (i.e., excess 0) such as the Gaussian, have been called meso-kurtic, see for instance [510], where it was reported that the distribution of the lengths of selected recruits (in Verona, 1875-1879) deviated stronger from meso-kurtosis than the distribution of all conscripts in that area. The kurtosis measures both the thickness of the tails and peakedness. For positive (negative) excess, the tails are heavier (less heavy) than for a normal distribution, and/or the distribution is less (more) peaked in the sense of 'squeezed along the y-axis', see for instance [476].
4. See [434] and [477] for a further discussion on various measures of skewness and kurtosis, among others based on distribution quantiles.

Definition 1.15.

(*1*) *The simultaneous distribution of two random variables X and Y is characterised by a probability measure $P_{X,Y}$ on $\mathbb{R}^2$ where $P_{X,Y}(A) = P\{\omega : \big(X(\omega), Y(\omega)\big) \in A\}$ denotes, for each measurable subset $A \subset \mathbb{R}^2$, the probability that the random point (X, Y) falls in A.*

(*2*) *The joint distribution function of X and Y is $F(x, y) = P\{X \leq x, Y \leq y\}$.*

(*3*) *The marginal distributions of X and Y are given by the distribution functions $F_X(x) = P\{X \leq x, Y < \infty\}$ and $F_Y(y) = P\{X < \infty, Y \leq y\}$, respectively.*

(*4*) *For an absolutely continuous distribution function, the joint probability density is $f(x, y) = \partial^2 F(x, y)/\partial x \partial y$, the marginal probability density of X is $f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$, and the conditional probability density of X given $Y = y$ is $f(x, y)/f_Y(y)$.*

Remark. The subscripts $_X$ and $_Y$ are just labels to distinguish the various distribution functions. It is recalled that *'X and Y are independent'* means

that $P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}$ for all non-pathological subsets A and B. It can be shown that this is equivalent to $F(x, y) = F(x)F(y)$, and, for continuous distributions, to $f(x, y) = f_X(x)f_Y(y)$.

Definition 1.16. *Let X and Y have a joint distribution with expectation $(\mu_X, \mu_Y) = \big(E(X), E(Y)\big)$. The covariance between X and Y is $cov\,(X, Y) = E(X-\mu_X)(Y-\mu_Y)$. The correlation coefficient between X and Y is $\rho(X, Y) = cov\,(X, Y)/\big(var\,(X)\, var\,(Y)\big)^{1/2}$.*

Theorem 1.2. *If X and Y are independent then they are uncorrelated (i.e., $\rho(X, Y) = 0$). (The converse does not hold true.)*

Exercise 1.13. Let $Y = X^2$, where X has a symmetric distribution with mean zero. (Hence, its skewness is also zero.) Prove that $\rho(X, Y) = 0$, while obviously X and Y are not independent.

Theorem 1.3. *For two random variables X and Y (whether they are independent or not)*

$$E(aX + bY) = aE(X) + bE(Y)\ . \tag{1.32}$$

If X and Y are independent, then

$$var\,(X \pm Y) = var\,(X) + var\,(Y)\ . \tag{1.33}$$

In general,

$$var\,(aX \pm bY) = a^2\, var\,(X) + b^2 var\,(Y) \pm 2\, a\, b\, cov\,(X, Y)\ . \tag{1.34}$$

Exercise 1.14. Prove that $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$. Recall that $\mathrm{E}(aX + bY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax + by) f(x, y)\, dx\, dy$. Prove also the second part of Theorem 1.3.

A mechanical analogue. The mechanical equivalent of a probability distribution is a mass distribution,[6] see Sect. 1.2. This entails the following correspondence:

probability	mechanics
expectation value μ	centre of gravity
second moment $\mathrm{E}(X^2)$	moment of inertia w.r.t. the origin
variance $\mathrm{var}\,(X)$	moment of inertia w.r.t. μ
$\mathrm{E}(X^2) = \mathrm{var}\,(X) + \mu^2$	Steiner's rule

A geometrical analogue. One can show, under weak regularity conditions, that

(1) $\mathrm{cov}\,(X, Y) = \mathrm{cov}\,(Y, X)$,

[6] For an historical development of the physical concept of mass, see [315].

(2) $\mathrm{cov}\,(X, aY_1 + bY_2) = a\,\mathrm{cov}\,(X, Y_1) + b\,\mathrm{cov}\,(X, Y_2)$,

(3) $\mathrm{cov}\,(X, Y) \leq \big(\mathrm{var}\,(X)\,\mathrm{var}\,(Y)\big)^{1/2}$,

(4) $\mathrm{var}\,(X) = 0 \iff X$ is deterministic.

Hence, $\mathrm{cov}\,(X, Y)$ has similar properties as an *inner product*. (In fact, it is the inner product between $X - \mathrm{E}X$ and $Y - \mathrm{E}Y$.) Using the inner product notation $\langle X, Y\rangle = ||X||\;||Y||\cos\phi$, where ϕ is the angle between the vectors X and Y, one can relate:

probability theory	geometry
$\mathrm{cov}\,(X, Y)$	$\langle X, Y\rangle$
$\mathrm{var}\,(X)$	$\|\|X\|\|^2$
$\rho(X, Y)$	$\cos\phi$

Note, however, that the space of random variables is infinite-dimensional. The state of affairs is more precisely expressed by: $\{X | X : \Omega \to \mathbb{R}, \mathrm{var}\,(X) < \infty\}$ is a *Hilbert space* with respect to the inner product $\langle X, Y\rangle = \mathrm{cov}\,(X, Y)$, if one considers equivalence classes of random variables that differ only a deterministic constant. This feature is useful to keep in the back of one's mind when making standard manipulations in practical calculations.

Exercise 1.15. Prove that $\mathrm{var}\,(X) = \mathrm{var}\,(X + a)$ and $\mathrm{cov}\,(X, Y) = \mathrm{cov}\,(X + a, Y + b)$ for all real numbers a and b, and relate this to the sentence about the Hilbert space.

Exercise 1.16. Extend the mechanical analogue to two-dimensional distributions.

Theorem 1.4. *Suppose X and Y are independent random variables with probability densities $f_X(x)$ and $f_Y(y)$, respectively. Then their sum $Z = X + Y$ has probability density*

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(u) f_Y(z-u)du = \int_{-\infty}^{+\infty} f_X(z-u) f_Y(u)du\,. \tag{1.35}$$

Proof. While drawing a picture, it is easily seen that

$$P\{Z < z\} = \int_{-\infty}^{z-x}\int_{-\infty}^{+\infty} f_X(x) f_Y(y)dxdy = \int_{-\infty}^{z-y}\int_{-\infty}^{+\infty} f_X(x) f_Y(y)dydx\,. \tag{1.36}$$

The result is obtained by differentiating with respect to z.

Remarks.

1. For brevity the notation $f_Z = f_X \otimes f_Y$ is used, $\otimes$ being called the convolution product. The space of absolutely integrable probability

density functions ('L^1') is a linear vector space with norm $||f|| = \int_{-\infty}^{+\infty} |f(x)|\, dx$. (No distinction is made between functions which differ between each other on a set of Lebesgue measure zero.) With the convolution product, it constitutes a normed ring and even a so-called (real) Banach algebra, see [568].

2. In Sect. 1.6, the characteristic function of a random variable is introduced as the Fourier transform of its probability density, entailing that the characteristic function of the sum of two *independent* random variables is the point-wise product of their characteristic functions, which is useful to establish theoretical results and, sometimes, also for practical calculations.

1.5 Parametric Families of Probability Distributions

A number of probability distributions, which are quite frequently used in statistics, are given in Table 1.2. The list is by no means exhaustive. All distributions in Table 1.2 happen to be unimodal, which means that each probability density possesses one maximum only. In this section, we simply explain the table and briefly describe distributions on the sphere. In the next section, a few practical examples and convenient defining properties of the main distributions are given. The chapter is concluded by discussing the unifying concept of an exponential family. A more extensive collection of families of distributions and further properties of the distributions discussed here, which are not described within the limits of this chapter, can be found in [378], Chap. 7 of [540], and [328, 330, 331, 395].

In addition to the skewness and (excess of) kurtosis, Table 1.2 also lists, for each entry, the moment generating function $\int_{-\infty}^{\infty} e^{sx} dF(x)$, which is the Laplace-transform of the corresponding (generalised) probability density $f(x) = F'(x)$. Together with the special functions that occur in Table 1.2, the moment generating functions are discussed further in Sect. 1.6. If the skewness and kurtosis are 'small', then one may approximate the corresponding distribution by a normal distribution with the same expectation value and the same variance. This can often be supported by the central limit theorem, which will be discussed in Sect. 1.7.

Exercise 1.17. Draw roughly the densities of $B(10, \frac{1}{2})$, $\mathcal{P}(5)$, $N(1, 3)$, $E(2)$, χ^2_2, χ^2_{20}. Interpret the columns of the skewness and kurtosis when various parameters go to infinity.

The distributions in Table 1.2 are all on the real line. It is also interesting to consider probability distributions on other topological spaces, such as for instance the surface of the p-dimensional hypersphere, $S^{p-1} = \{x : \sum_{i=1}^{p} x_i^2 = 1\}$, which can be expressed in $p-1$ polar coordinates, see Chap. 15 of [442]. Spherical distributions have many applications, obviously, but not exclusively

Table 1.2.

Density, expectation, variance, skewness, (excess of) kurtosis and moment generating function of some probability distributions.

name symbol	range E(X)	density var (X)	$\gamma_1(X)$	$\gamma_2(X)$	$M(s)$
Binomial $B(n,p)$	$\{0,1,\ldots,n\}$ np	$p_k=\binom{n}{k}p^k(1-p)^{n-k}$ $np(1-p)$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{1}{np(1-p)}-\frac{6}{n}$	$(1-p+pe^s)^n$
Neg. Binomial NB(n,p)	$\{0,1,2,\ldots\}$ $\frac{np}{1-p}$	$p_k=\binom{n+k-1}{n-1}p^k(1-p)^n$ $\frac{np}{(1-p)^2}$	$\frac{1+p}{\sqrt{np}}$	$\frac{1+4p+p^2}{np}$	$\left(\frac{1-p}{1-pe^s}\right)^n$
Poisson $\mathcal{P}(\mu)$	$\{0,1,2,\ldots\}$ μ	$p_k=e^{-\mu}\mu^k/k!$ μ	$1/\sqrt{\mu}$	$\frac{1}{\mu}$	$\exp\big(\mu(e^s-1)\big)$
Hypergeometric $H_{N,a,n}$	$\{\max(0,N-n-a),\ldots,\min(a,n)\}$ $\frac{an}{N}$	$p_k=\frac{\binom{n}{k}\binom{N-n}{a-k}}{\binom{N}{a}}$ $\frac{an(N-a)(N-n)}{N^2(N-1)}$	$\sqrt{\frac{(N-1)(N-2a)^2(N-2n)^2}{an(N-a)(N-n)(N-2)^2}}$	$\frac{N-1}{(N-2)(N-3)}(C_{a,n,N})$	$\frac{{}_2F_1(-n,-a,N-a-n+1;e^s)}{{}_2F_1(-n,-a,N-a-n+1;1)}$
Exponential $E(\lambda)$	$[0,\infty)$ λ	$\frac{1}{\lambda}e^{-x/\lambda}$ λ^2	2	6	$(1-\lambda s)^{-1}$
Chi-square χ^2_f	$[0,\infty)$ f	$\frac{1}{2^{f/2}\Gamma(f/2)}x^{\frac{1}{2}f-1}e^{-x/2}$ $2f$	$2\sqrt{2}/\sqrt{f}$	$\frac{12}{f}$	$(1-2s)^{-f/2}$
Student's t_f	$\mathbb{R}$ 0	$\frac{1}{\mathscr{B}(1/2,f/2)\sqrt{f}}\frac{1}{\sqrt{(1+x^2/f)^{f+1}}}$ $\frac{f}{f-2}$	0	$\frac{6}{f-4}$	$\frac{2^{(1-f/2)}}{\Gamma(f/2)}\,(\lvert s\rvert\sqrt{f})^{f/2}\,K_{\frac{f}{2}}\,(\lvert s\rvert\sqrt{f})$
Gamma $\Gamma_{f,g}$	$[0,\infty)$ fg	$\frac{1}{g^f\Gamma(f)}x^{(f-1)}e^{-x/g}$ fg^2	$2/\sqrt{f}$	$\frac{6}{f}$	$(1-gs)^{-f}$
Beta (I) $\mathcal{B}e_{f,g}$	$[0,1]$ $\frac{f}{f+g}$	$\frac{1}{\mathscr{B}(f,g)}x^{f-1}(1-x)^{g-1}$ $\frac{fg}{(f+g)^2(1+f+g)}$	$\frac{-2(f-g)}{2+f+g}\sqrt{\frac{1+f+g}{fg}}$	$\frac{6}{fg}\frac{f(f+1)(f-2g)+g(g+1)(g-2f)}{(f+g+2)(f+g+3)}$	${}_1F_1(f,f+g;s)$
F(isher) $F_{f,g}$	$[0,\infty)$ $\frac{g}{g-2}$	$\frac{1}{\mathscr{B}(f/2,g/2)}(\frac{f}{g})^{f/2}\frac{x^{(f/2-1)}}{(1+fx/g)^{(f+g)/2}}$ $\frac{g^2}{(g-2)^2}(\frac{f+2}{f}\frac{g-2}{g-4}-1)$	$2\sqrt{\frac{2}{f}}\frac{g+2(f-1)}{g-6}\sqrt{\frac{g-4}{g+f-2}}$	$\frac{12}{f}\frac{(g-2)(g-4)}{(g-6)(g-8)}(D_{f,g})$	${}_1F_1(\frac{f}{2},1-\frac{g}{2};-\frac{g}{f}s)$

Table 1.2 (cont.).

Density, expectation, variance, skewness, (excess of) kurtosis and moment generating function of some probability distributions.

name symbol	range E(X)	density var (X)	$\gamma_1(X)$	$\gamma_2(X)$	$M(s)$
Normal $N(\mu,\sigma^2)$	$\mathbb{R}$ μ	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ σ^2	0	0	$\exp(\mu s+\frac{1}{2}\sigma^2 s^2)$
Inv. Normal IN(μ,λ)	$[0,\infty)$ μ	$\frac{1}{\sqrt{2\pi x^3/\lambda}}e^{-\frac{\lambda}{2\mu^2}\frac{(x-\mu)^2}{x}}$ μ^3/λ	$3\sqrt{\mu/\lambda}$	$\frac{15\mu}{\lambda}$	$\exp\Big(\frac{\lambda}{\mu}(1-\sqrt{1-\frac{2\mu^2}{\lambda}s})\Big)$
Nm. Inv. Gauss. NIG(α,β)	$[0,\infty)$ $\frac{\beta/\alpha}{(1-(\beta/\alpha)^2)^{\frac{1}{2}}}$	$\frac{\alpha}{\pi}e^{(\sqrt{\alpha^2-\beta^2}+\beta x)}\frac{K_1(\alpha\sqrt{1+x^2})}{\sqrt{1+x^2}}$ $\frac{1}{\alpha(1-(\beta/\alpha)^2)^{\frac{3}{2}}}$	$\frac{3\beta/\alpha}{\alpha^{\frac{1}{2}}(1-(\beta/\alpha)^2)^{\frac{1}{4}}}$	$\frac{3\big(1+4(\beta/\alpha)^2\big)}{\alpha(1-(\beta/\alpha)^2)^{\frac{1}{2}}}$	$\exp\Big(\sqrt{\alpha^2-\beta^2}-\sqrt{\alpha^2-(\beta+s)^2}\Big)$
Beta Logistic $\mathcal{B}e\mathcal{L}o_{f,g}$	$\mathbb{R}$ $\psi(f)-\psi(g)$	$\frac{1}{\mathscr{B}(f,g)}\frac{e^{fx}}{(1+e^x)^{f+g}}$ $\psi'(f)+\psi'(g)$	$\frac{\psi''(f)-\psi''(g)}{\big(\psi'(f)+\psi'(g)\big)^{3/2}}$	$\frac{\psi'''(f)+\psi'''(g)}{\big(\psi'(f)+\psi'(g)\big)^2}$	$\frac{\Gamma(f+s)\Gamma(g-s)}{\Gamma(f)\Gamma(g)}$

The Gamma function $\Gamma(f)=\int_0^\infty t^{f-1}e^{-t}dt$ satisfies $\Gamma(f+1)=f\Gamma(f)$, with $\Gamma(\frac{1}{2})=\sqrt{\pi}$ and $\Gamma(1)=\Gamma(2)=1$. The Beta function $\mathscr{B}(f,g)=\int_0^1 x^{f-1}(1-x)^{g-1}dx$ can alternatively be defined as $\mathscr{B}(f,g)=\Gamma(f)\Gamma(g)/\Gamma(f+g)$, and $\psi(f)=(\partial/\partial f)\log\Gamma(f)$ is the digamma function. These special functions, see [732], are defined for any real (and even complex) f and g, with the exception of zero and the negative integers. See Sect. 2.2 for simple analytic approximations such as $\psi'(f)=(\partial/\partial f)\psi(f)\simeq\frac{1.06}{f+\frac{1}{2}}-\frac{0.06}{f}+\frac{1}{f^2}$ $(f>\frac{2}{3})$. In many practical applications, f and g can be restricted to the set of positive integers and are called degrees of freedom. In the table, the excess of kurtosis for the hypergeometric distribution contains the factor $C_{a,n,N}=3(\frac{N-2n}{N-n})+\frac{N^2\big(N(N+1)-6n(N-n)-6a(N-a)\big)}{an(N-a)(N-n)}$, and of the F–distribution the factor $D_{f,g}=\Big(\frac{g-2}{f+g-2}+\frac{f(5g-22)}{(g-2)(g-4)}\Big)$. The notation $X\sim N(\mu,\sigma^2)$ is used to indicate that X has a normal distribution with mean μ and variance σ^2.

so, in astronomy and geology. Compared with distributions on $\mathbb{R}^p$, they have their own intricacies. We discuss here in brief the simple, but interesting, special case of the unit circle $S^1 = \{(r,\theta)|r = 1, 0 \leq \theta < 2\pi\}$. The probability density of the *von Mises distribution* centered around an angle θ_0 is $\left(2\pi I_o(\kappa)\right)^{-1} e^{\kappa cos(\theta-\theta_0)}$. For $\kappa = 0$, the distribution is uniform. The larger the value of κ, the more the distribution is concentrated around θ_0. The expectation value of this distribution is a vector. Its direction is determined by the angle θ_0 and its magnitude by $A(\kappa) = I_1(\kappa)/I_0(\kappa)$, which is the ratio between two modified Bessel functions of the first kind, see [318, 670, 732]. For any integer $m = 0, 1, 2, \ldots$, $I_m(\kappa)$ is defined by

$$I_m(\kappa) = \frac{(\frac{1}{2}\kappa)^m}{\Gamma(m+\frac{1}{2})\Gamma(\frac{1}{2})} \int_0^\pi e^{\kappa\cos\theta} \sin^{2m}\theta d\theta \,. \tag{1.37}$$

For large values of κ, we have the asymptotic approximation

$$I_m(\kappa) \simeq \frac{1}{\sqrt{2\pi\kappa}}\, e^\kappa \left(1 - \frac{4m^2-1}{8\kappa}\right), \tag{1.38}$$

and hence

$$A(\kappa) \simeq 1 - \frac{1}{2\kappa}, \tag{1.39}$$

which is in practice already reasonably accurate for $\kappa \geq 3$. The variance of the von Mises distribution equals

$$\frac{1}{2\pi I_0(\kappa)} \int_{\theta_0-\pi}^{\theta_0+\pi} (\theta-\theta_0)^2 e^{\kappa\cos\theta} d\theta \tag{1.40}$$

and is approximated by $1/\kappa$. The skewness of the von Mises distribution is zero, because its probability density is symmetric around the direction determined by the angle θ_0.

Circular and p-dimensional spherical distributions have been used in [418] in the context of dipole orientations in a magnetic field, while in [708] circular distributions are used to describe the fractional part of atomic weights, see [58]. References on spherical distributions are [196, 441, 442].

1.5.1 Characterising Properties

Throughout this book, the abbreviation (i.i.d.), applied to a sequence of random variables $X_1, X_2, \ldots, X_n$, is used to indicate that the random variables are 'independent and identically distributed'. The *binomial distribution* occurs as the discrete probability distribution of the number of 'successes' or 'failures' in n independent Bernoulli experiments. A Bernoulli experiment has two possible outcomes: 'success' (or 'head', or 1, or ...) and 'failure' (or 'tail', or '0', or ...) with probability p and $1-p$, respectively.

Exercise 1.18. Play simplified 'fair' roulette, defined as

$$(p_{rouge}, p_{noir}, p_{banque}) = (18/37, 18/37, 1/37)\,. \tag{1.41}$$

Determine the probability that, if you play 10 times, you get at least 8 times rouge. Determine also the probability that you get at most 2 times rouge. Of course if p_{banque} would be zero and $p_{noir} = p_{rouge}$, then the two probabilities would be equal. Explain why now the first probability is lower than the second one.

The *negative binomial distribution* occurs as the (discrete) probability distribution of the number of failures k in a series of independent Bernoulli experiments which has been stopped at the nth success. As for the Poisson distribution, this number can take any integer value $0, 1, 2, 3, \ldots$, without the upper bound that occurs in the binomial distribution. Unlike as for the Poisson distribution, the variance of the negative binomial distribution can be varied independently from, but is always as least as large as, the expectation value. Therefore, this distribution has been used in situations where overdispersion is expected compared to the Poisson distribution. For instance, in numismatics, it has been employed as a model for the number of die varieties of which $1, 2, 3, \ldots$ coins have survived since antiquity [633]. The assumption of a binomial distribution allows one to estimate even the number of varieties of which no coin has actually been found. For further applications, approximations and methods of parameter estimation, the reader is also referred to [331].

The *Poisson distribution*, see [189,641], is the limit distribution of $B(n, p)$ when $n \to \infty$ and $p \to 0$, such that $np \to \mu$, $0 < \mu < \infty$. It has been applied in a variety of different contexts, such as the number of stars that can be observed within a certain solid angle, the number of disintegrations of a radioactive source during a certain time interval, and the number of deaths per regiment–year in Prussia due to horse-kick [707].

The *hypergeometric distribution* occurs when 'binomial sampling' is performed from finite populations without replacing the drawn objects. For instance, in quality control, when a lot contains N objects, of which a are defective, then the number of defectives k from a random sample of size n from this lot has a hypergeometric $H_{N,a,n} = H_{N,n,a}$ distribution. The experiment is conveniently summarised in a 2×2 table, see 1.3. The hypergeometric distribution has been used to estimate the total number N of fish in a pond from a capture-recapture experiment: A number a of fish is captured, marked and returned to the pond. After a suitable elapse of time, a number n of fish is recaptured, k of which turn out to be marked. Under simplifying assumptions concerning randomness, k has a $H_{N,a,n}$ distribution with expectation value $\frac{an}{N}$ and, hence, N can be estimated by $\frac{an}{k}$. The classical hypergeometric distribution can be generalised, among others by considering inverse sampling, in a similar way as described above for the negative binomial distribution. For more information, the reader is referred to [331].

Table 1.3. Two times two table: under random sampling from a finite population N (without replacement) with fixed marginals a and n, the integer k varies according to a $H_{N,a,n}$ hypergeometric distribution.

k	a-k	a
n-k	N-a-n+k	N-a
n	N-n	N

The *exponential distribution* is often used to describe a distribution of lifetimes, for instance of light-bulbs, or radioactive nuclei. One interesting property is that if $T \sim E(\tau)$, then $P\{T > t + a | T > a\} = P\{T > t\}$ for any positive number a, hence, the age of a particle does not influence its probability of survival. This characterising feature of the exponential distribution makes it less suitable to be used in animal lifetime studies. Another property, used in evaluating measurements of radioactive decaying materials, is the following: Assume n particles are decaying independently, with $E(\tau)$ distributed lifetimes. (Hence, the expected lifetime of each individual particle is τ.) Let $N(t)$ be the number of decays which one observes in the time interval $(a, a+t)$ for any $a \in \mathbb{R}$, which denotes the starting time of the measurement. If $n \to \infty$, $\tau \to \infty$, such that $\tau/n \to \tau_0$, then on average, one decay from the n particles is expected in a time interval of length τ_0, and the distribution of $N(t)$ tends to $\mathcal{P}(t/\tau_0)$. Hence, $N(t)$ is an estimate of $t/\tau_0 = nt/\tau$. (The expectation value as well as the variance of $N(t)$ are equal to $t/\tau_0 = nt/\tau$.) Therefore, one can estimate the number of radioactive particles if one knows the decay time, and the other way around. The parameter λ in Table 1.2 is the decay length of the exponential distribution. Its inverse is often called the intensity in statistical literature and there sometimes also denoted by λ, contrary to the convention used here. Using this statistical notation in the example of radioactive decay, we would have $N(t) \sim \mathcal{P}(\lambda t)$ with intensity $\lambda = 1/\tau_0$.

For other types of discrete distributions, the reader is referred to [331,378]. A diagram denoting their interrelationship in various limiting situations is given in Chap. 5 of Vol. 1 of [378].

The following characterising properties of the χ^2, t, $\mathcal{B}e$, F and $\mathcal{B}e\mathcal{L}o$ distributions can be viewed as simple formal, practical definitions. (We gloss over the precise domain of the parameters which in many cases has to be generalised from $\mathbb{N}$ to a subset of $\mathbb{R}$.) Alternatively, when the probability densities in Table 1.2 are taken as definitions, these properties are to be viewed as theorems. They can conveniently be derived by employing features of characteristic functions, see Sect. 1.6. A statistical motivation for considering these distributions is given in Sect. 2.4.

Definition 1.17. (χ^2 *distribution*) *If* $X_i \sim N(0,1)$ *(i.i.d.),* $i = 1, \ldots, n$ *(in words: the random variables* X_i *are independent and have a standard normal*

distribution) *then*

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2 \ . \tag{1.42}$$

Sometimes a two-parametric χ^2 family is introduced, where $\chi^2_{f,\sigma^2} = \Gamma_{f/2,2\sigma^2}$. However, in this text we adhere to the equivalent notation $Y \sim c\chi^2_n$ for $Y/c \sim \chi^2_n$. Hence, if $X_i \sim N(0,\sigma^2)$, $i = 1,\ldots,n$, are independent, then we write $\sum_{i=1}^n X_i^2 \sim \sigma^2\chi^2_n$ instead of $\sum_{i=1}^n X_i^2 \sim \chi^2_{n,\sigma^2}$.

Exercise 1.19. Consider a classical 3-dimensional Maxwellian (isotropic, noninteracting) gas. The velocity components of a single particle are denoted by V_x, V_y, V_z. According to this model, $V_x \sim N(0,\sigma^2)$ with $\sigma^2 = kT/m$. The components V_y and V_z have the same distribution and are independent. (In fact, the velocity components of all particles are considered to be independent.) What is the distribution of the kinetic energy of a single particle? Determine the expectation value, the variance and the skewness. Perform the same for the kinetic energy of n particles.

Definition 1.18. (*Gamma distribution*) *If* $Y_1, Y_2, \ldots, Y_f \sim E(\lambda) = \Gamma(1,\lambda)$, *with* Y_1, Y_2, ..., Y_f *independent, then*

$$\sum_{i=1}^{f} Y_i \sim \Gamma_{f,\lambda} \ . \tag{1.43}$$

Definition 1.19. (*Beta distribution*) *If* $Y_1 \sim \Gamma(f,\lambda) = \frac{\lambda}{2}\chi^2_{2f}$ *and* $Y_2 \sim \Gamma(g,\lambda)$, *with* Y_1 *and* Y_2 *independent, then*

$$\frac{Y_1}{Y_1 + Y_2} \sim \mathcal{B}e_{f,g} \ . \tag{1.44}$$

Definition 1.20. (*Student's t distribution*) *If* $X \sim N(0,1)$ *and* $Z \sim \chi^2_f$, *with* X *and* Z *independent, then*

$$\frac{X}{\sqrt{Z/f}} \sim t_f \ . \tag{1.45}$$

Definition 1.21. (*F distribution*) *If* $Y_1 \sim \chi^2_f$ *and* $Y_2 \sim \chi^2_g$, *with* Y_1 *and* Y_2 *independent, then*

$$\frac{Y_1/f}{Y_2/g} \sim F_{f,g} \ . \tag{1.46}$$

Definition 1.22. (*'Beta logistic' distribution*) *If* $X \sim \mathcal{B}e_{f,g}$, *then*

$$Y = \log\frac{X}{1-X} \sim \mathcal{B}e\mathcal{L}o_{f,g} \ , \tag{1.47}$$

which we call here the Beta logistic distribution.

The latter distribution derives its name from the logistic (or: logit) transformation $y = \log \frac{x}{1-x}$, which is is an easily invertible, differentiable bijection from $(0,1)$ to $(-\infty,+\infty)$. For $f = g = 1$, the Beta distribution is just the uniform distribution and the standard (or: uniform) logistic distribution is recovered, for which $f(y) = F(y)\big(1 - F(y)\big)$. For $f = g$ the distribution of Y is symmetric and when both f and g tend to infinity, $Y/\sqrt{\frac{1}{f} + \frac{1}{g}}$ tends to the standard normal. In [524] this class of distributions was used to describe dose levels in bio-assay. Obviously, $\frac{X}{1-X} \sim \frac{f}{g} F_{2f,2g}$, and $\frac{1}{2}(\log \frac{g}{f} + \mathcal{B}e\mathcal{L}o_{\frac{f}{2},\frac{g}{2}})$ is known as Fisher's $z_{f,g}$ distribution [199, 204]. In [337] the distribution of $Z \sim \mu + \sigma \log \frac{g}{f} \frac{X}{1-X}$ is called the generalised F distribution, and used as a flexible parametric model in survival analysis. For further properties, see also [329, 497, 656, 746].

By solving analytically $d/dx \log f_\theta(x) = 0$, one can easily derive that the modes of the χ^2_f, $\Gamma_{f,g}$, $\mathcal{B}e_{f,g}$, $F_{f,g}$ and $\mathcal{B}e\mathcal{L}o_{f,g}$ distributions are $f-2$, $g(f-1)$, $\frac{f-1}{f-1+g-1}$, $\frac{g}{f}\frac{f-2}{g+2}$, and $\log \frac{f}{g}$, respectively. The Pearson measures of skewness [508], invariant under scale transformations, are $\gamma_{KP,c} = c \times \frac{\mu - M}{\sigma}$, where M stands for the mode of the distribution, $\mu = \mathrm{E}(X)$, and c is some proportionality constant. The continuous distributions in Table 1.2 and experience from actual datasets[7] suggest 2 to be a reasonable practical convention and the interval $[1,3]$ to be a convenient range for c such that as a rule, but not always, $\gamma_{KP,c}$ tends to be numerically rather close to the moment-based skewness $\gamma_1(X)$, even though the ratio is undefined, of course, for any symmetric distribution. For the family of $\Gamma_{f,g}$ distributions, with $E(\lambda)$ and χ^2_f as special cases, $\gamma_1(X)/\gamma_{KP,2}(X) = 1$. For $\mathcal{B}e_{f,g}$, $F_{f,g}$, and $\mathcal{B}e\mathcal{L}o_{f,g}$, one obtains $\gamma_1(X)/\gamma_{KP,2}(X) = \frac{f+g-2}{f+g+2}$, $\frac{g+2}{g-6}$, and (approximately) $\frac{f^{-1}+g^{-1}}{(f+\frac{1}{2})^{-1}+(g+\frac{1}{2})^{-1}}$, respectively, which are close to 1 for reasonably large values of f and g. For the inverse normal distribution, the ratio $\gamma_1(X)/\gamma_{KP,1}(X)$ varies between 1 and 3, depending on $\frac{\lambda}{\mu}$. Discreteness of the probability distribution complicates the Pearson measure of skewness somewhat. For the Poisson distribution with an integer value of the parameter μ, $\gamma_{KP,c} = 0$. For the hypergeometric distribution $\gamma_1(X)/\gamma_{KP,2}(X) \simeq \frac{N+2}{N-2}$, see [331].[8] The reader is referred to [434] for an overview of various alternative measures of skewness for univariate distributions.

Exercise 1.20. Suppose a population of particles has an isotropic non-Maxwellian distribution with heavy tails, modeled by $V_x/\sigma \sim t_g$, where σ is a scale parameter. Derive that now the temperature, defined as the average energy per particle, equals $[g/(g-2)]\sigma$. What is in this case the distribution

[7] A single instance from a dozen in [508] is the length-breadth index of 900 (male and female) Bavarian skulls, excavated around 1880 [533].

[8] For convenience, the mode of the hypergeometric distribution, which is the largest integer less than $c = (a+1)(n+1)/(N+2)$ and both $c-1$ and c in case c is an integer, has been replaced here by $c - 1/2$.

of the kinetic energy per particle, and that of the average kinetic energy of n particles?

The *inverse normal distribution* derives its name, see [690], from the fact that its cumulant generating function (c.g.f., which is the logarithm of the moment generating function, see Sect. 1.6) is, modulo the sign, the inverse of the c.g.f., restricted to the interval $[0, \infty)$, of the Gaussian distribution, i.e.,

$$-K_{N_{(\mu,\sigma)}}(-K_{IN_{(\mu^{-1},\sigma^{-2})}})(s) = s \ .$$

Occurring, among others, as the distribution of the first passage time, with respect to a fixed boundary, of a Brownian motion with a constant drift, it has been investigated near the beginning of the twentieth century by Bachelier [25] (in relation to stock-market prices), and also (in relation to physical problems) by Schrödinger [596] and Tweedie [690]. Furthermore, it has been derived as the distribution of the (random) sample size in Wald's sequential probability-ratio test, see [719, Appendix 6]. The inverse Gaussian distribution, and its generalisation to a three parametric family which contains the χ^2_f distribution as a special case [32], possesses a number of other interesting properties as well, see for instance the monographs [333] and [608].

Exercise 1.21. We formally define $|\chi_1|^p$ to be the distribution of $|X|^p$ for $X \sim N(0,1)$ and $p \in \mathbb{R}\backslash\{0\}$. Derive that the probability density of $Y \sim |\chi_1|^p$ equals

$$f(y) = \sqrt{\frac{2}{\pi}}|p|^{-1}y^{\frac{1}{p}-1}e^{-\frac{1}{2}y^{\frac{2}{p}}} \tag{1.48}$$

and that the mth moment around the origin equals

$$\frac{\Gamma(\frac{1}{2}+\frac{mp}{2})}{\sqrt{\pi}}2^{\frac{mp}{2}}. \tag{1.49}$$

For mp an even integer, this equals $(mp-1)(mp-3)\cdots 1$, while for mp an odd integer, it is $(mp-1)(mp-3)\cdots 2\sqrt{\frac{2}{\pi}}$. Check that $\mathrm{E}Y^m = \mathrm{E}|x|^{mp}$. Look at the special cases $p=2$ and $p=-1$, the latter being somewhat more difficult. The skewness and kurtosis can be formally expressed in terms of the Γ function, but the expressions do not simplify very much, unless mp is an integer.

1.5.2 Normal Approximations

From Table 1.2, one can see that χ^2_f is, for some purposes, reasonably well approximated by $N(f, 2f)$ for $f > 100$, say. By making transformations, one can improve the symmetry and obtain more accurate normal approximations for moderately sized samples. For instance, it has been shown (before the advent of powerful computer software) that, to a good practical approximation,

$$(\chi^2_f/f)^{1/3} \simeq N\big(1 - \frac{2}{9f}, \frac{2}{9f}\big) \tag{1.50}$$

with skewness $\gamma_1 = 4.74 f^{-3/2} + o(f^{-3/2})$, instead of the usual $o(f^{-1/2})$, and kurtosis $\gamma_2 = \frac{-4}{9f} + o(f^{-1})$. Similarly, as has originally been shown by Fisher, the distribution of

$$Z_{f,g} = \frac{1}{2} \log F_{f,g} \tag{1.51}$$

is, to order $o(f^{-1})$ and $o(g^{-1})$, rather well approximated by $N\big(\frac{1}{2}(\frac{1}{g} - \frac{1}{f}), \frac{1}{2}(\frac{1}{g} + \frac{1}{f})\big)$. Obviously, $Z_{f,g} = \frac{1}{2}(\log \frac{g}{f} + \mathcal{B}e\mathcal{L}o_{\frac{f}{2},\frac{g}{2}})$, where $\mathcal{B}e\mathcal{L}o$ stands for a Beta-logistically distributed random variable, see Table 1.2. For more information and additional approximations, see, e.g., [328, 378].

1.5.3 Exponential Families

Except for the NIG (Normal Inverse Gaussian) and the F distribution, all distributions displayed in Table 1.2 belong to a special class of parametric distributions called exponential family. Mathematically they exhibit a particular structure, in that the first two derivatives of a normalising function $\psi(\theta)$ yield the first two moments of the associated random variables.[9] Under some regularity conditions, they are those distributions which admit estimation of the parameters by sufficient statistics. This is known as the Fisher–Koopman–Darmois–Pitman theorem. After Fisher's paper [201] which treats the case of a one-dimensional parameter θ, properties of multi-parameter exponential families have been derived, more or less independently, by Koopman [392], Pitman [516] and Darmois [133] during the late thirties. References to more recent literature are [28, 30, 34, 164, 191, 325, 395, 426], among others. For didactical reasons, we discuss the simple case of one parameter first.

Definition 1.23. *In canonical form, a (one-dimensional and univariate) exponential family, consists of probability distributions with (generalised) density*

$$f_\theta(x) = h(x) e^{\theta x - \psi(\theta)} \; . \tag{1.52}$$

In practice, we consider $h(x) = h_{ac}(x) + \sum_{i=1}^{\infty} h_i \delta(x - x_i)$, where h_{ac} is absolutely continuous with respect to the Lebesgue measure.[10] Since $\int_{-\infty}^{+\infty} f_\theta(x) dx = 1$, the normalisation constant (depending on θ) is equal to

$$\psi(\theta) = \log \int_{-\infty}^{+\infty} h(x) e^{\theta x} dx \; , \tag{1.53}$$

[9] The normalising function $\psi(\theta)$ is different from the digamma function $\psi(f) = (\partial/\partial f) \log \Gamma(f)$ used in Table 1.2.

[10] Somewhat more generally, we can identify $h(x)dx$ by $dF(x)$ with $F(x)$ a monotonic function or by $d\mu(x)$ with $\mu(x)$ a positive measure on $\mathbb{R}$.

which is naturally required to be less than infinity. One can derive directly (by using Hölder's inequality, see Sect. 1.9) that $\psi(\theta)$ is a convex function on its domain $\mathfrak{N}_h = \{\theta; \int_{-\infty}^{\infty} e^{\theta x} h(x) dx < \infty\}$, called the natural parameter space, which is a convex set (i.e., an interval, a point or an empty set) depending on h. Since we want to consider derivatives of the normalising function $\psi(\theta)$, we assume that $\mathfrak{N}(h)$ is a proper interval, which has a non-zero interior $\mathfrak{N}^0(h)$.[11] In fact, $\psi(\theta)$ is the log Laplace transform of $h(x)$, and by repeated differentiation one can directly derive that for any random variable $X \sim f_\theta(x)$ where $f_\theta(x)$ belongs to an exponential family, and $\theta \in \mathfrak{N}^0(h)$, the expectation value equals $\mathrm{E}(X) = \int x f_\theta(x) dx = \psi'(\theta)$ and $\mathrm{var}\,(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \psi''(\theta)$.

Remarks. It is noted that each fixed (generalised) density $h(x)$ generates an exponential family, hence exponential families constitute a semi-parametric class of distributions. An exponential family as defined above is usually called regular since (a) the interior of $\mathfrak{N}_h$ is a non-empty open set and (b) the support of $f_\theta(x)$ (i.e., for a classical function the closure of the set $\{x; f_\theta(x) \neq 0\}$) does not depend on θ. Counterexamples of the latter are given by $f_\theta(x) = I_{[0,\theta]}$, $f_\theta(x) = e^{-x} I_{[\theta,\infty]}$, etc. The parameter θ is called the natural parameter of the exponential family, and can be viewed as a transformation of another parameter, for instance $\psi'(\theta) = \mathrm{E}(X)$, which has often a more direct interpretation within a statistical model.

Exercise 1.22. Consider $X \sim \mathcal{P}(\mu)$, for which $P(X = k) = e^{-\mu}\mu^k/k! = e^{k \log \mu - \exp \log \mu}/k!$. In this case, $\theta = \log \mu$, $\psi(\theta) = e^\theta = \mu$ and $h(x) = \sum_{k=0}^{\infty} \delta(x - k)$. Derive that $\psi'(\theta) = \psi''(\theta) = \mu$, in agreement with Table 1.2. Analyse in a similar way $X \sim \mathrm{IN}(\mu, 1)$, i.e., the inverse Gaussian distribution with $\lambda = 1$.

Now we consider multi-parametric exponential families. They are fundamental in statistical estimation theory, however not a prerequisite to understand the next sections of this chapter on probability theory. Therefore, the busy reader may skip them at first instance, and return to them later. Mathematically they are in some sense a straightforward generalisation of one-dimensional exponential families. However, the aspect of reparameterisation is more pronounced in many applications. An interesting complication, albeit relevant in statistical practice, arises if the dimensionality of the natural parameter space $\boldsymbol{\Theta}$ is smaller than that of the random variables $\boldsymbol{X} = (X_1, \ldots, X_m)$.

Definition 1.24. *A (multi-dimensional and multivariate) exponential family consists of probability distributions with (generalised) density*

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = h(\boldsymbol{x}) e^{\sum_{i=1}^{m} \varphi_i(\boldsymbol{\theta}) t_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta})} , \tag{1.54}$$

where $\boldsymbol{x} \in \mathfrak{X} \subset I\!R^n$, $\boldsymbol{\theta} \in \Theta \subset I\!R^d$, $t_i \colon I\!R^n \to I\!R$, $\varphi_i \colon \Theta \to I\!R$ *and* $d \leq m$.

[11] A counter example is given by the Cauchy distribution for which $\pi^{-1} \int_{\infty}^{+\infty} e^{\theta x}(1 + x^2)^{-1} dx < \infty$ only if $\theta = 0$, and hence $\mathfrak{N}^0(h)$ is empty.

Remark. In Def. 1.54, $\boldsymbol{\varphi} = \big(\varphi_1(\boldsymbol{\theta}), \ldots, \varphi_m(\boldsymbol{\theta})\big)$ is called the canonical (multi-dimensional) parameter and $\boldsymbol{T} = \big(t_1(\boldsymbol{X}), \ldots, t_m(\boldsymbol{X})\big)$ the associated canonical sufficient statistic. Broadly speaking, T is called sufficient since, with regard to estimating $\boldsymbol{\varphi}(\boldsymbol{\theta})$, all statistically relevant information contained in $\boldsymbol{X} = (X_1, \ldots, X_n)$ is also contained in $\boldsymbol{T} = \big(t_1(\boldsymbol{X}), \ldots, t_m(\boldsymbol{X})\big)$, even if in typical situations m is much smaller than n. For $x \in \mathfrak{X} \subset \mathbb{R}$, the exponential family may be called univariate.

For regular exponential families, $d = m$ and $\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}^{-1}$ are one-to-one and differentiable mappings between

$$\mathfrak{M}_h = \big\{\boldsymbol{\theta}; \int_{-\infty}^{+\infty} e^{\sum_{i=1}^m \varphi_i(\boldsymbol{\theta}) t_i(\boldsymbol{x})} h(x) dx < \infty\big\} \tag{1.55}$$

and

$$\mathfrak{N}_h = \big\{\boldsymbol{\varphi}; \int_{-\infty}^{+\infty} e^{\sum_{i=1}^m \varphi_i(\boldsymbol{\theta}) t_i(\boldsymbol{x})} h(x) dx < \infty\big\} \,. \tag{1.56}$$

We require in that case that the convex set $\mathfrak{N}_h$ is not concentrated on a hyper-surface of a dimension smaller than m and has a non-empty interior (i.e., $\mathfrak{N}_h$ has a positive m-dimensional Lebesgue measure λ_m). Now, ψ can be considered as a function of $\boldsymbol{\theta}$ as well as of $\boldsymbol{\varphi}$. In terms of the natural parameters $\boldsymbol{\phi} = \boldsymbol{\varphi}(\boldsymbol{\theta})$, we have

$$(\partial/\partial\varphi_i)\psi(\boldsymbol{\phi}) = \mathrm{E}\big(t_i(\boldsymbol{X})\big) \tag{1.57}$$

and

$$(\partial/\partial\varphi_i \partial/\partial\varphi_j)\psi(\boldsymbol{\phi}) = \mathrm{var}\left(t_i(\boldsymbol{X}), t_j(\boldsymbol{X})\right) \,, \tag{1.58}$$

i.e., the gradient and the Hessian matrix of ψ with respect to $\boldsymbol{\varphi}$ at the point $\boldsymbol{\phi} = \boldsymbol{\varphi}(\boldsymbol{\theta})$ yield the vector of mean values and the covariance matrix of the associated sufficient statistics $\boldsymbol{t}(\boldsymbol{X})$, respectively. (In fact, higher order derivatives lead to higher order cumulants, see Sect. 1.6.) This property is very convenient for analytic calculations. It is also frequently applied in statistical mechanics, see for example [154, 248, 284, 291, 302, 683], where special exponential families denoted as (micro, grand) canonical ensemble, are utilized and where $Z = e^{\psi(\phi)}$ is called the partition function Z. (Such models have been investigated since Gibbs [227].)

Example. Consider $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. After taking logarithms, we can write

$$\begin{aligned} \log f_\theta(\boldsymbol{x}) &= -\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2) \,, \\ &= -\tfrac{1}{2}\sum_{i=1}^n \frac{x_i^2}{\sigma^2} + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \tfrac{1}{2}n\big(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2)\big) \,. \end{aligned} \tag{1.59}$$

Comparison with (1.54) provides $\big(\varphi_1(\boldsymbol{\theta}), \varphi_2(\boldsymbol{\theta})\big) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ as a pair of canonical parameters and $\big(T_1(X), T_2(X)\big) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ as associated sufficient statistic. The normalising function is $\psi(\boldsymbol{\theta}) = -\frac{n}{2}\big(\frac{\mu^2}{\sigma^2} - \log(\sigma^2)\big)$, whence $\psi(\boldsymbol{\phi}) = -n\big(\frac{\phi_1^2}{4\phi_2} - \log(-2\phi_2)\big)$. Differentiation of $\psi(\boldsymbol{\phi})$ with respect to the canonical parameters yields $E\sum_{i=1}^n X_i = n\mu$ and $E\sum_{i=1}^n X_i^2 = n(\mu^2 + \sigma^2)$. Notice that there is some freedom in the choice of canonical parameters and associated sufficient statistics. Consider likewise $X_1, \ldots, X_n \sim N(\mu, c\mu^2)$, where c is a fixed constant. Now the natural parameters corresponding to $\big(T_1(X), T_2(X)\big)$ are $\frac{1}{c}\big(\frac{1}{\mu}, -\frac{1}{2\mu^2}\big)$. This traces a parabola in (ϕ_1, ϕ_2) space and is called a *curved exponential family*, see [164, 363].

Exercise 1.23. Consider $X_1, \ldots, X_n$ to be independent, with $X_i \sim f_\theta(x) = h(x)e^{\theta x - \psi(\theta)}$. Prove that the distribution of the sum $S = \Sigma_{i=1}^n X_i$ also belongs to an exponential family with normalising function $e^{\tilde{\psi}(\theta)} = e^{n\psi(\theta)}$. (Hint: Use explicitly the convolution integral from Theorem 1.4, starting with $n = 2$, or use the properties of moment generating functions, described in Sect. 1.6.) Remark. This is sometimes called the *convolution property* of an exponential family.

We restrict now, for simplicity, the attention to the univariate case, $\underline{x} \in \mathfrak{X} \subset \mathbb{R}$, which covers a number of practical situations. The dependence of $\mathrm{var}\,(X)$ on the expectation value $\mu = E(X)$, and when there are many parameters, the dependence of the covariance matrix $\mathbf{var}\,(\boldsymbol{\mu}) = \big(\mathrm{var}\,(t_i(X), t_j(X))\big)_{i,j=1,\ldots,m}$ on $\boldsymbol{\mu} = E\big(t_1(X), \ldots, t_m(X)\big)$, is called the *variance function*. Under quite general regularity conditions, the variance function characterises a sub-family of distributions within the (semi-parametric) class of exponential families, see [428]. Therefore, it plays a prominent role in fitting distributions and in parameter estimation theory, among others in the context of generalised linear models, see [455]. Simple examples, which can directly be read from Table 1.2, are: normal: $\mathrm{var}\,(\mu) = \sigma^2$; Poisson: $\mathrm{var}\,(\mu) = \mu$; Gamma: $\mathrm{var}\,(\mu) = \mu^2/f$; binomial: $\mathrm{var}\,(\mu) = \frac{\mu}{n}(1 - \frac{\mu}{n})$; inverse normal: $\mathrm{var}\,(\mu) = \frac{\mu^3}{\lambda}$. Here the ancillary parameters σ^{-2}, f, n and λ all play a role similar to the 'sample size', at least for independent and identically distributed random variables.

1.6 Exponential Transformations

Exponential transformations are essentially the Laplace and Fourier transforms of the (generalised) probability density of a random variable (or the Laplace and Fourier–Stieltjes transform of its distribution function). It can be apprehended that therefore they provide a convenient theoretical frame and have multifarious, theoretical and practical, applications. There are various equivalent ways of introducing them. In the following definition we use

the notion of being the expectation value of the (parameterised) exponential of a random variable.

Definition 1.25. *The moment generating function of a random variable X is*

$$M(s) = Ee^{sX} \ . \tag{1.60}$$

Remarks.

1. The domain of the function M is $D = \{s \in \mathbb{C} \mid E|e^{sX}| < \infty\}$. It can be shown that D is a strip of the form $\{s|Re(s) \in I\}$, where I is some interval on $\mathbb{R}$ (possibly a single point), that $0 \in D$, and that M is an analytic function on the interior of this strip.
2. For $f = f_{ac} + \sum_k p_k \delta(x - x_k)$ we have

$$M(s) = \int e^{sx} f_{ac}(x)\, dx + \sum_k p_k e^{sx_k} \ . \tag{1.61}$$

Properties:

(1) As far as the moments of X exist, one can write

$$M(s) = 1 + s\,\mathrm{E}X + \frac{s^2}{2!}\mathrm{E}X^2 + \frac{s^3}{3!}\mathrm{E}X^3 + \cdots \ , \tag{1.62}$$

hence $\mu'_r = \mathrm{E}(X^r) = M^{(r)}(0)$, i.e., the moments of X are just the derivatives of the moment generating function at the origin.

(2) Expansion of $\log M(s)$ gives the *cumulants*: The function $K(s) = \log M(s)$ is called the *cumulant generating function.*

$$\log M(s) = \kappa_1 s + \kappa_2 s^2/2! + \kappa_3 s^3/3! + \cdots \ . \tag{1.63}$$

The cumulants can be expressed in the moments. Only the first few relations are simple:

$$\kappa_1 = \mu'_1 \ , \tag{1.64}$$
$$\kappa_2 = \mu'_2 - {\mu'_1}^2 = \mathrm{var}\,(X) \ , \tag{1.65}$$
$$\kappa_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2{\mu'_1}^3 \ , \tag{1.66}$$
$$\kappa_4 = \mu'_4 - 4\mu'_3\mu'_1 - 3{\mu'_2}^2 + 12\mu'_2{\mu'_1}^2 - 6{\mu'_1}^4 \ . \tag{1.67}$$

Unlike the corresponding moments, the cumulants $\kappa_2, \kappa_3, \ldots$ are invariant (while $\kappa_1 = \mu'_1$ transforms trivially, i.e., is equivariant) under location transformations, $\widetilde{X} = X + a$, while, similar to the moments, they are not invariant under scale transformations $\widetilde{X} = cX$. Therefore, the cumulants are traditionally called the *semi-invariants.* For more information about cumulant generating functions, also in a multivariate context, the reader is referred to [378, 454].

Exercise 1.24. Derive that the skewness $\mathrm{E}(X - \mathrm{E}X)^3/\mathrm{var}\,(X)^{3/2}$ equals $\kappa_3/\kappa_2^{3/2}$, and that the excess of the kurtosis $\mathrm{E}(X-\mathrm{E}X)^4/\mathrm{var}\,(X)^2 - 3$ equals κ_4/κ_2^2.

Exercise 1.25. Consider a gas with an anisotropic velocity distribution, modeled by $(V_x/\sigma_\perp) \sim t_f$, $(V_y/\sigma_\perp) \sim t_f$, and $(V_z/\sigma_\parallel) \sim t_g$, where $\perp$ and $\parallel$ are reminiscent to the notion 'parallel' and 'perpendicular' to the magnetic field in a plasma (the model is somewhat artificial). By using the technique of cumulant generating functions, derive the expectation value, variance, skewness and kurtosis of the distribution of the total energy $\mathrm{E} = \frac{1}{2}m(V_x^2+V_y^2+V_z^2)$ per particle, in the simple situation of an anisotropic Maxwellian gas ($f \to \infty$, and $g \to \infty$). For the more general case, with finite f and g, see below.

The moment generating functions of the Beta and Fisher distributions (see Table 1.2) are conveniently expressed in the *confluent hyper-geometric function*

$$ {}_1F_1(\alpha,\beta;z) = 1 + \frac{\alpha}{\beta}z + \frac{\alpha(\alpha+1)}{\beta(\beta+1)}\frac{z^2}{2!} + \cdots \,. \tag{1.68} $$

For the hypergeometric distribution, the moment generating function has been expressed in

$$ {}_2F_1(\alpha,\beta,\gamma;z) = 1 + \frac{\alpha\beta}{\gamma}z + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\frac{z^2}{2!} + \cdots \,. \tag{1.69} $$

In particular,

$$ {}_2F_1(\alpha,\beta,\gamma;1) = \frac{\Gamma(\gamma)\Gamma(\gamma-\alpha-\beta)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)} \,, \tag{1.70} $$

see [244, 385, 732]. The moment generating function of the t distribution involves the *modified Bessel function of the second kind*,[12] $K_\nu(z)$,

$$ K_\nu(z) = \frac{\pi}{2}\,\frac{I_{-\nu}(z) - I_\nu(z)}{\sin(\nu\pi)} \,, \tag{1.71} $$

where

$$ I_\nu(z) = \frac{1}{\Gamma(\nu+1)}\Big(\frac{z}{2}\Big)^\nu e^{-z}\ {}_1F_1(\nu+\frac{1}{2}, 1+2\nu, 2z) \tag{1.72} $$

stands for the *modified Bessel function of the first kind*, which has series expansion

$$ I_\nu(z) = \Big(\frac{z}{2}\Big)^\nu \sum_{k=0}^{\infty} \frac{(z/2)^{2k}}{k!\Gamma(\nu+k+1)} \,, \tag{1.73} $$

[12] Sometimes, somewhat unfortunately, K_ν is called the modified Bessel function of the third kind. Similar to J_ν and Y_ν for the Bessel equation, I_ν and K_ν are two linearly independent solutions of the modified Bessel equation $z^2y'^2 + zy' - (\nu^2 + z^2)y = 0$. For real arguments z, the Bessel functions of the first and second kind are real valued, while those of the third kind (Hankel functions) are complex valued.

while the following Laurent series expansion holds

$$e^{-\frac{1}{2}(t+\frac{1}{t})z} = \sum_{-\infty}^{+\infty} t^n I_n(z) \ . \tag{1.74}$$

Note that $K_\nu(z) = K_{-\nu}(z)$ and $I_\nu(-z) = (-1)^\nu I_\nu(z)$. For integer m, $I_{-m}(z)$ equals $I_m(z)$ and de l'Hôpital's rule has to be applied to (1.71). In that case the series expansion for $K_m(z)$ is somewhat more complicated. From the many integral representations for $K_\nu(z)$, we mention in particular

$$\begin{aligned} K_\nu(z) &= \tfrac{1}{2} \int_0^\infty t^{\nu-1} e^{-\frac{1}{2}z(t+\frac{1}{t})} dt \\ &= \frac{\Gamma(\frac{1}{2})}{\Gamma(\nu+\frac{1}{2})} \Big(\frac{z}{2}\Big)^\nu \int_1^\infty e^{-zt} (t^2-1)^{\nu-\frac{1}{2}} dt \ . \end{aligned} \tag{1.75}$$

For real z, the functions K_ν(z) are monotonically decreasing and tend for large z asymptotically to $K_\nu(z) \sim e^{-z}/\sqrt{z}$. A good intermediate approximation of $K_1(z)$, with a relative error less than 1.5×10^{-4} on the interval $1 \le z \le 12$, is

$$K_1(z) = 1.6362 e^{-z} z^{-0.700+0.0583\log(z)-0.0065\log^2(z)} \ . \tag{1.76}$$

For $m = f/2$ with f an odd integer, $K_m(z)$ is a *spherical Bessel function*, which can be expressed as a finite sum of elementary functions. The reader is referred to [179, 314, 436, 743] for further information.

The (standardised) normal inverse Gaussian distributions form a two-parametric family of distributions on $\mathbb{R}$, for which the skewness and kurtosis are rather simple functions of the parameters α and β. With two additional parameters, μ and δ for location and shape, respectively, a four-parametric family is obtained, $(X-\mu)/\delta \sim \mathrm{NIG}(\alpha, \beta)$, which has a flexibility comparable to that of the family $(\log X - \mu)/\delta \sim F_{f,g}$, and has a convolution property: The sum of independent, normal inverse Gaussian random variables, $X_+ = \sum_{i=1}^n X_i$, for which $(X_i - \mu_i)/\delta_i \sim \mathrm{NIG}(\alpha, \beta)$, is again a normal inverse Gaussian with the same shape parameters α and β and with $\mu_+ = \sum_{i=1}^n \mu_i$ and $\delta_+ = \sum_{i=1}^n \delta_i$. Normal inverse Gaussian distributions have been used, among others, in describing turbulence [211] and volatility in stock-markets [168], see [31].

Definition 1.26. *The characteristic function of a random variable X is*

$$C(t) = M(it) = Ee^{itX} \ . \tag{1.77}$$

Remarks.

1. The characteristic function of X is the Fourier transform of the density of X. Since a probability density is absolutely integrable, the characteristic function always exists for all real t. The characteristic function

determines the probability distribution uniquely. The convergence of a sequence of distribution functions $F_0, F_1, F_2 \ldots$ to a distribution function F is, under very weak regularity conditions, equivalent to the convergence of the associated characteristic functions. It can directly be derived that

(1) $C(0) = 1$,
(2) $|C(t)| \leq 1$ and $C(-t) = C^*(t)$ for all t,
(3) $Y = aX + b \Rightarrow C_Y(t) = e^{ibt} C_X(at)$.

2. If the density of X is concentrated on $\mathbb{R}^+ = [0, \infty)$, then it is especially convenient to use the one-sided Laplace transform $\varphi(s) = M(-s) = \mathrm{E}e^{-sx}$, which is defined in that case on (at least) $\{s \in \mathbb{C} | Re(s) \geq 0\}$. Restricting its domain to $\mathbb{R}^+$ one can avoid calculations in the complex plane.
3. If the density of x is concentrated on $\mathbb{N} = \{0, 1, 2, \ldots\}$, then it is expedient to use the (*probability-*) *generating function* or *z–transform* $G(z) = \mathrm{E}z^X = M(\ln z)$, which is defined on (at least) the unit disk $\{z \in \mathbb{C} | \, |z| \leq 1\}$. If $p_0, p_1, p_2, \ldots$ are the probabilities of a distribution on $\mathbb{N}$ then $G(z) = p_0 + p_1 z + p_2 z^2 + \cdots$. Given $G(z)$, one can recover these probabilities by

$$p_k = \frac{1}{k!} (\partial^k / \partial z^k) G(z) \Big|_{z=0} , \tag{1.78}$$

and obtain the moments by

$$\mu'_k = \big(z \frac{\partial}{\partial z}\big)^k G(z) \Big|_{z=1} . \tag{1.79}$$

The following property of characteristic functions is so important that it is stated as a theorem.

Theorem 1.5. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with characteristic functions $C_1(t), C_2(t), \ldots, C_n(t)$. Then $X = \sum_{i=1}^n X_i$ has characteristic function*

$$C(t) = \prod_{i=1}^{n} C_i(t) . \tag{1.80}$$

Exercise 1.26. Derive Theorem 1.5, using Def. 1.26.

Remark. The Fourier transform constitutes a ring isomorphism between $(L^1, \otimes)$ and $(L^1, \cdot)$, where L^1 is the space of all absolutely integrable probability density functions, $\otimes$ denotes the convolution product and $\cdot$ is the point-wise product of two functions. Some problems are elegantly solved by exploiting this ring isomorphism, see [568].

Exercise 1.27. Apply the probability generating function of $B(n,p)$ to prove rigorously the intuitively plausible result (why?) that if $X \sim B(n,p)$, $Y \sim B(m,p)$, and X and Y independent, then $X+Y \sim B(n+m,p)$. Determine the characteristic function of $N(\mu,\sigma^2)$, and prove along this way that if $X_1, X_2, \ldots, X_n$ are independent $N(\mu,\sigma^2)$, then $\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$.

1.7 Central Limit Theorem

We are now in a position to sketch the proof of the Central Limit Theorem, which plays a central role in large sample statistical inference. We discuss here the one-dimensional case only. For multivariate extensions, we refer to [607], and for the central limit theorem with values in more general spaces to [16].

Theorem 1.6. (*Central Limit Theorem*) *Let $X_1, X_2, \ldots$ be independent random variables having the same (arbitrary) probability distribution with (finite) expectation value μ and variance σ^2. Then, the distribution of $Z_n = (X_1 + \cdots + X_n - n\mu)/(\sigma\sqrt{n})$ tends for $n \to \infty$ to the standard normal, i.e., the distribution function of Z_n tends to $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt$.*

Loosely speaking, the central limit theorem tells us that the average, $\overline{X}_n$, of n independent and identically (but arbitrarily) distributed random variables (each having mean μ and variance σ^2) tends, for $n \to \infty$, to a Gaussian distribution (with mean μ and variance $\frac{\sigma^2}{n}$). The central limit theorem still holds when the conditions stated above are slightly relaxed. In addition, various 'Berry–Esseen type' error bounds have been derived between $\Phi(z)$ and the distribution function of Z_n as a function of the sample size n (sometimes also as a function of z). The reader is referred to [55,514,607] for an overview of various refinements, and further references. Applications in statistical mechanics are described in [284,382], among others. Some of the history of the central limit theorem, since DeMoivre [140] derived the Gaussian distribution as a limit of the binomial, can be found in [422,643].

Proof. (sketch) Let $C(t)$ be the characteristic function of $X_i - \mu$. Since $\mathrm{E}(X_i - \mu) = 0$ and $\mathrm{E}(X_i - \mu)^2 = \sigma^2$, we have

$$C(t) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2)\,, \tag{1.81}$$

where $o(t^2)$ denotes a quantity that, for $t \to 0$, tends faster to 0 than does t^2. The characteristic function of $\sigma\sqrt{n}Z_n$ equals $C^n(t)$, and hence that of Z_n equals

$$C^n\big(\frac{t}{\sigma\sqrt{n}}\big) = \big(1 - \frac{1}{2}\frac{t^2}{n} + o\,(\frac{t^2}{n})\big)^n\,. \tag{1.82}$$

For $n \to \infty$ the last expression tends to $e^{-\frac{1}{2}t^2}$, which is just the characteristic function of the standard normal.

This theorem, in its essence simple after the preceding preliminaries, explains in part the important role played by the normal distribution. On the one hand, physical phenomena which can, with some justification, be considered as the additive composition of random constituents, are approximately normally distributed, and on the other hand the normal distribution constitutes a limiting distribution of many other parametric distributions, such as the Poisson, Binomial, χ^2, etc., if their parameters related to the number of observations tend to infinity. In practice, the latter property can also be seen (if not proven, at least surmised) from the vanishing skewness and excess of kurtosis in Table 1.2. The central limit theorem does provide a rigorous proof. Finally, it is noted that the central limit theorem has also been investigated by stochastic simulation on computers. These generally confirm the (qualitative) resume that, in practice, the central limit holds, under quite broad conditions, already for moderate sample sizes (some dozens of observations) inasfar as the central part of the distribution is concerned. Convergence of the distribution of the tail is much more delicate in that (much) larger samples are required and in that the conditions for asymptotic normality to hold are more restricted.

1.8 Asymptotic Error Propagation

In the previous section we derived the limit distribution of the average of n identically distributed random variables. In the next chapter we shall discuss statistical estimation theory and shall see that in many cases an estimator, $\overline{X}_n$, of an unknown parameter μ can be written as 'nearly' an arithmetic average of 'approximately i.i.d.' variables.[13] Due to the central limit theorem, such an estimator $\overline{X}_n$ is a random variable with asymptotically (i.e., for sufficiently large sample sizes) a normal distribution and a variance that tends to zero when the sample size n tends to infinity. The mathematical notation $\overline{X}_n \sim AN(\mu, \sigma^2/n)$ is used to indicate that $(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}})$ converges to $N(0,1)$. In such cases, a simple and useful expression can be derived for the asymptotic distribution of a non-linear function, $f(\overline{X}_n)$, of this estimator. In a number of practical situations a more accurate asymptotic error propagation formula is needed, which is also given below. Obviously, the distribution of $f(\overline{X}_n)$ can be very complicated. The asymptotic results presented here are made for, and hold in practice only in, the situation that f changes 'smoothly' over the region that contains 'most' of the probability mass of the distribution of $\overline{X}_n$. Only the one-dimensional case is discussed. For the multi-dimensional situation, the reader is referred to the literature, especially [607, 658].

[13] A more precise formulation of the words 'nearly' and 'approximately' in this context is a subject of the area of *asymptotic statistics*. For the abbreviation i.i.d., see Sect. 1.5.

Theorem 1.7. (*a*) *If* $f'(\mu) \neq 0$, *then for* $\overline{X}_n \sim AN(\mu, \sigma_n)$ *with* $\sigma_n \to 0$ *for* $n \to \infty$, *asymptotically,*

$$Ef(\overline{X}_n) = f(E\overline{X}_n) = f(\mu) \tag{1.83}$$

and

$$var\ f(\overline{X}_n) = \big(f'(\mu)\big)^2 \sigma_n^2\ , \tag{1.84}$$

where $f'(\mu) = \left(\frac{\partial f}{\partial x}\right)_{x=\mu}$.

(*b*) *If the first* $p-1$ *derivatives of* $f(\mu)$ *are zero, while* $f^{(p)}(\mu) \neq 0$, *then the asymptotic distribution of* $f(\overline{X}_n) - f(\mu)$ *is* $(1/p!) f^{(p)}(\mu) \sigma^p \chi_1^p$, *where* χ_1^p *is the distribution of* Z^p *with* $Z \sim N(0,1)$, $p \in \mathbb{N}$.

Proof. We outline a proof of principle, without a precise description of the regularity conditions needed for the theorem to hold. For a more formal derivation (and a precise formulation of the type of asymptotic convergence) the reader is referred to Chap. 3 of [607]. Consider the Taylor series expansion

$$f(\overline{X}_n) - f(\mu) = f'(\mu)(\overline{X}_n - \mu) + \frac{f''(\mu)}{2!}(\overline{X}_n - \mu)^2 + \cdots\ . \tag{1.85}$$

By taking expectation values of both sides, after having raised them to the first and the second power, one gets

$$\mathrm{E} f(\overline{X}_n) - f(\mu) = \frac{f''(\mu)}{2!}\mathrm{E}(\overline{X}_n - \mu)^2 + \cdots \tag{1.86}$$

and

$$\mathrm{E}\big(f(\overline{X}_n) - f(\mu)\big)^2 = \big(f'(\mu)\big)^2 \mathrm{E}(\overline{X}_n - \mu)^2 + \cdots\ , \tag{1.87}$$

respectively. Since $\mathrm{E}(\overline{X}_n - \mu)^2 = \sigma_n^2 \to 0$ for $n \to \infty$, and the higher order terms tend to zero at least as fast as σ_n^2 for $n \to \infty$, we have the desired asymptotic result. In the more general case ($f'(\mu) = f''(\mu) = \cdots = f^{(p-1)}(\mu) = 0$, but $f^{(p)}(\mu) \neq 0$), the mth moment can be written as

$$\mathrm{E}\big(f(\overline{X}_n) - f(\mu)\big)^m = \Big(\frac{f^{(p)}(\mu)}{p!}\Big)^m \sigma_n^{mp}\ \mathrm{E} Z_n^{mp} + \cdots\ , \tag{1.88}$$

where $Z_n = \big(\frac{\overline{X}_n - \mu}{\sigma_n}\big)$ tends to the standard normal. The terms omitted tend to zero at least as fast as $\sigma_n^{(mp+1)}$.

Specialising to $p = 1$, we see that, asymptotically, all moments of $f(\overline{X}_n)$ are those of a normal distribution with mean zero and standard deviation $|f'(\mu)|\sigma_n$. The reader is invited to draw a picture in order to verify geometrically that $|f'(\mu)|$ is the dilation factor of the standard deviation due to the linear approximation of the function $f(x)$ around μ.

Inasfar as the moments uniquely determine the distribution[14] we have essentially derived the general enunciation of the theorem, at least for continuous random variables. For discrete random variables additional conditions are necessary to avoid degeneracies, for instance of the type $S \sim B(n,p)$ and $f(S/n) = \log(S/n)$, where S/n is an estimator of the parameter p. Now, there is a finite probability for S/n to be zero, for which the function f is infinite, and hence $\mathrm{E}f(S/n)$ is infinite. (We will not derive the precise conditions to avoid such degeneracies here.)

In practice, for moderate sample sizes, $f'(\mu)\sigma_n$, $f''(\mu)\sigma_n^2$ and $f'''(\mu)\sigma_n^3$ can be of comparable magnitude. In that case not only the leading term in the Taylor expansion should be retained, and the error propagation formula becomes more complicated. Quite generally,

$$\mathrm{E}\big(f(\overline{X}_n) - f(\mu)\big)^m = \mathrm{E}\Big(\sum_{j=1}^{\infty} \frac{(\overline{X}_n - \mu)^j f^{(j)}(\mu)}{j!}\Big)^m \,. \tag{1.89}$$

By multinomial expansion, the right-hand side can be written as

$$\lim_{p\to\infty} \sum_{j_1,\ldots,j_p} \frac{m!}{j_1!\cdots j_p!} \prod_{k=1}^{p} \Big(\frac{f^{(k)}(\mu)}{k!}\Big) \sigma_n^{kj_k} \nu_{j_1+2j_2+\cdots+pj_p} \,, \tag{1.90}$$

where $\nu_m = \mathrm{E}(\frac{\overline{X}_n-\mu}{\sigma_n})^m$. The summation is over all $j_1,\ldots,j_p$ from 0 to m such that $j_1+\cdots+j_p = m$. Obviously, the large number of terms can make this expression unwieldy in practice, though some help can be obtained from symbolic calculation packages such as Mathematica or NAG/AXIOM. An overview article on computer algebra in statistics is [379], see also [564].

Here, we present the first four central moments of $f(\overline{X}_n)$, calculated along the above line (using Mathematica) under the simplifying assumption that a second order Taylor expansion is already sufficient to describe the function $f(x)$ in roughly the range $(\mu - 3\sigma_n, \mu + 3\sigma_n)$, while $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ with $X_1,\ldots,X_n$ i.i.d. random variables with mean μ, variance σ^2, skewness γ_1 and excess of kurtosis γ_2, such that the cumulant generating function $K(t)$ of $\frac{\overline{X}_n-\mu}{\sigma_n}$ equals asymptotically, to order $O(n^{-3/2})$,

$$\frac{t^2}{2} + \frac{1}{6}\gamma_{1,n}t^3 + \frac{1}{24}\gamma_{2,n}t^4 \,, \tag{1.91}$$

with $\gamma_{1,n} = \gamma_1/\sqrt{n}$ and $\gamma_{2,n} = \gamma_2/n$.

[14] The moments often, but not always, determine the distribution function. They do so at least in the following situations: (a) if the moment generating function exists for S in an open strip around zero, (b) if the variables have a bounded range, and (c) under rather mild, 'Carleman–type' conditions, see, e.g., [375], Chap. 3 of [607], [190].

Theorem 1.8. *Under the just stated conditions, in particular if* $f'\sigma \simeq f''\sigma^2 \gg f^{(p)}\sigma^p$ *for* $p = 3, 4, \ldots$, *the first 4 moments of* $f(\overline{X}_n)$ *are asymptotically given by*

$$E\big(f(\overline{X}_n) - f(E\overline{X}_n)\big) = \frac{1}{2}f''\sigma^2 \,, \tag{1.92}$$

$$E\big(f(\overline{X}_n) - Ef(\overline{X}_n)\big)^2 = f'^2\,\sigma^2 + \gamma_{1,n}f'f''\,\sigma^3 + (\frac{1}{2} + \frac{1}{4}\gamma_{2,n})f''^2\,\sigma^4 \,, \tag{1.93}$$

$$\begin{aligned} E\big(f(\overline{X}_n) - Ef(\overline{X}_n)\big)^3 = {} & \gamma_{1,n}f'^3\,\sigma^3 + (3 + \frac{3}{2}\gamma_{2,n})f'^2f''\,\sigma^4 + 6\gamma_{1,n}f'\,f''^2\,\sigma^5 + \\ & + (1 + \frac{5}{4}\gamma_{1,n}^2 + \frac{3}{2}\gamma_{2,n})f''^3\,\sigma^6 \,, \end{aligned} \tag{1.94}$$

$$\begin{aligned} E\big(f(\overline{X}_n) - Ef(\overline{X}_n)\big)^4 = {} & (3 + \gamma_{2,n})f'^4\,\sigma^4 + 18\gamma_{1,n}f'^3\,f''\,\sigma^5 + \\ & + [15(1 + \gamma_{1,n}^2) + \frac{39}{2}\gamma_{2,n}]f'^2f''\,\sigma^6 + \\ & + \gamma_{1,n}(39 + \frac{35}{2}\gamma_{2,n})f'\,f''^3\,\sigma^7 + \\ & + [15(\frac{1}{4} + \gamma_{1,n}^2) + \frac{39}{4}\gamma_{2,n}]f''^4\,\sigma^8 \,, \end{aligned} \tag{1.95}$$

respectively, where f' *stands for* $(\partial f/\partial x)_{x=\mu}$ *and* f'' *for* $(\partial^2 f/\partial x^2)_{x=\mu}$.

This result can be viewed as a generalisation of the central limit theorem. Expressed in $\gamma_{1,n}$ and $\gamma_{2,n}$, all coefficients up to and including $O(n^{-3/2})$ have been retained. Only the first term in each of the expressions is needed if $f'(\mu)\sigma \gg f''(\mu)\sigma^2$, and only the last term if $f'(\mu)\sigma \ll f''(\mu)\sigma^2$. If the sample size is sufficiently large (and/or if γ_1 and γ_2 are sufficiently small), then the coefficients $\gamma_{2,n} = \gamma_2/n$, and $\gamma_{1,n} = \gamma_1/\sqrt{n}$ can be neglected.

Remark. If f is monotonic, one can numerically conveniently determine the quantiles of the distribution function of $f(\overline{X}_n)$, since either $P\{f(\overline{X}_n) < y\} = P\{\overline{X}_n < f^{-1}(y)\}$, or $P\{f(\overline{X}_n) < y\} = P\{\overline{X}_n > f^{-1}(y)\}$. For non-monotonic functions, this expression does not hold, but in regular cases one can split the function into monotonic parts, then determine the distribution of $f(\overline{X})$ for each part separately, and finally paste the results together.

Exercise 1.28. The Larmor radius ρ in a plasma is proportional to $\sqrt{T}/B$, where T is the local plasma temperature and B the magnetic field. Suppose T and B have been measured with a relative accuracy of 15% and 2%, respectively. Determine the error propagation for the local measurement of ρ, using a probabilistic model (1) on absolute scale, (2) on logarithmic scale.

1.9 Modes of Convergence

In this section, we describe convergence concepts of random variables, while striving not to more generality than random variables which map Ω into in

a metric space $\mathfrak{X}$, a special case being $\mathfrak{X} = \mathbb{R}^p$. The material in this section may assist the reader to understand more fully some details of Sect. 1.10 on conditional probabilities. It plays also a role in the theory of stochastic processes, a topic which is not treated in this book.

Definition 1.27. *Let $\mathfrak{X}$ be a metric space with distance function d.*[15] *A sequence of numbers converges to $x \in \mathfrak{X}$, if $d(x_n, x) \to x$ for $n \to \infty$, i.e., for every $\epsilon > 0$ there exists an integer N such that $d(x_n, x) < \epsilon$ for all $n > N$. We use the notation $x_n \to x$.*

Definition 1.28. *A sequence $x_1, x_2, \ldots$ is said to converge intrinsically and is called a Cauchy sequence if for every $\epsilon > 0$ there exists an integer N such that $d(x_n, x_m) < \epsilon$ for $n, m > N$.*

Remarks.

1. In general, the 'limit' of a Cauchy sequence does not need to belong to $\mathfrak{X}$. A metric space which contains the limits of all its Cauchy sequences is called *complete*.
2. By one of several, equivalent, constructions of the real numbers, the space $\mathbb{R}^p$ is a complete metric space in any metric $d_p(x, y) = (\sum_{i=1}^p |x_i - y_i|^p)^{1/p}$, $0 < p \leq \infty$. The space $C[0, 1]$ of continuous functions on the (compact) interval $[0, 1]$, with metric $d(f, g) = \int |f(x) - g(x)|^p dx$ is not complete with respect to point-wise convergence (for which $f_n \to f$ means that $f_n(x) \to f(x)$ for all $x \in [0, 1]$), since the limit function can be discontinuous. This 'anomalous' feature has led to various definitions of convergence on function spaces, stronger than the concept of point-wise convergence (e.g., uniform convergence, convergence in quadratic mean), and also to the consideration of various function spaces more general than $C[0, 1]$ and complete with respect to specific types ('modes') of convergence.

A feature of particular interest for real-valued random variables, where the domain is a probability space, is the fact that convergence is not required on 'negligible' sets which have probability zero. There are several modes of convergence and some delicate interrelationships between them, which are even more intricate for general measures. We only give an outline here, restricting attention to probability measures, while referring the reader to [167, 390, 504, 607] for extensions and further details. A rather strong convergence concept is uniform convergence, except on a set with probability zero. In the following, let $X_1, X_2, \ldots$ be a sequence of random variables on a probability space $(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$ which take values in $\mathbb{R}$, or sometimes more generally in a Polish space $\mathfrak{X}$, i.e., a complete metric space with a denumerable topological basis. We use the notation $L^p(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$ for the space of

[15] A distance function or metric d: $\mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ satisfies the following axiomatic properties: (a) $d(x, y) = d(y, x)$, (b) $d(x, y) \geq 0$, (c) $d(x, y) = 0 \Leftrightarrow x = y$, (d) $d(x, y) + d(y, z) \geq d(x, z)$, for any $x, y, z \in \mathfrak{X}$.

real-valued random variables for which $\mathrm{E}|X|^p < \infty$, $0 < p \leq \infty$. We consider the following modes of convergence, which are in part hierarchically ordered.

Definition 1.29.

(*1*) *For any fixed* $0 < p \leq \infty$, *the sequence of real valued random variables* $X_n \to X$ *in pth mean if* $X_1, X_2, \ldots$ *are elements of* $L^p(\Omega, \mathfrak{B}(\Omega), P)$ *and* $E(|X_n - X|^p) \to 0$ *for* $n \to \infty$;

(*2*) $X_n \to X$ *with probability one* (*w.p. 1*) *if* X_n *converges everywhere on* Ω, *except possibly on a set with probability measure zero, i.e., if* $P\{\omega; \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1$;

(*3*) $X_n \to X$ *in probability if the probability measure of the set on which* X_n *differs from* X *more than an amount* ϵ *tends to zero, i.e., if* $\lim_{n\to\infty} P\{\omega; d(X_n(\omega), X(\omega)) > \epsilon\} = 0$ *for all* $\epsilon > 0$.

We now express a convergence criterion in terms of different probability distributions $P_1, P_2, \ldots$ associated with the random variables $X_1, X_2, \ldots$

Definition 1.30.

(*4*) *Let* $P_1, P_2, \ldots$ *be a sequence of probability measures;* $P_n \to P$ *weakly* (*in the vague, i.e., weak-*, topology*) *if, for any bounded and continuous function* g, $\int g(\omega) dP_n(\omega) \to \int g(\omega) dP(\omega)$ *for some measure* P. *If the limiting measure* P *is also a probability measure, then the convergence is said to be complete. We say that* $X_n \to X$ *in distribution* (*synonym: in law*) *if the sequence of corresponding measures* P_n *converges completely to* P *in the above sense.*

Remarks.

1. Convergence in the space $L^\infty(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$ of functions which are 'essentially' bounded, is also called uniform convergence (w.p.1). It is the strongest type of convergence considered here, i.e., it implies the other types of convergence in Def. 1.29.
2. For $1 \leq p \leq \infty$, $(\mathrm{E}|X|^p)^{1/p}$ is a norm, and for $0 < p < 1$, $\mathrm{E}|X_1 - X_2|^p$ is a metric on $L^p(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$. (Note that the norm is a convex function, and that for $0 < p < 1$ the set with all elements having a constant distance to the origin in the just mentioned metric is concave; please draw a picture in $\mathbb{R}^2$.) The space $L^p(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$ is constructed such that it is complete in the norm (for $1 \leq p \leq \infty$) or in the metric (for $0 < p < 1$), in the sense that it contains all limits of sequences of random variables converging in pth mean in the sense of Cauchy.
3. By applying Hölder's inequality [280], see also [560],

$$\int_\Omega |XY| d\mathrm{P} \leq \left(\int_\Omega |X|^p d\mathrm{P} \right)^{1/p} \left(\int_\Omega |Y|^q d\mathrm{P} \right)^{1/q} \tag{1.96}$$

for $1 \leq p, q \leq \infty$ and $1/p + 1/q = 1$, which is based on convexity arguments and the positive difference between arithmetic and geometric

mean $(x+y)/2-(xy)^{1/2}>0$, and which implies Minkowski's triangle inequality [467]

$$\left(\int_\Omega |X+Y|^p d\mathrm{P}\right)^{1/p} \le \left(\int_\Omega |X|^p d\mathrm{P}\right)^{1/p} + \left(\int_\Omega |Y|^p d\mathrm{P}\right)^{1/p} \quad (1.97)$$

for $1 \le p \le \infty$, one can derive that if X_n converges in pth mean, it converges in qth mean for $q < p$. In particular, $L^\infty(\Omega, \mathfrak{B}(\Omega), \mathrm{P}) \subset L^2(\Omega, \mathfrak{B}(\Omega), \mathrm{P}) \subset L^1(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$.

4. Convergence with probability 1 is sometimes also called almost sure (a.s.) convergence or almost everywhere (a.e.) convergence. The important feature is that the sequence $X_1, X_2, \ldots$ converges point-wise everywhere, except on a 'negligible' set of probability measure zero. An equivalent criterion is the following. For any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}\big(\bigcup_{k=1}^{\infty} \{\omega; d(X_{n+k}(\omega), X(\omega)\big) > \epsilon\}\big) = 0\,, \quad (1.98)$$

see [167, 607], from which one can directly see that it implies convergence in probability. A simple example of a sequence convergent in probability, which does not converge w.p.1 is the following. Let $\Omega = [0,1]$, and consider the sequence of binary random variables assuming the value one on the interval $[k, k+1)/n$ and zero otherwise $(k = 0, \ldots, n; n = 1, 2, \ldots)$, i.e., $X_1 = I_{[0,1]}$, $X_2 = I_{[0,\frac{1}{2}]}$, $X_3 = I_{[\frac{1}{2},1]}$, $X_4 = I_{[0,\frac{1}{3}]}$, $X_5 = 1_{[\frac{1}{3},\frac{2}{3}]}$, etc. (Note that the probability mass on which the value of X_n differs from zero tends to zero, but moves around, such that the sequence fails to converge for every ω on the entire interval $[0,1]$.)

5. The following reverse implications hold, which either weaken the consequence or invoke an additional assumption: (a) If $X_1, X_2, \ldots$ converges in probability to X, then there exists a *subsequence* which converges to X w.p.1. (Hint: Consider a subsequence for which the distance to X decreases as $\epsilon/2^k$, $k = 1, 2, \ldots$, and apply (1.98)). (b) If the sequence $X_1, X_2, \ldots$ is either monotonic of dominated by X, i.e., $|X_n| < X$ a.e. for some X with $\mathrm{E}|X| < \infty$, then convergence in probability implies convergence in $L^p(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$, for $0 < p \le \infty$, and so does convergence w.p.1.

6. If $X_n \to X$ either w.p.1, in probability or in distribution, then $g(X_n) \to g(X)$ for any continuous function g. In particular, addition and multiplication of random variables are closed under these types of convergence. This property is frequently used in asymptotic statistics. For instance one can infer from $X_n \to X \sim N(0,1)$ that $X_n^2 \to Y \sim \chi_1^2$, and so on. For a proof, see for instance [607]. For an interesting chapter on the χ^2 distribution and its role in asymptotic statistics, the reader is referred to [474].

7. An equivalent condition for weak*–convergence is that $\mathrm{P}_n(A) \to \mathrm{P}(A)$ for any $A \in \mathfrak{B}(\Omega)$ such that $\mathrm{P}(\delta(A)) = 0$, where $\delta(A)$ is the topological boundary of A (Hint: consider $g(\omega)$ to be a close, continuous approximation of the indicator function $I_A(\omega)$). For real-valued random variables, two equivalent definitions of convergence in distribution are: (a) $F_n \to F$ for each x at which F is continuous (consider: $g(x) = I_{(-\infty,x)}$; the reverse implication is more complicated to derive and known as the second theorem of Helly–Bray, see [432, 607]), and (b) point-wise convergence of the associated characteristic functions to a continuous limit-function, see [232, 607]. An illustrative example which indicates that the condition $\mathrm{P}\big(\delta(A)\big) = 0$ is needed, is given by $F_n(x) = k/n$, $k, n \in \mathbb{N}$, where $k = 1, \ldots, n$ and $n \to \infty$, which converges in distribution to $F(x) = x$ on $[0, 1]$. Now, $P_n(\mathbb{Q}) = 1$ does not converge to $P(\mathbb{Q}) = 0$. Obviously, the topological boundary of $\mathbb{Q}$ is $[0, 1]$. For ease of exposition and because of its predominant role in asymptotic statistics, some results for convergence in distribution have been considered in Sects. 1.7 and 1.8.
8. An example of a sequence which converges in the weak*–topology, but does not converge completely is $F_n(x) = \Phi(x + n)$, where Φ denotes the cumulative standard normal distribution function. The issue is that the limit function is not a distribution function (i.e., the limit measure is not a probability measure), see [432]. In this situation we do not say that X_n converges in distribution to a random variable X. (Sometimes, par abus de langage, the expression 'X_n converges weakly to a random variable X' is used for convergence in distribution, but we shall avoid employing this terminology in this book.)

1.10 Conditional Expectation

In Section 1.4 the term absolutely continuous has been used as a shorthand expression for 'absolutely continuous with respect to the Lebesgue measure'. However, since the essential features do not depend on the particular features of the Lebesgue measure, the concept has been cast into a more general framework by Radon and Nikodym. A related topic is defining the conditional expectation $\mathrm{E}(Y|X)$ as an approximation of Y by a function of the random variable X. Such general definitions play a role in the theory of martingales, in stochastic processes, and in generalisations of the central limit theorem to situations where the random variables involved exhibit some form of dependence. One can compare this concept of conditional probability with disintegration of measures [102, 682, 739], which plays a more natural role in various statistical applications.

Definition 1.31. *Let $(\Omega, \mathfrak{B}(\Omega), P)$ be a probability space. A probability measure Q on Ω is called absolutely continuous with respect to P, if $P(A) = 0$ implies $Q(A) = 0$ for all $A \in \mathfrak{B}(\Omega)$. We use the notation $Q \ll P$.*

Example: Let $F = F_{ac} + \sum_i p_i x_i$, where $p_i = F(x_i^+) - F(x_i^-)$ is the jump of F at x_i, and $G = G_{ac}$, the subscript ac denoting absolute continuity with respect to the Lebesgue measure. If F_{ac} is strictly increasing where G_{ac} is strictly increasing, or, equivalently, the support of the measure associated with F_{ac} contains that of G_{ac}, then $G \ll F$, but not $F \ll G$.

Remark. Notice that the ordering $\ll$ is only partial since one can easily construct instances in which neither $\mathrm{P} \ll \mathrm{Q}$ nor $\mathrm{Q} \ll \mathrm{P}$.

Theorem 1.9. (*Radon–Nikodym*) *Let P and Q be two probability measures on $\big(\Omega,\mathfrak{B}(\Omega)\big)$ with $Q \ll P$. Then there exists a measurable function g on Ω such that for any $A \in \mathfrak{B}(\Omega)$,*

$$Q(A) = \int_A g(\omega) dP(\omega) \; . \tag{1.99}$$

Remarks.

1. The notation $g = d\mathrm{Q}/d\mathrm{P}$ is used, and g is called (a version of) the Radon–Nikodym (R–N) derivative. The Radon–Nikodym theorem holds, more generally, for $\sigma-$finite measures. For a proof see, e.g., [74, 167]. The integral in Theorem 1.9 does not change if g is changed on a 'negligible' set N with $\mathrm{P}(N) = 0$. Therefore, the R–N derivative is an equivalence class of integrable functions, i.e., $g \in L^1(\Omega, B(\Omega), \mathrm{P})$, and can be looked upon as a random variable. (By a mild abuse of notation we write dF/dG for the R–N derivative of the measures associated with the distribution functions F and G.)
2. If the probability measure Q is absolutely continuous with respect to P then, for any measurable function h and any $A \in \mathfrak{B}(\Omega)$,
$$\int_A h(\omega) d\mathrm{Q}(\omega) = \int_A h(\omega) g(\omega) d\mathrm{P}(\omega) \; . \tag{1.100}$$
3. As noticed in Sect. 1.2, see Remark 6, the linear functional notation $\mathrm{P}(g, A)$ can be used for $\int_A g(\omega) d\mathrm{P}(\omega)$.

Example. Let $F = F_{ac} + \sum_{i=1}^k p_i x_i$, and $G = G_{ac} + \sum_{i=1}^k q_i x_i$. Then $(dF/dG)(x) = f(x)/g(x)$ with $f = dF_{ac}/dx$ and $g = dG_{ac}/dx$ on $\mathbb{R} \setminus \{x_1, \ldots, x_k\}$ and $(dF/dG)(x_i) = p_i/q_i$ for $i = 1, \ldots, k$. The value of dF/dG on a set of Lebesgue measure zero *outside* the set $\{x_1, \ldots, x_n\}$ is immaterial.

For $X : \Omega \to \mathbb{R}$ one can consider the σ–algebra generated by X, i.e., by the sets $X^{-1}(\mathrm{O})$ where O runs through the open sets in $\mathbb{R}$.[16] Hence, we

[16] Being the intersection of all σ–algebras for which X is measurable, this is the smallest σ–algebra on Ω for which X is measurable.

can define $\mathrm{E}(Y|X)$ if we have a definition of $\mathrm{E}(Y|\mathfrak{D})$ for any sub σ–algebra $\mathfrak{D} \subset \mathfrak{B}(\Omega)$. Once we have $\mathrm{E}(Y|X)$, which has a use of its own, it is natural to define $\mathrm{P}(B|\mathfrak{D})$ by $\mathrm{E}(I_B|\mathfrak{D})$, and tempting, even though a precise identification has some equivocal aspects, to write $\mathrm{P}(B|A)$ for $\mathrm{E}(I_B|I_A)$, where as usual I_A stands for the indicator function of A for any $A \in \mathfrak{B}(\Omega)$. This was the route, introduced by [389] and perfected by [141], that led to a rather general concept of conditional expectation, albeit somewhat different than in Sect. 1.3. It is mentioned here, because it is often used in the theory of stochastic processes. For alternative approaches, more related to the traditional statistical usage of conditioning, the reader is referred to [102, 739], in which conditional probabilities are introduced as Markov kernels and 'disintegration measures', respectively, see also [682] and [339]. A flavour of the difference can be perceived as follows. With respect to the coarse σ–algebra with respect to the entire σ–algebra $\{\varnothing, A, A^c, \Omega\}$, I_A is a binary random variable assuming the values zero and one, with probability p and $1-p$, respectively. The disintegration measure approach considers $F_0(y) = P(Y < y|\omega \in A)$ and $F_1(y) = P(Y < y|\omega \in A^c)$, such that the marginal distribution function of Y equals $F(y) = pF_0(y) + (1-p)F_1(y)$. In the stochastic process approach, $\mathrm{E}(Y|I_A)$ is, in some specified sense, the best approximation of Y by a constant times the binary random variable I_A.

Definition 1.32. *Let* $Y\colon (\Omega, \mathfrak{B}(\Omega), P) \to \big(I\!R, \mathfrak{B}(I\!R)\big)$ *be a random variable with finite expectation, i.e.,* $EY = \int_\Omega Y dP < \infty$, *and let* $\mathfrak{D}$ *be a sub-algebra of* $\mathfrak{B}(I\!R)$. *Then, the conditional expectation of* Y *given* $\mathfrak{D}$, $E(Y|\mathfrak{D})$, *is defined as a random variable* $Z = E(Y|\mathfrak{D})$ *which is measurable with respect to* $\mathfrak{D}$ *and which satisfies*

$$\int_A Z dP = \int_A Y dP \quad \textit{for all} \quad A \in \mathfrak{D} \,. \tag{1.101}$$

Remarks.

1. If Y itself is $\mathfrak{D}-$measurable (or measurable with respect to a sub-algebra of $\mathfrak{D}$) then, obviously, $\mathrm{E}(Y|\mathfrak{D}) = Y$. If Y is measurable with respect to a σ–algebra finer than $\mathfrak{D}$, then $\mathrm{E}(Y|\mathfrak{D})$ averages Y over subsets in Ω which are elements of $\mathfrak{D}$. (For the definition of measurability, see Remark 1.6.)
2. For the conditional probability of the event B given the σ–algebra $\mathfrak{D}$, i.e., $\mathrm{P}(B|\mathfrak{D}) = \mathrm{E}(I_B|\mathfrak{D})$, one has the relation

$$\int_A \mathrm{P}(B|\mathfrak{D}) dP = \int_A I_B d\mathrm{P} = \mathrm{P}(A \cap B) \quad \text{for all} \quad A \in \mathfrak{D} \,. \tag{1.102}$$

Discussion. For any fixed B, $\mathrm{P}(B|\mathfrak{D})$ is the Radon–Nikodym derivative of the measure $Q_B(A) = \mathrm{P}(A \cap B)$ with respect to P. For $A = \Omega$ and $\mathfrak{D}$ generated by a partition of Ω, $\Omega = A_1 \cup \cdots \cup A_k$ with $A_i \cap A_j = \varnothing$ $(i, j = 1, \ldots, k)$, (1.102) resembles (1.18). In general, the precise relationship between $E(Y|X)$ as defined

above and $E(Y|X = x)$ as defined by probability densities is somewhat intricate. If (X, Y) has a two-dimensional distribution, absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$, then $E(Y|X = x)$ is the expectation of a random variable with probability density $f(y|x) = f(x, y)/\int f(x, y)dy$. Starting from this classical approach, there are two ways to define $E(Y|X)$: (a) by multiplying with the probability density of the conditioning variable X. Then, $E(Y|X)$ is a random variable with probability density $\int f(y|x)f(x)dx$, which is the marginal density of Y, i.e., in that case $E(Y|X) = Y$. Albeit unexpected, this makes some sense in view of the fact that X and Y are measurable with respect to the same σ–algebra on $\mathbb{R}$.[17] (b) by replacing in $f(y|x) = f(x, y)/\int f(x, y)dy$ the conditioning variable x by the random variable X, which can be quite complicated except in simple situations as below.

Example. For a bivariate normal distribution (see Chap. 2) centered at the origin, we get in this case $E(Y|X) = cX$ with $c = \mathrm{cov}\,(X, Y)/\mathrm{var}\,(X)$, which is a best unbiased predictor or Y in the sense that it has minimum variance $\mathrm{var}\,\big(\mathrm{E}(Y|X)\big) = \rho^2(X, Y)\mathrm{var}\,(Y)$ among all linear, and because of the properties of the normal distribution, even all unbiased estimators. This is an impressive feature in itself. However, in this case $\mathrm{E}(Y|X) = 0$ if X and Y are uncorrelated, even though the σ–algebras generated by X and Y are the same. One cannot conclude here that, according to Def. 1.32, $\mathrm{E}(Y|X)$ should be equal to Y, apparently in view of the fact that the $x-$axis and the $y-$axis do not constitute geometrically the same line.

From the above discussion, we retain that in the abstract approach of conditioning with respect to a σ–algebra, not relying on disintegration measures, the intuitive concept of a conditional distribution given a particular value of x of X has been replaced by a smoothing property, approximation of Y by (a function of) a variable X, measurable with respect to a coarser σ–algebra. This generalisation is, at least in part, 'backward compatible' with the classical approach of conditional densities and the random variables constructed from them by replacing the conditioning variables. To retain a certain distinction between the two concepts, we shall occasionally employ the term abstract conditional expectation for the conditional expectation as defined by (1.32). Some basic properties of (abstract) conditional expectations are as follows, which can be also used as defining properties, see, e.g., [515]:

Theorem 1.10. *Let X and Y be random variables with finite (absolute) expectation, i.e., $\int_\Omega |X|dP < \infty$, and similarly for Y. This means that X and Y are elements of the Banach space $L^1(\Omega, \mathfrak{B}(\Omega), P)$. Then*

(a) $E(aX + bY|\mathfrak{D}) = aE(X|\mathfrak{D}) + bE(Y\mathfrak{D})$ (linearity);

[17] If X is measurable with respect to the σ–algebra $\{(i, i + 1); i = \ldots, -1, 0, 1, \ldots\}$, then the $y-$axis has to be discretised in the same way and $E(Y|X)$ is the random variable Y_n whose distribution is the marginal distribution of the discretised two-dimensional distribution.

(*b*) *if* $X \le Y$ *then* $E(X|\mathfrak{D}) \le E(Y|\mathfrak{D})$, *with probability 1* (*monotony*);

(*c*) $\mid E(X|\mathfrak{D}) \mid \le E(|X| \; |\mathfrak{D})$ (*triangle inequality, convexity of modulus*);

(*d*) *if* $X_n \to X$ *w.p. 1 and either* $|X_n| \le Y$ *or* $X_n < X_{n+1}$ *then* $E(X_n|\mathfrak{D}) \to E(X|\mathfrak{D})$ *w.p. 1* (*dominated convergence, monotone convergence*);

(*e*) *if* X *is measurable with respect to* $\mathfrak{D}$ *and* $XY \in L^1(\Omega, \mathfrak{B}(\Omega, P)$, *then* $E(XY|\mathfrak{D}) = XE(Y|\mathfrak{D})$;

(*f*) *if* $\mathfrak{D}_1 \subset \mathfrak{D}_2$, *i.e., if* $\mathfrak{D}_1$ *is coarser than* $\mathfrak{D}_2$, *then* $E\big(E(X|\mathfrak{D}_2)|\mathfrak{D}_1\big) = E\big(E(X|\mathfrak{D}_1)|\mathfrak{D}_2\big) = E(X|\mathfrak{D}_1)$, *w.p.1* (*smoothing property*).

Remark. With respect to the inner product $(X, Y) = \int_\Omega XY d\mathrm{P}$, the random variables with finite second moment constitute a Hilbert space $L^2(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$ which is included in the Banach space $L^1(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$. The random variables that are, in addition, $\mathfrak{D}$–measurable form a (closed) linear subspace of $L^2(\Omega, \mathfrak{B}(\Omega), \mathrm{P})$, denoted by $L^2(\Omega, \mathfrak{D}, \mathrm{P})$. In this case, the conditional expectation $\mathrm{E}(X|\mathfrak{D})$ can be viewed as an orthogonal projection of X on $L^2(\Omega, \mathfrak{D}, \mathrm{P})$, since for any $A \in \mathfrak{D}$, the relations hold $\int_A X d\mathrm{P} = (X, I_A) = (X, \mathrm{P}_\mathfrak{D} I_A) = (\mathrm{P}_\mathfrak{D} X, I_A) = \int_A \mathrm{P}_\mathfrak{D} X d\mathrm{P}$, where $\mathrm{P}_\mathfrak{D}$ denotes the orthogonal projection operator on $L^2(\Omega, \mathfrak{D}, \mathrm{P})$, which is self-adjoint and acts as an identity on its range.

One naturally defines the (general) conditional variance of X given the σ–algebra $\mathfrak{D}$ as follows

Definition 1.33. *var* $(X|\mathfrak{D})$ *is a* $\mathfrak{D}-$*measurable random variable satisfying*

$$var\ (X|\mathfrak{D}) = E\big[\big((X - E(X|\mathfrak{D})\big)^2|\mathfrak{D}\big] = E(X^2|\mathfrak{D}) - 2[E(X|\mathfrak{D})]^2 + E[E(X|\mathfrak{D})^2]\,. \tag{1.103}$$

Exercise 1.29. Derive that $\mathrm{var}\,(X) = \mathrm{E}\big(\mathrm{var}\,(X|\mathfrak{D})\big) + \mathrm{var}\,\big(\mathrm{E}(X|\mathfrak{D})\big)$, and illustrate with the bivariate normal distribution that this is the breakdown of the total variation of X into the variation around the regression line (first term) and the variation described by the regression line (second term).

The reader is referred to [57, 414, 426, 744] for further theory, proofs and applications of conditional expectations in statistics, stochastic processes and communication theory.

2 Elements of Statistical Theory

ʿ Ομοίως δὲ καὶ ὁ ἐν τῷ Μένωνι λόγος ὅτι ἡ μάθησις ἀνάμνησις[†]

ARISTOTLE, 384–322 BC

2.1 Introduction

An idealised starting point of a statistical analysis in a probabilistic setting is a sequence of (vector valued) random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$: $(\Omega, \mathfrak{B}(\Omega), \mathscr{P}_\Omega) \to (\mathfrak{X}, \mathfrak{B}(\mathfrak{X}), \mathscr{P}_\mathfrak{X})$, where $\mathscr{P}_\Omega = \{\mathrm{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ is a family of probability distributions, defined on the space Ω and parameterised by $\boldsymbol{\theta}$, $\mathfrak{B}(\mathfrak{X})$ is a $\sigma-$algebra on $\mathfrak{X}$ and $\mathfrak{B}(\Omega)$ a corresponding $\sigma-$algebra of subsets on Ω, for which probabilities can be sensibly defined. We use the notation $P_{\boldsymbol{\theta}}(A) = P_{\boldsymbol{X},\boldsymbol{\theta}}\{A\} = \mathrm{P}_{\boldsymbol{\theta}}\{\omega; X(\omega) \in A\}$ for any subset $A \in \mathfrak{B}(\mathfrak{X})$. The measure $P_{\boldsymbol{X},\boldsymbol{\theta}}$ or, by a slight abuse of notation, $P_{\boldsymbol{\theta}} \in \mathscr{P}_\mathfrak{X}$ is a probability measure induced by $\mathrm{P}_{\boldsymbol{\theta}} \in \mathscr{P}_\Omega$. We often write $\mathscr{P}$ for $\mathscr{P}_\mathfrak{X}$ and sometimes simply identify Ω with $\mathfrak{X}$.

A number of statistical problems can be cast into the following form: $(Y_i, \boldsymbol{X}_i)$, where Y_i is an unobservable random variable and $\boldsymbol{X}_i$ is an observable random vector, and one is interested in the conditional distribution $\mathfrak{L}(Y_i|\boldsymbol{X}_i = \boldsymbol{x})$. In general, the unobservable part can be also a random vector, of course. It is noted, however, that in a regression analysis context, a number of separate univariate regressions can be employed instead of a single multivariate regression,[1] at least if one is not especially interested in the correlations between the elements of the (unobservable) vector of response variables $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$.

In other, more simple (univariate and one-sample) situations, $\boldsymbol{X}_i = X_i$ is observable for $i = 1, 2, \ldots$ and one is interested in characteristics of the probability distribution $\mathfrak{L}(X_i)$, such as the mean, the variance or some quantile $F^{-1}(p)$ for $p \in [0, 1]$, where F denotes the cumulative distribution function (see below).

In probability theory, a specific probability distribution $\mathfrak{L}(Y, \boldsymbol{X})$ is postulated and one investigates for instance $\mathfrak{L}(Y|\boldsymbol{X} = \boldsymbol{x})$, or, while postulating

[†] Similarly too with the theory in the Meno that learning is recollection (Analytica Priora, II, XXI)

[1] This expression is used in the sense that univariate regression yields the same point estimates and the same distribution of the regression coefficients as the point estimates and the marginal distributions in the corresponding multivariate regression, see Sect. 2.5.2.

$\mathfrak{L}(\boldsymbol{X})$, one asks for the distribution of a function $\boldsymbol{g}(\boldsymbol{X})$ of the original (basic) random variable $\boldsymbol{X}$. In this sense, probability theory can be viewed as a *deductive branch of science*: an extension of classical logic and a section of mathematics.

A distinctive feature of statistics is that the probability distribution which 'generated' the data is not assumed to be known. Instead, it is postulated to belong to a certain (non-singleton) sub-family of the family of all probability distributions. We denote this sub-family by $\mathscr{P}$. From the outcomes $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ of the observable random variables $\boldsymbol{X}_1, \boldsymbol{X}_2 \ldots$ one intends to derive information about the plausibility for various members of $\mathscr{P}$ to have generated the data, before one can proceed to estimate $\mathfrak{L}(Y|\boldsymbol{X} = \boldsymbol{x})$ or $\mathfrak{L}\big(\boldsymbol{g}(\boldsymbol{X})\big)$. This means that statistics essentially bears the character of an *inductive science*. With some justification one can call it even an *inductive meta-science*, as it constitutes a structuring element (framework as well as activity) on the interface between theory and experiment of (almost) any science. In inductive statistical inference one concentrates on estimating characteristic properties of families of (conditional) probability distributions, whereas in predictive inference, one is interested in predicting the possible outcomes of future events, for instance of $\boldsymbol{X}_{n+1}, \ldots, \boldsymbol{X}_{n+m}$ based on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. In this chapter, we mainly discuss inductive inference. For various aspects of predictive inference the reader is referred to [224, 359].

Traditionally, classical statistical inference is divided into the following areas (with rather intense interrelationships), which we will also discuss here: Point Estimation, Hypothesis Testing and the Construction of Confidence Regions, see [56, 424, 426, 610]. We shall describe in this chapter principally, and in a rather brief fashion, the classical approach to statistics, while for a discussion of so-called Bayesian analysis, based on inverse probability, we refer to the literature, especially to [50,53,81,138,323,358,496,555,582]. Since the eighties of the twentieth century a theory of Distributional Inference has been developed, which generalises the concept of confidence regions to confidence distributions, without necessarily adopting a (personalistic or formal) Bayesian interpretation of probability. More details on this approach can be found in [6, 358, 359, 573].

2.2 Statistical Inference

2.2.1 Point Estimation

We first give some simplified and stylised examples of statistical models. Let $X_1, X_2, \ldots, X_n$ be an independent random sample from $N(\mu, \sigma^2)$. This is formalised by the following identifications: $\Omega = \mathbb{R}, \mathfrak{B}(\Omega) = \mathfrak{B}(\mathbb{R})$, which are the 'non-pathological' ('Borel') sets of $\mathbb{R}, P_{\boldsymbol{\theta}} = N(\mu, \sigma^2)$, i.e., $\boldsymbol{\theta} = (\mu, \sigma^2) \in \Theta$ with $\Theta = \mathbb{R} \times \mathbb{R}^+$. We will refer to this model briefly as 'the case of a normal family' of distributions.

A practical physical example is given by the repeated measurements of the plasma density, based on Thomson scattering experiments, under stationary conditions, at a fixed radial position.[2]

In this section, we give a minimum number of formal definitions which are used in elementary statistics. Some other important concepts are explained without a formal definition. The standard normal family is used as a running example. In a certain sense, this can be considered as (starting) 'a game on the square meter', because of various rather small differences between alternative approaches that are usually not of any practical importance. The rationale is, however, on the one hand, that the normal distribution occurs rather frequently in physical applications (be it as an approximation in virtue of the central limit theorem) and furthermore that estimation theory of the two-parametric normal family is 'a quite simple example', albeit less trifling than, for instance, the one-parameter Poisson distribution or the exponential distribution. Finally, if the reader has worked through the normal example, he should be able to understand the computationally somewhat more complicated estimation problems of other two (and even more-) dimensional exponential families, a number of which have been listed in Table 1.2. As a single such example, the case of an independent sample from a $\Gamma_{f,g}$ distribution (of which χ^2_f is a special case) is discussed near the end of this section.

Let $\boldsymbol{g}\colon \Theta \to \mathbb{R}^m$ be some vector-valued function of the parameter $\boldsymbol{\theta}$. We can write $\boldsymbol{g}(\boldsymbol{\theta}) = \big(g_1(\boldsymbol{\theta}), \ldots, g_m(\boldsymbol{\theta})\big)$. Some examples, in the case of a normal family, are: $g_1(\boldsymbol{\theta}) = \mu, g_2(\boldsymbol{\theta}) = \sigma^2, g_3(\boldsymbol{\theta}) = \frac{\sigma}{\mu}$ (variation coefficient), $\boldsymbol{g}_4(\boldsymbol{\theta}) = (\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2})$, etc.

Definition 2.1. *An* estimator *of* $\boldsymbol{g}(\boldsymbol{\theta})$ *is a function* $\boldsymbol{T} = \boldsymbol{h}(X_1, \ldots, X_n)$ *of the random variables* $X_1, \ldots, X_n$*, intended to 'approximate'* $\boldsymbol{g}(\boldsymbol{\theta})$.[3]

Remark. Notice that an estimator is a random variable.

Definition 2.2. *An* estimate *is the* (*deterministic*) *outcome of an estimator in a particular experiment.*

Remark. A random variable which is a function of more basic random variables $X_1, \ldots, X_n$ is often called a statistic, see Chap. 2.1 of [424]. Hence, an estimator is a statistic designed to estimate an unknown parameter $\boldsymbol{\theta}$.

Notation. If $\boldsymbol{T} = \boldsymbol{h}(X_1, \ldots, X_n)$ is an estimator, we denote the associated estimate by $\boldsymbol{t} = \boldsymbol{h}(x_1, \ldots, x_n)$, where $x_1, \ldots, x_n$ are the values that have been

[2] The measurements have to be at discrete time points and sufficiently separated in time, so that the measured variations of the plasma can be considered to be independent. The variance σ^2 is then composed of (a) the measurement noise of the apparatus (laser-light source as well as detector) and (b) the spontaneous plasma fluctuations.

[3] The function $\boldsymbol{h}\colon \mathbb{R}^n \to \mathbb{R}^m$ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ which uniquely associates a point in $\mathbb{R}^m$ for each $(x_1, \ldots, x_n) \in \mathbb{R}^n$; $\boldsymbol{T}(\omega) = \boldsymbol{h}\big(X_1(\omega), \ldots, X_n(\omega)\big)$ is a random point in $\mathbb{R}^m$.

realised in the particular experiment.

Example. Consider $X_1, \ldots, X_n$ to be i.i.d. (i.e., independent and identically distributed), according to some distribution function F with expectation value μ and variance σ^2. In this case, $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is an estimator for μ, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ is an estimator for σ^2 and $\sum_{i=1}^n X_i / \sum_{i=1}^n (X_i - \overline{X})^2$ is a corresponding estimator for μ/σ^2.

Definition 2.3. *The* bias *of an estimator* $\boldsymbol{T}$ *of* $\boldsymbol{g}(\boldsymbol{\theta})$ *is*

$$E_{\boldsymbol{\theta}}\boldsymbol{T} - \boldsymbol{g}(\boldsymbol{\theta}) \,, \tag{2.1}$$

where $E_{\boldsymbol{\theta}}\boldsymbol{T}$ *denotes the expectation value of* $\boldsymbol{T}$ *when* $\boldsymbol{\theta}$ *is the true value of the parameter.*

Recall that if $P_{\boldsymbol{\theta}}$ has density $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, then

$$\mathrm{E}_{\boldsymbol{\theta}}\boldsymbol{T} = \int_{\mathbb{R}^n} \boldsymbol{h}(\boldsymbol{x}) f_{\boldsymbol{\theta}}(\boldsymbol{x}) d\boldsymbol{x} \,. \tag{2.2}$$

Definition 2.4. $\boldsymbol{T}$ *is an* unbiased *estimator of* $\boldsymbol{g}(\boldsymbol{\theta})$ *if* $E_{\boldsymbol{\theta}}\boldsymbol{T} = \boldsymbol{g}(\boldsymbol{\theta})$.

We now consider, for simplicity, a one-dimensional estimator T for θ.

Definition 2.5. *The* mean squared error (MSE) *of the estimator* $T = h(X_1, \ldots, X_n)$ *is*

$$MSE(T) = E_\theta\big(T - g(\theta)\big)^2 \,. \tag{2.3}$$

It is easily derived that, in the one-dimensional case,

$$\mathrm{MSE}(T) = \mathrm{var}\, T + \mathrm{bias}^2(T) \,, \tag{2.4}$$

where $\mathrm{var}\, T = \mathrm{E}_\theta (T - \mathrm{E}_\theta T)^2$.

Exercise 2.1. Derive formula (2.4). (*) Generalise this to a matrix expression for an m-dimensional statistic $\boldsymbol{T} = (T_1, \ldots, T_m)$.

Estimators for μ and σ^2 in the case of a normal family. (worked example) Consider $X_1, X_2, \ldots, X_n$ to be i.i.d. $N(\mu, \sigma^2)$. Then $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu})^2$ are familiar estimators for μ and σ^2, respectively. They are so-called sample–analogues of the expectation value and the variance, because they can be obtained by replacing in the expressions $\mu = \int x f(x) dx$ and $\sigma^2 = \int (x-\mu)^2 f(x) dx$ the density $f(x)$ by $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^n \delta(x - x_i)$. Nevertheless, X_1, $\frac{1}{2}(X_1 + X_2)$, etc., are also estimators for μ. Intuitively, even if somewhat guided by literary tradition, one prefers the estimator $\hat{\mu}$, however. What is the reason? Let us compare $\hat{\mu}$ and X_1 by calculating their expectation value and variance:

$$\begin{array}{cc} \mathrm{E}X_1 = \mu \;, & \mathrm{E}\hat{\mu} = \mu \;, \\ \mathrm{var}\, X_1 = \sigma^2 \;, & \mathrm{var}\, \hat{\mu} = \sigma^2/n \;. \end{array}$$

Exercise 2.2. Derive the above equalities.

So, both estimators have the pleasant property that their expectation value equals the parameter to be estimated (i.e., both estimators are unbiased). However, the variance of $\hat{\mu}$ is, for large n, considerably smaller than that of X_1, hence $\hat{\mu}$ is to be preferred.

From a certain point of view, it looks efficient to have an estimator which is unbiased and has minimal variance (of all possible, unbiased estimators). It can be proven (see for instance Chap. 2 of [426]) that in the case of a normal family, $\hat{\mu}$ has this property. Hence, it is called the uniformly minimum variance unbiased (UMVU) estimator. Let us now look at the estimator for σ^2. Is it unbiased? We calculate its expectation value.
Note that

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \tag{2.5}$$

can be written as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2\hat{\mu}X_i + \hat{\mu}^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2 \,. \tag{2.6}$$

Hence,

$$\mathrm{E}\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}X_i^2 - \mathrm{E}\hat{\mu}^2 \,. \tag{2.7}$$

This device, which is basically Steiner's rule, derived from Pythagoras' Theorem in $\mathbb{R}^n$ (see Chap. 1.4), is used rather frequently in this type of calculations.

Now, we know that

$$\mathrm{var}\, X_i = \sigma^2 = \mathrm{E}X_i^2 - \mu^2 \quad \text{(Steiner's rule)} \,; \tag{2.8}$$

$$\mathrm{var}\ \hat{\mu} = \frac{\sigma^2}{n} = \mathrm{E}\hat{\mu}^2 - \mu^2 \quad \text{(Steiner's rule)} \,. \tag{2.9}$$

Inserting these two equations into (2.7) we get

$$\mathrm{E}\hat{\sigma}^2 = \frac{1}{n}n(\sigma^2 + \mu^2) - (\frac{\sigma^2}{n} + \mu^2) = (1 - \frac{1}{n})\sigma^2 \,. \tag{2.10}$$

Therefore, $\hat{\sigma}^2$ is not unbiased, the bias $\mathrm{E}\hat{\sigma}^2 - \sigma^2$ being $-\frac{1}{n}\sigma^2$. For $n \to \infty$ the bias disappears, however, whence it is said that $\hat{\sigma}^2$ is asymptotically unbiased.[4] The corresponding unbiased estimator for σ^2 is $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$. It can be shown that this is the UMVU estimator for σ^2.

[4] In practice, this bias is not a serious point of concern if n is moderately large, but one has to take it into account, of course, if $n = 2$ or 3, for instance when 'duplicated measurements' are analysed.

It is useful to have some principles to obtain estimators that are not too difficult to compute and have a high degree of accuracy. One can be guided by

(1) *Intuition.* For example, by using the sample analogue of a parameter. The parameter μ denotes the expectation value $\mathrm{E}X = \int x f(x) dx$. The sample analogue is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, the corresponding estimator is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$.[5] Similarly, $\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$ is the sample analogue of the variance. More generally, it may be that some combinations of the parameter components of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ constitute the first, second, ..., mth moment of the distribution $F_{\boldsymbol{\theta}}$. Equating these combinations to the first m empirical moments, one gets m (usually non-linear) equations with m unknowns. Their solution is called the *moment estimator* of $\boldsymbol{\theta}$, and, hence, of the distribution $F_{\boldsymbol{\theta}}$. In a number of situations, for $m = 2$ or 3, the solution of the moment equations is possible analytically or by simple numerical methods. For relatively large m (≥ 4, say) the moment equations tend to be a rather unstable system in practice, unless the sample size is (very) large, in the order of several thousands of observations.

(2) *Statistical optimality theory.* For example, one can find an estimator that is UMVU, or that minimises some functional of a loss function. We shall not pursue the derivation of such estimators in this book, but refer the reader to the literature, in particular [191, 426, 538, 610]. Optimality theory for estimators is somewhat naturally linked with that of hypothesis testing, for which we refer to [139, 191, 204, 424, 584, 625].

(3) *The principle of maximum likelihood.* This is in some sense a compromise between (1) and (2). It has the advantage that it gives an algorithmic recipe to obtain estimators which have, in general, good large sample properties. They may not be optimal in the sense of (2), but under weak regularity conditions they are *asymptotically* (i.e., for $n \to \infty$) *unbiased*, and even *efficient.* This last property means that, in addition to being asymptotically unbiased, the estimators have asymptotically minimal variance. Some references on asymptotic inference are [34, 547, 607, 695].

We shall now briefly describe the method of maximum likelihood. Consider a statistical experiment $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$: $\big(\Omega, \mathfrak{B}(\Omega)\big) \to \big(\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n)\big)$, characterised by an unknown probability measure $\mathrm{P} = P_{\boldsymbol{X}}\boldsymbol{X}^{-1}$. It is supposed (though we can never be sure) that some specified family $\mathscr{P} = \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ contains the (unknown) probability measure $P_{\boldsymbol{X}}$. Given an outcome $\boldsymbol{x} = (x_1, \ldots, x_n)$ of this experiment, one can ask for the probability of get-

[5] The sample analogue is obtained by replacing f by $\frac{1}{n}\sum_{i=1}^{n}\delta(x - x_i)$, i.e., by the 'sampling distribution', which gives equal probabilities $1/n$ to the observations $x_1, \ldots, x_n$.

ting this outcome, for various values of $\boldsymbol{\theta}$. If this probability is high, it is said that the corresponding value of the parameter $\boldsymbol{\theta}$ is *likely* (given the outcome of the experiment). It is however noted that, according to the objectivistic interpretation, $\boldsymbol{\theta}$ is not a random variable: it simply has some (unknown) value. Hence, it makes no sense to speak of the *probability* that $\boldsymbol{\theta}$ assumes some specified value, say $\boldsymbol{\theta}_0$. The foregoing leads to the following definition:

Definition 2.6. *The* likelihood *of the parameter* $\boldsymbol{\theta}$ *at the observations* $\boldsymbol{x} = (x_1, \ldots, x_n)$ *is*

$$L_{\boldsymbol{x}}(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}\{\boldsymbol{X} = \boldsymbol{x}\} \tag{2.11}$$

if the distribution of $\boldsymbol{X}$ *is discrete, and*

$$L_{\boldsymbol{x}}(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{2.12}$$

if the distribution of $\boldsymbol{X}$ *is continuous. As usual,* $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ *denotes the joint probability density of* $\boldsymbol{X}$ *at the point* $\boldsymbol{x}$ *when* $\boldsymbol{\theta}$ *is the value of the unknown parameter.*

Remark. If $X_1, X_2, \ldots, X_n$ are independent with common density $f_{\boldsymbol{\theta}}(x)$, we have, obviously,

$$L_{\boldsymbol{x}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i)\ . \tag{2.13}$$

Definition 2.7. *The* maximum likelihood estimate *of the unknown true value,* $\boldsymbol{\theta}_0$*, of a parameter* $\boldsymbol{\theta}$ *is that value of 'the running variable'* $\boldsymbol{\theta}$ *which maximises* $L_{\boldsymbol{x}}(\boldsymbol{\theta})$*. This estimate is denoted by* $\hat{\boldsymbol{\theta}}_{ML}$*, or, if the property of being a maximum likelihood estimator is clear from the context, merely by* $\hat{\boldsymbol{\theta}}$*.*

Notation. A notational distinction between $\boldsymbol{\theta}_0$, the unknown true value of the parameter $\boldsymbol{\theta}$, and the parameter $\boldsymbol{\theta}$ itself or even the 'running variable' $\boldsymbol{\theta}$ is often not made.

Example. Consider the case of a normal family: $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$. What are the ML estimators of μ and σ^2? First, we write the likelihood function:

$$L_{\boldsymbol{x}}(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2}. \tag{2.14}$$

Since $L_{\boldsymbol{x}}(\mu, \sigma^2)$ and $l_{\boldsymbol{x}}(\mu, \sigma^2) = \log L_{\boldsymbol{x}}(\mu, \sigma^2)$ assume their maxima at the same value of (μ, σ^2), we seek the location of the maximum of

$$\log L_{\boldsymbol{x}}(\mu, \sigma^2) = -n \log(2\pi\sigma^2)^{\frac{1}{2}} - \frac{1}{2} \sum_{i=1}^{n} \big(\frac{x_i - \mu}{\sigma}\big)^2. \tag{2.15}$$

For each fixed σ^2, the ML estimator for μ can be found by solving

$$\frac{\partial}{\partial \mu} \log L_{\boldsymbol{x}}(\mu, \sigma^2) = 0 \,, \tag{2.16}$$

which leads in the normal case to

$$\frac{1}{\sigma} \sum_{i=1}^{n} \Big(\frac{x_i - \mu}{\sigma} \Big) = 0 \,. \tag{2.17}$$

Hence, we have the estimate $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and the estimator $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Since the estimator $\hat{\mu}$ happens not to depend on σ^2, it can be used for all σ^2, hence σ^2 need not be known. We now maximise (2.15) with respect to σ^2 for each value of μ:

$$\frac{\partial}{\partial \sigma^2} \log L_{\boldsymbol{x}}(\mu, \sigma^2) = -\frac{1}{2} \frac{n}{\sigma^2} + \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^4} = 0 \tag{2.18}$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 \,. \tag{2.19}$$

This is the ML estimate for σ^2 *if* μ *is known.*

Exercise 2.3. Show that in this case $\mathrm{E}\hat{\sigma}^2 = \sigma^2$, hence $\hat{\sigma}^2$ is unbiased!

If μ is not known, then one gets the global maximum of (2.15) by inserting the ML estimator $\hat{\mu}$, from (2.16), into (2.18). The resulting estimator $\frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2$ corresponds to the sample analogue, but is not unbiased. The important feature is, however, that it is a maximum likelihood estimator and, hence, enjoys the properties of (i) having a distribution which is asymptotically normal, (ii) being asymptotically unbiased and (iii) being efficient. The latter means that the estimator has, asymptotically, the least possible standard deviation. Since the bias tends to zero as $1/n$ and the standard deviation as $1/\sqrt{n}$ (see the application after Theorem 2.2), $\hat{\sigma}^2$ tends to σ^2 in probability (in fact in any of the several modes of convergence described in Defs. 1.29 and 1.30 from Chap. 1.9). Therefore, the estimator $\hat{\sigma}^2$ is called *consistent*, which is a somewhat stronger property than being asymptotically unbiased. These prominent features of maximum likelihood estimates have been essentially proven and advocated by Fisher, see [198, 204, 205]. Later the proofs have been refined and extended to more general circumstances. One can directly derive from the definition of maximum likelihood the following property which is so important that we formulate it as a theorem.

Theorem 2.1. *If* $\hat{\boldsymbol{\theta}}$ *is ML for* $\boldsymbol{\theta} \in \Theta$*, and* $\boldsymbol{g}$ *is a continuous function* $\Theta \to \mathbb{R}^p$ *then* $\boldsymbol{g}(\hat{\boldsymbol{\theta}})$ *is ML for* $\boldsymbol{g}(\boldsymbol{\theta})$*.*

Example. $\sqrt{\hat{\sigma}^2}/\hat{\mu}$ is the maximum likelihood estimator for the coefficient of variation σ/μ.

Remark. Note that the property of unbiasedness is *not* preserved under a non-linear transformation. However, in view of the asymptotic error propagation formula, approximate unbiasedness holds for large samples if $f''(\mu)\sigma^2 \ll 1$, see Chap. 1.8. For a discussion of the relationship between the likelihood function and, in a Bayesian context, posterior probabilities when flat prior distributions are used, see Question 13 in [358].

Exercise 2.4. Let $X_1, X_2, \ldots, X_n$ be i.i.d. with an exponential E (λ) distribution with density $f_\lambda(x) = \lambda^{-1} e^{-x/\lambda}$. Determine the ML estimator of λ, and calculate $\mathrm{E}\hat{\lambda}_{ML}$ and $\operatorname{var}\hat{\lambda}_{ML}$.

The maximum likelihood procedure has an important follow-up by the fact that the likelihood function contains information on the accuracy of the maximum likelihood estimator:

Theorem 2.2. *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ have a joint density $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is a p-dimensional parameter. Under weak conditions ($\hat{\boldsymbol{\theta}}_{ML}$ in the interior of the space Θ, and the likelihood being sufficiently differentiable), a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}_{ML}$ is given by $\boldsymbol{I}(\hat{\boldsymbol{\theta}}_{ML})^{-1}$, where*

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\frac{\partial^2 \log L_{\boldsymbol{x}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} = \left(-\frac{\partial^2 \log L_{\boldsymbol{x}}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right)_{p \times p}. \tag{2.20}$$

Proof. A heuristic derivation is as follows. Consider the fact that, in a Bayesian context, the normalised likelihood function, possibly multiplied with a prior distribution which we set for simplicity to be unity here, is sometimes used as a probabilistic expression of one's opinion. Applying a Taylor expansion of

$$l_{\boldsymbol{x}}(\boldsymbol{\theta}) = \log L_{\boldsymbol{x}}(\boldsymbol{\theta}) \tag{2.21}$$

around the ML estimator $\hat{\boldsymbol{\theta}}$:

$$l_{\boldsymbol{x}}(\boldsymbol{\theta}) = l_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}) + \left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} l_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}) \right)^t (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) + \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} l_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}) \right) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + R, \tag{2.22}$$

where the remainder, R, is neglected and the coefficient of the linear term, i.e., the derivative of the log-likelihood at $\hat{\boldsymbol{\theta}}$, vanishes, one can see that $e^{l_{\boldsymbol{x}}(\boldsymbol{\theta})}$ has, with respect to $\boldsymbol{\theta}$ the formal appearance of a normal probability density with covariance matrix $-\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} l_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}) \right)^{-1}$. For a formal proof in a frequentist context and a more precise formulation of the conditions, we refer to [695], Chap. 4 of [607] and Chap. 6 of [426].

Remark. $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ is called the *observed Fisher information matrix.* The theorem also holds if it is replaced by the expected Fisher information matrix $\mathrm{E}\, \boldsymbol{I}(\hat{\boldsymbol{\theta}})$. Usually, $\mathrm{E}\, \boldsymbol{I}(\hat{\boldsymbol{\theta}})$ is more useful for theoretical calculations, and $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ is more convenient for computer implementation.

Application. Consider the case of a normal family. From (2.16) one can see that

$$\frac{\partial}{\partial(\sigma^2)}\frac{\partial}{\partial\mu}\log L_{\boldsymbol{x}}(\mu,\sigma^2)\Big|_{\substack{\mu=\hat\mu\\ \sigma^2=\hat\sigma^2}} = 0\,, \tag{2.23}$$

hence $\hat\mu$ and $\hat\sigma^2_{ML}$ are (at least asymptotically) uncorrelated, and one can obtain the inverted Fisher information matrix by inverting just the two diagonal elements:

$$\hat{\mathrm{var}}\,(\hat\mu) = -\frac{\partial^2}{\partial\mu^2}\Big(\log L_{\boldsymbol{x}}(\mu,\sigma^2)\Big)^{-1}\Big|_{\substack{\mu=\hat\mu\\ \sigma^2=\hat\sigma^2}} = \Big(\frac{n}{\sigma^2}\Big)^{-1}\Big|_{\sigma^2=\hat\sigma^2} = \frac{\hat\sigma^2}{n}\,. \tag{2.24}$$

This is a familiar result. It is just the maximum likelihood estimator as well as the sample analogue estimator of $\frac{\sigma^2}{n} = \mathrm{var}\,\hat\mu$.
But now,

$$\begin{aligned}\hat{\mathrm{var}}\,(\hat\sigma^2) &= -\frac{\partial}{\partial\tau}\Big(-\frac{n}{2\tau}+\frac{\sum(X_i-\mu)^2}{2\tau^2}\Big)^{-1}\Big|_{\tau=\hat\tau} = \Big(\frac{-n}{2\tau^2}+\frac{n\tau}{\tau^3}\Big)^{-1}\Big|_{\tau=\hat\tau}\\ &= \frac{2\hat\tau^2}{n} = \frac{2\hat\sigma^4}{n}\,,\end{aligned} \tag{2.25}$$

where τ has been used as an abbreviation of σ^2. It is of course possible, but slightly more laborious, to find this by other means.

Exercise 2.5. Find $\hat{\mathrm{var}}\,\hat\sigma^2$ by using characteristic and/or cumulant generating functions.

Definition 2.8. *The* standard deviation *of an estimator* $\hat\theta$ *is* $\sigma(\hat\theta) = \sqrt{var\,(\hat\theta)}$.

At least for reasonably large samples, one estimates the standard deviation by $\hat\sigma(\hat\theta) = \sqrt{\hat{\mathrm{var}}\,(\hat\theta)}$. The precision $\sigma(\hat\theta)/\mathrm{E}(\hat\theta)$ of an estimator is in practice often estimated by $\hat\sigma(\hat\theta)/\hat\theta$.

Exercise 2.6. Consider the case of a normal family with $\mu = 5\sigma$. How large must the sample size n be in order to estimate μ with a precision of 1 %? How large must n be to estimate σ^2 with this precision?

Let us now look at the full distributions of $\hat\mu = \frac{1}{n}\sum_i X_i$ and $\hat\sigma^2_{ML} = \frac{1}{n}\sum_{i=1}^n (X_i-\hat\mu)^2$. Obviously, $\hat\mu \sim N(\mu, \frac{\sigma^2}{n})$. The distribution of $\hat\sigma^2_{ML}$ is somewhat more complicated. From Sect. 1.5 (see the χ^2 distribution) one can easily derive that $\sum_{i=1}^n \Big(\frac{X_i-\mu}{\sigma}\Big)^2 \sim \chi^2_n$, hence $\hat\sigma^2_{ML} \sim \frac{1}{n}\sigma^2\chi^2_n$. If μ is replaced by $\hat\mu$, it turns out that the χ^2 distribution 'loses one degree of freedom'. Hence, $\hat\sigma^2_{ML} \sim \frac{1}{n}\sigma^2\chi^2_{n-1}$, i.e., $\hat\sigma^2_{ML}$ is distributed as $\frac{1}{n}\sigma^2$ times a χ^2_{n-1} distributed random variable.

Exercise 2.7. Derive from this last result the exact variances of $\hat\sigma^2_{ML}$ and $\hat\sigma^2_{\mathrm{unbiased}} = \frac{n}{n-1}\hat\sigma^2_{ML}$. Which of the two is larger?

Exercise 2.8. Determine MSE $(\hat{\sigma}^2_{ML})$ and MSE $(\hat{\sigma}^2_{\text{unbiased}})$. Determine m such that $\hat{\sigma}^2_m = \frac{1}{n-m}\sum_{i=1}^n (X_i - \hat{\mu})^2$ has *minimal* mean squared error. Remark. Isn't the result, at first sight, a bit surprising? The moral of this little exercise is that one sometimes has to pay for the restriction to unbiasedness. The resulting estimator may have a larger MSE than is achievable without this restriction.

As a special illustration, we shall now look at maximum likelihood estimators for the parameters of a $\Gamma_{f,g}$ distribution. Let $X_1, X_2, \ldots$ be independent random variables with common distribution $\Gamma_{f,g}$. In this case,

$$n^{-1}\log L_{\boldsymbol{x}}(\boldsymbol{\theta}) = -f\log g - \log\Gamma(f) + (f-1)n^{-1}\Big(\sum_{i=1}^n \log x_i\Big) - g^{-1}n^{-1}\Big(\sum_{i=1}^n x_i\Big). \tag{2.26}$$

By comparing (2.26) with Def. 1.21 one can see that this is an exponential family with $(\theta_1, \theta_2) = (f-1, -g^{-1})$ as natural parameter and

$$(S_1, S_2) = \Big(n^{-1}\sum_{i=1}^n \log x_i, n^{-1}\sum_{i=1}^n x_i\Big) \tag{2.27}$$

as associated sufficient statistic. The normalising function equals

$$\psi(\theta) = -(\theta_1+1)\log(-\theta_2) + \log\Gamma(\theta_1+1)\ , \tag{2.28}$$

whence

$$\mathrm{E}S_1 = \frac{\partial}{\partial\theta_1}\psi(\boldsymbol{\theta}) = \log g + \frac{\partial}{\partial f}\log\Gamma(f) \tag{2.29}$$

and

$$\mathrm{E}S_2\ . \tag{2.30}$$

By subtracting the last two equations we get $\log \mathrm{E}S_2 - \mathrm{E}S_1 = \log \mathrm{E}\overline{X} - \mathrm{E}\overline{\log X} = \log f - \frac{\partial}{\partial f}\log\Gamma(f)$. Hence, the ML estimator for f satisfies

$$\log f - \frac{\partial}{\partial f}\log\Gamma(f) = \log\overline{X} - \overline{\log X}\ , \tag{2.31}$$

which can obviously also be obtained by direct differentiation of the log likelihood, (2.26). Note that, by convexity of the logarithm, $\log\overline{X}$ is always larger than $\overline{\log X}$. Equation (2.31) can be solved numerically to obtain $\hat{f}_{ML}$. Corresponding tables have been made. For an overview, see paragraph 7.2 of Chap. 17 in [328]. To a good practical approximation, with a relative error of less than 1% for $2/3 \le f \le 400$ (i.e., $0.0012 < R_n < 0.91$) and less than 4% for $0.5 \le f \le 2/3$ (i.e., $0.91 < R_n < 1.42$), (2.31) is inverted by

$$\hat{f}_{ML} \simeq 1/8 + \frac{1}{2R_n}\ , \tag{2.32}$$

where we have written R_n as an abbreviation for $\log \overline{X} - \overline{\log X}$, for n observations.[6] The maximum likelihood estimator of the scale parameter g is $\hat{g}_{ML} = \overline{X}/\hat{f}_{ML}$. By inverting the Fisher information matrix

$$n \begin{pmatrix} \frac{\partial^2}{\partial f^2} \log \Gamma(f) & \frac{1}{g} \\ \frac{1}{g} & \frac{f}{g^2} \end{pmatrix} \tag{2.33}$$

we obtain the following asymptotic covariance matrix,

$$\begin{pmatrix} \frac{\operatorname{var}(\hat{f})}{f^2} & \frac{\operatorname{cov}(\hat{f},\hat{g})}{fg} \\ \frac{\operatorname{cov}(\hat{f},\hat{g})}{fg} & \frac{\operatorname{var}(\hat{g})}{g^2} \end{pmatrix} = \frac{n^{-1} f^{-1}}{f \frac{\partial^2}{\partial f^2} \log \Gamma(f) - 1} \begin{pmatrix} 1 & -1 \\ -1 & f \frac{\partial^2}{\partial f^2} \log \Gamma(f) \end{pmatrix}. \tag{2.34}$$

The expression contains the trigamma function $\frac{\partial^2}{\partial f^2} \log \Gamma(f) = \psi'(f)$, which can in practice be approximated in a similar way, whence we obtain

$$\begin{pmatrix} \frac{\operatorname{var}(\hat{f})}{f^2} & \frac{\operatorname{cov}(\hat{f},\hat{g})}{fg} \\ \frac{\operatorname{cov}(\hat{f},\hat{g})}{fg} & \frac{\operatorname{var}(\hat{g})}{g^2} \end{pmatrix} = n^{-1} h(f) \begin{pmatrix} 1 & -1 \\ -1 & h_g(f) \end{pmatrix}, \tag{2.35}$$

where

$$h(f) = \frac{2(f + 0.25)}{f + 0.61} + C_h(f) \tag{2.36}$$

and

$$h_g(f) = f\psi'(f) = 1 + \frac{1}{f} - \frac{0.53}{f + \frac{1}{2}} + C_{h_g}(f)\,. \tag{2.37}$$

The correction terms[7] are practically small: For $2/3 < f < 500$, $|C_h(f)| < 5 \cdot 10^{-3}$ and $|C_{h_g}(f)| < 2.5 \cdot 10^{-3}$. The covariance matrix is estimated by substituting the maximum likelihood estimates for f and g. The maximum likelihood estimator $\hat{f}_{ML}$ can be compared with the moment estimator $\hat{f}_m = \frac{\overline{X}^2}{(n-1)^{-1} \sum_i (X_i - \overline{X})^2}$, which generally has a larger bias and a larger variance as well (see, e.g., [159]).

[6] This approximation, derived by the author while writing this book, was found by applying regression analysis using SAS/PROC REG [581]. It simplifies the classical, more accurate approximation due to Greenwood and Durand [247], the approximation by Thom [675], as well as the series expansion in [78], see Chap. 17 in [328]. A next order approximation is $\hat{f}_{ML} = 0.11582 + 0.499916/R_n + \sum_{j=1}^{4} 10^{-3} c_j \log^j(R_n)$, with $c_1 = -37.56$, $c_2 = -9.2$, $c_3 = -0.6$, $c_4 = 0.048$, and an absolute error less than $2.5 \cdot 10^{-3}$ for $0.5 < f < 400$, which is only slightly less accurate than the two-region Greenwood–Durand formula, see [78].

[7] They can be approximated by $C_h(f) = 10^{-3}\big(2.2 + 5.7/f - 0.4 \log f + 1.9 f^{-1} \log f - 5.6 f^{-1} \sin(\frac{\pi}{2} \log f) - 10.0 f^{-1} \cos(\frac{\pi}{2} \log f)\big) + R_h(f)$, and by $C_{h_g}(f) = 10^{-3}\big(-0.3 + 5.82/f - 0.016 \log f + 2.47 f^{-1} \log f - 1.1 f^{-1} \sin(\frac{\pi}{2} \log f) - 3.8 f^{-1} \cos(\frac{\pi}{2} \log f)\big) + R_{h_g}(f)$, respectively, with $|R_h(f)| < 2 \cdot 10^{-4}$ and $|R_{h_g}(f)| < 4 \cdot 10^{-4}$, for $0.5 < f < 400$.

As a side remark, it is noted that a convenient analytic approximation of the log gamma function, of a similar kind as above for its second derivative, is given by

$$\log\Gamma(f) = (f - \frac{1}{2})\log f - f + 0.9189 + 0.084 f^{-1} + C(f)\ , \qquad (2.38)$$

where $C(f) = -10^{-3}(2.2 f^{-2} + 0.74 f^{-3}) + R(f)$, with $|R(f)| < 1.6 \times 10^{-4}$ for $0.5 < f < 400$ and where $0.9189 \simeq 0.5\log(2\pi)$, see Binet's expansion in Chap. XII of [732]. A further approximation of $\log\Gamma_r(f) = \log\Gamma(f) - \big((f - \frac{1}{2})\log f - f + 0.5\log(2\pi)\big)$, useful for numerical purpose, is given by

$$\begin{aligned}\log\Gamma_r(f) = {} & 10^{-3}(-0.00220 + 23.824 f^{-1}\log(1 + 0.64 f^{-1}) + \\ & + 83.346 f^{-1} - 15.0105 f^{-2} + 0.938715 f^{-3}) + \\ & + 10^{-6}\big(2.42\cos(\frac{2.3\pi}{f + \frac{1}{4}}) - 3.68\sin(\frac{2.3\pi}{f + \frac{1}{4}})\big) + R_2(f)\ , \end{aligned} \qquad (2.39)$$

where the (oscillating) error term $R_2(f)$ satisfies $|R_2(f)| < 2 \times 10^{-7}$ for $0.5 < f < 400$. In the region $0 < f < 0.5$, the well-known formula $\log\Gamma(0.5 - f) = \log\frac{\pi}{\cos\pi f} - \log\Gamma(0.5 + f)$ can be applied. Equations (2.38) and (2.39), which improve upon the approximations to $\Gamma(f)$ in Chap. 6 of [552] and in Chap. 3.3 of [33], can be used to approximate the Beta function as well. In Table 2.1 below, we present simple approximations to the first four derivatives of $\log\Gamma(f)$, which can be used to approximate the expectation, variance, skewness and (excess of) kurtosis of the Beta logistic distribution, see Table 1.2.

Table 2.1. Simple analytic approximations for the scaled polygamma functions $f^k\psi^{(k)}(f) = f^k(\partial/\partial f)^{(k+1)}\log\Gamma(f)$ $(k = 1, 2, 3)$. The abbreviation flf stands for $f\log f$. The absolute errors are less than $\pm 2 \times 10^{-4}$ for $f\psi(f)$ and $f\psi'(f)$ and less than $\pm 10^{-3}$ for $f^2\psi''(f)$ and $f^3\psi'''(f)$.

$0 < f < 1$	$1 < f < 30$	$f > 30$
$f\psi(f) = f\psi(1+f) - 1$	$flf - \frac{1}{2} - \frac{0.117}{f} + \frac{0.0335}{f+\frac{1}{2}} + \frac{0.0175}{f^2}$	$flf - \frac{1}{2} - \frac{1}{12f}$
$f\psi'(f) = f\psi'(1+f) + \frac{1}{f}$	$1 + \frac{0.781}{f} - \frac{0.286}{f+\frac{1}{2}} + \frac{0.055}{f^2}$	$1 + \frac{1}{2f} + \frac{1}{6f^2}$
$f^2\psi''(f) = f^2\psi''(1+f) - \frac{2}{f}$	$-1 - \frac{2.15}{f} + \frac{1.168}{f+\frac{1}{2}} - \frac{0.033}{f^2}$	$-1 - \frac{1}{f} - \frac{1}{2f^2}$
$f^3\psi'''(f) = f^3\psi'''(1+f) + \frac{6}{f}$	$2 + \frac{5.11}{f} - \frac{2.14}{f+1/2} + \frac{1.21}{f^2} - \frac{0.4}{f^3}$	$2 + \frac{3}{f} + \frac{2}{f^2}$

Factorising the likelihood function

In various contexts, the likelihood function can be factorised, which allows to identify sufficient and ancillary statistics, two important concepts coined

and broached by R.A. Fisher.

Sufficiency. As stated before in Chap. 1, Sect. 1.5.3, in intuitive terms a sufficient statistic summarises all essential information about a parameter $\boldsymbol{\theta}$ which is present in the data, given a particular statistical model characterised by a family of probability distributions. This means that the distribution of the data given the sufficient statistic does not depend on $\boldsymbol{\theta}$, which can be viewed as a formal definition of sufficiency. A useful criterion is the factorisation lemma.

Theorem 2.3. (*factorisation criterion*) *Let* $\{P_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$ *be a family of probability distributions with densities* $f_{\boldsymbol{\theta}}$ *with respect to a* $\sigma-$*finite measure* λ (*in practice the Lebesgue measure or, for discrete distributions, the counting measure*), *and let* $X_1, \ldots, X_n$ *be i.i.d.* (*independent and identically distributed*) *with common distribution* $P_{\boldsymbol{\theta}}$. *Then,* $\boldsymbol{S}(X_1, \ldots, X_n)$ *is a sufficient statistic for the parameter* $\boldsymbol{\theta}$ *if there exist non-negative functions* $g_{\boldsymbol{\theta}}$ *and* $h(\boldsymbol{x})$, *such that*

$$l_{\boldsymbol{x}}(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto g_{\boldsymbol{\theta}}\big(\boldsymbol{s}(\boldsymbol{x})\big)h(\boldsymbol{x}) \; . \tag{2.40}$$

Proof. See [191] and Chap. 2 in [424], among others. In the latter book, by Lehmann, the theorem is formulated while paying more attention to measurability aspects. Examples where the factorisation criterion can be fruitfully applied are given by exponential families, see Chap. 1.

Ancillarity. Basically, an ancillary statistic, complementary to a sufficient statistic, is a summary of the data, the distribution of which does not depend on the parameter of interest $\boldsymbol{\theta}$. Ancillary statistics are used in compound models and in the context of conditional as well as, sometimes, of unconditional inference. An ancillary statistic is, for instance, the estimator of a random sample size, with a distribution independent of the single parameter of interest for any fixed sample size. They are especially useful when a (physically motivated) distinction can be made between the parameters of primary interest $\boldsymbol{\theta}_1$ and those of secondary interest $\boldsymbol{\theta}_2$, the latter ones also being called ancillary parameters or, depending on the context, 'nuisance parameters'. A formal definition is as follows (see, e.g., [546]).

Definition 2.9. *Let* $X_1, \ldots, X_n$ *be i.i.d. with common density* $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. *If*

$$\big(\boldsymbol{S}(X_1, \ldots, X_n), \boldsymbol{A}(X_1, \ldots, X_n)\big) \tag{2.41}$$

is sufficient for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ *and the likelihood function can be partitioned as*

$$l_{\boldsymbol{x}}(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2}(\boldsymbol{x}) \propto f_{\boldsymbol{\theta}_1}\big(\boldsymbol{s}(\boldsymbol{x})|\boldsymbol{a}(\boldsymbol{x})\big) f_{\boldsymbol{\theta}_2}\big(\boldsymbol{a}(\boldsymbol{x})\big) \; , \tag{2.42}$$

then $\boldsymbol{A}$ *is called ancillary to* $\boldsymbol{S}$ *for* $\boldsymbol{\theta}_1$, *in view of the fact that the second factor does not depend on* $\boldsymbol{\theta}_1$.

Remarks.

1. In view of the factorisation criterion, in the above situation, $\boldsymbol{S}$ is called sufficient for $\boldsymbol{\theta}_1$ conditional on $\boldsymbol{A}$ and in the presence of $\boldsymbol{\theta}_2$, see e.g. [124]. The statistic $\boldsymbol{A}$, while being ancillary for $\boldsymbol{\theta}_2$, is, by basic definition, sufficient for $\boldsymbol{\theta}_2$ since the first factor does not depend on $\boldsymbol{\theta}_2$.
2. Although an ancillary statistic does not provide direct information on the value of $\boldsymbol{\theta}_1$, it usually does give information on the precision with which $\boldsymbol{\theta}_1$ can be estimated by the corresponding conditional sufficient statistic, as is obvious from the instance of a random sample size. Serving as a 'precision index' is a generic property of an ancillary statistic, which holds in many instances, but not always, see, e.g., [91].

Example. (Slightly adapted from [124].) Consider a simple measuring device for a parameter θ_1 with precision θ_2, i.e., let $X_1, \ldots, X_n$ be uniformly distributed on the interval $(\theta_1 - \theta_2, \theta_1 + \theta_2)$. Now, $(\min_{i=1}^n X_i, \max_{i=1}^n X_i)$ is a sufficient statistic for (θ_1, θ_2). $A = \max_{i=1}^n X_i - \min_{i=1}^n X_i$ is independent of θ_1. Hence, A is an ancillary statistic for θ_1 and a sufficient statistic for θ_2. It makes sense to base any inference for θ_1 on $S = \frac{1}{2}(\max_{i=1}^n X_i + \min_{i=1}^n X_i)$ given the value of the value of A. The outcome of A gives information on the precision with which θ_1 can be estimated.

2.2.2 Hypothesis Testing

We consider again $X_1, \ldots, X_n$ as a series of independent and identically distributed random variables with common distribution function $F_{\boldsymbol{\theta}}$ for some $\boldsymbol{\theta} \in \Theta$ and fixed sample size n. Instead of estimating $\boldsymbol{\theta}$, one may test the hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ for a specific value $\boldsymbol{\theta}_0$ which is, for some reason or another, of special interest. One way to do this is to find a *test statistic*, i.e., a suitable function $T = h(X_1, \ldots, X_n)$ of the random variables which estimates θ, and to determine its probability distribution under the assumption that the hypothesis is true. Such a test statistic should 'destroy' as little information as possible from the original sample. As we have seen in Sect. 2.2, a formalisation of this concept is provided by the notion of a sufficient test statistic. If the outcome of the test statistic is 'far away' from the center of the distribution F_{θ_0}, which is in practice (at least for a unimodal probability distribution, see Chap. 1.5) equivalent to the combined statement that if, under repeated sampling, $P_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\{T \notin (t_1, t_2)\}$ is small for some suitable t_1 and t_2, and if the outcome t of T is outside the interval (t_1, t_2), then the plausibility of the null–hypothesis to hold is considered to be low,[8] and the

[8] More precisely stated: If $t \notin (t_1, t_2)$ then either the null–hypothesis does not hold, or one has to face a low probability for the event that the outcome of the test statistic T (under repeated sampling) is at least as 'extreme' as the presently observed outcome t. It is noted that the null–hypothesis itself is composed of two elements, each of which can be violated: (a) the common distribution function of $X_1, \ldots, X_n$ is $F_{\boldsymbol{\theta}_0}$, and (b) $X_1, \ldots, X_n$ are independent and identically distributed.

null–hypothesis is in practice rejected. The interval (t_1, t_2) is called the acceptance region and its complement $(t_1, t_2)^c$ the rejection region of the test statistic. It is sometimes a non-trivial issue to give a precise meaning of the expression 'at least as extreme'. For symmetric distributions where θ is a location parameter, the situation is clear since several different choices coincide. In that case, the quantity $P_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\{|T| > t\}$ is called, after Fisher [205], the p-value corresponding to the observed value t of the test statistic T. Fisher also suggested we reject H_0 if the p-value is smaller than a conventional cut-off value, such as 5% or 1%, which is called the *level* of the hypothesis test. This simple situation is treated in the following example.

Example. Consider the normal case and test H_0: $\theta = 2$, where σ^2 is supposed to be known. Since $\hat{\theta}$ is a good estimator for θ, it seems intuitively clear that

true state of nature / action taken	$\theta \in H_0$	$\theta \notin H_0$
H_0 rejected	α_θ	$1 - \beta_\theta$
H_0 accepted	$1 - \alpha_\theta$	β_θ

Fig. 2.1. The two types of error that can occur in 'binary' hypothesis testing, where the parameter space Θ is divided into two regions. The region where 'the null–hypothesis holds' and its complement.

$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a good one-dimensional reduction of the data.[9] Under H_0:

[9] It is in fact the canonical test statistic for testing hypotheses on θ in the normal case. Of course, one can equally well work with $T = \sum_{i=1}^{n} X_i$ or $\hat{\theta}^3$, etc., instead of with $\hat{\theta}$.

$$\hat{\theta} \sim N(2, \frac{\sigma^2}{n}) \Rightarrow \frac{\hat{\theta} - 2}{\sigma/\sqrt{n}} \sim N(0,1) \; . \tag{2.43}$$

Hence, H_0 is rejected at level α if

$$\frac{\hat{\theta} - 2}{\sigma/\sqrt{2}} \notin (-u_{\frac{\alpha}{2}}, u_{\frac{\alpha}{2}}) \; , \tag{2.44}$$

where u_α is a number such that $P\{X > u_\alpha\} = \alpha$ for $X \sim N(0,1)$, i.e.,

$$P\{X > u_\alpha\} = \frac{1}{\sqrt{2\pi}} \int_{u_\alpha}^{\infty} e^{-\frac{1}{2}x^2} dx \; . \tag{2.45}$$

For a normal distribution, the notation $\Phi(u_\alpha) = P\{X < u_\alpha\} = 1 - P\{X > u_\alpha\}$ is used, hence, $u_\alpha = \Phi^{-1}(1-\alpha) = -\Phi^{-1}(\alpha)$. A few selected points are: $(\alpha, u_\alpha) = (17\%,1)$, $(5\%,1.65)$, $(2.5\%,1.96)$, $(0.1\%,3)$. For simple analytic approximations to $\Phi(x)$, for $0 < x < 3$, and to $\Phi^{-1}(p)$, for $0.5 < p < 0.998$, see [429, 502, 618] and the Appendix in [350].

Notice that under repeated sampling the p-value itself is uniformly distributed on the interval $[0,1]$, and that, hence, in the situation that the null–hypothesis holds on the long run the fraction of times that the p-value is smaller than a predetermined, fixed level α is equal to α. This is the frequency interpretation of the level α of a test statistic T. Obviously, in classical statistics, the p-value should not be construed as the 'probability that the null–hypothesis is correct'. In Bayesian statistics, the (posterior) probability that θ_0 is the true value of the parameter θ is zero for (prior and well as posterior) probability distributions that are continuous, whereas in classical statistics the event $\theta = \theta_0$ does not have a frequency interpretation under repeated sampling. Although intuitively, a high p-value may indicate that some credence can be given to the null–hypothesis, this should not be done indiscriminately. For instance, the sample size plays a role. For small samples rather high p-values are easily obtained, even if H_0 does not hold, and the reverse holds true for large samples. It is a subject of current investigation to replace p-values by other estimators which express more adequately the 'epistemic' probability for the null–hypothesis to hold, see [358, 359, 401, 590], among others.

For asymmetric, but still unimodal, distributions, many different choices of t_1 and t_2 exist such that the sum of the two tail probabilities, $\int_{-\infty}^{t_1} f_{\theta_0}(t)dt + \int_{t_2}^{\infty} f_{\theta_0}(t)dt$ is equal to α. It is customary in that case to split the level of the test in equal parts for the left–hand and right–hand tails of the distribution, i.e., a two-sided test of level α is performed by taking as rejection region $(t_1, t_2)^c$, such that $P_{H_0}\{T < t_1\} = \alpha/2$ and $P_{H_0}\{T > t_2\} = \alpha/2$. This approach has the desirable feature that the procedure for testing H_0 at level α is equivariant under (monotonic) transformations of the test statistic, i.e., the critical region (t_1, t_2) for T transforms to the critical region $\big(g(t_1), g(t_2)\big)$

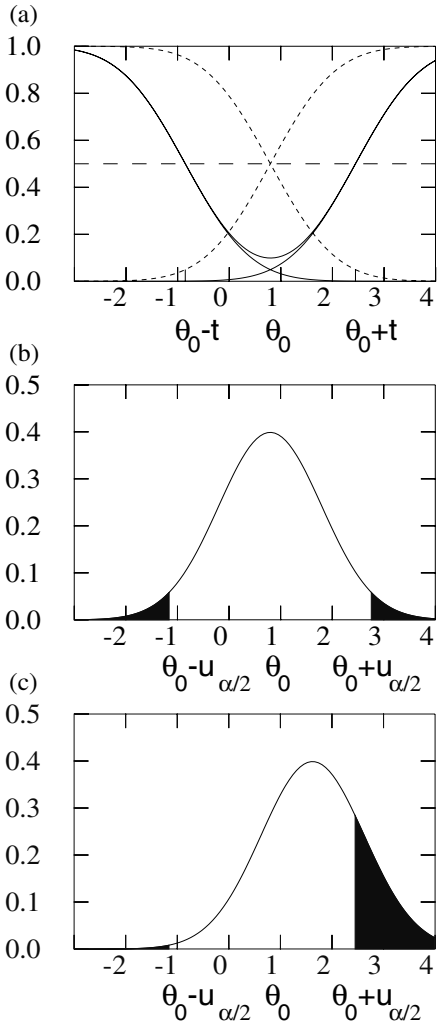

Fig. 2.2. (a) The U–shaped power function $P_\theta\{|T-\theta_0| > t\} = 1 + \Phi(\theta - t) - \Phi(\theta + t)$ and its constituents (for some t, $t = 1.65$) as a function of θ for a standard normally distributed test statistic T in the case of two-sided testing. For $\theta = \theta_0$, the power function is equal to α (error of the first kind). For $\theta \neq \theta_0$, the power function equals $1 - \beta_\theta$, where β_θ is the probability to make an error of the second kind. (b) The normal probability density function, centered around θ_0; the *black area* denotes the probability, α, to reject the null–hypothesis H_0: $\theta = \theta_0$ if in fact this null–hypothesis holds true (error of the first kind). (c) Shifted normal probability density function; the *white area* denotes the probability, β_θ, to accept the null–hypothesis H_0: $\theta = \theta_0$ if in fact the null–hypothesis does not hold (error of the second kind). The alternative hypothesis plotted here is A: $\theta = \theta_0 + t/2$ for some t, $t = 1.65$.

for any (monotonic) transformation g.[10] A viable alternative to the equal-tail area procedure is to plot the likelihood function $f_\theta(t)$ *as a function of* θ and reject H_0 if the likelihood ratio $f_\theta(t)/f_{\hat{\theta}}(t)$ is smaller than some critical value. This approach can be easily generalised to several dimensions and is equivariant under (monotonic) transformations. Moreover, it has gained popularity since numerical computations have been considerably facilitated with the advent of computing software. Nevertheless, in many dimensions, the 'curse of the dimension' still exists.

Let us now take up the historical thread again. Around 1930, Neyman & (Egon) Pearson (see [494]), extended the simple hypothesis-testing theory of Fisher, which was heavily inclined to looking at the distribution of T *only* under the null–hypothesis H_0. One should also look at what happens if $\boldsymbol{\theta} \notin H_0$. In some, even if limited, generality the possibilities are displayed in Fig. 2.1, from which one can notice that two types of error exist. For simplicity we now consider the special case that H_0 consists of 1 point only. Traditionally,

$$\alpha = P_{\boldsymbol{\theta} \in H_0}\{|T| > t\} \tag{2.46}$$

is used to denote the probability of rejecting H_0, when $\boldsymbol{\theta} \in H_0$ (*error of the first kind*), and

$$\beta_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta} \notin H_0}\{|T| < t\} \tag{2.47}$$

is used to denote the probability of accepting H_0, when $\boldsymbol{\theta} \notin H_0$ (*error of the second kind*). As an example, we plotted in Fig. 2.2 $P_\theta\{|T| > t\}$ as a function of $\theta \in \mathbb{R}$ for $T \sim N(\theta, 1)$, while considering the null–hypothesis H_0: $\theta = \theta_0$, and alternative A: $\theta \notin \theta_0$. This function is called the *power function of the test statistic* T.

In general, there is a trade-off between α (the error of the first kind) and β_θ (the error of the second kind), the latter one depending on θ. To illustrate this, the relationship between the two types of error is indicated in Fig. 2.3, where the one-sided testing problem is considered for shifted Cauchy distributions using a critical region (t_1, t_2) for rejecting the null–hypothesis.[11] In a slightly different concrete setting, the two types of error play a prominent role in ROC-curve and meta-analysis, see [166, 341, 481].

The Neyman–Pearson theory of hypothesis testing has been re-cast into a decision-theoretic framework by Wald [720]. In this approach, hypothesis testing is considered as a zero-sum two person game of Nature ("Player I") against the Statistician ("Player II"). This leads to attention for minimax

[10] The equivariance property is desirable since if T is a sufficient statistic for θ, then $g(T)$ is a sufficient statistic for $g(\theta)$. Furthermore, it is not shared by the (not customary and not actually recommended) procedure which determines t_1 and t_2 such that the interval (t_1, t_2) has minimum length, i.e., $f_{\theta_0}(t_1) = f_{\theta_0}(t_2)$, even though this procedure is more easily generalised to two or more dimensions.

[11] From the likelihood ratio criterion it follows that H_0 should be rejected for $(t-\theta)^2 < ct^2$ with c smaller than 1. Hence, in contrast to, for instance, the case of normal distributions, it is sensible to consider $t_2 < \infty$ here.

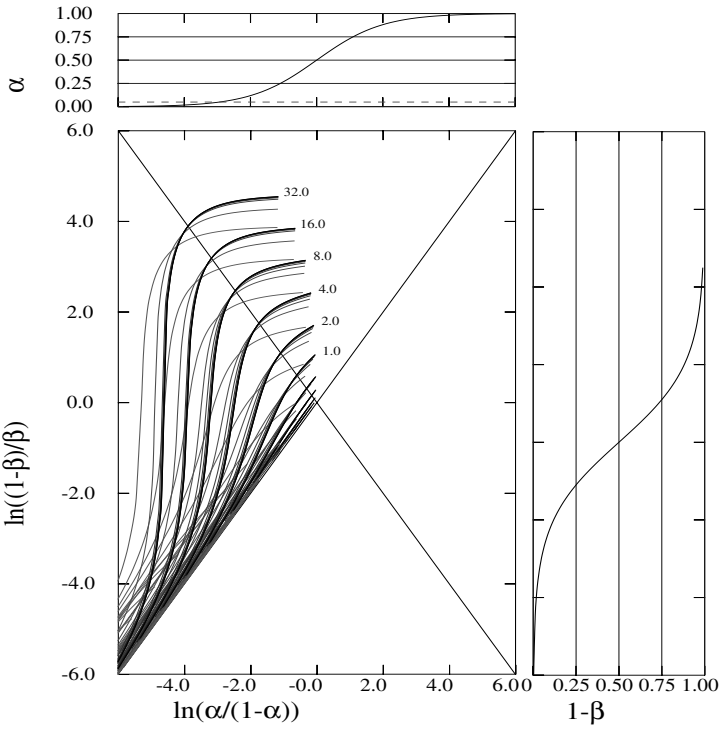

Fig. 2.3. The error of the first kind (on the horizontal axis) is plotted against the power function (1 minus the error of the second kind) on the vertical axis, after logit transformations which are shown in the two panel plots. The curves drawn are for the one-sided testing problem H_0: $\theta = 0$ against A: $\theta > 0$ for a standard Cauchy location problem where H_0 is rejected if the outcome of T falls within the interval (t_1, t_2). The characteristic curves are parametrically plotted using t_1 as a parameter, for various values of the (unknown) physical parameter θ, indicated by the numbers $1.0, 2.0, \ldots, 32.0$, and of t_2. If t_1 increases, both α and $1 - \beta$ decrease. The solid lines correspond to $t_2 = \infty$, the dotted lines to finite values of t_2. Discrimination performance is largest in the left-upper corner of the diagram and improves as θ increases. From the plot one can see that, interestingly, to obtain the 'most powerful' test at a fixed value of α (say 5%), indicated in the upper panel by a dotted line, and relatively small values of θ ($\theta \lesssim 8$), the upper bound of the rejection region, t_2, should be finite (in fact, not too far from the peak of the Cauchy distribution in the alternative). However, t_2 can be taken to be effectively infinite along the line $\alpha = \beta$, and also if one minimises the product of the 'odds of error' $\frac{\alpha}{1-\alpha}\frac{\beta}{1-\beta}$.

rules, which minimise the maximum loss (for the statistician) and are related to least favourable prior distributions in a Bayesian context, see also [191,358]. An extension to the three–decision situation, exemplified by A_{-1}: treatment A is recommended, A_{+1}: treatment B is recommended, and A_0: no recommendation is made, is discussed in Chap. 15 of [347]. A somewhat related area is sequential analysis, where the sample size is not fixed in advance, but experiments are performed consecutively and at each step a decision is made. For instance, A_{-1}: accept the null–hypothesis, A_1: accept the alternative hypothesis, and A_0: continue sampling. This approach too was originally developed by Wald [719]. At present an extensive literature exists. Some books on this subject are [239, 242, 621, 730].

2.2.3 Confidence Intervals

Using the framework of classical probability theory, a rather general theory of confidence regions was developed before the second World War, notably by J. Neyman [493], see also [191, 426]. For a Bayesian approach, see [81]. Some sort of compromise based on distributional inference is discussed in [359, 401]. Here we give a very simplified introduction in one dimension.

Definition 2.10. *A* confidence region *for* $\boldsymbol{\theta}$*, with* confidence coefficient $1-\alpha$*, is the set* $\{\boldsymbol{\theta}_0 \in \Theta | H_0$*:* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ *is not rejected at level* $\alpha\}$*.*

Remark. Loosely speaking, a confidence region consists of those parameters that are consistent with the observed outcome of the experiment. In regular cases, this region is an interval and is called a *confidence interval*.

Example. We want to determine a 95% confidence interval for μ in the normal case, based on n observations, where (a) σ^2 is known and (b) σ^2 is unknown. (a) The canonical test statistic is $\hat{\mu}$ (or a function thereof). H_0: $\mu = \mu_0$ is not rejected at level α if

$$-u_{\frac{1}{2}\alpha} \leq \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \leq u_{\frac{1}{2}\alpha} \; . \tag{2.48}$$

The set of all $\mu_0 \in \mathbb{R}$ that satisfy this condition, *given the observed value* $\hat{\mu}$, is

$$\left\{\mu_0 \,|\, \mu_0 \in \left(\hat{\mu} - \frac{\sigma}{\sqrt{n}} u_{\frac{1}{2}\alpha}, \hat{\mu} + \frac{\sigma}{\sqrt{n}} u_{\frac{1}{2}\alpha}\right)\right\} \; . \tag{2.49}$$

For $\alpha = 5\%$ we have approximately the confidence interval $(\hat{\mu} - 2\frac{\sigma}{\sqrt{n}}, \hat{\mu} + 2\frac{\sigma}{\sqrt{n}})$. (b) If σ is not known and $n \geq 15$, say, then we approximate the 95% confidence interval for μ by $(\hat{\mu} - 2\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + 2\frac{\hat{\sigma}}{\sqrt{n}})$. Note that this is $\left(\hat{\mu} - 2\sqrt{\hat{\mathrm{var}}\,(\hat{\mu})}, \hat{\mu} + 2\sqrt{\hat{\mathrm{var}}\,(\hat{\mu})}\right)$, and also that the length of the confidence interval tends to zero asymptotically as $\frac{1}{\sqrt{n}}$ for $n \to \infty$. If $n \leq 15$, say, then it is more accurate to use the exact distribution under H_0: $\mu = \mu_0$, i.e., $\frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$ (the t distribution with $n-1$ degrees of freedom). Hence, a 95% confidence interval is

$(\hat{\mu} - t_{n-1;\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{n-1;\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}})$, where $t_{n-1;\frac{\alpha}{2}}$ can be found in a table (at $\alpha = 5\%$, two-sided) of the t distribution.

Exercise 2.9. Consider the normal case with μ known, i.e., $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$, and construct a 95% confidence interval for σ^2 using the following dataset:

$$(0.5, 0.1, 0.7, 1.2, 1.8, 0.4, 0.6, 0.2, 1.5, 0.8, 0.9, 1.0) \tag{2.50}$$

Do the same if μ is not known. Use a table of the χ^2_f distribution.
For large n, one can approximate χ^2_n by $N(n, 2n)$ or χ^2_n/n by $N\big(1 - \frac{2}{9n}, \frac{2}{9n}\big)$, and dispose of the χ^2 table.

A graphical representation of confidence regions is given in Fig. 2.4.
Let $T = h(X_1, \ldots, X_n)$ be a test statistic with distribution function $F_\theta(t)$, and density $f_\theta(t), \theta \in \Theta = \mathbb{R}$. The null–hypothesis H_0: $\theta = \theta_0$ is rejected at (two-sided) level α if $T \notin (t_L, t_U)$, where $F_{\theta_0}(t_L) = \frac{\alpha}{2}$ and $F_{\theta_0}(t_U) = 1 - \frac{\alpha}{2}$. In Fig. 2.3 the (deterministic) boundaries t_L and t_U are plotted as a function of θ_0. (From now on we write θ instead of θ_0). From this graph one can read

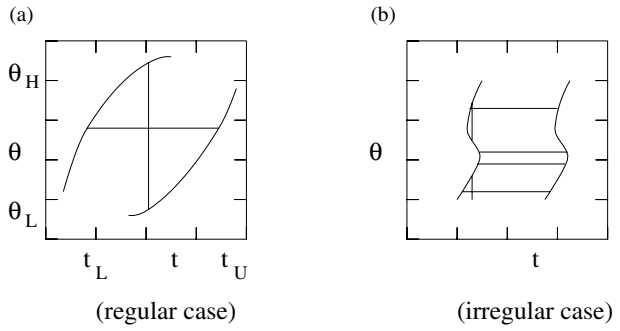

Fig. 2.4. Inversion of the test statistic. (a) Horizontally an acceptance region is drawn for the outcome of a test statistic T, and vertically a confidence interval for the parameter θ, both of which are connected in the regular situation. In the irregular case depicted in (b), although all acceptance regions are connected sets, some of the confidence regions consist of sets that are not connected. (Obviously, in two or more dimensions, rather complicated disconnected confidence regions are possible.)

off, for each outcome t of T, a $1 - \alpha$ confidence region (θ_L, θ_U) for θ, and for each θ a level-α acceptance region for T. The transformation of the *random variable* T into the *random interval* (θ_L, θ_U) is called: *inversion of the test statistic*. In objectivistic statistics, both the true value of the parameter θ and the acceptance regions (t_L, t_U) as a function of the running variable θ are always deterministic.

Exercise 2.10. Consider the number of deaths due to horse-kick in Prussian regiments, as given in Table 2.2. Assume a Poisson distribution. Estimate μ and give a 95% confidence interval. Compare the observed number of deaths with the fitted number of deaths. Remark. A qualitatively similar analysis can for instance be made for the number of tritium counts during a certain time interval, say 5 minutes, if this 'experiment' is repeated 280 times.

Table 2.2. The number of deaths due to horse-kick in the Prussian army, according to [707].

No of deaths	No of units
0	144
1	91
2	32
3	11
4	2
≥ 5	0

Legend: one unit = one 'regiment – year'

2.3 The $k(1, 2, \ldots)$-Sample Problem

In the previous chapter, we illustrated the statistical areas of Estimation, Hypothesis Testing and Confidence Interval Construction with the one-sample problem. Now we will extend the discussion to k samples, which is a statistical shorthand expression for: k samples, taken from k, possibly different, probability distributions. For didactic reasons we give the full discussion under two rather strict assumptions: normal distributions and *homoskedasticity* (i.e., equality of the variances). At first, we shall provide (only a few) references to literature that describes methods that can be applied if these two assumptions are not met. At least in homoskedastic situations, if *imputed normal distributions* are suspected (which means that, in each sample, part of the observations were drawn from a different distribution than the majority of observations), then it is useful to consider *robust methods* (see, e.g., [264, 567]), while for arbitrary, possibly clearly non-normal, distributions, the theory of rank statistics (also called *non-parametric statistics*) can be used (see [262, 282, 425, 619]). In *heteroskedastic* situations, the problem becomes more difficult (unless, of course, the ratio between the various variances is known, since then the problem can trivially be transformed into a homoskedastic problem). Good references are [466], Dijkstra's thesis [148],

and the didactic explanation in the 'original' BMDP Manual [150]. In practice, one tends to postulate homoskedasticity, which is more or less justified in case (for a reasonable sample size) the null-hypothesis H_0: $\sigma_1^2 = \cdots = \sigma_k^2$ is not rejected, which can be tested by using Bartlett's statistic [36, 37, 371] or Levene's statistic, see [90, 150], the latter being more robust against departures from normality. Based on the likelihood ratio principle, Bartlett's test considers the ratio between the arithmetic and geometric mean of the estimated variances in the k groups, the distribution of which is approximated by a suitable constant (depending on the sample sizes) times a χ^2 distribution.[12] Levene's statistic considers the ratio of the between-group variance and the within-group variance of the *absolute deviations* of the observations with respect to their group mean estimated in a more or less robust way, for instance by taking the sample median or the arithmetic average. Under the hypothesis of homoskedasticity, the distribution of Levene's statistic is asymptotically approximated by Fisher's F–distribution with suitable degrees of freedom.

All these more complicated excursions are presumably better appreciated if the simplified 'standard situation' is fully understood. So, let us consider the simple situation of two normal distributions with equal variances. We are interested in testing the hypothesis that the two expectation values are equal.

Example. Energy confinement times are measured for two types of discharges, e.g., with boronised or non-boronised walls, with the plasma current flowing in the 'positive' or in the 'negative' direction, etc., the other influences on confinement, as far as possible, being kept constant. We formulate the following mathematical *idealisation*:

The 2-Sample Problem

Let $X_{j,1}, X_{j,2}, \ldots, X_{j,n_j}$ be i.i.d. $N(\mu_j, \sigma^2)$, for sample j ($j = 1, 2$). For the sample sizes n_1 and n_2, we write n and m, respectively. We want to test H_0: $\mu_1 = \mu_2$ against the alternative A: $\mu_1 \neq \mu_2$ for (a) σ known, and (b) σ unknown. This is done as follows.

(a) We first have to look for a test statistic. Intuitively, $\widetilde{T} = \overline{X}_1 - \overline{X}_2$, or a function thereof, would be a sensible candidate. (As usual, $\overline{X}_j$ denotes the sample mean of the jth sample, $j = 1, 2$.) Since $\overline{X}_1 \sim N(\mu_1, \frac{\sigma^2}{n})$, and $\overline{X}_2 \sim N(\mu_2, \frac{\sigma^2}{m})$, with $\overline{X}_1$ and $\overline{X}_2$ independent, we have $\overline{X}_1 - \overline{X}_2 \sim N\big(\mu_1 - \mu_2, \sigma^2(\frac{1}{n} + \frac{1}{m})\big)$, hence, provided H_0: $\mu_1 = \mu_2$ holds,

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1) \; . \tag{2.51}$$

[12] This test is based on an asymptotic approximation which is at least suitable for normal parent distributions with sample sizes that are not too small.

From this follows that H_0 is rejected at level α if $T \notin (-u_{\frac{1}{2}\alpha}, u_{\frac{1}{2}\alpha})$ (two-sided), or $T \leq -u_\alpha$ (one-sided), or $T \geq u_\alpha$ (one-sided). It is recalled that one-sided testing is only allowed if there is prior evidence that μ is positive (negative).
(b) One expects that a statistic similar to T should be used, but now with σ replaced by an estimator $\hat{\sigma}$. Two estimators for σ^2 can immediately be seen:

$$\hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2 \quad \text{and} \quad \hat{\sigma}_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_{2i} - \overline{X}_2)^2 \,. \tag{2.52}$$

The question is now how to combine these two statistics to an accurate estimator for σ^2. Since $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are both unbiased estimators for σ^2, so is every linear combination $\hat{\sigma}_\lambda^2 = \lambda\hat{\sigma}_1^2 + (1-\lambda)\hat{\sigma}_2^2$. One can by differentiation directly derive that the linear combination that minimises $\operatorname{var}\hat{\sigma}_\lambda^2$ is obtained by choosing the weights λ and $1-\lambda$ inversely proportional to $\operatorname{var}(\hat{\sigma}_1^2)$ and $\operatorname{var}(\hat{\sigma}_2^2)$, respectively. Since $\hat{\sigma}_1^2 \sim \frac{\sigma^2}{n-1}\chi_{n-1}^2$, we know that $\operatorname{var}\hat{\sigma}_1^2 = \frac{2\sigma^4}{n-1}$ and $\operatorname{var}\hat{\sigma}_2^2 = \frac{2\sigma^4}{m-1}$, hence $\lambda = \frac{n-1}{n+m-2}$ and $1-\lambda = \frac{m-1}{n+m-2}$, so

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2 + \sum_{i=1}^{m}(X_{2i} - \overline{X}_2)^2}{n+m-2} \tag{2.53}$$

is a good estimator for σ^2, being the 'best linear' combination of $\hat{\sigma}_1^2$, and $\hat{\sigma}_2^2$. It is also uniformly minimum variance unbiased (UMVU) among *all* estimators, but this fact will not be proven here. The interested reader is referred to [426].

So, the canonical statistic is

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}} \,. \tag{2.54}$$

What is the distribution of T under H_0? For $n \to \infty$ or $m \to \infty$, it tends to the normal distribution (as $\hat{\sigma} \to \sigma$), but for finite n and m the distribution is not normal. Working at Guinness beer brewery, Gosset (writing under the pseudonym Student) derived, near the beginning of the twentieth century, a formula for the probability density of T. Since then, this is called the Student's t distribution, which has $n+m-2$ degrees of freedom in this case, see the denominator of (2.53).

Tables have been made of the t_f distribution function with f df (f degrees of freedom), for $f = 1, 2, 3 \ldots$. Here we repeat the formal definition from Chap. 1 (Def. 1.20), which is easy to retain and concordant to the historical development.

Definition 2.11. *If* $X \sim N(0,1)$ *and* $Z \sim \chi_f^2$, *then* $\frac{X}{\sqrt{Z/f}} \sim t_f$.

The probability density and moments of the t distribution can be found in Table 1.2 of Sect. 1.5.

Table 2.3. Lord Rayleigh's measurements (1894), (A) and (B) on the mass of atmospheric nitrogen (within a particular container at standard temperature and pressure) according to two methods (I: 'hot iron' and II: 'ferrous hydrate') to remove the oxygen, and (C) on the mass of nitrogen from nitric oxide, using method I.

(A) from air (method I)	(B) from air (method II)	(C) from NO (method I)
2.31017	2.31024	2.30143
2.30986	2.31010	2.29890
2.31010	2.31028	2.29816
2.31001		2.30182

Exercise 2.11. Check that in (2.54) T has a t_{n+m-2} distribution, using this formal definition.

Exercise 2.12. In Table 2.3 Lord Rayleigh's data [654] on the mass of nitrogen gas are shown.
(1) Test H_0: $\mu_A = \mu_B$ under the assumption that $\sigma_A^2 = \sigma_B^2$.
(2) Test H_0: $\mu_A = \mu_C$ under the assumption that $\sigma_A^2 = \sigma_C^2$.
(3) Estimate the fractional difference $(\mu - \mu_C)/\mu$, where μ corresponds to the average of the combined groups (A) and (B).[13]
(4) Analyse this problem with SAS (PROC TTEST) or another statistical package.

The $k(\geq 2)$-sample problem

Consider k independent, normally distributed, homoskedastic samples:

$$\begin{aligned} X_{11}, \ldots, X_{1n_1} &\sim N(\mu_1, \sigma^2) \\ &\cdots \\ &\cdots \\ &\cdots \\ X_{k1}, \ldots, X_{kn_k} &\sim N(\mu_k, \sigma^2) \ , \end{aligned} \tag{2.55}$$

where $n = n_1 + \cdots + n_k$ is the total sample size. Test H_0: $\mu_1 = \cdots = \mu_k$ against A: $\mu_i \neq \mu_j$, for at least one pair $(i, j), i, j = 1, \ldots, k$.
Method: Determine the k-sample means $\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$, and also the *grand-mean*

$$\overline{\overline{X}} = \frac{n_1 \overline{X}_1 + \cdots + n_k \overline{X}_k}{n_1 + \cdots + n_k} = \frac{1}{n} \sum_{i,j} X_{ij} \ . \tag{2.56}$$

[13] Some time after Lord Rayleigh's experiments, this intriguing difference was attributed to the presence of noble gasses in the air.

The total sum-of-squares can be broken down into 2 parts:

$$\sum_{i,j}(X_{ij}-\overline{\overline{X}})^2 = \sum_{i,j}(X_{ij}-\overline{X}_i)^2 + \sum_{i=1}^{k} n_i(\overline{X}_i-\overline{\overline{X}})^2 \quad \text{(Steiner's rule)}$$
$$= SS_{\text{within}} + SS_{\text{between}} \,. \tag{2.57}$$

SS_{within} and SS_{between} are statistically independent (which is not proven here). The second term tells us, roughly speaking, how far the outcomes $\overline{X}_1,\ldots,\overline{X}_k$ tend to violate the null hypothesis, and the first term indicates what variation can be expected due to chance (i.e., from 'random' fluctuations). In fact,

$$\hat{\sigma}^2_{\text{within}} = \frac{SS_{\text{within}}}{n-k} \sim \frac{\sigma^2}{n-k}\chi^2_{n-k} \,. \tag{2.58}$$

The degrees of freedom is equal to the number of observations minus the number of fitted parameters. If H_0 holds, then

$$\hat{\sigma}^2_{\text{between}} = \frac{SS_{\text{between}}}{k-1} \sim \frac{\sigma^2}{k-1}\chi^2_{k-1} \,, \tag{2.59}$$

whereas if H_0 does not hold, then $\hat{\sigma}^2_{\text{between}}$ tends to be larger than $\frac{\sigma^2}{k-1}\chi^2_{k-1}$. Hence,

$$T = \frac{\hat{\sigma}^2_{\text{between}}}{\hat{\sigma}^2_{\text{within}}} = \frac{SS_{\text{between}}/(k-1)}{SS_{\text{within}}/(n-k)} \tag{2.60}$$

seems to be a sensible test statistic. In fact, under

$$H_0{:}\mu_1 = \cdots = \mu_k \,, \qquad T \sim F_{k-1,n-k} \,, \tag{2.61}$$

where $F_{k-1,n-k}$ stands for the F distribution with $k-1$ and $n-k$ degrees of freedom. This follows directly from the formal definition of the F distribution, already mentioned in Sect. 1.5, and repeated here:

Definition 2.12. *If $Y_1 \sim \chi^2_f$ and $Y_2 \sim \chi^2_g$, with Y_1 and Y_2 independent, then*

$$\frac{Y_1/f}{Y_2/g} \sim F_{f,g} \,. \tag{2.62}$$

The critical values of the F distributions are tabulated, and are routinely provided in statistical and mathematical software packages, such as NAG, IMSL, SAS, S–PLUS and Mathematica.

Notation. The critical value of the $F_{f,g}$ distribution at the probability level α (also called the α fractile of the $F_{f,g}$ distribution) is denoted by $F_{f,g;\alpha}$.

Exercise 2.13. Consider the following (hypothetical) predictions for the confinement time in a Next Step tokamak device:

3.9 (±0.4), 3.5 (±0.5), 4.5 (±0.4), 5.5 (±1.0), 3.3 (±0.4), 2.9 (±0.3).

The numbers in brackets are one estimated standard deviation. Suppose that the predictions are based on independent datasets and that they are (approximately) normally distributed. (1) Test H_0: $\mu_1 = \cdots = \mu_k$, using the approximation that the given standard deviations are exact; (2) Test the same null-hypothesis under the assumption that the 'effective' sample sizes of these predictions are 25, 30, 40, 25, and 20, respectively; (3) Analyse this problem with SAS (PROC ANOVA) or another statistical program.

As we will see in the next section, the k-sample problem can be considered as a special case of the general linear model. Instead of testing against the general alternative in the k-sample problem, as discussed above, one can also postulate the model $\mu_j = \alpha + \beta_j$, $j = 1, \ldots, k$, and test H_0: $\beta = 0$ against $\beta \neq 0$, for instance. This we call for a moment 'a discrete type of regression'. It obviously leads to an adaptation of the test statistic T in (2.60). Analysing for instance the dependence of confinement time τ_E on current I_p, one can construct a k-sample problem by dividing the I_p axis into k intervals. If one now lets $k \to \infty$, we get in the limit a continuous regression model.

Table 2.4. Traditional names of statistical analysis for linear models with continuous response variables. Analysing the $1, 2, \ldots, k$-sample problem is a special case of one-way ANOVA. The abbreviation ANOVA stands for Analysis of Variance, and MANCOVA for Multivariate Analysis of Covariance.

No of dep. variables	*No of indep. variables*	*Type of indep. variables*	*Name of analysis*
1	1	discrete	ANOVA (one-way)
1	$p \geq 2$	discrete	ANOVA (multi-way)
1	1	continuous	Regression (single)
1	$p \geq 2$	continuous	Regression (multiple)
1	$p \geq 2$	mixed	ANCOVA
$q \geq 2$	1	discrete	MANOVA (one-way)
$q \geq 2$	$p \geq 2$	discrete	MANOVA (multi-way)
$q \geq 2$	1	continuous	Multivariate Regression (single)
$q \geq 2$	$p \geq 2$	discrete	Multivariate Regression (multiple)
$q \geq 2$	$p \geq 2$	mixed	MANCOVA

Historically, data analysis with intrinsically discrete, or discretised independent variables (such as in the k-sample problem) is called *analysis of variance* (ANOVA). If the independent variables are continuous it is called regression analysis, and if some of them are discrete and others continuous, it is called *analysis of covariance* (ANCOVA). If there is more than one response

variable, the prefix multivariate is used, see Table 2.4. All these models admit a general common formulation: *the general linear model*, sometimes abbreviated as GLM, which we discuss the the next section.

2.4 The General Linear Model

2.4.1 Scope

Consider an experiment consisting of n random observations $Y_1, \ldots, Y_n$ of a *response variable* Y satisfying

$$Y_i = a_0 + a_1 x_{i,1} + \cdots + a_p x_{i,p} + E_i \quad (i = 1, \ldots, n) \,, \tag{2.63}$$

where $x_1, \ldots, x_p$ are (deterministic) regression variables (also called *explanatory variables*, *independent variables* or *covariates*), $x_{i,1}$ denotes the value of the variable x at the ith observation, and $E_1, \ldots, E_n \sim N(0, \sigma^2)$ is a sequence of i.i.d. random disturbances.

Remarks.

1. It will be assumed that the variables $x_1, \ldots, x_p$ can be varied at will (within a certain operating regime of the experiment). Such an assumption is often realistic in physical experiments. It is assumed further that $x_1, \ldots, x_p$ are measured without random error, or at least with an error that is negligible with respect to $N(0, \sigma^2)$. This regression model is usually called *type I regression*.
2. A very similar regression theory can be derived under the assumption that the regressors are i.i.d. random variables, with a common multivariate normal distribution. This model assumption is frequently made in non-experimental empirical sciences (econometrics, sociology, epidemiology), and the regression model is called *type II regression*. By conditioning on the observed values of the random regressors in type II regression, one gets the formulae corresponding to type I regression.
3. Parts of the theory that follows still hold if the condition $E_i \sim N(0, \sigma^2)$ is relaxed to $E_1, \ldots, E_n$ i.i.d. with some (arbitrary) distribution function F satisfying $\mathrm{E}\, E_i = 0$ and $\mathrm{var}\, E_i = \sigma^2 < \infty$ $(i = 1, \ldots, n)$.

Example. (see also Exercise 4.1). Empirical scalings for the confinement time of an additionally heated plasma are often expressed in form of so called power laws. For instance,

$$\tau_E = c I^{a_1} B^{a_2} \langle n_e \rangle^{a_3} P^{a_4} \,, \tag{2.64}$$

where I stands for the plasma current, B for the magnetic field, $\langle n_e \rangle$ for the (volume averaged) electron density, and P for the total heating power. By taking logarithms, one gets

$$\log \tau_E = a_0 + a_1 \log I + a_2 \log B + a_3 \log\langle n_e \rangle + a_4 \log P \ , \tag{2.65}$$

with $a_0 = \log c$. The statistical regression model (2.63) is applicable if the measurement errors in τ_E are substantially larger than those in $B, I, \langle n_e \rangle, P$, and the relative error in τ_E, $\sigma(\log \tau_E) \simeq \frac{\sigma(\tau_E)}{\tau_E}$, is approximately constant.[14] Multiplicative errors for τ_E transform into additive errors for $\log \tau_E$. Additive errors for τ_E do so approximately as long as $\sigma\big(\log(\tau_E)\big) \ll 1$. This implies that, for sample sizes which are not very small, log-linear regression with additive errors leads to practically the same results as a number of alternative models in the case of empirical confinement time scalings, where $\sigma\big(\log(\tau_E)\big) \simeq 0.15$. (A worked example has been included in Sect. 4.4 of [354].)

Empirical confinement time scalings have been investigated in different contexts, and were stimulated by international cooperation. Just a few of the many papers and a brief indication of their main *statistical* features are: multiple regression analysis on a multi-tokamak L–mode database [369], [747], with an improved dataset (ITERL.DB1) [370]; regression analysis with interaction terms, linear restrictions, and errors-in-variables, based on a one and a two tokamak H–mode dataset [345, 346]; multi-tokamak H–mode dataset (ITERH.DB1) with principal component analysis [110]; improved ELMy H–mode dataset (ITERH.DB2) [678] with interaction models [360], additive power–law models [525], offset-linear or two-term scalings [122, 361, 662], and introduction to the analysis of catastrophe-type response functions [353, 355]; improved aspect-ratio scaling [349, 671], ITERH.DB3 dataset and application to ITER interval prediction [349, 350, 362, 663, 671].

Remarks.

1. Polynomial models also fall into the class of the general linear models. For example, suppressing the index i,

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1^2 + a_4 x_1 x_2 + a_5 x_2^2 \ , \tag{2.66}$$

   which is simply recoded as:

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + a_5 x_5 \ . \tag{2.67}$$

   Models containing products of individual regression variables are called *interaction models*.

2. The variables $x_1, \ldots, x_p$ can also be discrete. If $Y_i = a_0 + a_1 x_{i,1} + E_i$ with $x_{i,1} = 0$ for $i = 1, \ldots, n_1$ and $x_{i,1} = 1$ for $i = n_1 + 1, \ldots, n_2$, we have recovered the 2-sample problem. If $Y_i = a_0 + a_1 x_{i,1} + \cdots + a_k x_{i,k} + E_i$ with $x_{i,h} = 1$ for $\sum_{i=1}^{h-1} n_i + 1 \le i \le \sum_{i=1}^{h} n_i$ and 0 elsewhere, with $n_0 = 0$, and $h = 1, \ldots, k$, we have recovered the k-sample problem.

[14] Typical measurement errors for these quantities may be of the order 15% for τ_E, and $\le 1\%$, $\le 1\%$, $\simeq 5\%$, $\simeq 5\%$ for the other 4 variables, so the first assumption is not ideally fulfilled for $\langle n_e \rangle$ and P.

3. One can have different types of discrete variables acting simultaneously (ANOVA), for example: Confinement-time scalings with, simultaneously, a) Co- or Counter injection (of the neutral beams with respect to the direction of the plasma current); b) different isotopes for the neutral beams and the main plasma, i.e., $D^0 \to D^+$, $D^0 \to H^+$, $H^0 \to D^+$, or $H^0 \to H^+$ discharges[15]; c) wall carbonisation or no wall carbonisation (similarly for other wall-covering materials such as beryllium, boron, etc.); d) single null (SN) or double null (DN) magnetic configuration. Often such discrete variables just summarize, sometimes with considerable simplification, the values of underlying continuous variables. If, in addition to discrete variables, some continuous variables are used as regressors, the analysis is called ANCOVA, see Table 2.4.
4. In a number of situations, the response variable is discrete (e.g., 'success' or 'failure', 'alive' or 'dead') or can be made discrete by simplifying an underlying continuous response variable. For a binary response variable, the basic approach here is logistic regression, a particular case of the so-called generalised linear model, see Sect. 3.5, and related to two-group discriminant analysis, see Sect. 5.2.2. For further information on statistical methods of analysis in such cases, the reader is referred to [126, 455].

The above items 1 to 3, and in part also item 4, are covered by the general linear model.

2.4.2 Estimation of Regression Coefficients

The standard estimates for model (2.63) are based on the method of least squares, i.e., they consist of those values $\hat{a}_0, \ldots, \hat{a}_p$ that minimise

$$\sum_{i=1}^{n} \big(y_i - \sum_{h=0}^{p} a_h x_{i,h}\big)^2 \,. \tag{2.68}$$

Remark. In this formula, $x_{0,i} = 1$ for $i = 1, \ldots, n$. The coefficient a_0 is also called the intercept. It is stressed that least squares is a numerical approximation procedure which can also be applied outside a probabilistic context. Under probabilistic model assumptions, specific statistical properties of estimators based on least squares can be derived.

Exercise 2.14. Write down the probability density of $(Y_1, \ldots, Y_n)$ in (2.63) and derive from this that the maximum likelihood estimator (MLE) is iden-

[15] The symbol H is used for hydrogen and D for deuterium (both isotopes are ionised in the plasma). Superscript ^{0} is used to denote neutral particles, which are injected as highly energetic (neutral) beams into the plasma.

tical to the least-squares estimator (LSE). Obviously, this property does not hold any longer when the condition $E_i \sim N(0, \sigma^2)$ is dropped.[16]

In the following, we explain how explicit expressions for the least-squares estimators can be obtained (a) by differentiation, and (b) by geometrical arguments. In both cases, matrix notation is both efficient and illuminating. We conclude with a tensorial interpretation of the regression coefficients.

(a) by differentiation:

$$\frac{\partial}{\partial a_m} \sum_{i=1}^{n} \left(y_i - \sum_{h=0}^{p} a_h x_{i,h} \right)^2 = 0 \Rightarrow \tag{2.69}$$

$$-2 \sum_{i=1}^{n} \left(y_i - \sum_{h=0}^{p} a_h x_{i,h} \right) x_{i,m} = 0 \,, \tag{2.70}$$

for $m = 0, 1, \ldots, p$. This is a set of $p+1$ *linear* equations for the unknowns $a_0, \ldots, a_p$, which, under a usual regularity condition, can uniquely be solved by standard linear algebra. Its relatively stern appearance can be mitigated by interpreting $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \boldsymbol{y}$ and $\begin{pmatrix} x_{1,m} \\ \vdots \\ x_{n,m} \end{pmatrix} = \boldsymbol{x}_m$, $m = 0, \ldots, p$, as vectors in $\mathbb{R}^n$, and rewriting (2.70) as

$$\langle \boldsymbol{y}, \boldsymbol{x}_m \rangle - \sum_{h=0}^{p} a_h \langle \boldsymbol{x}_h, \boldsymbol{x}_m \rangle = 0 \quad (m = 0, 1, \ldots, p) \,, \tag{2.71}$$

where $\langle \boldsymbol{y}, \boldsymbol{x}_m \rangle = \sum_{i=1}^{n} y_i x_{i,m}$ denotes the inner product between $\boldsymbol{y}$ and $\boldsymbol{x}_m$. These $p+1$ equations are called the *normal equations* for $a_0, \ldots, a_p$.[17] Note that the matrix of second derivatives of (2.68) with respect to $\boldsymbol{a}$ is positive definite, hence there is a unique minimum satisfying (2.71), the contours of constant sum-of-squares (as a function of $\boldsymbol{a}$) being ellipses.

(b) by geometrical arguments:
Equation (2.63) can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{E} \,, \tag{2.72}$$

where

[16] Gauß and Laplace derived around 1800 the normal probability density by *postulating* that MLE = LSE, see also [645].

[17] Since they are of the form $\sum_h C_{mh} a_h = b_m$ or $\boldsymbol{C}\boldsymbol{a} = \boldsymbol{y}$, with $C_{mh} = \langle \boldsymbol{x}_h, \boldsymbol{x}_m \rangle$ and $b_m = \langle \boldsymbol{y}, \boldsymbol{x}_m \rangle$, the regularity condition is of course that the matrix C is non-singular, i.e., $\det C \neq 0$, in which case the explicit solution is $\hat{\boldsymbol{a}} = \boldsymbol{C}^{-1}\boldsymbol{y}$.

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix}, \ \boldsymbol{E} = \begin{pmatrix} E_1 \\ \vdots \\ E_n \end{pmatrix} \tag{2.73}$$

and the vector of regression coefficients equals

$$\boldsymbol{a} = \begin{pmatrix} a_0 \\ \vdots \\ a_p \end{pmatrix} . \tag{2.74}$$

Minimising $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{a}\| = \langle \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{a}, \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{a} \rangle$ as a function of $\boldsymbol{a} \in \mathbb{R}^{p+1}$ amounts to projecting $\boldsymbol{Y}$ on the subspace V_{p+1} of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{X}$, see Fig. 2.5. Denoting this projection by $\boldsymbol{Y}_P = \boldsymbol{X}\hat{\boldsymbol{a}}$, a necessary and sufficient condition is that $\boldsymbol{Y} - \boldsymbol{Y}_P = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{a}}$ is perpendicular to each column of $\boldsymbol{X}$, i.e., $\boldsymbol{X}^t(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{a}}) = \boldsymbol{0}$, hence $\hat{\boldsymbol{a}}$ satisfies

$$\boldsymbol{X}^t\boldsymbol{Y} = \boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{a}} . \tag{2.75}$$

These are the normal equations. They have the explicit solution

$$\hat{\boldsymbol{a}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} , \tag{2.76}$$

provided $\boldsymbol{X}^t\boldsymbol{X}$ is invertible.

Exercise 2.15. Read and re-read the two derivations. Try to find out whether you understood everything.

2.4.3 Geometrical Interpretation

Let us repeat the geometrical argument as a prelude to the tensorial interpretation: We search $\boldsymbol{Y}_P = \boldsymbol{X}\hat{\boldsymbol{a}} \in V_{p+1}$ such that we can decompose $\boldsymbol{Y} = \boldsymbol{Y}_P + \boldsymbol{E}$ with $\boldsymbol{E} \perp V_{p+1}$, which means $\boldsymbol{X}^t\boldsymbol{E} = \boldsymbol{0}$. Hence $\boldsymbol{X}^t\boldsymbol{Y} = \boldsymbol{X}^t\boldsymbol{Y}_P = \boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{a}}$, which are the normal equations. The elements of $\hat{\boldsymbol{a}}$ are (by definition) the *contravariant tensor components* of the vector $\boldsymbol{Y}_P$ in the covariant base $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_p\}$,[18] whereas the elements of $\boldsymbol{X}^t\boldsymbol{Y}_P$ are the *covariant tensor components* of $\boldsymbol{Y}_P$ in this base, which has *metric tensor* (matrix of inner products) $\boldsymbol{X}^t\boldsymbol{X}$.[19] The covariant (contravariant) tensor components are obtained from the contravariant (covariant) components simply by multiplying them with the (inverse) metric tensor. This is the tensorial interpretation of (2.75) and (2.76). If the base vectors $\boldsymbol{x}_0, \dots, \boldsymbol{x}_p$ are considered to have unit length, then the covariant (contravariant) tensor components of $\boldsymbol{Y}_P$ correspond to the perpendicular (parallel) projections of $\boldsymbol{Y}_P$ on the base vectors.

[18] To fix the convention, $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_p\}$ are assumed to be covariant base vectors.
[19] Since $\boldsymbol{E} \perp V_{p+1}$, the covariant components of $\boldsymbol{Y}$ are the same as those of $\boldsymbol{Y}_P$.

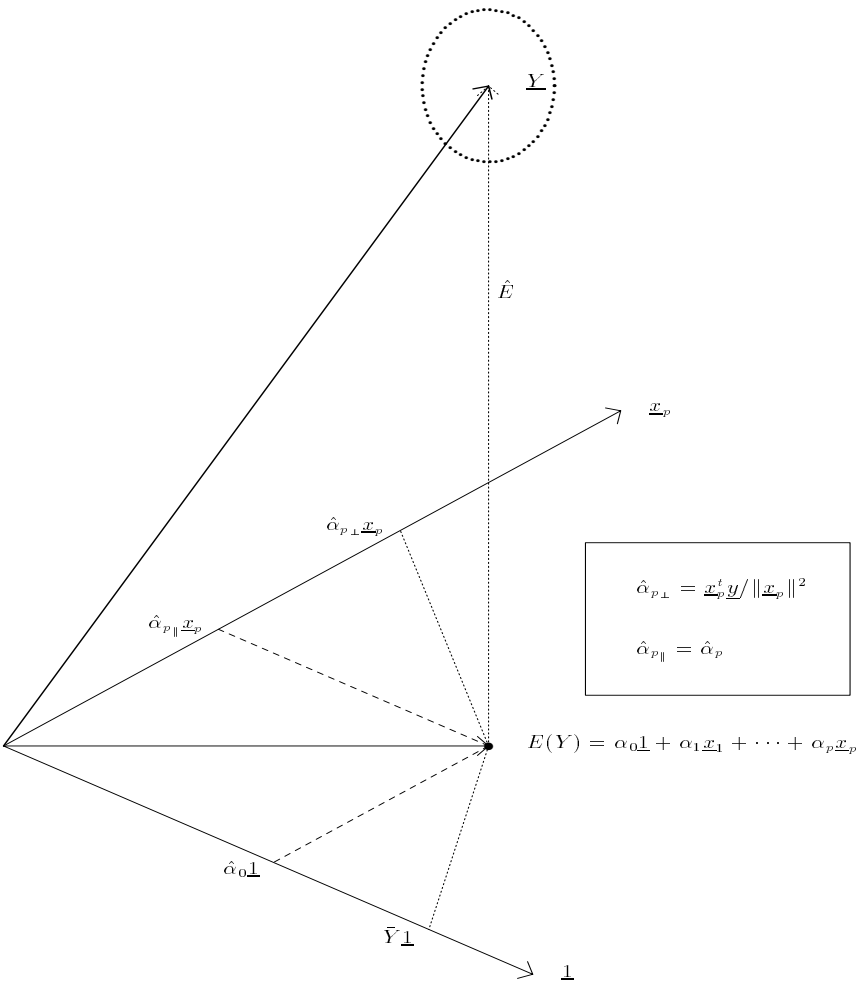

Fig. 2.5. Geometrical interpretation of least-squares regression

Theorem 2.4. *If, under the assumptions of model (2.63), the design matrix* $\boldsymbol{X}$ *is of full rank, then the LS estimator for* $\boldsymbol{a}$*, given by (2.76), has the following distribution:*

$$\hat{\boldsymbol{a}} \sim N_{p+1}(\boldsymbol{a}, (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2) \; . \tag{2.77}$$

It is recalled that $N_p(\mu, \boldsymbol{\Sigma})$ *stands for the multivariate normal distribution with expectation value* $\mu \in I\!R^p$ *and* $p \times p$ *covariance matrix* $\boldsymbol{\Sigma}$*.*

Proof. $E\hat{\boldsymbol{a}} = E(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t E\boldsymbol{Y} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{a} = \boldsymbol{a}$, hence $\hat{\boldsymbol{a}}$ is unbiased, and we have proven the first part. For the second part

we use the following property of the multivariate normal distribution: If $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then (a) for any p-dimensional vector $\boldsymbol{c}$, $\boldsymbol{c}^t\boldsymbol{Y} \sim N_1(\boldsymbol{c}^t\boldsymbol{\mu}, \boldsymbol{c}^t\boldsymbol{\Sigma}\boldsymbol{c})$ and, more generally, (b) for any $q \times p$ matrix $\boldsymbol{C}$, $\boldsymbol{C}\boldsymbol{Y} \sim N_q(\boldsymbol{C}\boldsymbol{\mu}, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^t)$. We note that (a) can be proven with characteristic functions, and that (b) follows directly from (a). Applying this property with $\boldsymbol{C} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$ gives the desired result.

Remarks.

1. If the assumption $E_i \sim N(0, \sigma^2)$ is relaxed to $E_i \sim F$ where F is an arbitrary distribution function with $\mathrm{E}(E_i = 0)$, $\mathrm{var}\,(E_i = \sigma^2) < \infty$, then $\hat{\boldsymbol{a}}$ does not need to be normally distributed. Nevertheless, $\hat{\boldsymbol{a}}$ is still unbiased ($E\hat{\boldsymbol{a}} = \boldsymbol{a}$) and has still the covariance matrix $(\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2$. It can be shown that in this situation, $\hat{\boldsymbol{a}}$ has minimal variance under all *linear* estimators of $\boldsymbol{a}$, i.e., is the *Best Linear Unbiased Estimator* (*BLUE*). This means that there does not exist a linear estimator with a smaller covariance matrix than $(\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2$ under the ordering

$$\boldsymbol{\Sigma}_1 \leq \boldsymbol{\Sigma}_2 \quad \text{if and only if} \quad \boldsymbol{x}^t\boldsymbol{\Sigma}_1\boldsymbol{x} \leq \boldsymbol{x}^t\boldsymbol{\Sigma}_2\boldsymbol{x} \text{ for all } \boldsymbol{x} \,.$$

   Moreover, under some regularity conditions, by virtue of the central limit theorem, the estimator $\hat{\boldsymbol{a}}$ is asymptotically (i.e., for $n \to \infty$) normal. If $E_i \sim N(0, \sigma^2)$, then $\hat{\boldsymbol{a}}$ is normal for every sample size n, and is UMVU, i.e., has *uniformly minimum variance* among *all* estimators of $\boldsymbol{a}$. A proof of the latter property is based on theorems of Rao–Blackwell and Lehmann–Scheffé. The first of these two theorems states that minimal variance is achieved among all unbiased estimates by an appropriate function of a sufficient statistic T and that this function can be obtained by taking the conditional expectation of some unbiased estimate with respect to T.[20] The second theorem tells us that such an unbiased statistic with minimal variance is unique if T is complete, in the sense that, for any $f \neq 0$, $\mathrm{E}f(T) = 0$ does not have a solution except of course for the trivial solution $T \equiv 0$. For further information, see [65, 191, 426, 427, 534].
2. The variance σ^2 of the errors can be estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - Y_{fit,i})^2 \,, \tag{2.78}$$

   where $Y_{fit,i} = \sum_{h=0}^{p}\hat{a}_h x_{h,i}$. This estimator can geometrically be interpreted as the squared distance between $\boldsymbol{Y}$ and its projection $\boldsymbol{Y}_P$, divided by the degrees of freedom:

[20] The theorem of Rao-Blackwell considers the situation that at least one unbiased estimator exists. If this is the case, then any unbiased estimator will do. The proof of this theorem is based on the decomposition of $\mathrm{var}\,(X)$ as in Excercise 1.29, where now $\mathfrak{D}$ is the σ–algebra generated by T.

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\|\boldsymbol{Y} - \boldsymbol{Y}_P\|^2 \,, \tag{2.79}$$

see Fig. 2.5.

Theorem 2.5. *Under model (2.63),*

$$\hat{\sigma}^2 \sim \frac{1}{n-p-1}\sigma^2\chi^2_{n-p-1} \,. \tag{2.80}$$

Hence, $E\hat{\sigma}^2 = \sigma^2$ *and var* $\hat{\sigma}^2 = 2\sigma^4/(n-p-1)$.

Theorem 2.6. *Under model (2.63), each linear combination of the regression coefficients is normally distributed:*

$$\boldsymbol{c}^t\hat{\boldsymbol{a}} \sim N(\boldsymbol{c}^t\boldsymbol{a}, \boldsymbol{c}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{c}\sigma^2) \,, \quad \boldsymbol{c} \in I\!R^{p+1} \,. \tag{2.81}$$

Application. This can be used to construct *confidence bands* for a regression line.

Example. $Y_i = a_0 + a_i x_i + E_i$, or equivalently, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{E}$ with $\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $\boldsymbol{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$. Hence, $\hat{\boldsymbol{a}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y}$. By taking the origin such that $\sum_{i=1}^n x_i = 0$ one can derive that $\hat{a}_0 = \overline{y} - \hat{a}_1\overline{x}$, with $\hat{a}_1 = S_{xy}/S_{xx}$, where $S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y})$. Note that $\hat{a}_1$ is equal to the ratio of the the empirical covariance between $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ and the empirical variance of $\boldsymbol{x} = (x_1, \ldots, x_n)$. Also, $\text{var}\,(\hat{\boldsymbol{a}}) = \begin{pmatrix} n & 0 \\ 0 & \sum_i x_i^2 \end{pmatrix}^{-1}$ σ^2, hence $\text{var}\,(\hat{a}_0 + \hat{a}_1 x) = \frac{1}{n} + \frac{x^2}{\sum_i x_i^2}\sigma^2$ (Theorem 2.6 has been used with $\boldsymbol{c} = (1, x)$.). For each x, a 95% confidence interval for $Y(x)$ is given by $\hat{a}_0 + \hat{a}_1 x \pm t_{n-1;.05}(\hat{\text{var}}\,(\hat{a}_0 + \hat{a}_1 x))^{1/2}$. For n large enough, $t_{n-1;.05} \simeq 2$.

Exercise 2.16. Work out the details of this last example, and plot the confidence band as a function of x. How do you interpret the minimum width of the confidence band?

Exercise 2.17. Consider the data:

$$(y, x) = (2.1, 1), (7.2, 5), (3.4, 2), (6.4, 4), (8.0, 6), (5.6, 3)$$

(1) Fit the models $y = a_0 + a_1 x$ and $y = a_0 + a_1 x + a_2 x^2$ by hand, or by using, e.g., PROC REG in SAS or the function *lm* in S–PLUS. (2) Fit the same models using the interactive matrix language SAS/IML. What is the estimate of σ^2 if the quadratic model is correct? (3) Test H_0: $a_2 = 0$ against A: $a_2 \neq 0$ using a t statistic (with how many degrees of freedom?). Interpret this t statistic geometrically in $I\!R^n$. Draw an approximate 95% confidence band under the assumption that the linear model is correct (using, e.g., SAS/GRAPH or SAS/INSIGHT).

2.4.4 Linear Parameter Restrictions

The method of least squares described above has been generalised to cover the following situations (in any combination), each of which can occur in practice: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{E}$, where (a) $\operatorname{rank}(\boldsymbol{X}) < p+1$, (b) $\boldsymbol{E} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$ with $\boldsymbol{\Sigma}_e$ possibly singular, and (c) explicit parameter restrictions $\boldsymbol{R}\boldsymbol{a} = \boldsymbol{r}$ ($\boldsymbol{r} \in \mathbb{R}^q$, $1 \le q \le p+1$), where $\boldsymbol{R} \in \mathbb{R}^q \times \mathbb{R}^{(p+1)}$ is a matrix describing q linear parameter restrictions. If there are only a few parameters, one can treat parameter restrictions 'by hand' through reparameterising the model.

If there are many parameters, this becomes unwieldy. Alternatively, one can use the method of restricted minimisation using *Lagrange multipliers* as a general numerical tool (implemented, e.g., in NAG and SAS). Here, we want to look at some explicit formulae for estimators under *linear restrictions* in the general linear model, which are derivable by using Lagrange's method of undetermined multipliers as an analytic tool. We follow the treatment in [435], see also [144, 270, 686] for some discussion.

First, we give a simple expression that is directly derivable from the method of Lagrange multipliers and valid in case the design matrix is not singular and $\boldsymbol{\Sigma}_e = \sigma^2 \boldsymbol{I}$.

Theorem 2.7. *The least-squares estimator for $\boldsymbol{a}$ in the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{E}$ under the restriction $\boldsymbol{R}\boldsymbol{a} = \boldsymbol{r}$ satisfies*

$$\begin{pmatrix} \boldsymbol{X}^t\boldsymbol{X} & \boldsymbol{R}^t \\ \boldsymbol{R} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^t\boldsymbol{Y} \\ \boldsymbol{r} \end{pmatrix} , \tag{2.82}$$

where $\boldsymbol{\lambda}$ is a Lagrange multiplier.

Especially in more general situations, explicit expressions for the least-squares estimator become somewhat lengthy, even in matrix language. A notational (and conceptual) reduction is provided by the concept of (Moore–Penrose) *generalised inverse*, $\boldsymbol{X}^-$, of a matrix $\boldsymbol{X}$, which exists for any $n \times (p+1)$ matrix $\boldsymbol{X}$, see [48, 49, 539], and is defined by the following properties:

$$\boldsymbol{X}\boldsymbol{X}^-\boldsymbol{X} = \boldsymbol{X} , \tag{2.83}$$

$$\boldsymbol{X}^-\boldsymbol{X}\boldsymbol{X}^- = \boldsymbol{X}^- , \tag{2.84}$$

$$(\boldsymbol{X}^-\boldsymbol{X})^t = \boldsymbol{X}^-\boldsymbol{X} , \tag{2.85}$$

$$(\boldsymbol{X}\boldsymbol{X}^-)^t = \boldsymbol{X}\boldsymbol{X}^- . \tag{2.86}$$

Note that the matrix product is associative, whence $(\boldsymbol{X}\boldsymbol{X}^-)\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^-\boldsymbol{X})$. The above properties state that $\boldsymbol{X}^-\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^-$ are orthogonal projections (i.e., symmetric and idempotent) in $\mathbb{R}^{(p+1)}$ and $\mathbb{R}^n$, respectively. If $\boldsymbol{X}$ is non-singular, then the generalised inverse $\boldsymbol{X}^-$ equals $(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$, and the least-squares estimator, (2.76), can concisely be written as $\hat{\boldsymbol{a}} = \boldsymbol{X}^-\boldsymbol{Y}$. We first consider generalisation (a) only. If $\boldsymbol{X}$ is singular, then the least-squares solutions for $\boldsymbol{a}$ are not unique. They form a hyper-plane of dimension $p+1-$

rank($\boldsymbol{X}$) in parameter space. It can be derived that for such a case the set of all solutions is given by

$$\boldsymbol{a} = \boldsymbol{X}^{-}\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{X}^{-}\boldsymbol{X})\boldsymbol{b}\,, \tag{2.87}$$

where $\boldsymbol{b}$ runs over $\mathbb{R}^{(p+1)}$, while the minimum sum-of-squares equals

$$\boldsymbol{Y}^{t}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^{-})\boldsymbol{Y}\,. \tag{2.88}$$

Notice, however, that the least-squares estimator for $\boldsymbol{X}\boldsymbol{a}$ is still unique! In fact, for any matrix $\boldsymbol{C}$ for which each *row* can be written as a linear combination of the *rows* of $\boldsymbol{X}$, the vector of parameters $\boldsymbol{C}\boldsymbol{a}$ has the unique LS estimate $\boldsymbol{C}\boldsymbol{X}^{-}\boldsymbol{Y}$. Such vectors are called *estimable functions* of $\boldsymbol{a}$, and the condition on $\boldsymbol{C}$ will be called the (unique) *estimability condition*. We now want to generalise this by considering, in addition, generalisation (b) with the 'model consistency' condition that the observed data $\boldsymbol{Y}$ can be generated by $\boldsymbol{X}$ and $\boldsymbol{E}$. (If $\boldsymbol{\Sigma}_e$ is regular, this is satisfied.) In that case, one can derive

Theorem 2.8. *For any $\boldsymbol{C}$ satisfying the estimability condition, the generalised LS estimator for $\boldsymbol{C}\boldsymbol{a}$ (for $\boldsymbol{\Sigma}_e$ known) equals*

$$\boldsymbol{C}\hat{\boldsymbol{a}} = \boldsymbol{C}(\boldsymbol{X}^{t}\boldsymbol{\Sigma}_0^{-}\boldsymbol{X})^{-}\boldsymbol{X}^{t}\boldsymbol{\Sigma}_0^{-}\boldsymbol{Y}\,, \tag{2.89}$$

where $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_e + \boldsymbol{X}^{t}\boldsymbol{X}$. The covariance matrix of $\boldsymbol{C}\hat{\boldsymbol{a}}$ equals

$$\boldsymbol{var}(\boldsymbol{C}\hat{\boldsymbol{a}}) = \boldsymbol{C}\big((\boldsymbol{X}^{t}\boldsymbol{\Sigma}_0^{-}\boldsymbol{X})^{-} - \boldsymbol{I}\big)\boldsymbol{C}^{t}\,. \tag{2.90}$$

If the hyper-plane in $\mathbb{R}^{p}$ generated by the columns of $\boldsymbol{X}$ is contained in the hyper-plane in which the errors $\boldsymbol{E}$ (with their expectation value translated to the origin) 'live', then the expressions can be simplified to

$$\boldsymbol{C}\hat{\boldsymbol{a}} = \boldsymbol{C}(\boldsymbol{X}^{t}\boldsymbol{\Sigma}_e^{-}\boldsymbol{X})^{-}\boldsymbol{X}^{t}\boldsymbol{\Sigma}_e^{-}\boldsymbol{Y}\,, \tag{2.91}$$

and

$$\boldsymbol{var}(\boldsymbol{C}\hat{\boldsymbol{a}}) = \boldsymbol{C}(\boldsymbol{X}^{t}\boldsymbol{\Sigma}_e^{-}\boldsymbol{X})^{-}\boldsymbol{C}^{t}\,, \tag{2.92}$$

respectively.

Remark. If this last condition is not fulfilled, then the singularity of $\boldsymbol{\Sigma}_e$ induces one or more implicit restrictions on the admissible parameters $\boldsymbol{a}$, so a nice aspect of the general formulation is that it can be used to cover situation (c) as well! Any explicit linear restrictions $\boldsymbol{R}\boldsymbol{a} = \boldsymbol{r}$ can be modeled by extending the design matrix with new 'observations' (the rows of $\boldsymbol{R}$), and the corresponding vector $\boldsymbol{Y}$ with $\boldsymbol{r}$, which are new 'response observations' with infinite precision that extend $\boldsymbol{\Sigma}_e$ by just adding for each restriction a zero row and zero column. (This forces the extension of $\boldsymbol{\Sigma}_e$ to be singular.)

Hence, (2.89) and (2.90) are powerful explicit formulae for *restricted linear regression* problems. They are useful in practice provided one can easily

and routinely calculate generalised inverses. Such facilities are provided by SAS/IML, NAG, S–PLUS, IMSL and MATLAB.

In general, adding constraints improves the condition of the regression problem, but it also introduces some bias to the regression estimates (inasmuch a constraint is not satisfied by the underlying data). Hence, adding constraints is related to ridge regression, see Chap. 3. This can well be seen from the following special case of Theorem 2.8:

Theorem 2.9. *If the constraints entirely remove any strict collinearity, i.e., if rank*$(\boldsymbol{X}^t \| \boldsymbol{R}^t) = p + 1$*, while the rows of* $\boldsymbol{R}$ *are independent of those of* $\boldsymbol{X}$*, and* $\boldsymbol{\Sigma}_e$ *is invertible, then the restricted least-squares estimate of* $\boldsymbol{a}$ *is*

$$\hat{\boldsymbol{a}} = \boldsymbol{G}_R^{-1}(\boldsymbol{X}^t \boldsymbol{\Sigma}_e^{-1} \boldsymbol{Y} + \boldsymbol{R}^t \boldsymbol{r}) \,, \tag{2.93}$$

and has covariance matrix

$$\boldsymbol{var}(\hat{\boldsymbol{a}}) = \boldsymbol{G}_R^{-1} - \boldsymbol{G}_R^{-1} \boldsymbol{R}^t \boldsymbol{R} \boldsymbol{G}_R^{-1} \,, \tag{2.94}$$

where $\boldsymbol{G}_R = (\boldsymbol{X}^t \boldsymbol{\Sigma}_e^{-1} \boldsymbol{X} + \boldsymbol{R}^t \boldsymbol{R})$.

Remark. The notation $(\boldsymbol{X}^t \| \boldsymbol{R}^t)$ is used to denote the juxtaposition of two matrices which have the same number of rows.

Exercise 2.18. Consider the data of Exercise 2.17: Estimate the model $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ under the restriction $a_1 = a_3$ (1) by using SAS / PROC IML, (2) by using SAS / PROC REG, (3) 'by hand' or by using S–PLUS (function *lm*), after reparameterisation.

In practice, when $\boldsymbol{\Sigma}_e$ is unknown, one can insert standard (ML or unbiased) estimators of $\boldsymbol{\Sigma}_e$ into (2.93) to (2.94). This is called the method of *generalised least squares*. Such estimators tend to suffer from overfitting, however, and more stable, albeit not unique estimators may be obtained in a number of situations by applying unweighted (or simply weighted) least squares instead. In any case, it may pay off to consider estimation theory for regression with (unknown) special covariance matrices $\boldsymbol{\Sigma}_e$ that have a reduced number of parameters and that are tailored to the practical situation at hand. Such models are investigated in multivariate analysis [376, 403] and in time series analysis [80, 377].

2.5 Introduction to Multivariate Analysis

Multivariate analysis is an important area of statistics, with many branches, extensive theoretical developments and numerous practical applications. A few selected monographs are [14, 181, 207, 290, 332, 376, 402–404, 442, 480, 681]. Here we will only give a first introduction by analysing the bivariate normal density, formulating an important theorem for the multivariate normal distribution and discussing the principle, as well as some applications, of principal component analysis.

2.5.1 Bivariate Normal Density

The probability density of $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$f(\boldsymbol{x}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \quad (\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^p) \,. \tag{2.95}$$

As the simplest, yet illustrative, multivariate example, we will investigate more closely the *bivariate normal distribution*. The study of this distribution is interesting for a number of reasons. In kinetic plasma theory, it occurs for instance as the velocity distribution of the electrons parallel and perpendicular to the magnetic field. Its covariance matrix is

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \,, \tag{2.96}$$

where ρ is the correlation coefficient between X_1 and X_2, hence, $|\boldsymbol{\Sigma}| = \sigma_1^2\sigma_2^2(1-\rho^2)$ and

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \sigma_1^{-2} & -\rho\sigma_1^{-1}\sigma_2^{-1} \\ -\rho\sigma_1^{-1}\sigma_2^{-1} & \sigma_2^{-2} \end{pmatrix} \,. \tag{2.97}$$

The last identity is easily seen by writing $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$, and applying Kramer's rule to obtain $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1}$. The probability density of the bivariate normal distribution, in a coordinate system with the origin at $\boldsymbol{\mu} = \boldsymbol{0}$, can be written as

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{1}{1-\rho^2}\left(\frac{x_1^2}{\sigma_1^2} - \frac{2x_1x_2\rho}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right)} \,. \tag{2.98}$$

For fixed parameters ρ, σ_1^2 and σ_2^2, the contours of constant probability density are the ellipses

$$\frac{1}{1-\rho^2}\Big(\frac{x_1^2}{\sigma_1^2} - \frac{2x_1x_2\rho}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\Big) = c \,. \tag{2.99}$$

The reader is referred to Fig. 3.1 for a picture of such a probability contour. A useful parametric representation of these ellipses is

$$\begin{cases} x_1(t) = k\sigma_1 \sin\omega t \,, \\ x_2(t) = k\sigma_2 \sin(\omega t + \varphi) \,, \end{cases} \tag{2.100}$$

with $k = \sqrt{c}$ and $\cos\varphi = \rho$. This gives an engineering interpretation of the correlation coefficient ρ of the bivariate normal distribution. Notice that $x_2(t) = k\sigma_2(\rho\sin\omega t + \sqrt{1-\rho^2}\cos\omega t)$, so ρ is the part of $x_2(t)$ that is 'in

phase' with $x_1(t)$. It is interesting to note that the conditional distribution of X_2, given $X_1 = c$, is also normal. The expectation value $\mathrm{E}(X_2|X_1 = c)$ is obtained by projecting the center of the ellipse onto the line $x_1 = c$ *along the line conjugate to* $x_1 = c$. The curve $\mathrm{E}(X_2|X_1 = c)$, where c runs through $\mathbb{R}$, is the 'Locus of vertical tangential points' in Fig. 3.1. Since Galton it is called the regression line of X_2 on X_1.[21] The tangent point of the ellipse that touches the vertical line $x = c > 0$ is $(\sqrt{c}\sigma_1, \sqrt{c}\rho\sigma_2)$, which one can see immediately by taking $\omega t = \frac{\pi}{2}$, for which $\sin \omega t$ is maximal. Hence, the regression line is $x_2 = \rho\frac{\sigma_2}{\sigma_1}x_1$. Note that this gives also another interpretation of ρ: it is the slope of the regression line in standardised coordinates (where $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1$).

Naturally, one can also regress X_1 on X_2. This gives the line conjugate to $x_2 = c$, which is $x_2 = \frac{\sigma_2}{\sigma_1}\frac{1}{\rho}x_1$, and which is called the 'Locus of horizontal tangential points' in Fig. 3.1.

The two regression lines are essentially different. Unless $\sigma_1 = \sigma_2$, the second line cannot simply be obtained by interchanging x_1 and x_2, just as the regression $Y = a + bx + E, E \sim N(0, \sigma^2)$ is essentially different from $X = c + dy + E, E \sim N(\sigma, \sigma^2)$, i.e., $\hat{b}$ is not equal to $1/\hat{d}$ in general.

Exercise 2.19. Derive the formula $x_2 = \frac{\sigma_2}{\sigma_1}\frac{1}{\rho}x_1$, for the regression of X_1 on X_2.

The two other important lines in Fig. 3.1 are the *principal* (major and minor) *axes*. They are the eigenvectors (more precisely the 1-dimensional eigenspaces) corresponding to the largest and smallest eigenvalue of $\boldsymbol{\Sigma}$, respectively.[22] For instance, if $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, i.e., if $\sigma_1^2 = \sigma_2^2 = 1$, then one has $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = (1+\rho)\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = (1-\rho)\begin{pmatrix} 1 \\ -1 \end{pmatrix}$, hence the principal axes are $x_2 = x_1$ and $x_2 = -x_1$. The principal axes define a new orthogonal coordinate system.

Exercise 2.20. Show by direct calculation that the principal axes of (X_1, X_2) in the general bivariate situation (i.e., with $\boldsymbol{\Sigma}$ arbitrary) do not coincide with the axes obtained from back-transforming the principal axes of the standardised variables $(X_1', X_2') = (X_1/\sigma_1, X_2/\sigma_2)$, i.e., do not coincide with $x_2 = \pm(\sigma_2/\sigma_1)x_1$. In fact, the slopes of the genuine principal axes of (X_1, X_2)

[21] Note that this is regression with a stochastic regressor variable, customarily called regression II (see Chap. 3.3). If $Y = a + bx + E$, $E \sim N(0, \sigma^2)$, then trivially $E(Y|X = x) = EY = a + bx$.

[22] An eigenvector $\boldsymbol{x}$ of a matrix $\boldsymbol{\Sigma}$ and its associated eigenvalue λ are defined by the equation $\boldsymbol{\Sigma}\boldsymbol{x} = \lambda\boldsymbol{x}$, while $\boldsymbol{x} \neq 0$. The eigenvalues can be calculated in principle from the equation $|\boldsymbol{\Sigma} - \lambda\boldsymbol{x}| = 0$. For a symmetric matrix all eigenvalues are real numbers and the eigenvectors corresponding to different eigenvalues are othogonal. For numerical aspects, the reader is referred to [649,735] and to the documentation of NAG, IMSL and MATLAB (see Sect. 6.1), among others.

depend for $\sigma_1 \neq \sigma_2$ on σ_1, σ_2 *and* ρ. This lack of equivariance (or, in somewhat inaccurate physical terminology, lack of invariance) is an important characteristic of principal component analysis.

Let $(X_{1,j}, X_{2,j})$, $j = 1, \ldots, n$, be a sample of size n from a bivariate normal distribution. The product-moment correlation coefficient

$$R = \frac{\sum_j (X_{1,j} - \overline{X}_1)(X_{2,j} - \overline{X}_2)}{\sqrt{\sum_j (X_{1,j} - \overline{X}_1)^2 \sum_j (X_{2,j} - \overline{X}_2)^2}} \,, \tag{2.101}$$

can be interpreted as the cosine of the angle between $\boldsymbol{X}_1 = (X_{1,1}, \ldots, X_{1,n})$ and $\boldsymbol{X}_2 = (X_{2,1}, \ldots, X_{2,n})$ in $\mathbb{R}^n$, see Sect. 1.4. For $\rho = 0$, the distribution of R^2 is $\mathcal{B}e_{\frac{1}{2},\frac{n}{2}-1}$, see [404, 655], hence $\log \frac{R^2}{1-R^2} \sim \mathcal{B}e\mathcal{L}o_{\frac{1}{2},\frac{n}{2}-1}$. From Sect. 1.5 we recall that $\mathcal{B}e\mathcal{L}o_{\frac{1}{2},\frac{n}{2}-1} = 2Z_{1,n-2} - \log(n-2) = \log \frac{1}{n-2} F_{1,n-2}$, whence $U = \frac{R^2}{1-R^2} \sim \frac{1}{n-2} F_{1,n-2}$ and $\sqrt{U} \sim \frac{1}{\sqrt{n-2}} t_{n-2}$. For ρ different from zero, the distribution of r is rather complicated and has been derived by R.A. Fisher, see [13, 135, 197, 205, 289], who also noted that a particularly simple formula is obtained after a variance-stabilising transformation, the variance of $Z = \frac{1}{2} \ln \frac{1+R}{1-R}$ being to a good approximation $1/n$, while Z exhibits notably less skewness than R.[23] For non-normal bivariate distributions or if outliers are present, a rather robust correlation coefficient is based on ranks rather than on the original observations, see [1, 262, 397, 619]. For two practical tutorial articles, with some historical remarks, see [754] and [558].

2.5.2 Multivariate Normal Density

The formulae and the geometric intepretation in Sect. 2.5.1 can be generalised to more dimensions. We discuss here only some algebraic aspects in relation to regression analysis. The geometric generalisation is not intrinsically difficult: One has to consider $(p+q)$-dimensional ellipses, p-dimensional hyperplanes, conjugate to q-dimensional ones, etc. For background on the geometry of linear and quadratic forms, the reader is referred to [51, 231, 319]. Again, the multivariate normal distribution is taken as a paradigmatic example, often suitable to some approximation and possessing a simple analytic structure. Some of this structure also holds for elliptical non-normal distributions, such as the multivariate t–distribution, see [183] and Chap. 44.3 in [395]. Other continuous multivariate distributions often have a more complicated relation to their conditional and marginal densities. For a rather comprehensive overview, the reader is referred to [395]. A basic theorem is the following:

[23] The asymptotic bias, variance, skewness and excess of kurtosis of Z are approximately, see [289], $\frac{\rho}{2n}$, $\frac{1}{n} + \frac{4-\rho^2}{2n^2}$, $\frac{\rho^3}{n^{3/2}}$, $\frac{2}{n}$. These are to be compared with those for R: $-\frac{\rho(1-\rho^2)}{2n}$, $(1-\rho^2)^2(\frac{1}{n} + \frac{11\rho^2}{2n^2})$, $-6\frac{\rho}{\sqrt{n}}(1 + \frac{-30+77\rho^2}{12n})$, $\frac{6}{n}(12\rho^2 - 1)$, which are also taken from [289]. For $\rho = 0$, one can derive that $Z \sim \frac{1}{2}\mathcal{B}e\mathcal{L}o_{\frac{n}{2}-1,\frac{n}{2}-1} = Z_{n-2,n-2}$, which is Fisher's z distribution.

Theorem 2.10. *Let* $(\boldsymbol{X}, \boldsymbol{Y}) \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, with mean* $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ *and covariance matrix*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{21} \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22} \end{pmatrix} . \tag{2.102}$$

Then the conditional distribution of $\boldsymbol{Y}$ *given* $\boldsymbol{X} = \boldsymbol{x}$*,* $\mathfrak{L}(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x})$ *is also normally distributed, with mean*

$$\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) \tag{2.103}$$

and variance

$$\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} . \tag{2.104}$$

Proof. We have seen in the proof of Theorem 2.4 that for $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and two matrices $\boldsymbol{A} \in \mathbb{R}^{q_1 \times p}$ and $\boldsymbol{B} \in \mathbb{R}^{q_2 \times p}$, $\boldsymbol{AX} \sim N_{q1}(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^t)$ and $\boldsymbol{BX} \sim N_{q2}(\boldsymbol{B\mu}, \boldsymbol{B\Sigma B}^t)$. One can similarly derive that $\operatorname{cov}(\boldsymbol{AX}, \boldsymbol{BX}) = \boldsymbol{A\Sigma B}^t$, which reflects the bilinear property of the covariance matrix. From this property it follows by direct calculation that the random variables $\boldsymbol{X}$ and $\boldsymbol{Y} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}$ are uncorrelated and hence, since they are normally distributed, also independent, the mean and the variance of the latter being $\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, respectively. The basic steps are the following: write $\boldsymbol{X} = \boldsymbol{A}[\boldsymbol{X}^t||\boldsymbol{Y}^t]^t$ and $\boldsymbol{Y} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} = \boldsymbol{B}[\boldsymbol{X}^t||\boldsymbol{Y}^t]^t$ with $\boldsymbol{A} = [\boldsymbol{0}||\boldsymbol{I}]$ and $\boldsymbol{B} = [\boldsymbol{I}|| - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}]$. Since $\boldsymbol{X}$ and $\boldsymbol{Y} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}$ are independent, the conditional distribution of $\boldsymbol{Y} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{x}$ is, for any $\boldsymbol{x}$, equal to the marginal distribution of $\boldsymbol{Y} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}$, for which we have just derived the mean and the variance. $\mathfrak{L}(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x})$ is just the conditional distribution $\mathfrak{L}(\boldsymbol{Y} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}|\boldsymbol{X} = \boldsymbol{x})$ shifted by an amount $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{x}$. This shift affects the mean (in an obvious way), but not the variance.

By looking at the separate components of (2.103) one can immediately derive that a univariate regression of Y_j on $\boldsymbol{X}$, while neglecting $(Y_1, \ldots, Y_{j-1})$ and $(Y_{j+1}, \ldots, Y_q)$, yields the same point estimates as is given by the jth component of the multivariate regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$. Similarly, the variance of the regression coefficents in the univariate regression is the same as the variance of their marginal distribution in the multivariate regression, see (2.104).

For $q = 1$ and $p = 1$, the two-dimensional result in Sect. 2.5.1 is recovered, the slope of the regression line being $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} = \rho\frac{\sigma_2}{\sigma_1}$. Comparing the above regression analysis with that in Sect. 2.4, where the regression variables were fixed, we can make the following identification: $\hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1} = \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{XY}$. Hence, the point estimates of the regression coefficients are the same in both cases. Notice that for convenience we have chosen the coordinate system at the center of the multivariate distribution here, and that $n^{-1}(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \hat{\boldsymbol{\Sigma}}_{11}$ is the maximum likelihood estimator of $\boldsymbol{\Sigma}_{11}$. Furthermore, $\boldsymbol{\Sigma}_{22}$ corresponds to the (total) variance of Y around its (true, global) mean value and $\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ to σ^2 in Sect. 2.5.1, which is the variance of E, or, equivalently, the variance of Y around the (true) regression

plane $\boldsymbol{a}^t\boldsymbol{x}$. The dispersion matrix of $\hat{\boldsymbol{a}} = \hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1}$ for stochastic regression variables is rather complicated and related to the Wishart distribution, see, e.g., [14, 161, 442].

An interpretative pitfall in multiple regression analysis can occur if one plots the residuals of the response variable against the observed values y, or against a function thereof. In that case, an apparent trend appears, even in the case that the linear model $Y = a_0 + a_1x_1 + \cdots + a_px_p + E$ is known to be correct. To illustrate this point, we consider Fig. 3.1, which was made more than a century ago. We regress the children's height (on the horizontal axis) against the mid-parents' height (vertical). We must adjust ourselves from our present habit to plot the response variable vertically. The regression line (obtained by minimising the horizontal sum of squares) is given by the locus of horizontal tangential points.[24] If the residuals (the observed children's heights minus those fitted by the regression line) are plotted against the observed values of the response variable (children's height), then a positive correlation appears.[25] Children larger than the average, quite naturally, happen to have large residuals (measured horizontally), while the opposite holds true for children smaller than the average. An interpretative mistake occurs if this correlation is used to adjust the regression coefficient! This somewhat delusory correlation does not appear if the residuals are plotted against the regression variable (or a function thereof): in our example, if the children's residuals are plotted against the mid-parents' height, or, almost equivalently, against the children's height *as predicted* (*by linear regression*) *from the mid-parents' height.*

2.5.3 Principal Components and their Application

A versatile technique with interesting theoretical ramifications and many practical applications is *principal component analysis*. We give here a brief introduction.

Definition 2.13. *The* (*empirical*) *principal components of a point P in $\mathbb{R}^p$ with respect to a data cloud $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ are the coordinates of P, obtained by parallel projection, along the principal axes of the concentric family of ellipses that approximate this cloud in some optimal sense.*

Since, in the standard situation, the principal axes are orthogonal, the principal components are also the orthogonal projection of P on the principal axes of the approximating ellipse.

[24] We use the property here that the distribution of point estimates of the regression coefficients in type I regression, with fixed regression variables, is practically the same as in type II regression, with a stochastic regressor and a bivariate normal distribution. This property has been investigated at various occasions and can, to a fair extent, also be verified by stochastic simulation.

[25] We do not present the corresponding plot here, but the feature should be clear from Fig. 3.1.

Definition 2.14. *Let the probability density of the random vector* $\boldsymbol{X} = (X_1, \ldots, X_p)$ *be constant on each member of a concentric family of ellipses. Then the principal components* (*p.c.'s*) *of the random vector* $\boldsymbol{X}$ *are the* p *random variables obtained by projecting* $\boldsymbol{X}$ *on the principal axes of these ellipses.*

Note that the principal components are linear combinations of the original random variables.

Exercise 2.21. Derive that, for $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$, the principal components are $Y_1 = \frac{X_1+X_2}{\sqrt{2}}$ and $Y_2 = \frac{X_1-X_2}{\sqrt{2}}$.

Before we give some applications of principal component analysis (PCA), we consider again the empirical principal components, derived from a random sample of size n.

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the usual estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \tag{2.105}$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^t = \frac{1}{n-1}\boldsymbol{S}\,, \tag{2.106}$$

where $\boldsymbol{S}$ is the *matrix of* (centered) *sum-of-squares and cross-products*. If $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ have an arbitrary multivariate distribution (with finite expectation E $\boldsymbol{X}$ and finite variance $\boldsymbol{\Sigma}$), then $\overline{\boldsymbol{X}}$ and $\frac{1}{n-1}\boldsymbol{S}$ are still sensible estimators for E $\boldsymbol{X}$ and $\boldsymbol{\Sigma}$, respectively, but their distribution is more difficult. Asymptotically, i.e., for $n \to \infty$, $\overline{\boldsymbol{X}} \to N_p(E\boldsymbol{X}, \frac{1}{n}\boldsymbol{\Sigma})$.

Principal component analysis has been introduced in statistics by [509] and [288]. Many monographs and articles have been devoted to this topic. We just mention [243, 444], [442], and [101], which are suitable for introduction, and [14] for asymptotic theory. If one relaxes the condition of orthogonality between the principal components, one can construct common principal components of two distributions, see [207, 595]. Other types of generalisation are (a) non-linear principal component analysis, where one considers principal components of nonlinear transformations of the original variables (see, e.g., [54]), (b) principal curves, corresponding to orthogonal non-linear regression, see [267], and in some sense also (c) principal points, which are related to cluster analysis, see [206].

Definition 2.15. *The empirical principal components and empirical principal axes, based on a sample* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, *are obtained by using* $\frac{1}{n-1}S = \hat{\boldsymbol{\Sigma}}$ *instead of* $\boldsymbol{\Sigma}$ *in the definition of the principal components and principal axes.*

Definition 2.16. *Multivariate regression is regression analysis with more than one response variable. The response variables may be correlated, having, say, a q-dimensional multivariate normal distribution. The independent variables may either consist of fixed values* (*regression type I*)*, or they may be random variables with, say, a p-dimensional normal distribution* (*regression type II*)*.*

The two-sample problem, the k-sample problem and regression analysis can be developed analogous to and as an extension of the univariate situation. This extension includes hypotheses testing and the construction of confidence regions. For this, the reader is referred to any good book on multivariate analysis, for instance [442].

Some applications of principal component analysis are

(i) *Data reduction:*
Consider a cloud of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$. Let $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ be the eigenvalues of $\boldsymbol{S} = \sum_{i=1}^n (x_i - \overline{x}_i)(x_i - \overline{x}_i)^t$ with corresponding unit eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$. If the last few eigenvectors are much smaller than the first r ones, not much information is lost by projecting the points of the cloud onto the space spanned by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$, i.e., by restricting attention to the first r principal components. An analogous situation holds for $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ i.i.d. with some probability distribution in $\mathbb{R}^p$ and a covariance matrix $\boldsymbol{\Sigma}$ that has (nearly) rank r.

(ii) *Regression on principal components:*
Sometimes, regression of a response variable Y on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is reduced to regression of Y on a subset of all principal components of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. The purpose of this is to avoid ill-conditioning of the regression. This procedure should be used with caution, however, because the principal directions of the independent variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ do not at all need to coincide with the directions in X-space along which the response variable Y varies most.

(iii) *Regression collinearity diagnostics:*In multiple regression analysis, in contrast to univariate analysis, correlations between the regression variables are taken into account, and transformed into correlations and standard deviations between the estimated regression coefficients. In several practical situations, however, the regression is hampered by multi-collinearity aspects, sometimes called ill-conditioning. This means that the data ranges of the variables are so small and/or the correlations between them so large that the regression coefficients can only be estimated with large standard deviations, and are unstable with respect to small variations of the data. Moreover, by the usual formulae, the standard deviations are estimated unreliably. (From the discussion below, one can see that both the data ranges and the correlations do play a role.)

In order to diagnose whether such a situation occurs, consider

$$\mathbf{var}\,(\hat{\boldsymbol{a}}) = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2\ . \tag{2.107}$$

The variance, σ^2, of the response variable y is inflated by $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$, which can be studied by an eigenvalue decomposition of the (symmetric) matrix $\boldsymbol{X}^t\boldsymbol{X}$ or, equivalently, of

$$\boldsymbol{S} = \sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^t = \sum_{j=1}^{p}\lambda_j^2\boldsymbol{u}_j\boldsymbol{u}_j^t\ , \tag{2.108}$$

where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ are p mutually orthogonal eigenvectors. In this representation, the inverse matrix of $\boldsymbol{S}$ equals

$$\boldsymbol{S}^{-1} = \sum_{j=1}^{p}\lambda_j^{-2}\boldsymbol{u}_j\boldsymbol{u}_j^t\ . \tag{2.109}$$

Hence, geometrically speaking, the variance inflation is large in the eigendirections with small eigenvalues λ_k, and estimation of the regression coefficients is unstable if in such directions the measurement errors in the regression variables are not small with respect to the ranges of the available data, see Fig. 2.6, where elliptical contours are drawn corresponding to

$$\boldsymbol{x}^t\boldsymbol{S}^{-1}\boldsymbol{x} = c \tag{2.110}$$

(in the space of the regression variables) and to

$$(\hat{\boldsymbol{a}} - \boldsymbol{a})^t\boldsymbol{S}(\hat{\boldsymbol{a}} - \boldsymbol{a}) = d \tag{2.111}$$

(in the dual space of the regression coefficients), for some constants c and d. A more sophisticated plot of this type can be found in Fig. 1 of [346]. Eigenvalue analysis of $\boldsymbol{X}^t\boldsymbol{X}$ corresponds to singular-value decompostion (see, e.g., [435,539]) of the singular matrix $\boldsymbol{X}$, and is intimately related to the statistical method of principal component analysis, discussed in Sect. 2.5.3. For some further aspects of collinearity analysis, of which only the conceptual idea has been outlined here, the reader is referred to [358] and Sect. 3.4.1.

(iv) *Estimation in functional relationship models:*
We restrict attention to two dimensions: Consider the model $\eta = a\xi + b$, where η and ξ are the true underlying variables, both of which can only be observed with a random error, see Fig. 2.7. Hence, observable are $Y = \eta + E_2$, with $E_2 \sim N(0, \sigma_2^2)$, and $X = \xi + E_1$, with $E_1 \sim N(0, \sigma_1^2)$. One wants to estimate the unknown parameters $a, b, \xi_1, \ldots, \xi_n, \sigma_1^2, \sigma_2^2$ from a sample $(x_1, y_1), \ldots, (x_n, y_n)$.[26]

[26] Note that for $\sigma_2 \gg \sigma_1$, we have ordinary regression of Y on X, and for $\sigma_1 \gg \sigma_2$ we have ordinary regression of X on Y. For obvious reasons, functional relationship models are also called errors–in–variable models.

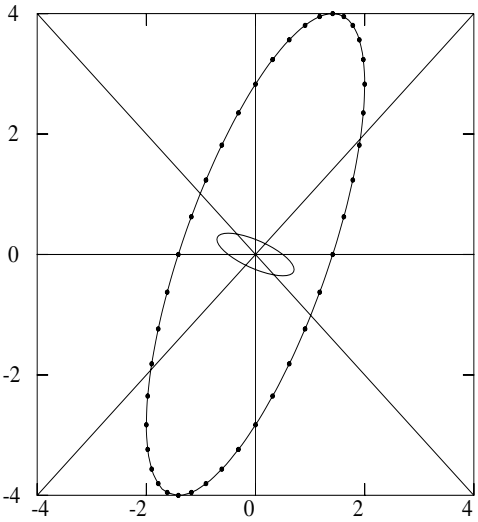

Fig. 2.6. Elliptical contours corresponding to $\boldsymbol{x}\boldsymbol{S}^{-1}\boldsymbol{x} = c$ and $\hat{\boldsymbol{a}}\boldsymbol{S}\hat{\boldsymbol{a}} = d$, respectively. The components of $\boldsymbol{x}$ are plotted along the same axes as those of $\hat{\boldsymbol{a}}$. The first contour may correspond to an ellipse fitted to the data and the second contour, for a linear regression model, to a confidence region of the regression coefficients. The ellipses have the same principal axes and are dual to each other. The latter means that the roles of the first and of the second ellipse are interchangeable. All translations have been suppressed in the figure.

It can be shown that, if $\sigma_1 = \sigma_2$, the ML estimator of the slope a corresponds to the first principal axis. If $\sigma_1 = k\sigma_2$ with k known, one can use PCA after rescaling the axes. The general problem, where all $n+4$ parameters are unknown, turns out to be difficult, because of an *identifiability problem*, which means that not all parameters can be simultaneously estimated. By differentiating the likelihood of the observations with respect to the parameters, a stationary point is found which turns out to be a saddle point! In fact, either σ_1 or σ_2 or a function of σ_1 and σ_2 such as the ratio $\kappa = \sigma_1/\sigma_2$ has to be known beforehand, or determined from other experiments, or estimated from replicated data. Otherwise, one cannot distinguish between the two extreme regression lines, obtained by minimising the vertical sum-of-squares and by

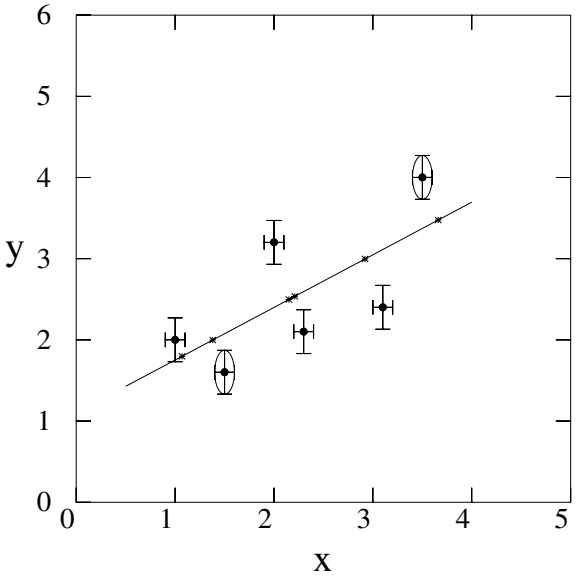

Fig. 2.7. Regression line in a two-dimensional functional relationship model. The markers on the regression line, which are the projections of the data points in the metric of the (elliptical) contours of equal error probability, are estimates of the true values (ξ, η) corresponding to each measurement, provided the linear regression model with elliptical contours of constant measurement error holds.

minimising the horizontal sum-of-squares, respectively, see Fig. 3.1.

Remark. The indetermination strictly exists only for normally distributed errors. For non-normal errors, the parameters can be estimated from the empirical moments (including the moments of order > 2). These estimates are, however, rather inaccurate in practice, unless the error distribution deviates substantially from normal, and one has a large sample size. A plasma-physical analogue is the determination of $l_i/2$ from the *Shafranov shift*, which is only possible if the plasma cross section deviates from a circle.

A particular aspect of regression analysis with errors–in–variable models is that the distribution of the regression coefficients is substantially more complicated than for ordinary regression. Inverse regression is a special case, see Sect. 3.5.4. Asymptotic expressions for the covariance matrix of these regression coefficients can be found, among others in [214] and [277,430]. However,

the appropriate approach seems to be that the regression coefficients are expressed in projective space and distributions on the sphere are used for the normalised regression coefficients. The regression plane is described by

$$a_0\eta + a_1\xi_1 + \cdots + a_p\xi_p = c\ , \tag{2.112}$$

normalised by $\sum_{k=0}^{p} a_k^2 = 1$, instead of by the usual normalisation $a_0 = -1$. In a two-dimensional, but heteroskedastic, setting this has been illustrated in [342]. The following application is sometimes viewed as a generalisation of principal component analysis.

(v) *Factor analysis model:*
In general: The 'true' values in $\mathbb{R}^p$ are constrained to lie in a hyper-plane with dimension $h \leq p-1$. The errors (possibly in all variables) are assumed to be independent, preferably with a normal distribution. Note the following special cases: If $h = p-1$, we have the previous case of a functional relationship model. If $h = p-1$ and there is only an error in the 'response variable', we have univariate multiple regression. If $h < p-1$ and there are errors only in the $p - h \geq 2$ 'response variables', we have multivariate regression. We will not present the estimation theory in this case, but refer the reader to the literature, [12, 15, 214, 277, 296, 430, 611].

As one would expect, the analysis, with its various ramifications, is centered around an eigenvalue analysis of $\frac{1}{n-1}\boldsymbol{S}$ 'in the metric' defined by $\boldsymbol{\Sigma}_e$, where $\boldsymbol{S}$ is the matrix of (corrected) sums-of-squares and cross-products of the data, and $\boldsymbol{\Sigma}_e$ is the covariance matrix of the errors. In practice this can be done by calculating the eigenvalues and eigenvectors of $\frac{1}{n-1}\boldsymbol{\Sigma}_e^{-1}\boldsymbol{S}$.

As indicated in Exercise 2.20, the lack of equivariance has its practical consequences. To discuss a concrete situation, suppose one is interested in confinement time scalings with I_p, B_t, $\langle n_e \rangle$ and P_L as regression variables, see the example from Sect. 2.4.1 and Exercise 7.1. Restricting attention to simple power–law scalings, principal component analysis can be performed on the 4 regression variables to analyse the collinearity (see Sect. 3.5 of the regression variables, and on all 5 variables to estimate the regression plane (on log scale) in an errors–in–variables model. The analysis can be based, after transformation on logarithmic (or possibly some other) scale, on
(a) the covariance matrix of the data;
(b) the correlation matrix of the data;
(c) the metric of the covariance matrix, $\boldsymbol{\Sigma}_e$, of the measurement errors;
(d) the feasibility region, i.e., the effectively possible operating range, in (say) one machine.

Cases (b), (c), and (d) correspond to PCA of the covariance matrix after a further ('secondary') transformation of the data. (In case (b) such that the standard deviation of the data becomes unity, in case (c) such that $\boldsymbol{\Sigma}_e$ becomes the unit matrix, and in case (d) such that the feasibility region becomes a unit cube or unit sphere.)

The results of these four analyses will not be equivalent. Which choice is most appropriate is influenced by the purpose of the investigation, and by the type and the quality of the measurements. Sometimes, more than one analysis is useful to highlight different aspects of the data. We try to give some rough practical guidelines.

Choice (a) can be used if the units on each scale are more or less comparable. With respect to the other cases, it has the advantage that the units do not depend on the uncertainties and variability of the secondary transformation.

Choice (b) is reasonable in those situations that the units along the axes of the variables are totally incomparable, for instance, one variable being expressed in eV and some other one in m, and the data can be considered to be a representative sample from some population.

Since the correlation matrix, by its construction, does not reflect the amount of variation of the data on a physically relevant scale, it should be avoided to use PCA of the correlation matrix to investigate collinearity related to prediction reliability, see also Sect. 3.5.2.

Choice (c) is indicated when a reasonably reliable estimate of the error covariance matrix is available and the purpose of the investigation is to fit errors–in–variable models as an improvement and/or as a check of ordinary least squares regression.

Choice (d) is a sensible approach, for a feasibility region that has some reasonably convex shape, when the purpose of investigation is to compare (for a given number of observations) the statistical condition and the prediction accuracy of a specific regression with the condition and accuracy that can be obtained when using the total operating range of the machine.

Exercise 2.22. Using SAS, S–PLUS or MATLAB, apply principal component analysis to fit a simple power law model to the (hypothetical) ASDEX Upgrade confinement data (see Exercise 7.1), the density limit data (see Exercise 7.8) or a dataset from your own experience. Take the measurement errors in the regression variables into account and compare the results with ordinary least squares regression.

3 Applied Linear Regression

3.1 Introduction

In this section we revisit the subject of regression analysis while focusing attention to topics of an applied nature. Inevitably there will be some overlap with Chap. 2. In effect, a good understanding of Sect. 2.4 is most useful, but not a prerequisite, for following the present chapter. While shall still strive for an accurate style of narrative, we will no longer adhere to the formal style of 'definition, theorem, proof' which has been employed in the first two chapters. At times, the points at issue are stated rather briefly while references are made to the literature. For actual applications based on concrete datasets the reader is referred to Chap. 7.

Origin of the term 'Regression'

More than one hundred years ago, the British anthropologist and statistician Sir Francis Galton [229] investigated the dependence of the height of children on the height of their parents.[1] Figure 3.1 comes from one of Galton's publications on inheritance [218], see also [219]. More than 1000 families had been included in the study.

Looking at the straight line which constitutes the 'Locus of horizontal tangential points', which are the tangential points of horizontal lines with the concentric family of similarly shaped ellipses, one of which approximates the data distribution by fitting the (two) first and (three) second moments, one can see that taller parents have taller children. However, the slope of this line is larger than 1. This means that the increase of the height (with respect to the average) of the children is smaller than the increase of the height of the (mid-)parents. So, from generation to generation, there seems at least to be a downward tendency in the mean height of children. To describe this tendency, Galton used the term *Regression*. The reader should be aware that the origin of the word regression is, hence, based on an apparent, albeit not a real, effect. For a recent re-analysis of the dataset, see [712].[2]

[1] In some statistics courses, it is retained that from his own pocket Galton offered prizes for the collection of high quality datasets.

[2] In this article, the dataset is analysed for both sexes separatedly, and a nonlinearity in the regression is detected.

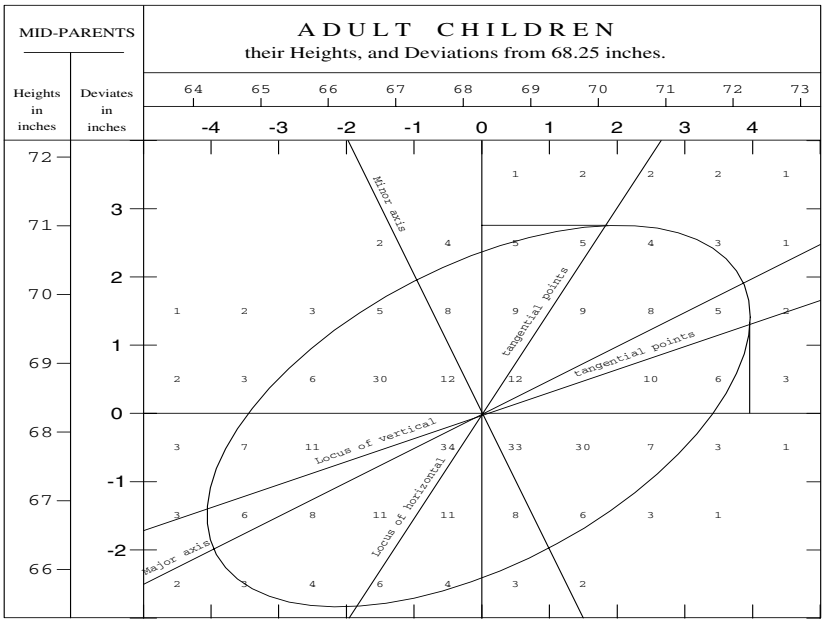

Fig. 3.1. Plot by Sir Francis Galton [218], published in 1886, depicting the correlation between the height of children and the average height of their parents. All female heights have been multiplied by 1.08. The ellipse describing the data-approximating distribution is centered around the global mean values of the children's and mid-parent's heights. The printed numbers give the number of observations in a corresponding square.

Regression Analysis

In mathematical analysis, a *function:* $f\colon X \to Y$ is defined as a many-to-one mapping, or, equivalently, as a binary relation (i.e., a subset of the Cartesian product $X \times Y$) with the special property that it assigns to each value $x \in X$ one and only one value $y \in Y$. In statistics, the corresponding concept is the *regression of* Y *on* X, where Y is a random variable, to be estimated for each realised value of X. We distinguish between three frameworks (or 'models') of regression analysis:

(1) X is exactly determined, while the errors in Y have a (normal) distribution;

(2) X and Y have together a multivariate (normal) distribution;

(3) there are errors in both X and Y.

The first two cases are covered by standard regression analysis, and the third case by errors–in–variable models.

Important European contributions to applied regression analysis during the previous century, embedded into the statistical discipline at large [327, 396], were made, among others, by:

- F. Galton and K. Pearson around 1900
- R.A. Fisher 1920–1950
- J. Neyman and E. Pearson 1930–1960
- D.R. Cox 1960–2000

Modern statistical computer packages make it easy to apply regression analysis even on rather complex problems. In order to study a certain problem, the first prerequisite is pertinent scientific knowledge. However, some additional statistical understanding is helpful to: (1) avoid misinterpreting the empirical contents of experimental data; (2) avert treating wrong models or wrong estimates of the parameters as true. Biographies of leading statistical personalities who were born before the twentieth century can be found in [327].

3.2 Estimation and Hypothesis Testing

3.2.1 Linear Regression Models

The following two models of (linear) regression are called *type I regression* and *type II regression*, respectively.

Regression I

$Y = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p + E\,, \qquad E \sim N(0, \sigma^2)\,,$

where:

Y is the dependent variable (response variable);

$\alpha_0, \alpha_1, \ldots, \alpha_p$ are the *regression parameters* (α_0 is the *intercept*);

$x_1, x_2, \ldots, x_p$ are the independent variables (*regressors*);

E is a random error (e.g., random measurement error).

Sometimes, a regression model exists between Y' and $x'_1, \ldots, x'_p$ which can be written as a linear model between Y and $x_1, \ldots, x_p$, where $x_j = \varphi_j(x'_j)$ and $Y = \varphi(Y')$ for monotonic, for instance logarithmic, transformations $\varphi_1, \ldots, \varphi_p$ and φ. It is noted that, in general, additivity of the errors is not preserved under such non-linear transformations.

Regression II

$(Y, X_1, \ldots, X_p)$ is a $(p+1)$-dimensional random vector from a $(p+1)$-dimensional normal distribution $N(\mu, \Sigma)$, where the conditional expectation $\mathrm{E}(Y|x_1, \ldots, x_p)$ and the (conditional) variance $\mathrm{var}\,(Y|x_1, \ldots, x_p)$ are to be estimated. It is assumed that the (conditional) probability density

$f_Y(y|X_1 = x_1, \ldots, X_p = x_p)$ is symmetric for each value of $\boldsymbol{x} = (x_1, \ldots, x_p)$. Because of this symmetry, the mean, median and modus of $f_Y(y|\boldsymbol{x})$ coincide.

Both models provide estimation formulae that can be easily converted into each other. For a third model, with errors in both X and Y, special multivariate methods are required, and the reader is referred to Chap. 2. Here, we restrict attention to type I regression. As observations of the response variable we consider n outcomes of Y with different (independent) realisations $e_1, \ldots, e_n$ of a random variable E.

A principal objective of regression analysis is to find 'optimal' estimates for the regression parameters. There are several optimality criteria. One of these criteria requires that an 'optimal' estimator should be *uniformly minimum variance unbiased* (UMVU), which means that the estimator

(1) is unbiased, i.e., its expected value is equal to the estimated parameter,

(2) has 'minimal variance', i.e., the determinant of its covariance matrix is minimal.

For n observations we can write

$$Y_i = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi} + E_i \quad (i = 1, \ldots, n) \,, \tag{3.1}$$

or in matrix form

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_p \end{pmatrix} + \begin{pmatrix} E_1 \\ \vdots \\ E_n \end{pmatrix} \,, \tag{3.2}$$

or

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{E}, \qquad \boldsymbol{E} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2) \,. \tag{3.3}$$

The model, concisely described by (3.3), implies the following standard assumptions:

(0) The linear model is 'correct';

(1) $\boldsymbol{x} = (x_1, \ldots, x_p)$ is measured without errors (or with errors that can be practically neglected);

(2) $E_1, \ldots, E_n$ are independent random variables;

(3) $E_i \sim N(0, \sigma^2)$.

Remarks.

1. In general, σ^2 is unknown.
2. Since E_i is the only random variable in the expression

$$\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p + E_i \,,$$

we have

$$\mathrm{E}(Y_i) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p \,,$$

$$\mathrm{var}\,(Y_i) = \mathrm{var}\,(\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p + E_i) = \mathrm{var}\,(E_i) = \sigma^2\ , \quad (3.4)$$

and

$$Y_i \sim N(\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p, \sigma^2)\ . \quad (3.5)$$

3. If $\boldsymbol{E}$ is multivariate normally distributed, then also $\boldsymbol{Y}$ is multivariate normally distributed.
4. In Sect. 3.5 we discuss aspects of regression analysis when these standard assumptions are relaxed.

3.2.2 Maximum Likelihood and Least-Squares Estimators

Maximum Likelihood estimate (ML)
We consider the following simple example of linear regression:

$$Y_i = \alpha_0 + \alpha_1 x_i + E_i \quad (i = 1, \ldots, n)\ . \quad (3.6)$$

If the errors are independent and have the same probability distribution, then the likelihood $L = L_{\boldsymbol{y}}(\alpha_0, \alpha_1)$ is given by

$$\begin{aligned} L &= f_{\alpha_0,\alpha_1}(Y_1 = y_1, \ldots, Y_n = y_n) \\ &= \prod_{i=1}^{n} f(y_i - \alpha_0 - \alpha_1 x_i) \\ &= f^n(E_i)\ . \end{aligned} \quad (3.7)$$

The maximum likelihood estimate maximises

$$\begin{aligned} (\hat{\alpha}_0, \hat{\alpha}_1)_{ML} &= \max_{\alpha_0,\alpha_1} L \\ &= \max_{\alpha_0,\alpha_1} \log(L) \\ &= \max_{\alpha_0,\alpha_1} \sum_{i=1}^{n} \log\, f(y_i - \alpha_0 - \alpha_1 x_i)\ . \end{aligned} \quad (3.8)$$

Least-Squares estimate (LS)
Minimising

$$(\hat{\alpha}_0, \hat{\alpha}_1)_{LS} = \min_{\alpha_0,\alpha_1} \sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 x_i)^2$$

is equivalent to maximising

$$(\hat{\alpha}_0, \hat{\alpha}_1)_{LS} = \max_{\alpha_0,\alpha_1} \big(-\sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 x_i)^2\big)\ . \quad (3.9)$$

Comparing (3.8) with (3.9), we see that the least squares and maximum likelihood estimators coincide, i.e., "LS = ML", if

$$f(z_i) = e^{-z_i^2} \ , \quad \text{where} \quad z_i = \text{error}_i = y_i - \alpha_0 - \alpha_1 x_i \ . \tag{3.10}$$

Conclusion. In linear regression, if the errors are independent and normally distributed, then the LS estimator equals the ML estimator of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$.

In general, an LS estimator has three advantages over an ML estimator:

(1) in a linear model it allows estimation of the regression parameters through an explicit matrix expression and does not require numerical minimisation;

(2) it has a direct geometrical interpretation, see Sect. 2.4;

(3) it tends to be more robust than an ML estimator against misspecification of the error distribution.

On the other hand, ML estimators may be more 'efficient' if the family of the error distributions is known.

3.2.3 Method of Least Squares

In general, we look for an estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$, such that

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2 = \hat{\boldsymbol{e}}^t \hat{\boldsymbol{e}} \tag{3.11}$$

is minimised, where $\hat{y}_i = \hat{\boldsymbol{\alpha}}^t \boldsymbol{x}_i$ is the fitted value of y_i, and $\hat{e}_i$ is an *empirical residual*. Hence, we minimise the sum-of-squares of the residuals $\hat{e}_i$.

In case of linear regression, see (3.3), differentiation with respect to $\hat{\boldsymbol{\alpha}}$,

$$\Big(\frac{\partial \hat{\boldsymbol{e}}^t \hat{\boldsymbol{e}}}{\partial \boldsymbol{\alpha}}\Big)_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}} = \boldsymbol{0} \ , \tag{3.12}$$

yields the following *normal equations* (c.f. [604])

$$\boldsymbol{X}^t \boldsymbol{X} \boldsymbol{\alpha} = \boldsymbol{X}^t \boldsymbol{Y} \ . \tag{3.13}$$

If $\boldsymbol{X}^t \boldsymbol{X}$ is invertible,[3] then a unique solution exists:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{Y} \ . \tag{3.14}$$

The estimator $\hat{\boldsymbol{\alpha}}$ has following properties:

[3] The matrix $\boldsymbol{X}^t \boldsymbol{X}$ is called invertible if $\boldsymbol{X}$ is nonsingular, i.e., if there does not exist a linear dependence between the columns of $\boldsymbol{X}$, which means that not any regression variable can be written as a linear combination of the other ones.

(1) It is unbiased

$$\boldsymbol{E}(\hat{\boldsymbol{\alpha}}) = [(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t]\boldsymbol{E}(\boldsymbol{Y}) = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\alpha} = \boldsymbol{\alpha}\ . \tag{3.15}$$

(2) The covariance matrix of $\hat{\boldsymbol{\alpha}}$ is

$$\mathbf{var}\,(\hat{\boldsymbol{\alpha}}) = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2\ . \tag{3.16}$$

It can be proved that among all unbiased estimators, the estimator $\hat{\boldsymbol{\alpha}}$ has a 'minimal' covariance matrix, see Sect. 2.4.3.

3.2.4 Geometrical Interpretation

Let us look again at the graphical representation in Fig. 2.5.

The regressors $\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ define a $(p+1)$-dimensional subspace $V_X \subset \mathbb{R}^n$. Under the assumption that $n > p+1$, we have

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\alpha}} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y}\ . \tag{3.17}$$

The matrix

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t \tag{3.18}$$

is symmetric and idempotent, i.e.,

$$\boldsymbol{P}^t = \boldsymbol{P}, \qquad \boldsymbol{P}^2 = \boldsymbol{P}\ . \tag{3.19}$$

In other words, $\boldsymbol{P}$ is an orthogonal (since $\boldsymbol{P}^t = \boldsymbol{P}$) projection (since $\boldsymbol{P}^2 = \boldsymbol{P}$) matrix of $\boldsymbol{Y}$ on V_X. It is remarked that omission of an important regressor leads to a biased LS estimator.

Derivation of the Least-Squares (LS) estimator

A derivation of the LS estimator and its properties has been given in Sect. 2.4. Here, we recall and re-interpret some of the essential aspects at a leisurely pace, while occasionally using a somewhat different formulation.

If the assumptions of the model (3.3) are satisfied, then $\mathrm{E}(\boldsymbol{E}) = \mathbf{0}$, and $\mathrm{E}(\boldsymbol{Y})$ is a linear combination of $\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ in V_X. Therefore, we try to find an estimator for $E(\boldsymbol{Y})$ in V_X:

$$\hat{\boldsymbol{Y}} = \hat{\alpha}_0 + \hat{\alpha}_1\boldsymbol{X}_1 + \cdots + \hat{\alpha}_p\boldsymbol{X}_p\ . \tag{3.20}$$

The LS estimator minimises $||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||^2 = ||\hat{\boldsymbol{E}}||^2$. In a Euclidean space, this is an orthogonal projection of $\boldsymbol{Y}$ on V_X, which implies that $(\boldsymbol{Y} - \hat{\boldsymbol{Y}})$ is orthogonal to each column of $\boldsymbol{X}$.

In algebraic notation:

$$(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^t\boldsymbol{X} = \mathbf{0}\ , \tag{3.21}$$

or, after transposition,

$$\boldsymbol{X}^t(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \mathbf{0}\ , \tag{3.22}$$

whence we have

$$\boldsymbol{X}^t\hat{\boldsymbol{Y}} = \boldsymbol{X}^t\boldsymbol{Y} \ . \tag{3.23}$$

Substitution of

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\alpha}} \ , \tag{3.24}$$

yields

$$\boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{\alpha}} = \boldsymbol{X}^t\boldsymbol{Y} \ . \tag{3.25}$$

These are the normal equations, see (3.13). Hence, we derived, in a geometric way, (3.14) for the estimator $\hat{\boldsymbol{\alpha}}$:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} \ . \tag{3.26}$$

Further geometrical interpretations

The *regression coefficients* are the *contravariant components* (parallel projections) of the vector (1-dimensional tensor) $\hat{\boldsymbol{Y}}$ in the covariant base $\{\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p\}$. The numbers $\mathbf{1}^t\hat{\boldsymbol{Y}}, \boldsymbol{X}_1^t\hat{\boldsymbol{Y}}, \ldots, \boldsymbol{X}_p^t\hat{\boldsymbol{Y}}$ are the *covariant components* of tensor $\hat{\boldsymbol{Y}}$ in this base. Note that if the base vectors $\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ are unit vectors, then the covariant components are exactly the perpendicular projections. The matrix $\boldsymbol{X}^t\boldsymbol{X}$ is the *metric tensor* of $\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$. The covariant components of a vector are converted into the contravariant components by multiplication with the inverse of the metric tensor.

The vector $\mathbf{1}$ multiplied with the *mean value* $\bar{u}$ of a vector $\boldsymbol{U}$, $\overline{\boldsymbol{U}} = \bar{u}\mathbf{1}$, is the orthogonal projection of $\boldsymbol{U}$ on the vector $\mathbf{1}$.

The *centered vector* $\widetilde{\boldsymbol{U}} = \boldsymbol{U} - \overline{\boldsymbol{U}}$ is the vector of differences between $\boldsymbol{U}$ and its orthogonal projection on $\mathbf{1}$.

The *simple correlation* between two regression variables $\boldsymbol{X}_1, \boldsymbol{X}_2$ is equal to

$$\operatorname{corr}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \cos\ \varphi(\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_2) \ , \qquad \widetilde{\boldsymbol{X}}_i = \boldsymbol{X}_i - \overline{\boldsymbol{X}}_i \ . \tag{3.27}$$

The concept of a partial correlation generalises that of a simple correlation.

The *partial correlation* between two regression variables $\boldsymbol{X}_1, \boldsymbol{X}_2$, *given* $\boldsymbol{X}_3$

$$\operatorname{corr}\big(\boldsymbol{X}_1, \boldsymbol{X}_2 | (\mathbf{1}, \boldsymbol{X}_3)\big) \ ,$$

is the correlation between the residuals of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, after $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have been predicted as well as possible by a linear combination of $\mathbf{1}$ and $\boldsymbol{X}_3$. In other words, when $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are projected on the subspace spanned by $(\mathbf{1}, \boldsymbol{X}_3)$, and these projections are denoted by $\boldsymbol{X}_{1,p}$ and $\boldsymbol{X}_{2,p}$, then the partial correlation equals the cosine of the angle between $\boldsymbol{X}_1 - \boldsymbol{X}_{1,p}$ and $\boldsymbol{X}_2 - \boldsymbol{X}_{2,p}$.

A basic property is the following: simple/partial correlation coefficients are the simple/partial regression coefficients for the centered variables with length (i.e., standard deviation) one.

The *degrees of freedom* associated with a sum-of-squares is the number of dimensions in which the corresponding vector is 'free to move'. In the situation above, we are to assign the following degrees of freedom:

$$\boldsymbol{Y} \to n\,, \qquad \hat{\boldsymbol{Y}} \to p+1\,, \qquad \boldsymbol{E} \to n-p-1\,. \tag{3.28}$$

3.2.5 Distribution Theory of Linear Regression Estimators

Variance of the LS estimator

Here we will derive an expression for the variance of the $(p+1)$-dimensional vector of regression coefficients, as well as of the n-dimensional vector of predicted values $\hat{\boldsymbol{Y}}_{pred}$.

We recall the following property of the variance, derived in Chap. 1, Theorem (1.3):

$$\mathrm{var}\,(U_1+U_2) = \mathrm{var}\,(U_1) + \mathrm{var}\,(U_2) + 2\mathrm{cov}\,(U_1,U_2)\,. \tag{3.29}$$

If U_1 and U_2 are independent $(\mathrm{cov}\,(U_1,U_2)=0)$, then

$$\mathrm{var}\,(U_1+U_2) = \mathrm{var}\,(U_1) + \mathrm{var}\,(U_2)\,.$$

This can be generalised to the variance of a linear combination of the components of a random vector:

$$\mathrm{var}\,(\boldsymbol{c}^t\boldsymbol{Y}) = \boldsymbol{c}^t[\mathbf{var}\,(\boldsymbol{Y})]\boldsymbol{c} \tag{3.30}$$

and even to the covariance matrix of a number of such linear combinations

$$\mathbf{var}\,(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}[\mathbf{var}\,(\boldsymbol{Y})]\boldsymbol{A}^t\,. \tag{3.31}$$

In particular, $\hat{\boldsymbol{\alpha}}$ is a linear function of $\boldsymbol{Y}$:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y}\,.$$

Quite generally,

$$\mathbf{var}\,(\hat{\boldsymbol{\alpha}}) = [(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t][\mathbf{var}\,(\boldsymbol{Y})][(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t]^t\,.$$

Under the assumption that $\mathbf{var}\,(\boldsymbol{Y}) = \boldsymbol{I}\sigma^2$, we get

$$\mathbf{var}\,(\hat{\boldsymbol{\alpha}}) = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2\,. \tag{3.32}$$

Analogously, one can derive:

$$\mathbf{var}\,(\hat{\boldsymbol{Y}}) = \boldsymbol{X}[\mathbf{var}\,(\hat{\boldsymbol{\alpha}})]\boldsymbol{X}^t = \boldsymbol{P}\sigma^2\ , \tag{3.33}$$

$$\mathbf{var}\,(\hat{\boldsymbol{Y}}_{pred}) = \mathbf{var}\,(\boldsymbol{Y}) + \mathbf{var}\,(\hat{\boldsymbol{Y}}) = (\boldsymbol{I}+\boldsymbol{P})\sigma^2\ , \tag{3.34}$$

$$\mathbf{var}\,(\hat{\boldsymbol{E}}) = (\boldsymbol{I}-\boldsymbol{P})\sigma^2\ . \tag{3.35}$$

In (3.34) $\hat{\boldsymbol{Y}}_{pred}$ represents the prediction of a single new observation at the point $(x_1,\ldots,x_n)$.

Note that $\mathbf{var}\,(\hat{\boldsymbol{Y}}) + \mathbf{var}\,(\hat{\boldsymbol{E}}) = \boldsymbol{I}\sigma^2$, because $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{E}}$ are geometrically orthogonal or, equivalently, statistically uncorrelated. Since a variance cannot be negative, it follows from (3.35) that $0 < p_{ii} < 1$ for each diagonal element p_{ii} in $\boldsymbol{P}$. This entails

$$\mathrm{var}\,(\hat{Y}_i) \le \mathrm{var}\,(Y_i)\ . \tag{3.36}$$

The practical interpretation is that fitting a continuous model gives a more precise prediction of the mean value of Y_i than does any single observation. (This occurs because the former takes into account the 'surrounding' observations as well.)

Summary of distribution formulae in linear regression

According to the standard assumptions made for linear regression model, errors in the response variable are normally distributed. Since linear combinations of normally distributed random variables are again normally distributed, we get, using the formulae above,

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\alpha}, \boldsymbol{I}\sigma^2)\ , \tag{3.37}$$

$$\hat{\boldsymbol{\alpha}} \sim N(\boldsymbol{\alpha}, (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2)\ , \tag{3.38}$$

$$\hat{\boldsymbol{Y}} \sim N(\boldsymbol{X}\boldsymbol{\alpha}, \boldsymbol{P}\sigma^2)\ , \tag{3.39}$$

$$\hat{\boldsymbol{Y}}_{pred} \sim N(\boldsymbol{X}\boldsymbol{\alpha}, (\boldsymbol{I}+\boldsymbol{P})\sigma^2)\ , \tag{3.40}$$

$$\hat{\boldsymbol{E}} \sim N(\boldsymbol{0}, (\boldsymbol{I}-\boldsymbol{P})\sigma^2)\ . \tag{3.41}$$

3.2.6 Confidence Regions and Testing of Hypotheses

Confidence Regions

Case 1. Confidence region for a single unknown parameter.

The above outlined distribution theory of linear regression estimators can be used to construct, for a certain model and for a certain data set, a confidence interval for the unknown parameter α_j,

$$\hat{\alpha}_j \pm t_{f;\alpha} s_{\hat{\alpha}_j}\ , \tag{3.42}$$

where $s_{\hat{\alpha}_j}$ is an estimate of the standard deviation of $\hat{\alpha}_j$, and $t_{f;\alpha}$ is some constant, depending on the degrees of freedom f, and on the significance

level α (or 'confidence level' $1-\alpha$). In classical (objectivistic) statistics, the unknown parameter has a fixed (non-random) value, whereas a confidence interval is a random quantity.

Generally speaking, a confidence region for an unknown (estimated) parameter U can be defined as the set of all possible values of u, for which the null–hypothesis $U = u$ is not rejected. A 95% confidence (random) interval means that, if the model conditions are met, the unknown parameter is located in the interval with 95% probability. Increasing of the confidence level from 95% to, e.g., 99%, as well as reducing the number of observations, enlarges the interval.

Case 2. Confidence regions for a number of unknown parameters.
We assume that $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{Y}}$ have multidimensional normal distributions. A confidence region defined for the regression coefficients is

$$\{\boldsymbol{\alpha}; (\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}})^t (X^t X)(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}) \le (p+1)s^2 F_{p+1,n-p-1;\alpha}\} \; . \tag{3.43}$$

The left-hand side is a positive definite quadratic form of $\boldsymbol{\alpha}$, the quantities $\hat{\boldsymbol{\alpha}}$, $\boldsymbol{X}^t\boldsymbol{X}$, $(p+1)s^2$ and $F_{p+1,n-p-1;\alpha}$ being known. Hence, the inequality can be represented by an ellipsoid in $\mathbb{R}^{p+1}$, centered around $\hat{\boldsymbol{\alpha}}$, as is shown in Fig. 3.2.

Case 1 (cont.).
As we have seen in case 1 above, a confidence interval for a single regression parameter α_j is given by $\hat{\alpha}_j \pm t_{f;\alpha} s_{\hat{\alpha}_j}$. We can now make some further identifications: t is the critical value of a t-distribution with $f = n-p-1$ degrees of freedom and significance level α, while $s_{\hat{\alpha}_j}$ is the usual estimate of $\sigma_{\hat{\alpha}_j}$, i.e.,

$$s^2_{\hat{\alpha}_j} = c_{jj} s^2 \; , \tag{3.44}$$

where c_{jj} is the jth diagonal element of $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$.

We now construct confidence intervals for a linear (but unknown) response function. At any $\boldsymbol{x} = (1, x_1, \ldots, x_p) \in \mathbb{R}^{p+1}$, where $\boldsymbol{x}$ is a row from $\boldsymbol{X}$, the confidence interval is

$$\hat{Y}(\boldsymbol{x}) \pm t_{f;\alpha} s_{\hat{Y}(\boldsymbol{x})} \; , \tag{3.45}$$

where

$$\begin{aligned} \hat{\boldsymbol{Y}}(\boldsymbol{x}) &= \boldsymbol{x}^t \hat{\boldsymbol{\alpha}} \; , \\ s_{\hat{Y}(\boldsymbol{x})} &= \sqrt{\boldsymbol{x}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}^t}\, s \; , \\ s^2 &= \frac{1}{n-p-1} \sum_{k=1}^{n} (y_k - \boldsymbol{x}_k^t \hat{\boldsymbol{\alpha}})^2 \; . \end{aligned}$$

For example, in the simple case of $Y = \alpha_0 + x\alpha_1$ we have

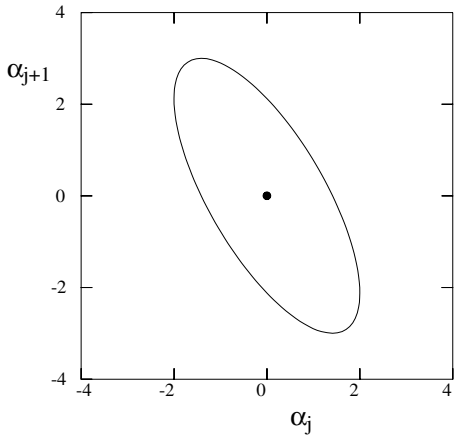

Fig. 3.2. A $(p+1)$-dimensional confidence region for the unknown parameter α in a multiple regression model, intersected with the (α_j, α_{j+1})-plane. For normally distributed errors, the shape of the region is an ellipse. For large samples, the shape is approximately an ellipse under rather weak distributional assumptions, by virtue of the central limit theorem. The center of the ellipse indicates the point estimate $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \ldots, \hat{\alpha}_p) \in \mathbb{R}^{p+1}$.

$$s^2_{\hat{\alpha}_0} = s^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{k=1}^n (x_k - \overline{x})^2} \right], \tag{3.46}$$

$$s^2_{\hat{\alpha}_1} = \frac{s^2}{\sum_{k=1}^n (x_k - \overline{x})^2}, \tag{3.47}$$

$$s^2_{\hat{Y}(x)} = s^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{k=1}^n (x_k - \overline{x})^2} \right]. \tag{3.48}$$

From the last formula one can directly infer that the variance of $\hat{Y}$ reaches its minimum at $x = \bar{x}$, which can also be seen in Fig. 3.3. Naturally, $s^2_{\hat{Y}(0)}$ equals $s^2_{\hat{\alpha}_0}$.

Testing of Hypotheses

There exists a close relationship between estimating confidence intervals and testing hypotheses.

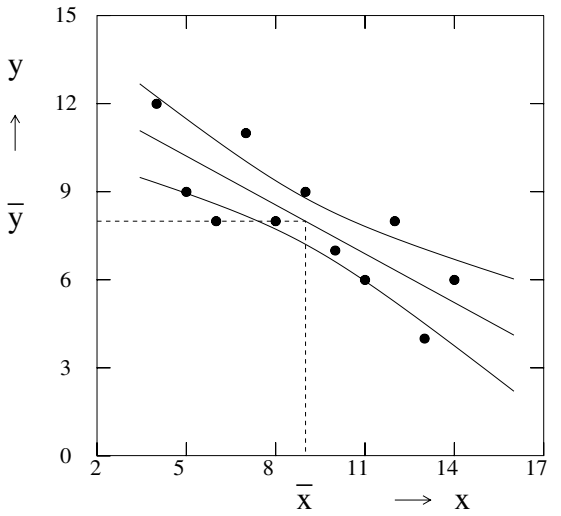

Fig. 3.3. Illustration of a (non-simultaneous) confidence band in simple regression. The (artificial) observations are indicated by solid dots. The smooth curves, which are two branches of an hyperbola, enclose a confidence region. For a single fixed x_0, $\big(y_L(x_0), y_H(x_0)\big)$ is a confidence interval for $E(Y|x_0)$, while, by inversion, for a single fixed y_0, $\big(x_L(y_0), x_H(y_0)\big)$ is a confidence interval for $x(y_0)$.

Case 1. We consider a single regression parameter and test the null hypothesis

$$H_0\colon\ \alpha_j = 0\ , \tag{3.49}$$

(for $i \neq j$, α_i may assume any value) against the alternative

$$A\colon\ \alpha_j \neq 0\ . \tag{3.50}$$

Note that $\alpha_j = 0$ means there does not exist a linear correlation between $\boldsymbol{X}_j$ and $\boldsymbol{Y}$ when all other variables are held constant. If the null hypothesis is true, then

$$\frac{\hat{\alpha}_j}{s_{\hat{\alpha}_j}} = \frac{\hat{\alpha}_j - 0}{\sqrt{\widehat{\mathrm{var}}(\hat{\alpha}_j)}} = T \tag{3.51}$$

has a (central) t-distribution with $(n - p - 1)$ degrees of freedom. For large n this quantity approaches a normal distribution, see Sect. 1.5. According to

standard two-sided testing theory, the null hypothesis $\alpha_j = 0$ is not rejected if $|T| < t_{n-p-1;0.05}$, see Fig. 3.4. This test is called an unbiased test, see [424], in the sense that the power function, for any parameter α_j in the alternative, is larger than the error of the first kind. This corresponds to the U-shaped curve in Fig. 2.2 (a) assuming its minimum for $\theta = \theta_0$. In Fig. 3.4, the probability densities are depicted for Student's t distributions with 1, 2, 3, 5 and 30 degrees of freedom. The critical values, $t_{df;0.05}$ for df=1, 2, 3, 4, 5, 10, 30, ∞ are 12.7, 4.3, 3.18, 2.78, 2.57, 2.76, 2.0, 1.96, respectively.

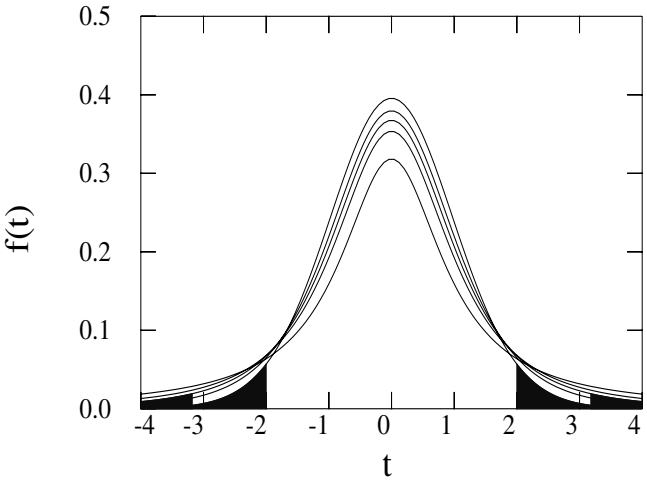

Fig. 3.4. Student's t probability densities $f_{df}(t)$ with df = 1, 2, 3, 5, and 30 degrees of freedom. The first curve, with the broadest tail and smallest maximum, corresponds to the Cauchy (Lorentz) distribution. For df = 30, the distribution is close to the standard normal, $N(0, 1)$. The intervals $(-\infty, -t_{df;0.05})$ and $(t_{df;0.05}, \infty)$ are indicated by the black areas under the distribution curves (drawn for df = 3 and df = 30) and constitute the rejection regions for the outcome of a random variable ('test statistic') $T = \hat{\alpha}_j / SD(\hat{\alpha}_j) \sim t_f$ when the null–hypothesis H_0: $\alpha_j = 0$ is tested against A: $\alpha_j \neq 0$ at the (two-sided) significance level $\alpha = 0.05$. The standard deviation of $\hat{\alpha}_j$, for an approximately normal distribution, has been estimated by $SD(\hat{\alpha}_j)$, based on df degrees-of-freedom (i.e., on df 'effective observations').

Case 2. A linear combination of regression parameters.
The null hypothesis is

$$H_0\text{: } \boldsymbol{K}^t\boldsymbol{\alpha} = \boldsymbol{m}\ , \tag{3.52}$$

and a possible alternative hypothesis is

$$A\text{: } \boldsymbol{K}^t\boldsymbol{\alpha} \neq \boldsymbol{m}\ . \tag{3.53}$$

A concrete example is

$$\boldsymbol{\alpha}^t = (\alpha_0,\ \ \alpha_1,\ \ \alpha_2,\ \ \alpha_3)\ , \tag{3.54}$$

with

$$H_0\text{:} \qquad \alpha_1 - \alpha_2 = 2\ ,\ \ \alpha_1 + \alpha_2 = 3\alpha_3\ ,\ \ \alpha_0 = 10\ , \tag{3.55}$$

which we can write as

$$H_0\text{:} \quad \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 1 & -3 \\ 1 & 0 & 0 & 0 \end{pmatrix} \boldsymbol{\alpha} = \begin{pmatrix} 2 \\ 0 \\ 10 \end{pmatrix}\ . \tag{3.56}$$

It is natural to take $\boldsymbol{K}^t\hat{\boldsymbol{\alpha}}$ as an estimator for $\boldsymbol{K}^t\boldsymbol{\alpha}$. Since $\hat{\boldsymbol{\alpha}} \sim N(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$, we have

$$\boldsymbol{K}^t\hat{\boldsymbol{\alpha}} \sim N(\boldsymbol{K}^t\boldsymbol{\alpha}, \boldsymbol{K}^t\boldsymbol{\Sigma}\boldsymbol{K})\ , \tag{3.57}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2.$$

If H_0: $\boldsymbol{K}^t\boldsymbol{\alpha} = \boldsymbol{m}$ (i.e., $\boldsymbol{K}^t\boldsymbol{\alpha} - \boldsymbol{m} = \boldsymbol{0}$) holds, then

$$\boldsymbol{K}^t\hat{\boldsymbol{\alpha}} - \boldsymbol{m} \sim N\big(\boldsymbol{0}, \boldsymbol{K}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{K}\sigma^2\big)\ . \tag{3.58}$$

The contours of constant probability density of $\boldsymbol{K}^t\hat{\boldsymbol{\alpha}}$ form ellipses in $\mathbb{R}^{r(\boldsymbol{K})}$, where $r(\boldsymbol{K})$ is the column rank of the matrix $\boldsymbol{K}$ (or, equivalently, the row rank of $\boldsymbol{K}^t$). This can be understood as follows, see also Fig. 3.2. The rows of the matrix $\boldsymbol{K}^t$ span an $r(\boldsymbol{K})$–dimensional subspace[4] of $\mathbb{R}^p$. In fact, the lines associated with the rows of $\boldsymbol{K}^t$ constitute a non-orthogonal, linear coordinate system of this subspace. The $(p+1)$–dimensional ellipse, a two-dimensional section of which is depicted in Fig. 3.2, projected onto this subspace yields the distribution of $\boldsymbol{K}^t\hat{\boldsymbol{\alpha}}$, since $\boldsymbol{K}^{(j)}\hat{\boldsymbol{\alpha}}$, with $\boldsymbol{K}^{(j)}$ the jth column of $\boldsymbol{K}$, is proportional to the projection of $\hat{\boldsymbol{\alpha}}$ on the line associated with $\boldsymbol{K}^{(j)}$, defined as $c\boldsymbol{K}^{(j)}$, $c \in \mathbb{R}$. Finally, in case the columns of $\boldsymbol{K}$ are geometrically plotted as orthogonal vectors in $\mathbb{R}^{r(\boldsymbol{K})}$, rather than as non-orthogonal vectors in

[4] In terms of linear algebra: Each of the $r(\boldsymbol{K})$ independent linear equations determines a p–dimensional subspace. The intersection of these is a $p - r(\boldsymbol{K}) + 1$ dimensional subspace. The $(p+1)$–dimensional confidence ellipse is projected onto the $r(\boldsymbol{K})$ dimensional subspace dual to, and since each linear equation can be viewed as an inner product in a Euclidean space also orthogonal to, the $p - r(\boldsymbol{K}) + 1$ dimensional intersection space.

$\mathbb{R}^p$, the property of the contours being elliptical is preserved. The symmetric matrix $\boldsymbol{K}^t\boldsymbol{\Sigma}\boldsymbol{K}$ defines a norm (and hence a metric) in $\mathbb{R}^{r(\boldsymbol{K})}$:

$$||\boldsymbol{Z}||^2_{\boldsymbol{K}^t\boldsymbol{\Sigma}\boldsymbol{K}} = \boldsymbol{Z}^t(\boldsymbol{K}^t\boldsymbol{\Sigma}\boldsymbol{K})^{-1}\boldsymbol{Z} \ . \tag{3.59}$$

The square of the distance between $\boldsymbol{K}^t\hat{\boldsymbol{\alpha}} - \boldsymbol{m}$ and the point $\boldsymbol{0}$ in $\mathbb{R}^{r(\boldsymbol{K})}$, in the metric $\boldsymbol{K}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{K}$, is

$$Q = (\boldsymbol{K}^t\hat{\boldsymbol{\alpha}} - \boldsymbol{m})^t[\boldsymbol{K}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}(\boldsymbol{K}^t\hat{\boldsymbol{\alpha}} - \boldsymbol{m})\sigma^2 \ . \tag{3.60}$$

It can be proven that the random quantity Q has a $\chi^2_{r(\boldsymbol{K})}$ distribution, where $r(\boldsymbol{K})$ is the rank of the matrix $\boldsymbol{K}$ (see, e.g., [442]).

In practice, the parameter σ^2 is unknown and its estimate s^2 is used instead. The estimator s^2 has a $\chi^2_{n-p-1+r(\boldsymbol{K})}$ distribution. Hence, according to the definition of the F distribution, the quantity

$$F = \frac{Q/r(\boldsymbol{K})}{s^2} \tag{3.61}$$

has an F distribution with $r(\boldsymbol{K})$ and $n - p - 1 + r(\boldsymbol{K})$ degrees of freedom, and the null hypothesis H_0: $\boldsymbol{K}^t\boldsymbol{\alpha} = \boldsymbol{m}$ should be rejected at level α if

$$F > F_{r(\boldsymbol{K}),n-p-1+r(\boldsymbol{K});\alpha} \ , \tag{3.62}$$

see Fig. 3.5.

3.3 Model Selection and Validation

3.3.1 Motivation

The computer revolution has – to an almost dramatic extent – increased the practical possibilities to fit, in a semi-automatic or even fully automatic way, a large number of different models to a particular dataset. Apart from the advantages, which are obvious, this development seems to roost also a somewhat dangerous side-effect. The 'time-honoured' approach to fit, as far as possible, only well thought-off models that were considered to be physically plausible, or at least to be desirable to fit because of very specific reasons, is tending to fade away.

Despite the technicalities described in this section about model selection, it is hoped that the reader perceives that a sensible use of these increased facilities should not forestall the necessity to (re-)think about the (original) investigative purposes. Evidently, they also imply the need to understand the background, as well as the statistical consequences, of various variable–selection algorithms that are used by various statistical packages (see Chap. 6 for an overview). An implicit purpose of this section is to enable the reader

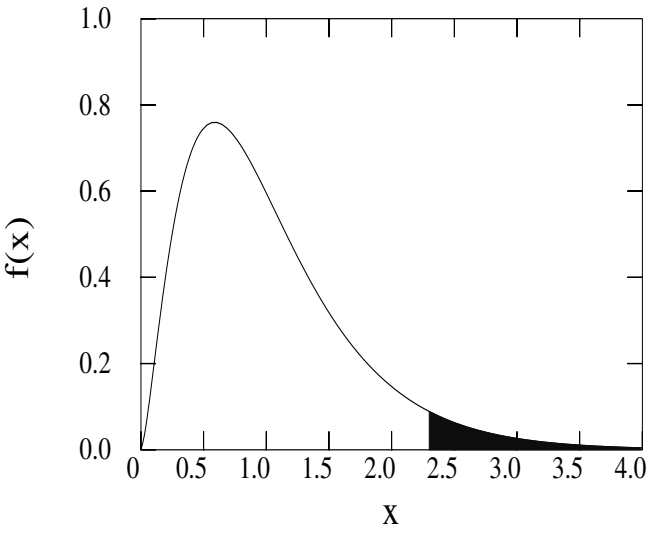

Fig. 3.5. The probability density $f(x)$ of the $\mathrm{F}_{5,\,100}$ distribution. The black area denotes a 5% rejection interval $(2.3, \infty)$ for the corresponding test statistic.

to exploit the new technology without unknowingly slipping into possibly perilous pitfalls, sometimes partly masked by an 'embarras de richesse'.

All the properties of the least-squares estimator discussed above are based on the assumption that the model is correct, i.e., that

(1) all important variables are taken into account;

(2) the linear model in the (transformed) dependent and independent variables (i.e., X, X^2, $1/X$, $\log X$, $\log Y$, ...) is correct;

(3) the errors are independent and (approximately) normally distributed.

In the so-called top-down approach we start with a rather large set of all candidate variables $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\}$, and select, according to certain criteria, a subset of relevant variables

$$\{\boldsymbol{x}_{s(1)}, \ldots, \boldsymbol{x}_{s(r)}\}\ ,\ \{s(1), \ldots, s(r)\} \subset \{1, \ldots, p\}\ ,\ s(1) < s(2) < \cdots < s(r)\ , \tag{3.63}$$

which is considered to be appropriate for the final regression equation.

Reasons for eliminating some of the candidate variables are:

(1) Regression equations containing less variables are in general simpler to handle, while many variables usually imply a higher computing expenditure. However, along with the performance of computer hardware and

software, the number of variables that can conveniently be handled has drastically increased.

(2^A) Since each variable is tainted with errors, the gain in information given by an additional variable can be smaller than the loss of certainty caused by the errors introduced by the new variable.

(2^B) A large number of variables is more apt to lead to ill conditioning and hence to increased uncertainty in the parameter estimates.

(3) Sometimes, one wants to compare a given regression model with theoretical considerations, which, for instance, predict that the response variable is not influenced by some of the regression variables.

On the other hand, eliminating an important variable (i.e., a variable that, when all other variables are held constant, has a non-negligible effect on the response variable) increases the bias in the remaining regression coefficients.

The emphasis in the model selection depends on the goal of the regression analysis. In general, one can distinguish between the following purposes:

(1) Estimation of the regression parameters $\alpha_0, \alpha_1, \ldots, \alpha_p$;

(2) Estimation of the variance of the random errors σ^2;

(3) Estimation of standard deviations of the estimated regression parameters;

(4) Testing of hypotheses about the regression parameters;

(5) Prediction of the response variable (interpolation and extrapolation);

(6) Outlier detection;

(7) Determination of the goodness-of-fit (which implies, in addition to estimating σ^2, checking the validity of the model).

Consider, for example, the case where one is interested in a precise estimation of the parameters. Then, the statistical goal is to obtain a small covariance matrix of $\hat{\boldsymbol{\alpha}}$ and, at the same time, a small bias in $\hat{\boldsymbol{\alpha}}$. Note that by using fewer variables, one gets a larger bias. On the other hand, by using more variables, one gets a larger variance and possibly a poor conditioning (collinearity).

In another situation, the purpose of the analysis can be to obtain a good prediction for the response variable. In that case, the goal is to achieve small prediction errors, i.e., a small variance and a small bias of $\hat{\boldsymbol{Y}}$.

In the second case one is willing to accept uncertainties in linear combinations $\boldsymbol{c}^t\hat{\boldsymbol{\alpha}}$ of the elements of $\hat{\boldsymbol{\alpha}}$ that are not important for the prediction of $\boldsymbol{Y}$. Such uncertainties could be unacceptable in the first case.

3.3.2 Selection of Variables

Because of non-orthogonality between regression variables, the value of a certain regression parameter depends on which variables have been included in the model, see the geometrical interpretation in Sect. 2.4.3.

In order to find the (nominally) best fitting subset of regression variables, theoretically the only way is to compare all possible subsets. In practice this concept is often not realistic, since there are $2^p - 1$ different, non-empty subsets of a given set of all candidate variables $\{x_1, \ldots, x_p\}$, which may be inconveniently large even in a computerised environment. Furthermore, the danger of overfitting the data is prominent in such an approach. Therefore, many computer algorithms have been designed to automate the process of finding relevant subset of variables without using the 'brute force' method of fitting all subsets.

Forward Selection

This algorithm starts with one variable. In each step, the variable that causes the largest decrease of the 'Residual Sum-of-Squares' $\big(SS(Res)\big)$ is added into the model. In the first step, this is the variable that has the highest correlation with the response variable, and in following steps it is the variable that has the highest partial correlation with the response variable, given all variables that are already included in the model.

The stopping rule is based on

$$SL = \frac{\textit{Reduction of } SS(Res) \textit{ caused by the new variable}}{s^2 \textit{ from the model including the new variable}} \,. \tag{3.64}$$

Utilising invariance under affine transformations, see [191,424], one can derive that

$$SL = \big(\frac{\hat{\alpha}_j}{s_{\hat{\alpha}_j}}\big)^2 \,, \tag{3.65}$$

and the stopping rule is equivalent to applying some t-test of significance about the new variable. The name SL stands for 'selection level'. The forward selection will be stopped if $SL < F_{\alpha_{enter}}$ for each new candidate variable, where $F_{\alpha_{enter}}$ is the critical value of the F-distribution, with appropriate degrees of freedom, corresponding to a right tail probability α_{enter}.

Backward Elimination

The algorithm starts with all candidate variables included in model. In each step, the variable whose elimination increases $SS(Res)$ least of all, is excluded from the model.

The stopping rule is based on

$$SL = \frac{\textit{Increase of } SS(Res) \textit{ caused by elimination of variable}}{s^2 \textit{ from the model including the variable}} \,. \tag{3.66}$$

Backward elimination will be stopped if $SL > F_{\alpha_{remove}}$ for each variable in the current model.

It should be noted that through addition or elimination of a variable the 'importance' of other variables can change. Therefore, in many practical cases one uses

Stepwise Selection

This algorithm combines forward selection with backward elimination.

Variables are added to the model until no variable passes the α_{enter} criterion any more. Subsequently, variables are removed until no variable passes the α_{remove} criterion. Then variables are again considered to be included, and so on, until none of the variables *in the model* passes α_{remove} and none of the variables *outside the model* passes α_{enter}.

Usually, one applies the algorithm with $F_{\alpha_{enter}} \geq F_{\alpha_{remove}}$ in order to avoid that soon in the process a variable that has just been added will be eliminated in the next step, which is a criterion to halt the step-wise procedure in order not to provoke an infinite loop. For instance, the default values in SAS / PROC REG, when the STEPWISE option is used, are $\alpha_{enter} = 0.15, \alpha_{remove} = 0.15$, see [578, 581].

Goodness–of–Fit Criteria in Subset Selection

Several criteria have been defined to compare the overall goodness–of–fit between various subsets. We consider five of these.

1. *Coefficient of Determination*

$$R^2 = \frac{SS(Regr)}{SS(Total)} \,, \tag{3.67}$$

which specifies the proportion of the total sum–of–squares which is 'explained' by the regression variables. By Pythagoras' theorem in $\mathbb{R}^n$, we have

$$SS(Total) = SS(Regr) + SS(Res) \,. \tag{3.68}$$

If a relevant variable is added to the model, then the value of R^2 increases.

2. *Residual Mean Square* s^2

$$s^2 = MS(Res) = \frac{SS(Total) - SS(Regr)}{n - (p+1)} \,, \tag{3.69}$$

which is an unbiased estimator of σ^2 if all relevant variables are included in the model. The expectation value of $MS(Res) = s^2$ is larger than σ^2 if a relevant variable is missing.

3. *Adjusted Coefficient of Determination*

$$R^2_{adj} = 1 - \frac{MS(Res)}{MS(Total)} = 1 - \frac{(1-R^2)(n-1)}{n-(p+1)} \,, \tag{3.70}$$

which uses, in contrast to R^2, the mean sum–of–squares and therefore is more suitable than R^2 to compare models with different number of parameters. If a relevant variable is added to model then the absolute value of R^2_{adj} increases.

4. *Mallow's C_p Statistic*

$$C_{p,k} = \frac{SS_k(Res)}{s^2} + M(k+1) - n \;, \tag{3.71}$$

where s^2 is an estimator of σ^2 from the model containing all p variables plus intercept, and M is a suitable constant. If a subset containing k variables is approximately correct, then both $SS_k(Res)/(n-k-1)$ and $s^2 = SS_p(Res)/(n-p-1)$ are reasonable estimators for σ^2, and

$$C_{p,k} \simeq (M-1)(k+1) \;. \tag{3.72}$$

If an important variable is missing, then $SS_k(Res)$ estimates a quantity that is somewhat larger than $(n-k-1)\sigma^2$ (which one can see from the geometrical interpretation), and $C_{p,k}$ is larger than $(M-1)(k+1)$. One minimises $C_{p,k}$ as a function of k. The larger the value of M, the smaller the number of variables selected.
The choice of which value of M should be used depends on which aspects of the regression one wants to optimise, as well as on the true (but unknown) values of the regression parameters and on whether or not *selection bias*, see Sect. 3.4.3, is effectively avoided. Simulation experiments presented in [463] (with a few hundred observations and up to 40 variables) suggest a practical choice of $M = 1.5 - 2.0$ if selection bias is avoided and $M = 2.0$ if selection bias is not avoided and (most of) the true regression parameters deviate about 1 standard deviation from their estimates. The evaluation criterion considered in [463] is the minimisation of the mean squared error of prediction (MSEP) over the region where the data are available. Special caution is needed if one wants to predict future observations outside this region.

5. *Akaike's Criterion*
Based on general considerations about information distances, in [4] it is proposed to minimise

$$-2\big(\log L_{\boldsymbol{x},k}(\boldsymbol{\theta}) - kf(n)\big) \tag{3.73}$$

as a function of $\boldsymbol{\theta}$, where $L_{\boldsymbol{x},k}(\boldsymbol{\theta})$ is the likelihood of a regression model with k parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, and where $f(n)$ is a slowly varying function of the sample size, e.g., $f(n) = \ln(n)$ or $\ln\ln(n)$.
Akaike's article (with an introduction by DeLeeuw) has been included in the first volume of [396]. One can directly derive (see, e.g., the appendix of [463]) that for $f(n) = 1$ and $n \gg k$, Akaike's criterion is roughly equivalent to Mallow's C_p with $M = 2$ and to $F_{\alpha_{enter}}$ close to 2.

A comparison of various selection criteria in the context of non-linear growth curve models can be found in a contribution of Otten et al. in [186].

3.3.3 Model Validation

Automatic procedures for selecting a regression model are directed to finding the best possible fit of the dataset at hand.

The next step toward a reliable model is the *model validation* procedure. This can be carried out by checking the regression equation with another, independent dataset. Validation is important, above all if one applies automated model selection, because of the danger of *overfitting* and *selection bias* on one hand and *omission bias* on the other hand.

Over-fitting occurs if the response variable can be described by, say, k regressors, but the data have been fitted with $p \gg k$ variables. Selection bias occurs if the response variable seems to be explainable by k regressors, but in fact these k regressors have been heavily selected from a considerably larger set of p regressors. Omission bias occurs if important regression variables are not included in the model. Ideally, both types of bias should be avoided.

At times it is impossible to obtain an independent dataset. If the available dataset is sufficiently large, then one can use the data for both estimation and validation: the dataset is divided into two or more representative parts. One of these subsets, usually called the training set, is used for estimation, while the other subset is used for validation. It is important to realise that data for the model validation have to come from the population for which one intends to make predictions.

Further reflection about sailing between these two dangers in regression analysis (as in ancient times between Scylla and Charybdis) can be found in the first cryptic issue discussed in [358].

3.4 Analysis under Non-Standard Assumptions

3.4.1 Behaviour of Mean Values

Model Assumptions

Assumptions about the *mean–value structure* of the model relate to:

(1) the form of the model (linear, quadratic, ...);

(2) which regression variables belong to the model.

Assumptions about the *error structure* of the model have to be made about:

(1) whether the errors have a (multivariate) distribution at all;

(2) whether the errors are independent (or not);

(3) whether the errors can be considered to have been generated by a specific (class of) probability distribution(s).

One must distinguish between the variation in the response variable described by the mean–value structure and the variation in the response variable described by the error structure. In *repeated measurements designs* this is easily possible. In *single measurement designs*, this is much more difficult, and strong additional restrictive assumptions about the errors (e.g., that they are normally or exponentially distributed, etc.) need to be made. This can be based on comparable experiments carried out before, or on theoretical considerations based on error propagation. When analysing experimental data it is important to distinguish between systematic measurement errors and 'random' measurement errors (see, e.g., [528]). Once detected, systematic measurements can be included in the mean–value structure and be corrected for. However, their detection usually implies considerable experimental effort.

Examples.

(1) (mean–value structure:) one knows from former experiments, that in a tokamak the plasma energy is approximately proportional to the plasma current;

(2) (error structure:) count rates in neutral particles detectors are well described by a Poisson distribution.

Quite often, many physical experiments are performed according to a single measurement design.

In Sect. 3.3 it was assumed: $\boldsymbol{E} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$, i.e., the errors are independent, identically distributed and additive. The least-squares technique, particularly suited under these assumptions, is often called the *Ordinary Least Squares* (OLS) technique. In practice, these assumptions are met only approximately, and often they are seriously violated.

Merits of a Linear Model

A linear model is the starting point for many analyses. Also, numerous nonlinear dependencies can be approximately described using a linear model.

Often, one tries to choose a 'simple model', where simplicity means:

(1) A model with as few parameters as possible is simpler than a model with many parameters.

(2) A linear model is simpler than a nonlinear one.

An important advantage of a linear model is that there exists an explicit expression for the regression parameters

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} \ ,$$

such that the numerical minimisation of the sum–of–squares (with problems such as choosing of an appropriate starting point, convergence evaluation, and the occurrence of secondary minima) is not required.

Examples of linear models are:

1. *univariate polynomial model*

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_p x^p + E \; ; \tag{3.74}$$

2. *multivariate polynomial model*

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_{11} x_1^2 + \alpha_{22} x_2^2 + \alpha_{12} x_1 x_2 + E \; ; \tag{3.75}$$

3. *Fourier series model*,

$$y = \alpha_0 + \alpha_1 \cos t + \cdots + \beta_1 \sin t + \cdots + \beta_p \sin pt \; . \tag{3.76}$$

Using specific transformations, some nonlinear models can be approximately converted to a linear model.

Collinearity

The situation when two or more independent variables are nearly linearly dependent is called *collinearity*. It can cause severe problems in *ordinary least squares regression*. Algebraically speaking, collinearity means that the design matrix X is almost singular. The fact that in case of collinearity the regression coefficients react very sensitively on small changes in the independent variables, is expressed by saying that the determination of the regression coefficients is *unstable*.

In case of experimental data, the sources of collinearity are:

(1) *an over-parameterised model*;

(2) *physical dependencies*; For instance, in plasma physics, for Ohmically heated discharges the relation $P_{heat} = U_{loop} I_p$ holds, where $P_{heat} \simeq P_L$ is the power which heats the plasma (see Sect. 2.4 and Exercise 7.2) and U_{loop} is the plasma loop voltage. Since U_{loop} is approximately constant, this leads to collinearity between P_{heat} and I_p.

(3) *too little variation in the regression variables*. This means collinearity with the unit vector $\mathbf{1}$.

Collinearity can be detected by

(1) inspection of results from the regression analysis: 'unreasonable' estimates for the regression coefficients, large standard errors, no significant coefficients for clearly important regressors, sensitivity to small disturbances in the regression variables; all this can occur, despite a 'good fit' to the data, as indicated by a small root mean squared error (RMSE);

(2) eigenanalysis of $\widetilde{\boldsymbol{X}}^t \widetilde{\boldsymbol{X}}$, where $\widetilde{\boldsymbol{X}} = \boldsymbol{X} - \overline{\boldsymbol{X}}$.

There does not exist a standard, i.e., for each case satisfactory, measure of collinearity. Some applicable countermeasures against deception in a collinear regression problem are:

(1) to add *new* or *more* data (with more variation in some directions);

(2) to apply *biased regression*, e.g., ridge regression or principal component regression.

Eigenanalysis of $\widetilde{\boldsymbol{X}}^t\widetilde{\boldsymbol{X}}$

The $(p\times p)$ matrix $\widetilde{\boldsymbol{X}}^t\widetilde{\boldsymbol{X}}$ is decomposed as a weighted sum of $(p\times p)$ matrices of rank 1:

$$\widetilde{\boldsymbol{X}}^t\widetilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^t = \sum_{j=1}^{p} \lambda_j^2 \boldsymbol{u}_j \boldsymbol{u}_j^t \,, \tag{3.77}$$

where $\boldsymbol{\Lambda} = Diag(\lambda_1^2, \ldots, \lambda_p^2)$ is a matrix of eigenvalues of $\widetilde{\boldsymbol{X}}^t\widetilde{\boldsymbol{X}}$, and $\boldsymbol{U} = (\boldsymbol{u}_1|\cdots|\boldsymbol{u}_p)$ is a matrix of eigenvectors of $\widetilde{\boldsymbol{X}}^t\widetilde{\boldsymbol{X}}$. The eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ are orthogonal, hence, they define a new orthogonal coordinate system in $\mathbb{R}^n$. Then, one defines:

$$\boldsymbol{W} = \widetilde{\boldsymbol{X}}\boldsymbol{U} \,. \tag{3.78}$$

The orthogonal columns of $\boldsymbol{W}$, which we denote by $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_p$, are called *principal components*. They are projections of the data on the principal axes of the ellipse

$$\boldsymbol{x}^t(\widetilde{\boldsymbol{X}}^t\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{x} = c, \quad c > 0 \,, \tag{3.79}$$

in $\mathbb{R}^p$, and are linear combinations of the original regression variables, see also Sect. 2.5.1. Principal components have some important properties:

(1) they are orthogonal;

(2) being linear combinations of the regression variables, they define the same subspace V_X as the regression variables do;

(3) $||\boldsymbol{w}_j||^2 = \langle \boldsymbol{w}_j; \boldsymbol{w}_j \rangle = \lambda_j^2$=Variance$(\boldsymbol{w}_j)$.

The total variance of the dataset is

$$S_D = \frac{1}{n-1} Tr(\widetilde{\boldsymbol{X}}^t\widetilde{\boldsymbol{X}}) = \sum_{j=1}^{p} \mathrm{var}\,(\boldsymbol{x}_1) + \cdots + \mathrm{var}\,(\boldsymbol{x}_p) = \sum_{j=1}^{p} \lambda_j^2 \,. \tag{3.80}$$

Usually, the eigenvalues are arranged in decreasing order: $\lambda_1^2 > \lambda_2^2 > \cdots > \lambda_p^2$. For detection of collinearity the smallest eigenvalues are most important.

A criterion for assessment of collinearity is: the dispersion of the data in direction $\boldsymbol{w}_j$ is sufficient if the error of the regression variables in that direction is at least 3–5 times smaller than $s_{\boldsymbol{w}_j}$. An applied discussion of collinearity in the context of confinement time scalings has been given (by the present author) in Sect. 3 of [110]. For further information, the reader is also referred to [47, 157, 358, 473], which describe general statistical aspects, and to [346, 349, 747], oriented to applications in plasma physics.

3.4.2 Behaviour of Variances

Non-normality of Error Distribution

If errors are not normally distributed, then the least-squares estimator is no longer a maximum likelihood estimator, and confidence regions and hypothesis tests based on OLS are biased, since the test statistics t, χ^2, F demand that the underlying random variable ist normally distributed.

Nevertheless, an OLS procedure is still practicable (1) if deviations from normality are not too large, or (2) if the sample size is large enough (because of the central limit theorem). The point estimates of the regression parameters are often less affected by non-normality than the confidence region.

Non-normality can be detected (1) by residual analysis and *normal probability plots*, and (2) by formal test statistics. Residual analysis is a simple, but powerful, method for detecting several inadequacies in the model. In particular, it can be used for detecting non-normality. Some aspects of non-normality can be detected by univariate test statistics, such as those based on sample analogues of the skewness and kurtosis, see below and Sect. 1.4, Chap. 2.2 of [607], and Sect. 27.7 of [131]. A practical book on normality testing is [674].

An applicable countermeasure against non-normality is a transformation of the response variable.

Theoretical and empirical (estimated) residuals

The *theoretical residuals* $\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}$ correspond to random errors in the response variable. Under standard assumptions, theoretical residuals are independent and have equal variances. If they are, in addition, normally distributed, then we write:

$$\boldsymbol{E} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2) \; . \tag{3.81}$$

The *empirical residuals* are the differences between observed and predicted values of the response variable:

$$\hat{e}_i = y_i - \hat{y}_i \quad (i = 1 \ldots, n) \; . \tag{3.82}$$

Even if $\boldsymbol{E} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$, the empirical residuals are not independent and do not have equal variances. In fact,

$$\hat{\boldsymbol{E}} \sim N(\boldsymbol{0}, (\boldsymbol{I} - \boldsymbol{P})\sigma^2) \; . \tag{3.83}$$

The off-diagonal elements in the variance matrix $(\boldsymbol{I} - \boldsymbol{P})\sigma^2$ are not zero and are not equal. The difference between theoretical and estimated residuals is particularly large in case of overfitting.

It is often more informative to look at estimated residuals which are normalised in some way, such as

$$\hat{r}_i = \frac{\hat{e}_i}{s\sqrt{1 - v_{ii}}} \; , \tag{3.84}$$

which are called the *standardised residuals*. As usual, the RMSE, s, estimates σ, and v_{ii} is the ith diagonal element of $\boldsymbol{P}$, such that $s_{\hat{e}_i} = s\sqrt{1 - v_{ii}}$ is estimated standard deviation of $\hat{e}_i$, and all standardised residuals have unit variance.

In order to reduce the dependence between estimated residuals, one sometimes also uses *studentised residuals*

$$\hat{r}_i^\star = \frac{\hat{e}_i}{s_{(i)}\sqrt{1 - v_{ii}}} \,, \tag{3.85}$$

where $s_{(i)}$ denotes the RMSE computed without observation Y_i.
Studentised residuals have a t_{n-p-2} distribution, but they remain somewhat dependent.

Normal Probability Plot
We check whether the residuals come from a normal distribution $N(\mu, \sigma^2)$ by plotting the ordered residuals $\hat{e}_{(1)} \leq \hat{e}_{(2)} \leq \cdots \leq \hat{e}_{(n)}$ against $\Phi^{-1}(p_1) \leq \cdots \leq \Phi^{-1}(p_n)$, where

$$p_i = \frac{i - 3/8}{n + 1/4} \,. \tag{3.86}$$

The quantiles of a standard normal distribution $N(0, 1)$ are

$$Q(p_i) = u_{(i)} = \Phi^{-1}(p_i) \,. \tag{3.87}$$

If the residuals have a $N(\mu, \sigma^2)$ distribution, then

$$E\,\hat{e}_{(i)} = \mu + \sigma u_{(i)} \,, \tag{3.88}$$

so the ranked expected residuals should fall on a straight line.

More generally, the procedure can be used to test whether residuals stem from a specific distribution or from a class of distributions with unspecified location parameter μ and scale parameter σ.

Sample skewness and kurtosis
The sample skewness

$$\hat{\gamma}_1 = \frac{m_3}{m_2^{3/2}} \,, \tag{3.89}$$

where m_2 and m_3 denote empirical moments around the mean, is a measure of the asymmetry of the distribution. For a normal distribution, the sample skewness has expectation 0 and variance $\frac{6}{n}\big(1+O(\frac{1}{n})\big)$. Hence, for large sample sizes, the assumption of a normal distribution is rejected at the 5% level if $|\hat{\gamma}_1| > 2\sqrt{\frac{6}{n}}$.

In simple terms, the sample excess of kurtosis

$$\hat{\gamma}_2 = \frac{m_4}{m_2^2} - 3 \,, \tag{3.90}$$

measures the tendency of the distribution to be flat or peaked, see also Sect. 1.4. For a normal distribution, the sample excess of kurtosis has expectation 0 and variance $\frac{24}{n}\big(1 + O(\frac{1}{n})\big)$. Hence, for large sample sizes, the assumption of a normal distribution is rejected at the 5% level if $|\hat{\gamma}_2| > 2\sqrt{\frac{24}{n}}$.

In [132] the empirical distributions of the sample skewness and kurtosis for sample sizes between 20 and 200 are discussed. The asymptotic approximations are discussed in [131].

Shapiro–Francia test

In [612] a test function is developed, which is directly connected with a normal probability plot. Let $\boldsymbol{u} = (u_{(1)}, \ldots, u_{(n)})$ be the vector of an order statistic from a standard normal distribution. Then

$$W' = \frac{(\hat{\boldsymbol{e}}^t \boldsymbol{u})^2}{(\hat{\boldsymbol{e}}^t \hat{\boldsymbol{e}})(\boldsymbol{u}^t \boldsymbol{u})} \,, \tag{3.91}$$

which is the square of the correlation coefficients between $\hat{\boldsymbol{e}}$ and $\boldsymbol{u}$, is a suitable test statistic. The hypothesis of normality is rejected, if W' is small. The distribution of W' under the hypothesis of normality has been tabulated (see, e.g., [543]). The Shapiro–Francia method works reliably for a moderately large number of observations, say at least 30.

Heterogeneous Variances

In case of heteroskedastic errors, i.e., if the errors have unequal variances, the (unweighted) LS estimator does not have anymore minimal variance under all unbiased estimators, and the inaccuracies in the estimated regression parameters are wrongly estimated by ordinary least squares (OLS). Heterogeneous variances can be detected (1) by making residual plots, and (2) by applying statistical tests (see, e.g., [120,148,466]). Applicable countermeasures against heterogeneous variances are (1) applying variable transformations, and (2) *weighted regression.*

Correlated Errors

In case of correlated errors, an OLS estimator does not have minimal variance under all unbiased estimators, and variances of parameters based on OLS are not correctly estimated. Therefore, hypothesis tests and estimated confidence regions based on OLS are invalid. A positive correlation implies that the estimated variance of the theoretical residuals is too small. Hence, the confidence regions are too small, and null hypotheses, that hold in reality, are rejected with probability larger than the nominal level α (e.g., 5%).

A negative correlation implies that the estimated variance of the theoretical residuals is too large and hence the corresponding confidence regions are too large.

Correlated errors, which occur typically in time series analysis, can be detected by:

(1) residual plots of $\hat{e}_i$ against $\hat{e}_{i-1}$, or of $\hat{e}_i$ against $\hat{e}_{i-k}, k = 1, 2, 3, \ldots$;

(2) a test statistic (sample analogue) for the autocorrelation

$$R = \frac{\sum_{i=2}^{n} (\hat{e}_i \, \hat{e}_{i-1})}{\sum \hat{e}_i^2} \; ; \tag{3.92}$$

(3) Durbin–Watson's test statistic

$$D = \frac{\sum_{i=2}^{n} (\hat{e}_i - \hat{e}_{i-1})^2}{\sum \hat{e}_i^2} \; . \tag{3.93}$$

Remark. The value of D is always positive. To a good approximation, $D = 2 - 2R$. The critical values of D are tabulated (see, e.g., [543]). The null hypothesis '$\rho = 0$' is rejected, if ρ is sufficiently far from 0, or, equivalently, if d is sufficiently far from 2.

Applicable countermeasures against correlated errors are:

(1) using a procedure for fitting time–series models (for instance, an autoregressive integrated moving average (ARIMA) model);

(2) applying variable transformations, and, if the structure of the covariance matrix is known;

(3) applying weighted regression.

Contaminated Data

An *outlier* is an observation that, compared with other observations, strongly deviates from the postulated model. This does not mean that an outlier must strongly deviate from the fitted regression plane. In other words, an outlier is not the same as an *observation with a large empirical residual*. A data point can be an outlier, but still have an inconspicuous residual.

An *influential observation* is an observation, that, in comparison to other observations, has a strong influence on the estimation of the regression parameters. This influence can be defined in two ways: while removing the observation from the dataset, or while shifting this observation in X-space or in the Y direction. In OLS regression, each observation has the same weight, but does not have the same influence on the regression. An observation that is far away from other datapoints, in one or more X directions, is generally influential.

An influential outlier with a small empirical residual is to be called a *malignant outlier*. Handling of malignant outliers can sometimes be based

on procedures that are based on *robust regression*. For an introduction, see, e.g., [264].

A few single outliers can be detected by graphical residual analysis, for smaller datasets to be based on robust regression, see Sect. 3.5.3, and their influence can be assessed by comparing regression results with and without such groups of observations. It must be remarked, however, that in multivariate problems, outliers can considerably be masked (see, e.g., [149]).

In principle, all aspects of a fit can be influenced by an observation y_i. Some of the important aspects, which we shall briefly discuss now are: the vector $\hat{\boldsymbol{\alpha}}$ as a whole, single parameters $\hat{\boldsymbol{\alpha}}_j$, a fitted value $\hat{y}_k$, the estimated covariance matrix.

The influence of the observation y_i on the regression parameter $\boldsymbol{\alpha}$ can be expressed by *Cook's D* [120]:

$$D_i = \frac{(\hat{\boldsymbol{\alpha}}_{(i)} - \hat{\boldsymbol{\alpha}})^t (\boldsymbol{X}^t \boldsymbol{X})(\hat{\boldsymbol{\alpha}}_{(i)} - \hat{\boldsymbol{\alpha}})}{(p+1)s^2} = \frac{\hat{r}_i^2}{(p+1)} \Big(\frac{v_{ii}}{1 - v_{ii}} \Big) \,, \tag{3.94}$$

where $\hat{\boldsymbol{\alpha}}_{(i)}$ is an estimate of $\boldsymbol{\alpha}$, based on a regression fit without the observation y_i.

In other words, D_i describes the shift in $\hat{\boldsymbol{\alpha}}$ provoked by removing the observation y_i. From (3.94) one can see that the value of D_i is large if the standardised residual $\hat{r}_i$ is large, and also if the datapoint is far from the centre of the data ellipsoid. Therefore, Cook's D measures a mixture of both outlier and influence properties.

It is noted that one does not need to repeat the whole regression in order to estimate $\boldsymbol{\alpha}_{(i)}$, because

$$\hat{\boldsymbol{\alpha}}_{(i)} = \hat{\boldsymbol{\alpha}} - \frac{(\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_i \hat{e}_i}{1 - p_{ii}} \,, \tag{3.95}$$

where $p_{ii} = (\boldsymbol{P})_{ii}$ with the projection matrix $\boldsymbol{P}$ defined by (3.18).

The influence of the observation y_i on the precision of the parameter estimation can be expressed by the covariance ratio

$$C_V = \frac{\det(s^2_{(i)}[(\boldsymbol{X}^t_{(i)} \boldsymbol{X}_{(i)})^{-1}])}{\det(s^2[(\boldsymbol{X}^t \boldsymbol{X})^{-1}])} \,, \tag{3.96}$$

which is the ratio of the determinants of the covariance matrices without and with the ith observation. A value $C_V \simeq 1$ implies that the observation y_i hardly influences the precision, a value $C_V > 1$ means that y_i increases the precision, and $C_V < 1$ says that the precision is decreased by the observation y_i.

3.4.3 Applicable Techniques

Transformation of Variables

There are various reasons to consider making transformations of the scale on which the original data have been measured: simplification of the model,

making the error variances in the response variable more homogeneous, and improving the normality of the error distribution. One can transform the regressors and/or the response variable.

Here are some examples of transformations:

(1) *Logistic model* (when β is known) with multiplicative errors:

$$Y = \frac{\beta}{1+\gamma e^{-\boldsymbol{\alpha}^t \boldsymbol{x}}} E \qquad \to \ln(\frac{\beta}{Y} - 1) \qquad \Rightarrow \quad Y^\star = \ln\gamma - \boldsymbol{\alpha}^t \boldsymbol{x} + \ln E \,, \tag{3.97}$$

(2) *Exponential model* with multiplicative errors:

$$Y = \gamma e^{\boldsymbol{\alpha}^t \boldsymbol{x}} E \qquad \to \ln \qquad \Rightarrow \quad Y^\star = \ln\gamma + \boldsymbol{\alpha}^t \boldsymbol{x} + \ln E \,, \tag{3.98}$$

(3) *Inverse polynomial model*:

$$Y = \frac{x}{\beta + \gamma x + E} \qquad \to \frac{1}{Y} \qquad \Rightarrow \quad Y^\star = \gamma + \frac{\beta}{x} + \frac{1}{x} E \,, \tag{3.99}$$

(4) *Multiplicative model* (*'power–law'*):

$$Y = e^{\alpha_0} x_1^{\alpha_1} \cdots x_p^{\alpha_p} E \quad \to \ln \quad \Rightarrow \quad Y^\star = \alpha_0 + \sum_{j=1}^{p} \alpha_j \ln(x_j) + \ln E \,. \tag{3.100}$$

Consider the last model. To meet the standard regression assumptions, errors should be additive after transformation. However, they do not need to be exactly multiplicative in the original model. For example, if they are originally additive, then it is sufficient that they are small in comparison with $E(Y)$, because for $u \ll 1$,

$$\log(1+u) \simeq u \qquad \big(u \equiv e_i / E(y_i)\big) \,. \tag{3.101}$$

This situation occurs, for example, for log-linear scalings of the the thermal energy of a plasma.

In [79], Box and Cox propounded a family of transformations with which one can pursue all above mentioned purposes simultaneously. The *Box–Cox transformation family* is defined as

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda (GM(\boldsymbol{y}))^{\lambda-1}} & \text{for} \quad \lambda \neq 0 \,, \\ GM(\boldsymbol{y}) \ln(y_i) & \text{for} \quad \lambda = 0 \,, \end{cases} \tag{3.102}$$

where $GM(\boldsymbol{y})$ is the geometric mean of the observations $y_1, \ldots, y_n$. It serves to make the RMSEs comparable for different values of λ (the RMSEs for different λ have different units!).

By carrying out ordinary least-squares (OLS) regression for the model

$$\boldsymbol{Y}^{(\lambda)} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{E} \,, \qquad \mathbf{var}\,(\boldsymbol{E}) = \boldsymbol{I}\sigma^2 \,, \tag{3.103}$$

say, for $\lambda = -2, -1.5, \ldots, 2$, one gets the estimated regression parameters

$$\hat{\boldsymbol{\alpha}}_{\lambda} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y}^{(\lambda)} \,, \tag{3.104}$$

and the residual sum–of–squares

$$SS_{(\lambda)}(Res) = (\boldsymbol{Y}^{(\lambda)})^t(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}^{(\lambda)} \,. \tag{3.105}$$

It can be shown, see [120], that these estimated residual sums of squares are approximately equally scaled. In general, it is reasonable to use values of λ, for which $SS_{(\lambda)}(Res)$ is not too far from its minimum over λ. Large sample theory on linear regression models with a Box–Cox transformed response variable is discussed in [106], see also [572] for an overview until the last decade of the previous century.

Weighted Regression

The OLS method assumes that errors in the response variable are independent and normally distributed with unknown, homoskedastic variance σ^2, i.e.,

$$\mathbf{var}\,(\boldsymbol{E}) = \boldsymbol{I}\sigma^2 \,. \tag{3.106}$$

Consider the model

$$\boldsymbol{E} \sim N(\boldsymbol{0}, \boldsymbol{W}\sigma^2) \,, \quad \mathbf{var}\,(\boldsymbol{E}) = \boldsymbol{W}\sigma^2 \neq \boldsymbol{I}\sigma^2 \,. \tag{3.107}$$

We will now derive formulae for parameter estimation in this situation, to which we refer as the *generally weighted regression model*. Suppose, we have a positive definite covariance matrix $\boldsymbol{W}$. Then, there exists a unique $n \times n$ matrix $\boldsymbol{C}$, which is a square root of $\boldsymbol{W}$.
Properties of $\boldsymbol{C}$ are:

$$\boldsymbol{C}^t\boldsymbol{C} = \boldsymbol{C}\boldsymbol{C}^t = \boldsymbol{C}^2 = \boldsymbol{W}^{-1} \,, \tag{3.108}$$

$$\boldsymbol{C}^{-1}(\boldsymbol{C}^{-1})^t = \boldsymbol{W} \,. \tag{3.109}$$

Starting with the model

$$\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\alpha}} + \boldsymbol{E} \,, \tag{3.110}$$

and pre-multiplying by $\boldsymbol{C}$:

$$\boldsymbol{C}\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{C}\boldsymbol{E} \,, \tag{3.111}$$

we can write

$$\boldsymbol{Y}^{\star} = \boldsymbol{X}^{\star}\boldsymbol{\alpha} + \boldsymbol{E}^{\star} \,. \tag{3.112}$$

Notice that

$$\begin{aligned}\mathbf{var}\,(\boldsymbol{CE}) &= \boldsymbol{C}(\boldsymbol{W}\sigma^2)\boldsymbol{C}^t \\ &= \boldsymbol{C}[\boldsymbol{C}^{-1}(\boldsymbol{C}^{-1})^t)]\boldsymbol{C}^t\sigma^2 \\ &= \boldsymbol{C}\boldsymbol{C}^{-1}(\boldsymbol{C}^{-1})^t\boldsymbol{C}^t\sigma^2 \\ &= \boldsymbol{I}\sigma^2\ . \end{aligned} \tag{3.113}$$

Hence, the OLS assumptions are met for (3.112), so one can carry out standard regression for this model.

On the original scale, we get the following results for the *generalised least squares* (GLS) estimate:

$$\hat{\boldsymbol{\alpha}}_{GLS} = (\boldsymbol{X}^t\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{W}^{-1}\boldsymbol{Y}\ ; \tag{3.114}$$

$$\mathrm{E}(\hat{\boldsymbol{\alpha}}_{GLS}) = \boldsymbol{\alpha}\ ; \tag{3.115}$$

$$\mathbf{var}\,(\hat{\boldsymbol{\alpha}}_{GLS}) = (\boldsymbol{X}^t\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\sigma^2\ . \tag{3.116}$$

In practice, the main problem in weighted regression lies in estimating accurately the elements of the matrix $\boldsymbol{C}$ or, equivalently, $\boldsymbol{W}$. In case of repeated measurements one can estimate the weights from the variations of repeated measurements around their mean values. In other situations one has to use an iterative method, for instance in the following way:

a) one starts with, e.g., $\widetilde{\boldsymbol{W}} = \boldsymbol{I}$;

b) regression is carried out with $\boldsymbol{W} = \widetilde{\boldsymbol{W}}$;

c) one gets a new $\widetilde{\boldsymbol{W}}$ using residuals from step b;

d) the steps b and c are iterated until the estimate of $\boldsymbol{W}$ does not change any more.

The special case of *simply weighted regression* is obtained, if $\boldsymbol{W}$ is diagonal, $W_{ij} = \sigma^2 W_{ij} = \sigma^2 w_j^2 \delta_{ij}$, i.e., if the errors are independent.

Ridge Regression

Ridge regression is a *biased regression method*, of which we outline the main idea. More information is given in [543] and [727].

The *mean squared error* (*MSE*) of an estimator is defined as (see Sect. 2.2):

$$\mathrm{MSE}_\theta(\hat{\theta}) = E_\theta(\hat{\theta} - \theta)^2 = \mathrm{var}_\theta(\hat{\theta}) + \mathrm{bias}_\theta^2(\hat{\theta})\ , \tag{3.117}$$

where

$$\mathrm{var}_\theta(\hat{\theta}) = E_\theta[\hat{\theta} - E_\theta(\hat{\theta})]^2\ , \tag{3.118}$$

$$\mathrm{bias}_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta\ . \tag{3.119}$$

If $\tilde{\theta}$ is unbiased (i.e., $E_\theta(\tilde{\theta}) = \theta$) then

$$\mathrm{MSE}_{\theta}(\hat{\theta}) = \mathrm{var}_{\theta}(\hat{\theta}) \; . \tag{3.120}$$

The method of least squares leads to an unbiased estimator with a variance that is minimal among the variances of all unbiased estimators, see Sect. 2.4.3. However, in case of collinearity this minimum can be unacceptably large. By giving up the demand for unbiasedness, one enlarges the class of possible estimators, and there is a chance to find an estimator with a smaller variance than the variance of the LS estimator, for each (reasonable) value of θ. A simple sketch of the situation is shown in Fig. 3.6.

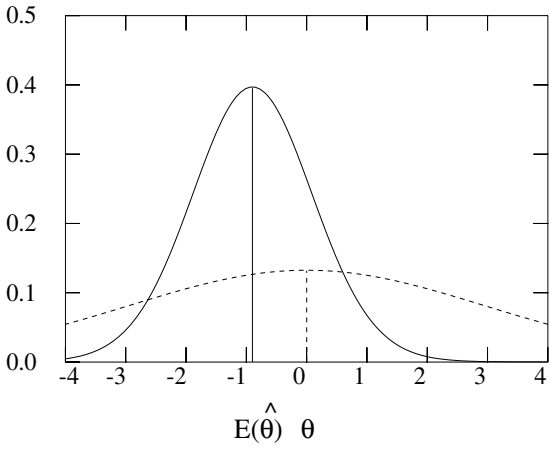

Fig. 3.6. Illustrating the concept of a biased estimator. The dotted curve shows the probability distribution of an unbiased estimator for θ, which has a larger variance than the biased estimator with expectation value $E(\hat{\theta})$ different from θ.

In ridge regression one usually works with a centered and scaled matrix of independent variables, $\boldsymbol{Z}$. Consequently, in the model

$$\boldsymbol{Y} = \mathbf{1}\alpha_0 + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{E} \; , \tag{3.121}$$

α_0 can be estimated independently from the other α's, because $\mathbf{1}$ is orthogonal to $\boldsymbol{Z}$.

A ridge regression estimator for $\alpha_1, \ldots, \alpha_p$ is

$$\hat{\boldsymbol{\alpha}}_{R} = (\boldsymbol{Z}^t\boldsymbol{Z} + k\boldsymbol{I})^{-1}\boldsymbol{Z}^t\boldsymbol{Y} \; , \tag{3.122}$$

with

$$\mathbf{var}\,(\hat{\boldsymbol{\alpha}}_{R}) = (\boldsymbol{Z}^t\boldsymbol{Z} + k\boldsymbol{I})^{-1}(\boldsymbol{Z}^t\boldsymbol{Z})(\boldsymbol{Z}^t\boldsymbol{Z} + k\boldsymbol{I})^{-1}\sigma^2 \ . \tag{3.123}$$

By expansion of the matrix inversions, one gets, equivalently,

$$\mathbf{var}\,(\hat{\boldsymbol{\alpha}}_{R}) = \boldsymbol{Z}^t(\boldsymbol{Z}\boldsymbol{Z}^t)^{-1}(k\boldsymbol{Z}\boldsymbol{Z}^t + \boldsymbol{I})^{-2}(\boldsymbol{Z}\boldsymbol{Z}^t)^{-1}\boldsymbol{Z}\sigma^2 \ . \tag{3.124}$$

In practice, it is often sensible to consider replacing $k\boldsymbol{I}$ by $k\boldsymbol{Z}_{(1)}^t\boldsymbol{Z}_{(1)}$, where, $\boldsymbol{Y} = \mathbf{1}\alpha_0 + \boldsymbol{Z}_{(1)}\boldsymbol{\alpha} + \boldsymbol{E}$ is suspected to be a reasonable model for the data, based on prior theoretical or experimental evidences.

A Bayesian interpretation (see, e.g., [727]), is that $\hat{\boldsymbol{\alpha}}_R$ can be viewed as the posterior estimate of $\boldsymbol{\alpha}$ corresponding to the prior estimate $\boldsymbol{\alpha} \sim N(0, k^{-1}\boldsymbol{I})$.

The ridge regression is carried out for different values of k small with respect to the diagonal of $\boldsymbol{Z}$, e.g., 0.005, 0.01, ..., 0.02. The goal is to adopt a value of k which (approximately) minimises the mean squared error. Usually, the behaviour of the regression coefficients as a function of k is plotted. One chooses a value of k for which the regression parameters begin to stabilise. The bias increases monotonically with k, but the variance decreases. Initially, the MSE decreases, and after some steps begins to increase, albeit the value for k where this occurs is in practical situations not known.

Robust Regression

Estimations based on minimising the sum of squared deviations are sensitive to 'data contamination', because the residuals are squared. *Robust methods* are less sensitive to outliers, and are, hence, also more suitable to detect outliers.

A very simple example of a robust estimator is the sample median, which is less sensitive to outliers than the arithmetic sample mean. A more general class of robust estimators is formed by so-called *M–estimators*, which minimise the sum of a certain function of the residuals. (M stands for 'Maximum–likelihood–like'.)

For estimating the mean value of a distribution, the quantity

$$\sum_{i=1}^{n} \rho\big((y_i - \mu)/\sigma\big) \ , \tag{3.125}$$

and for linear regression

$$\sum_{i=1}^{n} \rho\big((y_i - \boldsymbol{x}_i^t\boldsymbol{\alpha})/\sigma\big) \tag{3.126}$$

is minimised. Here, the quantity σ is a scale factor.

A special class of M–estimators uses

$$\rho(z) = |z|^p \ . \tag{3.127}$$

In the very special case of $\rho(z) = z^2$, with any constant σ, one gets the OLS estimator $\hat{\boldsymbol{\alpha}}_{LS}$, which minimises $\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^t\boldsymbol{\alpha})^2$.

For $1 \leq p < 2$ one gets estimators, which are more robust than the corresponding OLS estimators. The special case, that corresponds to $p = 1$, is sometimes called the *minimum absolute deviation* (*MAD*) *estimator*.

Choosing

$$\rho(z) = \begin{cases} z^2/2 & (|z| \leq c) \ , \\ c|z| - c^2/2 & (|z| > c) \ , \end{cases} \tag{3.128}$$

where c is a constant, one gets another class of robust estimators. By using a proper value of c, say 1 or 2, the large residuals get less weight, and, hence, the estimation is not so much dominated by outliers as it is in the case of OLS.

Many robust methods are now routinely implemented in statistical software packages such SAS and S–PLUS, see Sects. 6.2 and 6.3. For regression problems with many variables they can be quite computer intensive. For such problems, a fast practical compromise is given by the TRim and DElete method propounded in [149]. Further references for robust methods are [292, 334, 421, 567]. They have a wide range of applicability. For instance, in a plasma-physical context to the analysis of edge temperature profiles, where data can be heavily tainted with outliers from the Thomson scattering diagnostic, see [155] for a clear discussion based on a Bayesian approach.

Scientifically, robust analysis should be considered as a preliminary stage toward a more complete analysis of a corrected dataset.

3.4.4 Inverse Regression

Inverse regression occurs if one wants to estimate the value(s) of x for which the response variable y assumes a fixed value y_0. This is also called the *calibration problem*, and is a special case of the errors–in–variable model combined with a linear restriction on variables, see Sect. 2.4.4.

We consider here only the very simple model

$$Y = \alpha_0 + \alpha_1 X + E \ . \tag{3.129}$$

It is noted that if $(\hat{\alpha}_0, \hat{\alpha}_1)$ is an unbiased least-squares estimator for (α_0, α_1), then, unfortunately,

$$\hat{x}_0 = \frac{y_0 - \hat{\alpha}_0}{\hat{\alpha}_1} \ . \tag{3.130}$$

is not an unbiased estimator of

$$x_0 = \frac{y_0 - \alpha_0}{\alpha_1} \ . \tag{3.131}$$

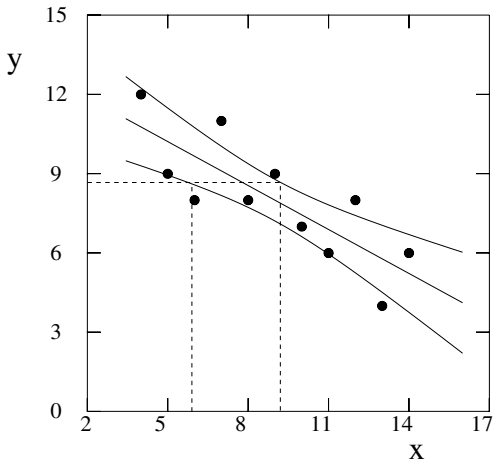

Fig. 3.7. Illustrating the concept of inverse regression based on the same (artificial) data as in Fig. 3.3. Inversion of the confidence band at a fixed position y_0 gives a 95% confidence region for x_0. However, inversion of the regression line at y_0 does *not* give an unbiased estimate of x_0.

If the variation coefficient of $\hat{\alpha}_1$, $SD(\hat{\alpha}_1)/\alpha_1$, is small and y_0 is close to $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, then the bias is not large. In other cases it can be considerable. A first order bias-adjusted estimator for x_0 is

$$\hat{\hat{x}}_0 = \bar{x} + \frac{y_0 - \bar{y}}{\hat{\alpha}_1 + \frac{SS_x^{-1} SS_y}{n\hat{\alpha}_1}} \,, \tag{3.132}$$

where $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$, and $(\bar{x}_0, \bar{y}_0)$ is the center of gravity of the data. This bias correction is reasonably accurate as long as $n\hat{\alpha}_1^2 \gg SS_x^{-1} SS_y$, or equivalently, $R^2 \gg 1/n$, where $R = SS_{xy}/\sqrt{SS_x SS_y}$ is the empirical correlation coefficient. The distributions of $\hat{x}_0$ and $\hat{\hat{x}}_0$ can be approximated by applying asymptotic error propagation, see Sect. 1.8, and/or by simulation.

For confidence bands the situation is fundamentally different from that for the (point-estimated) regression lines. A 95% confidence region for x_0 is

given by $\{x|H_0\text{: } y(x) = y_0 \text{ is not rejected}\}$, i.e., by inverting the confidence band at a (fixed) value y_0, see Figs. 3.3 and 3.7. Note that this region can consist of one or two intervals containing infinity or even the whole real line.

For further information, the reader is referred to Chap. 3 of [465] and Chap. 5 of [466].

3.5 Generalisations of Linear Regression

3.5.1 Motivation

So far we discussed the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{E}\,, \tag{3.133}$$

which is characterised by a linear dependence on the unknown parameters and by additive, normally (or symmetrically) distributed errors with constant variance. As previously observed, see Sect. 2.4, the linearity of the regression model does not necessarily mean that the response variable has to depend linearly on the predictor variables. For instance, polynomial models are linear regression models. In the standard regression context where the regression variables are considered without error, linearity of the model means linearity in the parameters, i.e.,

(1) each (additive) term contains only one parameter;
(2) each parameter appears only as a multiplicative constant (on the independent variable).

We will discuss here two generalisations of the linear regression model and just touch upon some of the other extensions.

(1) *Generalised linear model*:
The expected value of the response variable is not a linear combination of the independent variables, but a function of this linear combination. The restrictions of normality (or symmetry) and additivity of the errors are dropped.
(2) *Nonlinear regression*:
The restriction of linearity in the parameters is dropped. The errors are considered to be normally distributed and additive on some scale.

3.5.2 Generalised Linear Models

In a linear model the density of the response variable can be written as

$$f_Y(y) = f_E(y - \boldsymbol{X}\boldsymbol{\alpha})\,. \tag{3.134}$$

In a generalised linear model (GLM), originally investigated in [488], the density of the response variable is written as

$$f_Y(y) = f(y, \boldsymbol{X}\boldsymbol{\alpha}) \; . \tag{3.135}$$

The expectation of the response variable Y depends on the regression variables only through the quantity $\boldsymbol{X}\boldsymbol{\alpha}$, but the deviations are not necessarily additive. The other generalisation is that the errors do not need to be normally distributed, but are assumed to belong to an exponential family of distributions, see Sect. 1.5.3. As a consequence, the response variable can be limited to its naturally restricted range, while in the classical linear model the range of the response variable is, in general, not restricted. For example, by using a generalised linear model, one is able to fit positive or even binomially distributed response variables.

The enhanced flexibility of generalised linear models is a clear benefit. On the other hand, it can also lead to results that are less robust, and the interpretation of which is more complicated, than for simple regression.

Mean Value and Error Structure

We consider the following class of statistical distributions, which is usually called the class of generalised linear models:

$$f_{\theta,\phi}(y) = \exp\Big(\frac{y\theta - c(\theta)}{\phi} + h(y,\phi)\Big) \; , \tag{3.136}$$

and, for multidimensional $\boldsymbol{y}$ and $\boldsymbol{\theta}$,

$$f_{\boldsymbol{\theta},\phi}(\boldsymbol{y}) = \exp\Big(\frac{\boldsymbol{y}^t\boldsymbol{\theta} - c(\boldsymbol{\theta})}{\phi} + h(\boldsymbol{y},\phi)\Big) \; . \tag{3.137}$$

The functions $c(\cdot)$ and $h(\cdot)$ are assumed to be known. If ϕ is a known constant, we have an *exponential family*.

A generalised linear model for a random variable Y is characterised by the following properties:

(1) the probability distribution of each Y_i, $f_{Y_i}(y) = f(y, \eta_i)$, belongs to an exponential family;

(2) linear combinations of the explanatory variables (regressors) are utilised as a linear predictor η, i.e., for observation i we have

$$\eta_i = \sum_{r=1}^{p} x_{ir}\alpha_r \; ; \tag{3.138}$$

(3) the expected value of each observation can be expressed by a known, invertible and differentiable, function of its linear predictor

$$E(Y_i) = \mu_i = g^{-1}(\eta_i) \; , \tag{3.139}$$

where (by historical convention) g^{-1} is called the *inverse link* function.

For a *canonical link* function g, the linear predictor $\eta = g(\mu)$ corresponds to the parameter θ of the exponential family. Hence, in that case, we have for the expectation value

$$\mu_i = \frac{dc}{d\eta_i} = g^{-1}(\eta_i) \;, \tag{3.140}$$

and for the variance function

$$\operatorname{var} Y_i = V(\mu_i) = \frac{d^2c}{d\eta_i^2} = \frac{d}{d\eta_i}\big(\mu_i(\eta_i)\big) = \frac{1}{d\eta_i/d\mu_i} = \frac{1}{g'(\mu_i)} \;. \tag{3.141}$$

A few commonly used families of error distributions and their corresponding canonical link functions are: normal: $g(\mu) = \mu$; Poisson: $g(\mu) = \log \mu$; Gamma: $g(\mu) = 1/\mu$; binomial: $g(\mu) = \log[(\mu/(1-\mu)]$; inverse Gaussian: $g(\mu) = 1/\mu^2$. In general, an arbitrary link function can be combined with an arbitrary variance function.

Remark. A generalised linear model should not be confounded with a linear model obtained by transforming the response variable. In the latter case we have

$$E\big(g(Y_i)\big) = \sum_{r=1}^{p} x_r \alpha_r \;, \tag{3.142}$$

whereas in a generalised linear model

$$g\big(E(Y_i)\big) = \sum_{r=1}^{p} x_r \alpha_r \;. \tag{3.143}$$

In practice, however, transformed linear and corresponding generalised linear models often tend to give rather similar results. For a worked example see [354].

As in the case of classical linear models, fitted generalised linear models can be summarised by parameter estimates, their standard errors and goodness–of–fit statistics. One can also estimate confidence intervals for (combinations of) parameters, test various hypotheses, or attempt some type of distributional inference [359]. Except for models with canonical link functions, exact distribution theory is not available, and normally, one uses specific iterative estimation procedures which have been implemented in GLIM and afterwards also in other statistical software packages such as SAS and S–PLUS, see, e.g., [181, 208, 455, 495, 578, 581].

Statistical Inference

Usually a variant of the maximum likelihood method is used for estimating $\boldsymbol{\alpha}$ in (3.135). Since explicit expressions for an ML estimator generally do

not exist, such estimates must be calculated iteratively. It can be shown that the asymptotic properties of $\boldsymbol{\alpha}$ depend on the response distribution only through the mean function μ and the variance function $V(\mu)$. Hence, in order to estimate $\hat{\boldsymbol{\alpha}}$ one does not need the full likelihood corresponding to the distribution of the exponential family, but only the 'canonical part' depending on these two parameters. This method of estimating is called the *quasi-likelihood* method [725]. In order to estimate the dispersion parameter ϕ, methods other than those based on maximum likelihood are often used, for instance the 'method of moments', see Sect. 2.2.1.

The counterpart of the variance in a linear model is, in the case of a generalised linear model, the so-called *deviance*. In general, it can be expressed as twice the difference between the maximum attainable log likelihood and the log likelihood of the model under consideration:

$$D(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}) = 2\phi\{l(\boldsymbol{Y}; \boldsymbol{Y}, \phi) - l(\boldsymbol{Y}; \hat{\boldsymbol{\mu}}, \phi)\} \; . \tag{3.144}$$

The deviance is a non-negative quantity. It equals zero only if the model fits the data perfectly, i.e., $\hat{\boldsymbol{\mu}} = \boldsymbol{Y}$. If the assumed model is true, the deviance is distributed approximately as $\phi\chi^2_{n-p}$, where n is the number of observations and p is the number of fitted parameters. In general however, as is the case with the usual residual sum–of–squares, the distribution of the deviance depends on the 'true' vector of mean values $\boldsymbol{\mu}$ and differs asymptotically from the χ^2distribution.

From (3.144) one can see that maximisation of the likelihood means minimisation of the deviance, so the maximum likelihood method provides the best fit according to the deviance criterion. For further information about generalised linear models, the reader is referred to [194, 272, 455].

3.5.3 Nonlinear Regression

In a nonlinear regression model, the expectation value of the response variable can be expressed as a (nonlinear) function of the independent variables and the parameters. The model is:

$$E\, Y_i = \mu_i = \mu(\boldsymbol{x}_i^t, \boldsymbol{\theta}) \; , \tag{3.145}$$

where $\boldsymbol{x}_i^t$ is the ith row vector of the design matrix of the observations and $\boldsymbol{\theta}$ is the vector of unknown parameters.

The observation Y_i has a probability distribution $f\big(\mu(\boldsymbol{x}_i^t, \boldsymbol{\theta}), y\big)$ with respect to its mean value.

Usually (but not necessarily) the errors

$$E_i = Y_i - \mu(\boldsymbol{x}_i^t, \boldsymbol{\theta}) \tag{3.146}$$

are assumed to be independent and approximately normally distributed. Multiplicative errors proportional to the expectation value of the response vari-

able are roughly equivalent with homoskedastic, additive errors on a logarithmic scale, while constant multiplicative errors lead to additive errors with (known) heteroskedasticity, see Sect. 4.4 of [354] for a further discussion.

Some examples of nonlinear models, with multiplicative or additive errors in cases (1) to (4), are:

(1) the *exponential decay model*

$$\mu_i = \beta e^{-\alpha x_i} \,, \tag{3.147}$$

the vector of parameters $\boldsymbol{\theta}^t$ being (β, α);

(2) the *Weibull model*

$$\mu_i = \beta e^{-(x_i/\sigma)^\gamma} \,, \tag{3.148}$$

with $\boldsymbol{\theta}^t = (\beta, \sigma, \gamma)$; see, e.g., Chap. 14 of [543].

(3) the *generalised Gamma model*

$$\mu_i = \frac{\sqrt{f_1/f_2}}{\Gamma(\sqrt{f_1 f_2}) g^{f_1}} x_i^{f_1 - 1} e^{-(x_i/g)\sqrt{f_1/f_2}} \,, \tag{3.149}$$

with $\boldsymbol{\theta}^t = (f_1, f_2, g)$; f_1 and f_2 are shape parameters and g is a scale parameter. Here, the model for μ_i corresponds to the shape of a probability density rather than of a tail probability considered in the previous two cases. For $f_1 = f_2 = f$ the Gamma distribution is recovered, and for $f_1 f_2 = 1$ the Weibull distribution. The half normal distribution corresponds to $f_1 = \sqrt{2}$, $f_2 = (2\sqrt{2})^{-1}$ and the log normal distribution to $f_1 f_2 \to \infty$. The shape parameter f_2 describes the tail behaviour of the distribution. For further information, see [328, 337].

(4) the *logistic growth model*

$$\mu_i = \frac{\beta}{1 + \gamma e^{-\alpha x_i}} \,, \tag{3.150}$$

with $\boldsymbol{\theta}^t = (\beta, \gamma, \alpha)$;
The nonlinear logistic model should not be confused with the *logistic regression model*, which is a special case of the generalised linear model, cf. [727].

(5) *multinomial models with polytomous regression.* In a $k \times q$ contingency table with k rows ('classes'), q columns ('responses') and fixed class marginals, each response vector $\boldsymbol{Y}_i$ has a multinomial distribution:

$$\boldsymbol{Y}_i \sim M(n_i; p_{i,1}, \ldots, p_{i,q}) \quad (i = 1, \ldots, k) \,, \tag{3.151}$$

with

$$P\{\boldsymbol{Y}_i = \boldsymbol{y}_i\} = \frac{n_i!}{y_{i,1}! \cdots y_{i,q}!} p_{i,1}^{y_{i,1}} \cdots p_{i,q}^{y_{i,q}} \,, \tag{3.152}$$

and constraints $\sum_j^q p_{i,j} = 1$, $\sum_j^q y_{i,j} = n_i$ and $\sum_i^k n_i = n$. A particular class of regression models is given by

$$g(p_{i,j}) = X_{i,1}\alpha_{1,j} + \cdots + X_{i,m}\alpha_{m,j} + \theta_i + \phi_j \quad (i = 1, \ldots, k; j = 1, \ldots, q) \tag{3.153}$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ (with $\boldsymbol{X}_h = (X_{1,h}, \ldots X_{k,h})^t$) are m manifest covariates, such as age, dose of a toxin, plasma impurity content, etc., θ_i is a latent trait represented by a random effect with, e.g., $\theta_i \sim N(\theta, \sigma^2)$, and ϕ_j describes an item-specific null-effect (i.e. $g(p_{i,j})$ at zero values of $\boldsymbol{X}_1, \ldots \boldsymbol{X}_m$ and θ_i). In the absense of covariates and for a logistic link function $g(p_{i,j}) = \log(p_{i,j})/\log(1 - p_{i,j})$, the parameters θ and σ^2 of the latent trait θ, in the presence of ϕ_j, can be calculated in closed form, see [455, 575], and the model is called the classical Rasch model in item response theory, see [69, 195, 541]. In the absense of any random effect, we have the classical logistic regression model, which is called polytomous for $q > 2$. The responses can be of a nominal nature (e.g., the type of ELMs in a plasma discharge), ordinal (e.g., the severity of a plasma disruption) or have an ambivalent or hybrid character (e.g. colour or bidirectional strength of affinity). For ordinal responses, a logistic odds model [455] might be used for the cumulative probabilities $p_{i,h} = \sum_j^h p_{i,j}$ and, more generally, methods of order-restricted inference apply, see [5, 556, 623, 625]. For other link functions than the logistic, such as $g(p) = \Phi^{-1}(p)$, $g(p) = \ln\big(-\ln(p)\big)$, or quantit [121] $g_\nu(p) = \int_{0.5}^p \big((1 - |2z - 1|)^{\nu+1}\big)^{-1} dz$ ($\nu > -1$), the (quasi) likelihood equations are non-linear, see [455, 479, 686]. For a computer implementation of classical logistic regression [126, 479], see NAG / GLIM [208] and SAS / PROC LOGISTIC [581]. Latent variable models are described in [66, 127, 334], among others. In a classical regression context (with real-valued response variables $Y_1, \ldots, Y_q$) a fairly large number of these are incorporated in SAS / PROC CALIS [581]. An application with teratological data is described in [574].

(6) *offset-linear scalings*, for the thermal energy of a fusion plasma. The thermal energy of a plasma is decomposed into two additive parts, the scaling of each of which is approximated by a simple power law, i.e.,

$$W_{th} = W_o + W_l \ , \tag{3.154}$$

where

$$W_o = a_0 I_p^{a_I} B_t^{a_B} \overline{n}_e^{a_n} M^{a_M} R^{a_R} a^{a_a} \kappa^{a_\kappa} P_{L'}^{a_P} \tag{3.155}$$

models the offset term, and

$$W_l = b_0 I_p^{b_I} B_t^{b_B} \overline{n}_e^{b_n} M^{b_M} R^{b_R} a^{b_a} \kappa^{b_\kappa} P_{L'}^{b_P} \tag{3.156}$$

models the linear term. The vector of unknown parameters is

$$\boldsymbol{\theta}^t = (a_0, a_I, \ldots, a_P, b_0, b_I, \ldots, b_P) \ . \tag{3.157}$$

The quantities $I_p, B_t, \overline{n}_e, M, R, \kappa, \varepsilon, P_{L'}$ are plasma current, magnetic field, plasma density, isotope mass, three geometrical quantities (major radius R, minor radius a, elongation κ), and the net absorbed power by the plasma, respectively, see also Sect. 4.1, and CASE I in Sect. 7.2. For a practical account of fitting such a type of scaling to the second version of the standard dataset of the ITER H–mode confinement database ('ITERH.DB2'), the reader is referred to [361].

As in the case of linear models, least squares or, now not equivalently, some variant of maximum likelihood, is regularly used to estimate the parameters in a non-linear model. Here, we have only a brief look at the method of least squares.

The least-squares estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{LS}$, is the vector of parameters that minimises the sum of squared residuals

$$SS_{res}(\boldsymbol{\theta}) = \sum_{i=1}^{n} [Y_i - f(\boldsymbol{x}_i^t, \boldsymbol{\theta})]^2 \ . \tag{3.158}$$

The normal equations have the form

$$\frac{\partial SS_{res}(\boldsymbol{\theta})}{\partial \theta_j}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\sum_{i=1}^{n} [Y_i - f(\boldsymbol{x}_i^t, \hat{\boldsymbol{\theta}})] \Big[\frac{\partial f(\boldsymbol{x}_i^t, \hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_j}\Big] = 0 \ , \tag{3.159}$$

for $j = 1, \ldots, p$.

The second bracket in (3.159) contains the partial derivative of the functional form of the model. The partial derivatives of a nonlinear model are functions of the parameters, hence, the normal equations are nonlinear.

In general, finding non-linear least-squares estimates requires the application of iterative minimisation procedures, and can be marred (especially in problems with many parameters and/or ill-conditioning) by complications such as the choice of suitable starting values, multiple solution branches, interpretation problems in case of non-convergence, etc.

Several software packages, for instance SAS, NAG, S–PLUS and the Optimisation Toolbox for use with MATLAB provide flexible nonlinear minimisation procedures with a wide range of algorithmic fine-tuning options to cope with convergence problems. MathWork's Statistics Toolbox provides a graphical interface to explore non-linear regression surfaces.

Many distributional properties of OLS estimators that hold in linear models, also hold asymptotically (i.e., for sufficiently large sample sizes) for nonlinear regression estimators. This stems from the fact that the non-linear model can be linearised around the true value of the regression parameter vector, and the size of the confidence regions shrinks with increasing sample size.

If the errors are independently normally distributed, then $\hat{\boldsymbol{\theta}}$ is approximately normally distributed,

$$\hat{\boldsymbol{\theta}} \;\sim\; N[\boldsymbol{\theta}, (\boldsymbol{F}^t\boldsymbol{F})^{-1}\sigma^2] \;, \tag{3.160}$$

where the matrix $\boldsymbol{F}$,

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x}_1^t,\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial f(\boldsymbol{x}_1^t,\boldsymbol{\theta})}{\partial\theta_2} & \cdots & \frac{\partial f(\boldsymbol{x}_1^t,\boldsymbol{\theta})}{\partial\theta_p} \\ \frac{\partial f(\boldsymbol{x}_2^t,\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial f(\boldsymbol{x}_2^t,\boldsymbol{\theta})}{\partial\theta_2} & \cdots & \frac{\partial f(\boldsymbol{x}_2^t,\boldsymbol{\theta})}{\partial\theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(\boldsymbol{x}_n^t,\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial f(\boldsymbol{x}_n^t,\boldsymbol{\theta})}{\partial\theta_2} & \cdots & \frac{\partial f(\boldsymbol{x}_n^t,\boldsymbol{\theta})}{\partial\theta_p} \end{bmatrix} , \tag{3.161}$$

is an $(n \times p)$ matrix of partial derivatives at all n data points, which is equal to the design matrix $\boldsymbol{X}$ in case of a linear model.

In contrast to $\boldsymbol{X}$, the matrix $\boldsymbol{F}$ depends on $\boldsymbol{\theta}$. In computational practice, the estimates of $\boldsymbol{F}$ and σ^2 at $\hat{\boldsymbol{\theta}}$, i.e.,

$$\hat{\boldsymbol{F}} = \boldsymbol{F}(\hat{\boldsymbol{\theta}}) \;, \tag{3.162}$$

and

$$s^2 = \frac{SS_{res}(\hat{\boldsymbol{\theta}})}{n-p} \tag{3.163}$$

are used in (3.160).

The residual sum-of-squares, $SS_{res}(\hat{\boldsymbol{\theta}})$, is approximately χ^2_{n-p} distributed. Theoretical and practical aspects of fitting non-linear models are discussed in [100, 217, 293, 543, 605], among others.

3.5.4 Other Extensions

To conclude this section, we mention a few extensions of linear regression models that have become in vogue because of their flexibility and in view of the fact that they can be efficiently implemented in computer programs. In fact, many of them are indeed routinely provided by commercial statistical packages such as S–PLUS and SAS/INSIGHT.

(1) In *local regression* (see, e.g., [700]) nonlinear data are fitted locally, i.e., any point at the curve depends only on the observations at this point and on some specified neighbouring points. The estimated response function is called *smooth*, because it is less variable than the originally observed response variable. The procedures producing such types of fit are sometimes called *scatter–plot smoothers* [268].

(2) In regression with a *penalty functional* (see, e.g., [181]) an overparameterised model is regularised by minimising the sum-of-squares plus a constant λ times a roughness penalty for the response function, i.e.,

$$SS_{res}(\boldsymbol{\theta}) + \lambda L(\boldsymbol{\theta}) \ , \tag{3.164}$$

is minimised, where for each $\boldsymbol{\theta}$, $L(\boldsymbol{\theta})$ is a functional of the approximating function $f(x, \boldsymbol{\theta})$ and its derivatives. This is akin to the regularisation method for ill-posed problems developed by Tichonov in the early forties, see [680]. Several forms of the roughness functional are possible. We restrict attention to one-dimensional approximation problems. Consider

$$L(\boldsymbol{\theta}) = \int \sqrt{1 + \big(f'(x, \boldsymbol{\theta})\big)^2} dx \tag{3.165}$$

and

$$L(\boldsymbol{\theta}) = \int \big(f''(x, \boldsymbol{\theta})\big)^2 dx \ . \tag{3.166}$$

In the first case, the functional penalises the arc length of the approximating curve, and in the second case it penalises some average curvature. Obviously, the first functional rather than the second should be used if a polygon (i.e., a continuous, piecewise linear function) is fitted to the data. In both cases, a straight line is fitted if $\lambda \to \infty$. If a parabola rather than a linear function is preferred as a limiting response function (e.g. when fitting plasma profiles, see Chap. 4), then one should rather use a functional like

$$L(\boldsymbol{\theta}) = \int \big(f'''(x, \boldsymbol{\theta})\big)^2 dx \ . \tag{3.167}$$

For approximating functions that are sufficiently often differentiable, it seems sensible to choose a penalty functional, adapted to the specific problem at hand, from the following class (*Sobolev–norm penalty functions*)

$$L(\boldsymbol{\theta}) = \int \sum_{k=p_l}^{p_u} \big(f^{(k)}(x)\big)^2 c^{2k} dx \ , \tag{3.168}$$

where p_l and p_u determine a sensible range of the expansion and c is a constant which indicates the relative penalisation strength of the various derivatives. The value of c depends on 'scaling' of the problem. In practice, for a properly scaled problem (the standard deviations of the data in x and y direction being of order 1) one can use $c = 1$. In a robust version, the 2-norm may be replaced by a p-norm with p, say, between 1.2 and 1.5. The larger the values of λ and of c, the smoother the fit. For both λ and c sufficiently large, a polynomial of order $p_l - 1$ is fitted. The degree of smoothness (i.e., the trade-off between variance and bias) is partly a matter of taste and dependent on the specific situation. Cross-validation or Akaike-type criteria [4] can be applied that minimise the expected mean squared error of prediction (estimated in some way) as a (semi-)automatic procedure to determine a reasonable value of λ and possibly also for c.

(3) In an *additive model*,

$$Y = \alpha + \sum_{i=1}^{n} f_i(X_i) + E \,, \tag{3.169}$$

the expectation value consists of an additive combination of smooth functions of the predictors. The standard linear regression model is a special case of an additive model.

(4) A *generalised additive model* (see, e.g., [268]) extends the generalised linear model by replacing its linear predictor with an additive combination of smooth functions of the predictors, i.e., the mean–value structure is written as

$$g(\mu_i) = \eta_i = \alpha + \sum_{i=1}^{p} f_i(X_i) \,. \tag{3.170}$$

The non-parametric functions f_i are estimated from the data using some smoothing method, for instance a scatter–plot smoother.

(5) Abstracted from their biological origin [153, 286, 374, 698], *neural networks* [606] are often used as an adaptive method to fit multiple response surfaces. From a statistical point of view (see, e.g., [35, 577]) they constitute a rather flexible class of statistical regression models and can be viewed as an extension of multivariate splines, generalised additive models, etc. The increased flexibility offers new possibilities in applied regression modelling. These possibilities contain a benefit and a danger. The benefit is that more realistic representations of the actual behaviour of p-dimensional data clouds can be found, the danger being that of 'overfitting' [358]. Consequently, the validity of such 'data-driven models' with respect to new situations may easily be overestimated, unless special precautions are taken. At present this involves in most cases rather elaborate simulations of the experimental environment.

(6) Regression with *catastrophe-type response functions* is a special area. The roots of catastrophe, or perestroika, theory[5] tracing back to description of phase transitions in thermodynamics by Maxwell, Gibbs and van der Waals, see [18], R. Thom and E.C. Zeeman have stimulated, among others, the development, since the late sixties of the previous century, of a local classification theory of catastrophe manifolds, which are the stationary points of certain non-linear dynamical systems [255,489,704], for which the time derivative of the response variable can be written as the gradient of a potential, see Chap. 9 of [521], and also [18]. Such mappings can be related to various plasma-physical theories which utilise

[5] Perestroika originally meaning restructuring, either one of these terms (coined by Thom and Arnold, respectively) is used to denote a qualitative, sometimes 'abrupt' or 'discontinuous', change of a system induced by a small change of its parameters, on which otherwise, 'generically', it depends continuously.

the language of dynamical systems (see, e.g., [305]). Suitably parameterised [355], they can, in principle, also be used to describe simultaneously, for instance, the scaling of the plasma energy in L(ow)-mode and H(igh)-mode confinement regimes. The flexibility of these models holds the middle ground between standard multivariate (non-linear) regression and multivariate non-parametric regression models. Fitting such classes of non-linear multiple-valued response functions has not yet received a concentrated attention in statistics. However, it is an interesting area of investigation, also in view of Thom's sceptical remark on p. 134 of [676], albeit characterised by a fair amount of modeling and computational complexities. For further background information on this topic, see [18, 88, 112, 230, 297, 306, 353, 355, 417, 498, 521, 677].

4 Profile Analysis

4.1 Introduction: Fusion Context

Thermonuclear fusion research constitutes an interesting and important area in physics with both fundamental and technological aspects. There are several approaches. Confining the plasma in a toroidal magnetic field is a very promising one. Briefly formulated, the goal is here to confine a plasma (a high temperature gaseous mixture of electrons and ions) in such a way that, at a suitably high ion temperature T_i, the product of plasma density n and confinement time τ_E is sufficiently elevated to maintain a self-burning plasma, in which deuterium and tritium ions fuse yielding highly energetic helium ions and neutrons, see [350, 729].

Fusion processes take also place in the sun [63]. In a power plant the excess energy of the fusion reaction products has to be converted into electrical energy. Carbon fibre composites (CFC's), tungsten ('Wolfram') and beryllium are considered as plasma-facing components, see [188, 283, 490, 532, 672], each in a different area of the vessel and of the divertor region, which removes heat from the plasma and exhausts various plasma impurities. Beryllium also can serve as an absorber and a multiplier of neutrons [627]. Tritium, which is one of the two fuel components, is scarcely available in nature and therefore has to be produced ('bred') from lithium, which in the form of brine and ore is abundantly available [145]. The breeding process captures neutrons that are produced by the primary fusion reaction between deuterium and tritium. Deuterium can be produced by electrolysis of ordinary water followed by liquifaction to separate the hydrogen from deuterium, see for instance [283]. In [84, 632] more details on the fuel cycle are provided, while in [384] a currently remote option of an advanced fuel cycle based on helium-3 is discussed. This requires temperatures which are about a factor five higher than for D–T reactions, and is motivated by the perspective of generating heat with a reduced production of neutrons. Several concepts of a 'blanket' with suitable heat removal and breeding capacities have been developed, which (in form of modular cassettes) can be tested in the presently planned Next–Step ITER device [301, 672].[1] Using stainless steel or, possibly, a vanadium-based alloy

[1] The acronym ITER ('Road' in Latin) stands for International Toroidal Experimental Reactor, a tokamak designed in international cooperation and incepted

as structural material, water-cooled lithium–lead eutectic, or helium-cooled ceramic lithium compounds contained in a pebble–bed are currently being considered as a tritium producing material [367,410,531,627].[2] The blanket also serves to shield various components outside the vessel from the flux of neutrons [43]. Subsequently, a stepwise heat-exchange procedure is applied, admitting the coolant of the last step to propel turbines generating electricity. For the construction of such a power plant there is still a considerable way to go, requiring parallel technical and physical developments [410]. The efforts in research and development are motivated by the fact that eventually an important source of energy would become available for mankind, constituting a relatively environment-friendly alternative to power plants based on fossil fuels, solar energy or nuclear fission. For environmental and safety aspects of fusion power generation, the reader is referred to [237,238,256,443,530,531, 570,672,685].

The tokamak[3] is one of the toroidal devices developed for fusion and is characterised by the fact that the plasma itself carries an electric current, which is in the order of several MA for the larger present-day experiments. This current heats the plasma and, while producing a poloidal magnetic field, contributes to its thermal confinement. The poloidal magnetic field is to modify the toroidal magnetic field produced by external currents in such a way that large scale (macroscopic) instabilities of the plasma are avoided. [4] As

(after the INTOR project) on the verge of the new era induced by Gorbachov and Reagan, see [84]. The physics basis of the 1998 ITER design has been described in [671]. Subsequently a redesign of size and shape has been made to achieve a reduction of costs while largely maintaining the physical and technological objectives of the device, see [21,95,350,483,615]. While still bearing some resemblance to the Next European Torus (NET) [669], in the mid-eighties planned as successor of the Joint European Torus (JET) which operates in England, the present, matured design of the next step international tokamak experiment is sometimes called ITER–FEAT, where the second acronym stands for Fusion Energy Amplification Tokamak, see [23,350,485,614,672].

[2] An alternative approach, propounded already some time ago, is based on a self-cooling liquid blanket material such as elemental lithium or lithium-beryllium fluoride, see [472,529].

[3] The name tokamak, a Russian acronym for 'toroidalnaya kamera c magnetnymi katuschkami', i.e. a 'toroidal vessel with magnetic coils', was exported to the West in [152] and is due to its originators Tamm and Sacharov, see [84,335,482,571,665]. (A list of Russian keyword translations is included at the end of this book.) For readers in command of the Russian language, it is interesting to study [482], which describes the state of the art in fusion-oriented tokamak research around 1980 with an outline of its history.

[4] This induces a lower limit on the 'safety factor', which is, for a circular cross-section, $q_{eng} \sim (B_t/\overline{B}_p)(a/R)$ with B_t the toroidal magnetic field on the axis and $\overline{B}_p$ the average poloidal magnetic field, a the minor and R the major plasma radius, and which implies an upper limit on the plasma current $I_p \sim a\overline{B}_p$ for a given toroidal field B_t and machine size.

for any current ring, the plasma current experiences a radially expanding 'hoop-force' which is counterbalanced by interaction of the plasma current with the poloidal magnetic field that is produced by external poloidal-field coils. Figure 4.1 presents a schematic overview of the situation. Tokamaks of various sizes have been built in various countries, in the past mainly in Europe, Japan, Russia, and the USA. With the objective in mind to prepare the physical basis for Next-Step experiments such as ITER, and eventually for an environmentally and economically viable fusion-power delivering device, a fruitful collaboration has been established between representatives of the various tokamak teams, as exemplified by the ITER (presently ITPA) expert groups and the ITER Physics Basis Document [671]. The internet site http://efdasql.ipp.mpg.de/HmodePublic/ contains pictures of tokamaks collaborating in the area of plasma confinement analysis. Results of the investigations are published in journals such as Nuclear Fusion, Plasma Physics and Controlled Fusion, Physics of Plasmas, Plasma Physics Reports, Journal of Plasma and Fusion Research, Physical Review Letters, Review of Scientific Instruments, Fusion Engineering and Design, Fusion Science and Technology, Journal of Nuclear Materials, Fusion Energy (proceedings of the biennial IAEA conference), among others. Other types of toroidal devices are stellarators and reversed field pinches. In a reversed field pinch (RFP), the plasma is radially contained between a toroidal magnetic field near the axis and a similarly large poloidal magnetic field in the outside region. Characteristically, the toroidal magnetic field reverses its sign inside the plasma. The reader is referred to [210,500,588,728]. In stellarators a rather complicated set of magnetic coils is used, which obviate the need of a plasma current and allow for continuous operation, see [630,717] and, for an overview of plasma transport, also [212,433,652]. First devised and constructed in Princeton (around 1950), 'American' stellarators constituted an alternative approach to the 'Russian' tokamaks, see [84] for an historic review. At present, after successful operation of Wendelstein 7–AS[5], see [178,313,446,453,576,715,742] among others, the W7–X stellarator is being built in Germany, which has approximately the same physical dimensions as the Large Helical Device (LHD) operating at NIFS, Japan. Somewhat in parallel, also in Germany, a conceptual stellarator plasma–ignition experiment, of type Helias,[6] has been developed, see [276,741].

One of the problems is that of confinement: the high temperature plasma (about 10^8 degrees Kelvin) tends to lose its energy, due to radiation,[7] and heat transport (conduction and convection), which is related to (micro-) instabil-

[5] The acronym AS stands for advanced stellarator and fortuitously coincides with the shorthand signature of one of its founding fathers [592].

[6] The name Helias stands for helical-axis advanced stellarator [249].

[7] Broadly speaking, three types of radiation can be distinguished: Bremstrahlung, arising from transitions of electrons in the electric field of an ion, cyclotron radiation (or magnetic Bremstrahlung), due to electrons gyrating in the magnetic field,

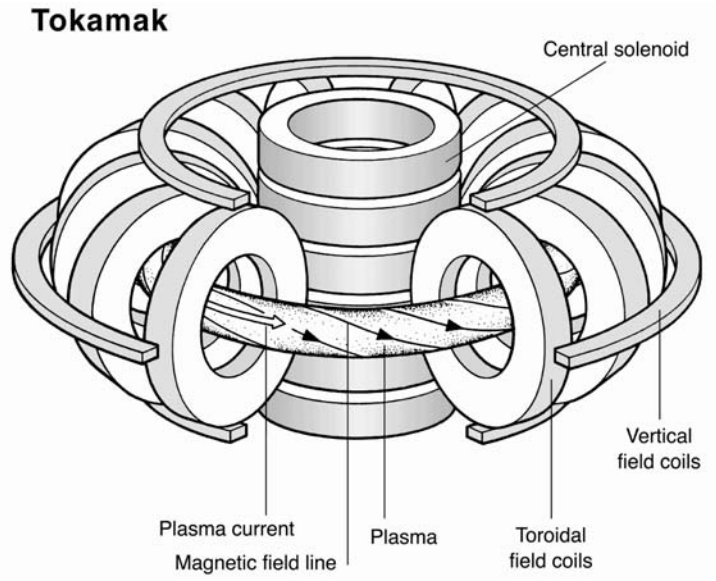

Fig. 4.1. Schematic view of a plasma confined in a tokamak (drawing reproduced by courtesy of IPP Garching). One can distinguish the toroidal field coils, the poloidal field coils and the central solenoid. The arrows on the plasma column indicate the direction of the plasma current and of the magnetic field lines. To give an impression of the size: For the ASDEX tokamak, which has a circular cross-section, the major radius R is 1.65 m and the minor radius a is 0.4 m. For comparison, the major radii of JET, of the next step ITER FEAT tokamak, and of a traditionally projected 3 GW fusion power plant are 3.0 m, 6.2 m, and some 8–10 m, respectively.

ities and various other, not yet fully understood, processes, such as plasma turbulence. The exponential decay time of the total plasma energy $W(t)$, in the hypothetical situation that no heating of the plasma (neither from external nor from internal sources) would occur, is called the energy confinement time τ_E. In the presence of heating, τ_E is defined by the differential equation

$$\frac{dW}{dt} = P - \frac{W}{\tau_E} \,, \tag{4.1}$$

where P denotes the total heating power of the plasma by internal or external sources. It is noticed that in transient situations W and τ_E are functions of time. The global confinement time τ_E depends to a large extent on controllable plasma parameters, such as the plasma current I_p, the magnetic field B_t, the absorbed heating power P, the electron density n_e and obviously also

and line radiation (or impurity radiation), which arises from transitions between different energy levels of electrons that are bound to an ion, see [70, 82, 350, 632].

on the size of the plasma, characterised by the major radius R, the minor radii a and b, and the triangularity δ, see [176].

Although such a description of global confinement is useful for engineering purposes, for instance in extrapolating present-day experiments to future reactor-type experiments, physical understanding is enhanced by relating the confinement, and other aspects of plasma transport, to the temperature and density profiles of the electrons and ions. To a good approximation, toroidal symmetry exists and the plasma density and temperature are constant on flux–surfaces, which, according to classical electrodynamics, are nested tori on which the magnetic field lines reside. Hence, in tokamaks with a circular cross-section, such profiles can be described as a function of one spatial coordinate r only ($-a < r < +a$). On the one hand, in stationary situations where $dW/dt = 0$, τ_E can be expressed as an integral over such profiles:

$$\tau_E = cP^{-1} \int_{-a}^{+a} \big(n_i(r)T_i(r) + n_e(r)T_e(r)\big) r dr \ , \tag{4.2}$$

for some (specific) constant[8] c. On the other hand, several types of micro-instabilities which affect the heat transport and hence also the confinement time are influenced by the gradients of the temperature and density profiles (which may be coupled [42]). A considerable amount of theoretical effort has been devoted to describing these relationships. Because of the complexity of the system, there exists a variety of models, based on different physical assumptions, see [303, 306, 485, 671].

On the empirical side, direct measurements of these profiles at discrete radial positions have become available since the advent of so-called 'active' diagnostics, which rely on the interaction of diagnostic (either neutral-particle or laser-light) beams with the plasma. Physicists are interested in 'smoothing' these empirical profiles, in order to be able to use them in interpretative computer codes as well as to compare them with predictions from theory. Obviously, an accurate empirical representation of the dependencies of these profiles on the basic plasma parameters is an important intermediate step in the understanding of plasma transport and in plasma discharge control. As individual profile measurements exhibit often considerable scatter, it is of importance to clearly express the statistical (in-)accuracy obtainable from a series of profile measurements.

Publications on statistical analysis of plasma profiles are comparatively rare, see for instance [343], and [451]. A particular type of semi-parametric profile analysis is used in [299, 600]. In this chapter, we give a systematic overview and discussion of the statistical techniques to analyse plasma profiles and their associated global quantities, while restricting attention to the framework of parametric profile representation.

[8] For a circular cross-section, $c = \frac{3}{2} \times 1.6 \times 2\pi$ if the temperatures are expressed in eV (1 eV corresponds to approximately 11600 degrees Kelvin) and the densities in 10^{19} particles per m^3.

The statistical methods of analysis are also of interest for their own sake, as their applicability exceeds the plasma-physical context in which they are described here. Some specific examples are decay curves of radioactivity in a mixture of substances, see Chap. 5 of [672], decay of glomerular filtration rate in kidney patients, see Chap. 14 of [347], and analysis of children's growth–curves [519, 535]. Several other interesting medical examples in this context are discussed in [405]. Therefore it is hoped that by studying this chapter, the reader will be able to employ these methods also in other contexts, where the influence of various factors on (time-dependent) growth curves or (space-dependent) profiles are at stake.

In Sect. 4.2, we describe the analysis of a single sample of n unsmoothed profiles, for a fixed value of the plasma variables I_p, B_t, etc., by standard multivariate statistical analysis. In this approach, each empirical profile, consisting of temperature or density measurements at p different radial locations, is considered as one observation in $\mathbb{R}^p$.

In Sect. 4.3, we consider various continuous representations for the true underlying profile, and we discuss a number of relevant error structures.

In Sect. 4.4, the influence of plasma variables, such as I_p, B_t, etc. on the profiles is modelled, and a mathematical definition of profile invariance is given.

In Sect. 4.5, various methods of estimating the free parameters in the models of Sects. 4.3 and 4.4 are reviewed and discussed.

In Sect. 4.6, several statistical tests for discriminating between different concrete profile representations are given, paying attention the the mean–value structure as well as the error structure.

In Sect. 4.7, confidence and prediction bands for the underlying profiles are constructed, as well as confidence intervals for the derived volume-averaged quantities.

We outline here the physical context of the analyses in the rest of this chapter, by providing a simplified and idealised physical picture of tokamak operation, in terms of an 'input-output' system [489, 518]. This description may contain some new and rather unfamiliar terms for readers not familiar with plasma physics. While they must grasp some essential information only, in order to improve their understanding of the context of this chapter, they are encouraged to look for further information at the physical description of the datasets in Chap. 7 and the references to the literature cited there.

For each plasma discharge, which consists, under stationary conditions, of a number of physically 'equivalent' time-slices, called (multivariate) observations in the sequel, a set of global engineering plasma parameters have been selected, and subsequently refined, by the tokamak design group, by the physicists setting the campaign-period goals, and by the experiment leader. All these parameters we call input variables, and we shall denote them by $\boldsymbol{x} = (x_1, \ldots, x_w)$. For Ohmic discharges, the main groups of input variables are: (1) the toroidal magnetic field B_t, the safety factor q_{eng}, the total cur-

rent I_p (for a fixed geometry, any two of these three variables determine the third one[9] and the line-averaged electron density $\overline{n}_e$; (2) the horizontal minor radius a, the vertical minor radius b, the major radius R, and the vacuum magnetic-field ripple; (3) the plasma 'cross-sectional shape', which we use in a rather broad physical sense and can be characterised by discrete variables, such as the configuration type (limiter, single-null divertor, double-null divertor), as well as by continuous variables such as elongation $\kappa = b/a$, triangularity δ, and the distance between the separatrix and the wall; (4) the main ion species mixture (hydrogen, deuterium, etc.); (5) the addition of impurity ions to enhance (edge) radiation; (6) the relative sign of the direction of the plasma current and the magnetic field; (7) the direction of the toroidal magnetic field (clock-wise or counter-clockwise as seen from the top), which is a binary variable which influences the direction of the ion drift ('towards' or 'away from' the lower X-point) related to the magnetic field gradient; (8) the wall material and its coverage, for instance by carbon, beryllium, boron, or tungsten, achieved by applying special conditioning discharges. For plasmas with neutral beam heating, additional engineering variables play a role, such as the input power, angle of injection, beam particles being co- or counter-injected with respect to the plasma current, energy per injected particle, and the species of the injected particles. For radio–frequency heated plasmas, a specification is needed of input power, the location and shielding of the antenna, the spectra of the injected microwaves, etc. Heating by radio-frequency waves is at present performed by three main categories, with different technological characteristics, abbreviated by ICRH, ECRH and LH, which stand for ion cyclotron resonance heating, electron cyclotron resonance heating and (heating by) lower hybrid (waves), respectively. The frequency of the injected electromagnetic waves is chosen such as to produce resonance with electron cyclotron, ion cyclotron or lower hybrid waves inside the plasma. In addition to neutral beam injection, also pellet injection can be used to refuel the plasma. In that case, additional input variables are the size, frequency, angle of injection, species and velocity of the pellet.

The output variables that we will be principally interested in here are the temperature and density (and hence pressure) profiles, as well as the derived local transport coefficients. Other output variables are the profiles of current, impurity density and radiation, the loop voltage, or, equivalently, plasma resistivity, the toroidal rotation, the (poloidal or toroidal) plasma beta, the energy confinement time, the Shafranov shift as well as other moments of the magnetic field, and the characteristics of edge localised modes (ELMs) (frequency, amplitude) and sawteeth (frequency, amplitude, and inversion radius). By radially integrating the fitted ion and electron pressure profiles one obtains an estimate of the total plasma energy (W_{kin}). This can be compared with two other methods of measuring the total stored energy, which

[9] Using the units T, MA and m, $q_{eng} = 5\frac{B_t}{I_p}\frac{Area}{\pi R}$, where $Area$ denotes the plasma cross-sectional area, which is equal to πab for an elliptical cross-section.

utilise a diamagnetic loop (W_{dia}) and a full plasma equilibrium reconstruction (W_{mhd}), respectively. Under stationary conditions, $\frac{dW}{dt} = 0$, the ratio between plasma energy and input power is the global energy confinement time τ_E. The plasma profiles contain more detailed information about the plasma transport than do the global quantities W and τ_E, but the attainable experimental accuracy is somewhat lower.

In this chapter we will describe statistical plasma profile analysis, while focusing attention to the evaluation of measurements from the YAG–laser Thomson scattering system [562] as operated at the circular ASDEX device [668], now rebuilt as HL–2A in Chengdu, China. Although the statistical framework is especially oriented toward evaluation of electron temperature and density profiles, as measured by the Thomson scattering diagnostic, with only rather minor adaptations as required in any concrete case, it can also be used for measurements based on other active diagnostic methods.

The circular cross–section of the ASDEX device simplifies the mapping between the spatial radial coordinates and the physical flux–surface coordinates. This mapping depends on physical models implemented in plasma equilibrium codes, see [449, 450], and is not the primary object of our statistical analysis. With some modification, much of what follows can also be applied to tokamaks with an elongated and somewhat triangular cross-section, such as JFT–2M, ASDEX Upgrade, DIII–D, and JET, which are geometrically closer to the design of next-step tokamaks such as ITER [23].

4.2 Discrete Profile Representations

In this section and in the next, we consider data at single point in plasma parameter space, consisting of n separate observations of a plasma profile, such as temperature, at p distinct radial points, $T_j(r_l)$ with $1 \leq j \leq n$ and $1 \leq l \leq p$. Let the variable Y denote the logarithm of the profile variable (density or temperature). The n observed profiles can be represented by the basic data matrix:

$$\begin{pmatrix} Y_1(r_1) & \dots & Y_1(r_p) \\ \vdots & \ddots & \vdots \\ Y_n(r_1) & \dots & Y_n(r_p) \end{pmatrix}. \tag{4.3}$$

All variables depend on time. We will restrict, however, our attention to steady state plasmas and assume that the final state is independent of the past history of the discharge. We further note that in a stationary phase all plasma profiles evolve more or less periodically, the periodicity being dictated by the sawtooth crashes. Our data will consist of n distinct time points.

Each profile measurement has essentially three sources of deviations from the 'ideal profile': (1) deterministic (systematic) errors, (2) random errors due to measurement noise, (3) fluctuations due to variations in the plasma.

Systematic errors arise either from overlooking some physical effect, or from idealisation of the measuring process. For Thomson scattering experiments, these assumptions typically include perfectly collimated measurement devices, physical localization of the scattering, neglect of other nonresonant scattering mechanisms, and the assumption that the electron velocity distribution is Maxwellian. For some further background on Thomson scattering, see also [143]. Often the size of such errors can be roughly estimated, but it is usually very difficult to determine them more precisely, so that the resulting bias can only rarely be corrected for. (In fact, the 'easier' types of systematic errors are already accounted for in the standard evaluation of the diagnostic.) The random errors due to measurement noise may often be estimated by an error-propagation analysis.

A special difficulty lies in characterising the condition of the wall. Aside from discrete categories such as gettered/ungettered and wall carbonisation, it is clear that the immediate past history of the device influences impurity recycling from the wall. Only to a rough approximation one may assume that the influence of the wall on the plasma can be parameterised by a single variable such as Z_{eff}, as measured by Bremsstrahlung [639].

In Ohmic discharges, the primary source of the intrinsic plasma variation is the $m = 1$ sawtooth activity. The perturbations about the mean profile from sawteeth are highly correlated in space, since the profiles are flatter immediately after a sawtooth crash. This spatial correlation of the fluctuations should be taken into account in a careful profile analysis.

The profile fluctuations are generally temporally as well as radially correlated. The profiles of the ASDEX Thomson scattering experiments are sampled with a fixed 60 Hz frequency. We assume that this frequency is incommensurable with the sawtooth period, and hence that the sampling times are quasi-randomly distributed with respect to the sawtooth phase $(t - t_{crash})/\tau_{saw}$. This assumption of quasi-randomness with respect to the sawtooth phase allows us to neglect, for simplicity, the temporal correlation.

The spatial correlation is conveniently described by viewing the jth profile measurement $\boldsymbol{Y}_j = \big(Y_j(r_1), \ldots, Y_j(r_p)\big)$ as one (random) vector of observations with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Assuming also normality, which can sometimes, if necessary, be relaxed somewhat, we write:

$$\boldsymbol{Y}_j \sim N_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) , \qquad j = 1, \ldots, m \, . \tag{4.4}$$

Note that the systematic errors are not described by the covariance matrix $\boldsymbol{\Sigma}$. One can formally decompose

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{noise} + \boldsymbol{\Sigma}_{plasma-variation} \, . \tag{4.5}$$

The number of free parameters is p for the μ and $p(p+1)/2$ for $\boldsymbol{\Sigma}$. If the total number of observations, np is sufficiently large (at least several times the number of free parameters) these parameters may be estimated by using the standard maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\hat{\boldsymbol{\mu}} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{Y}_j, \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{m}\boldsymbol{S} = \frac{1}{m}\sum_{j=1}^{m}(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}})^t \,. \tag{4.6}$$

The matrix $\boldsymbol{S}$ is sometimes called 'the matrix of residual sums of squares and sums of cross products', which we shall abbreviate by 'the residual SSCP matrix'. $\boldsymbol{\Sigma}_{plasma-variation}$ can only be estimated if one has an independent estimate of $\boldsymbol{\Sigma}_{noise}$ from the error-propagation analysis. A reasonable assumption for the ASDEX Thomson scattering experiment is that $\boldsymbol{\Sigma}_{noise}$ is diagonal.

A special feature of the Nd:YAG–laser Thomson scattering diagnostic at ASDEX, see [561, 562], is that 10 of the 16 channels are, to some reasonable approximation, located symmetrically with respect to the horizontal midplane, see Fig. 4.2. While idealising the situation somewhat, this affords the following possibility to test for up-down asymmetry for circular double null plasmas with $z = 0$. In that case, radial measurement positions that do not have a symmetrical counterpart are dropped from the analysis. The remaining spatial locations are divided into two groups, corresponding to the upper and lower part of the plasma. The temperature measurement vector is partitioned accordingly: $(\boldsymbol{\mu}^t = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)^t)$, and the null-hypothesis to be tested is $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. If the profiles are up-down symmetric, then $\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$ has a multivariate normal distribution, whose covariance matrix is easily derived from the partitioned covariance matrix of $\boldsymbol{Y}$. Inserting the usual estimate for this covariance matrix, we find that

$$T = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^t(\boldsymbol{S}_{11} - \boldsymbol{S}_{12} - \boldsymbol{S}_{21} + \boldsymbol{S}_{22})^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) \tag{4.7}$$

is distributed as $p/(m-p)$ times the $F_{p,m-p}$ distribution. (Here, $\boldsymbol{S}_{11}$ stands for the residual SSCP matrix corresponding to $\boldsymbol{Y}_1$, $\boldsymbol{S}_{12}$ for the residual SSCP matrix 'between' $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, and so forth.)

Physically one expects that the electron temperature is constant on flux surfaces. Hence, for symmetric plasma discharges, the true, underlying radial profile should be up-down symmetric and an observed asymmetry should be due to an asymmetry in the measuring process. Supposing that the two types of asymmetry ('physical', and 'measurement error') are additive, the physical asymmetry of the density profile may be estimated by subtraction under the assumption that the density and the temperature profile exhibit the same total amount of asymmetry arising from measurement error, apart from possibly an additional asymmetry factor that can be estimated by other means.

4.3 Continuous Profile Representations

To obtain a physically meaningful continuous profile representation, a preliminary mapping of the physical measurement points r'_l to the corresponding flux radii, r_l must be performed. For relatively low beta, large aspect-ratio

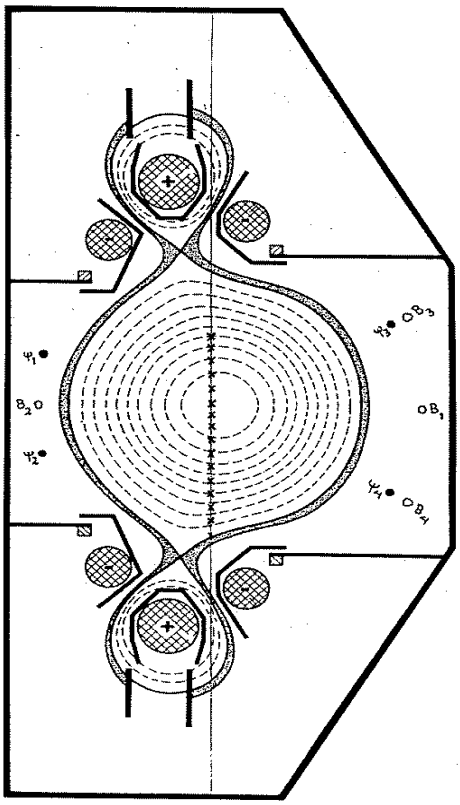

Fig. 4.2. Cross section of the ASDEX tokamak, drawing by courtesy of IPP Garching. The dotted contours are flux surfaces of the plasma equilibrium. The crosses denote the measurement positions of the 16 YAG–laser channels, 10 of which are approximately symmetrically located with respect to the midplane. Furthermore, one can see the separatrix (last closed flux–surface) with the X-points near the divertors and the structures in the two divertor regions. Note the up-down symmetry. The positions (outside the plasma) of the magnetic field coils and of the poloidal flux loops are denoted by $B_1, \ldots, B_4$ and $\psi_1, \ldots, \psi_4$, respectively, see also Sect. 7.8, which contains a similar cross section of ASDEX Upgrade.

devices, a simple calculation based on the Shafranov shift is sufficient. For smaller aspect ratios, a nonlinear correction (for instance based on function parameterisation) has to be used [83, 452].

We will not consider such corrections here, nor the possible errors associated with the flux–surface mapping.

For a complete specification of a statistical model for plasma profiles, we must specify both its mean–value structure (i.e., give a description of how the

mean values of the profiles depend on the input variables) and its covariance structure (i.e., present a stochastic model for the errors in the observations).

4.3.1 Mean Value Structures

For every discrete set of profile measurements, $\{T(r_l),\ l = 1, \ldots, p\}$, we seek a continuous representation of the profile. This has several advantages. The profile may be described by a relatively small number of coefficients. Furthermore, it facilitates comparison between profiles that are measured at different sets of radial points. Finally, smoothness is imposed in the belief that the profiles are in diffusive equilibrium.

Various transformations of the plasma profile may be considered. A natural transformation is to take logarithms of the temperature, density and pressure profiles. Minimising the error in the fit on logarithmic scale corresponds to minimising the relative error as opposed to the absolute error. If the relative error is more nearly constant over the database than the absolute error, then on logarithmic scale, in regressing the response variable (temperature, etc.) against the input variables, one can apply simple unweighted regression. Often, however, the absolute error in the experimental measurements increases whereas the relative error decreases with increasing value of the response variable. In that case, a weighted regression is needed on either scale. A logarithmic fit has the advantage that no restriction on the regression parameters is needed to ensure that the fitted temperature and density profiles are always positive. Furthermore, after logarithmic transformation, the concept of profile consistency can be defined as additivity of the radial dependence and the plasma–parameter dependence of the profiles.[10] Some additional conveniences in using a logarithmic scale come into effect when a power–law type scaling can be postulated. Firstly, a power–law scaling is transformed into a linear regression model. This facilitates the estimation procedure. Furthermore, taking logarithms makes the scaling law dimensionless: The usual transition from dimensional variables to dimensionless variables, see, e.g., [116], by taking products and quotients, corresponds to a linear transformation on logarithmic scale, possibly with reduction of the dimensionality by linear constraints. Finally, since on logarithmic scale the pressure is just the sum of the density and the temperature, one can easily obtain the full regression information by multivariate regression of only two of these three variables against the plasma parameters.

We restrict our attention to profile models with radial parametric dependencies of the form:

$$\mu(r) = \sum_{h=1}^{p'} \alpha_h f_h(r) \ , \tag{4.8}$$

[10] The notion of 'additivity' or 'non-interaction' is a well-known statistical concept, appearing in various contexts, see for instance Sect. 2.4, and, for a discussion with binary response variables, [52, 733].

where $f_1(r), \ldots, f_{p'}(r)$ are basis functions. For p radial measurement positions $r_1, \ldots, r_p$, we can write

$$\boldsymbol{\mu} = \boldsymbol{X}_{rad}\boldsymbol{\alpha} \,, \tag{4.9}$$

where $\boldsymbol{\mu} = \big(\mu(r_1), \ldots, \mu(r_p)\big)^t$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p'})^t$ is a vector of regression coefficients, and

$$\boldsymbol{X}_{rad} = \begin{pmatrix} f_1(r_1) & \ldots & f_{p'}(r_1) \\ \vdots & \ddots & \vdots \\ f_1(r_p) & \ldots & f_{p'}(r_p) \end{pmatrix} \tag{4.10}$$

is the radial design matrix. We now consider three possible sets of basis functions for continuous representations of the plasma profiles.

4.3.2 Polynomials

Clearly polynomials constitute a simple representation. Assuming symmetry and analyticity of the plasma profiles, we expand in even polynomials in r. Expanding the logarithm of the temperature, the representation is

$$T(r) = T_0 \exp\big(\sum_{n=1}^{\infty} a_n r^{2n}\big) \,. \tag{4.11}$$

The basis functions $(1, r^2, r^4, r^6, \ldots)$ have the property that the higher order polynomials are essentially localised near the outer edge of the plasma. Only a few terms are fitted in practice. It should be noted that these simple basis functions are highly non-orthogonal and hence their estimated coefficients are rather ill-determined and have relatively large standard deviations. This need not be a strong inconvenience if the emphasis is on predicting the profiles, and one is not especially interested in the regression coefficients themselves. The estimated coefficients of orthogonal polynomials are close to being statistically independent, at least if the radial positions are 'reasonably' located, and if, for ordinary least squares regression, the measurement errors are independent and radially uniform.

4.3.3 Perturbation Expansion

We write $T(r)$ as a Gaussian profile multiplied with a polynomial expansion in r. Allowing for asymmetry,

$$T(r) = T_o \exp(-d_o r^2)\big(1 + \sum_{n=1}^{\infty} b_n r^n\big) \,. \tag{4.12}$$

Such a representation may work conveniently whenever the plasma profiles differ little from Gaussians, in which case one hopes for a rapid convergence. Another possibility is to express the expansion by Hermite polynomials.

$$T(r) = T_o \exp(-c_o r^2)\big(1 + \sum_{n=1}^{\infty} a_n H_{n,c_0}(r)\big) \ , \tag{4.13}$$

where $H_{n,c_0} = e^{c_0 r^2}(-d/dr)^n e^{-c_0 r^2}$.

For each c_0 we have a complete set of basis functions. The choice of c_0 affects how many terms are needed. A convenient procedure is to estimate c_0 by fitting a Gaussian function to the temperature profile, and subsequently estimate the polynomial coefficients $a_1, a_2, \ldots$ for fixed c_0. Both steps can be carried out by applying linear regression. For any fixed number of coefficients $a_1, \ldots, a_m$, an optimal value of c_0 can be determined by minimising the residual sum of squares with respect to c_0. The correlation between the estimated polynomial coefficients will be reduced if the polynomials are orthogonalised on the finite interval [0,1], i.e when generalised Hermite polynomials are used. (Whether this orthogonalisation will be worth the effort in practice is another question.)

For any fixed value of c_0, the coefficients $a_1, a_2, \ldots$ are linearly related to the moments

$$m_n = \int r^n T(r) dr \ , \qquad n = 1, 2, \ldots \tag{4.14}$$

of the temperature distribution, see the Appendix in this chapter. The profiles are symmetric if the odd coefficients of the Hermite expansion, or, equivalently, the odd moments of the temperature, are zero. As the number of radial measurement positions is limited, only the lower moments can be tested. A convenient way is to test, e.g., $a_1 = a_3 = a_5 = 0$, directly from the asymptotic normal distribution of the fitted regression coefficients.

The kurtosis, K_u, is the normalised fourth central radial moment of the plasma profile. For a symmetric profile (for which $m_1 = 0$), it is defined [709] as

$$K_u = \frac{m_4/m_0}{(m_2/m_0)^2} \ . \tag{4.15}$$

The kurtosis, which is 3 for a Gaussian profile, is a measure of the 'broadness' or of lack of concentration near the axis of the profile, see Sect. 1.4. Expressed in the Hermitian regression coefficients, the excess of kurtosis, $K_u - 3$, equals $\frac{6a_4 - 3a_2^2}{(a_2 + \frac{1}{4c_0})^2}$. In our situation, the kurtosis may for instance be used in scaling studies of density–profile effects on the energy confinement time.

4.3.4 Splines

The physical domain of the plasma is subdivided into a number of regions and low order polynomials are used to represent the profile, i.e., the logarithm of the temperature, density or pressure, in each region. As a general spline representation, we consider

$$\mu(r) = \begin{cases} \varphi_0(r) & \text{for } 0 \le r < r_1 \quad \text{(Inner Region)} \\ \varphi_0(r) + \varphi_1(r) & \text{for } r_1 \le r < r_2 \quad \text{('1-2' Region)} \\ \varphi_0(r) + \varphi_1(r) + \varphi_2(r) & \text{for } r_2 \le r < r_3 \quad \text{('2-3' Region)} \\ \quad \cdots & \\ \varphi_0(r) + \varphi_1(r) + \varphi_2(r) + \cdots + \varphi_l(r) & \text{for } r_l \le r \le 1 \quad \text{(Outer Region)} \end{cases} \tag{4.16}$$

where $\varphi_0(r) = \mu(0) + (-1)^s a_0 r + b_0 r^2 + c_0 r^3$ and $\varphi_j(r) = b_j(r - r_j)^2 + c_j(r - r_j)^3$ for $j = 1, 2, \ldots, l$.

The representation is understood to be on the interval $[0, 1]$, after the negative radial positions have been reflected. The linear term on $\varphi_0(r)$ is intended to describe a possible, simple asymmetry. Hence $s = 0$ for, say, the positive radial positions and $s = +1$ for the negative radial positions. Clearly, $2b_j$ represents the jump of the second derivative, and $6c_j$ the jump of the third derivative at the jth knot. If the spline is supposed to be twice continuously differentiable, then the coefficients b_j are constrained to be zero. We will call this a 'second-order spline'. For a so-called Hermitian spline, the coefficients b_j are arbitrary. The order of continuity imposed at a knot position will be called the 'continuity index' of the knot, and the regions between the knots ('0-1', '1-2', etc.) will be called the 'knot regions'. Note that the spline coefficients $\mu(0), a_0, b_j, c_j$ for $j = 1, 2, \ldots, l$ occur as linear parameters in this representation, which permits a simple direct estimation. The spline model is intended to represent only the global radial profile behaviour. Hence, the discrepancies between the observed and predicted smooth profile are to be attributed to model misspecification, experimental error and plasma profile fluctuations. The above model is quite general. Decisions have to be made about four (interrelated) aspects of a further model specification.

(1) Choosing the number of knots;

(2) choosing the continuity indices of the knots;

(3) choosing the polynomial degrees in the knot regions;

(4) choosing the knot positions.

We use the following notation (which reflects the first 3 aspects). A 2.3.2 spline model has two knots, continuity up to the first derivative at these knots, and a 2^{nd}, a 3^{rd} and a 2^{nd} degree polynomial in the 3 regions, respectively. A second-order spline with the same number of knots and the same polynomial degrees, but second-order continuity, is denoted by 2:3:2, etc. Note that requiring a continuity index of k for a knot between two polynomials of degree less or equal to k, corresponds to removal of the knot.

(1) The choice of the number of knots is influenced by the number of available radial measurement positions, the noise level, and the number of profiles available. Knots are needed where sudden changes in the second derivative of the profile are expected. Usually, one should have at least one measurement position between two knots to avoid collinearity problems. The lower the noise level, and, more or less equivalent, the larger the number of profiles

that are simultaneously fitted, the larger the number of knots that can be used. The maximal number of knots is fixed by the requirement that the number of fitted parameters should not exceed the total number of distinct radial positions. For the 16 channel ASDEX YAG–laser measurements, 2 to 5 knot spline models have been extensively investigated [451].

(2) If one believes that diffusion must make the profiles twice differentiable, one can impose second-order splines. We note that given the sharp discontinuity of the profiles after a sawtooth crash, one should probably not assume a high degree of smoothness near the sawtooth inversion layer. Hence, requiring first order continuity near the sawtooth inversion radius, and second-order continuity away from the inversion radius seems to be a plausible choice.

(3) Traditionally, one uses third degree polynomials. For regularisation, i.e., to avoid 'unphysical' wild behaviour, a boundary condition at the edge (for instance, $\mu''(1) = 0$) may be imposed. This reduces the effective degree in the outer region by 1. Second order splines with this boundary condition are historically called 'natural splines'. The following interrelationship between the various aspects of model specification is noted. If one wants to keep the same flexibility, then increasing a continuity constraint should be accompanied by increasing a nearby polynomial degree or the number of knots. If, for example, a 2.3.2 model is considered to more or less adequate, except for its discontinuity of the second derivative, then natural alternatives would be 'natural' 3:4:2, or 3:3:3:2, or 2:3:3:3:2 splines. Obviously, increasing the number of knots complicates the problem of finding the most suitable knot positions.

(4) A physical approach is to divide the tokamak into three regions, a sawtooth region, a "confinement" region and an edge region. A natural choice would be to choose r_a near the sawtooth inversion radius. The exact sawtooth inversion radius, a regressed approximation, or the simple empirical relation $r_{inv} = 1/q_{eng}$ may be used. The other knot positions may be chosen such that the measurement positions are roughly equally distributed in the various regions. This 'natural choice' is, however, open to the objection that the 'built-in' correlation with q_{eng} may complicate the task of determining the full dependence of the profile shape on q_{eng}. Another approach is to consider knot positions $r_1, r_2, \ldots$ as (nonlinear) free parameters which are estimated by numerically minimising the residual sum of squares. Some preference should be expressed, however, for manually varying the knot positions, using the physical considerations mentioned above, and examining the sensitivity of the fit on the choice of the knots, by investigating the residuals (in particular the residual sum of squares) for various alternative fits.

An elaboration of the 2.3.2 Hermitian spline model, applied to ASDEX profile analysis, was given in [347]. In [451] more extensive investigations are made using up to 5 knot, second-order splines. As we shall see in Sect. 4.6, one can formally apply an F–statistic to test, on the basis of n experimental

profiles, whether the addition of an extra knot or, more generally, decreasing the continuity index at some radial position(s), leads to a statistically significant reduction of the residual sum of squares. The significance level of such an F test, as well as the precise fit of the final model, depend, however, on the assumed error structure. For a discussion on adaptive procedures to fit polynomial splines in one and more dimensions, which use automatic knot selection based on essentially the F–statistic, and on other smoothing methods, such as those based on penalty functionals, see Sect. 3.5.4, the reader is referred to [285, 650].

4.3.5 Error Structures

Combining the mean–value structure, described by (4.9) with the assumption of a multivariate normal distribution of the deviations, as described by (4.4), we write

$$\boldsymbol{Y}_j = \boldsymbol{X}_{rad}\boldsymbol{\alpha} + \boldsymbol{E}_j \;, \qquad j = 1, \ldots, m \;, \tag{4.17}$$

where $\boldsymbol{E}_j \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{X}_{rad}$ is given by (4.10). We will consider the following models with respect to the error structure:

Model I: $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$
Model II: $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{W}_d + \boldsymbol{X}_{rad} \boldsymbol{\Lambda}\, \boldsymbol{X}_{rad}^t$,
where $\boldsymbol{W}_d$ is a known diagonal matrix, and $\boldsymbol{\Lambda}$ is an arbitrary $p' \times p'$ covariance matrix,
Model III: $\boldsymbol{\Sigma}$ arbitrary.

Model I is the simplest case, and corresponds to the usual assumption made in justifying the use of ordinary least-squares regression. This simple model for the error structure may not be a good approximation to reality, however.

Model III is the most general of the above three models. It has $\frac{1}{2}p(p+1)+p'$ free parameters, which may be a rather large number to estimate in practice. Hence, there is some need for a realistic model of the error structure, but with less degrees of freedom than model III.

Model II may be such a model with $\frac{1}{2}p'(p'+1)+1+p'$ free parameters. It has the following physical interpretation: $\sigma^2 \boldsymbol{W}_d$ represents the independent measurement errors (arising from detector noise, etc.), whose relative magnitude can sometimes be estimated (independently from the profile data) by analysing the propagation of errors of the measurement process. The second term in Model II represents global variations of the plasma profile, which, by energy and particle conservation, cannot be independent. It is noted that II may looked upon as a random coefficient model, which can be written as

$$\boldsymbol{Y}_j = \boldsymbol{X}_{rad}\boldsymbol{A} + \boldsymbol{E}_j \;, \qquad j = 1, \ldots, m \;, \tag{4.18}$$

where $\boldsymbol{A} \sim N_{p'}(\boldsymbol{\alpha}, \boldsymbol{\Lambda})$, and $\boldsymbol{E}_j \sim N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{W}_d)$. So, correlated large-scale plasma variations are modelled by assuming that the underlying profiles vary

globally, on a suitable timescale, according to an arbitrary multivariate normal distribution of the regression parameters.

Models I, II, and III are special cases of the covariance structure

$$\boldsymbol{\Sigma} = \sum_{i=1}^{f} \theta_i \boldsymbol{G}_i \ . \tag{4.19}$$

Estimation theory for the model given by (4.16) with the covariance structure expressed by (4.18) has been developed in [11].

The flux–surface mappings, and hence the radial design matrices $\boldsymbol{X}_{rad}$ are not precisely the same for the various profiles. Hence, a refinement of (4.17) is

$$\boldsymbol{Y}_j = \boldsymbol{X}_{rad,j}\boldsymbol{\alpha} + \boldsymbol{E}_j \quad (j = 1, \ldots, m) \ . \tag{4.20}$$

As we will see in the next section, this model formulation is also suitable to describe the more general situation where the plasma profiles depend on the plasma variables.

The above described models of random errors are primarily adequate for describing statistical measurement noise. With some stretch of their original function, they may, however, also be used to describe quasi-random measurements, such as those associated with sawtooth activity. To see this, we present a simple model for the sawtooth activity inside the sawtooth inversion radius r_{inv}. We assume that the temperature is flat just after the sawtooth crash and grows linearly between the crashes. Thus $T(r,t) = T(0) + a_2(r_{inv}^2 - r^2)(\frac{t}{\tau})$ for $r \leq r_{inv}$. Assuming that the time sampling frequency is incommensurate with the sawtooth period, we have the random coefficient model

$$T(r,t) = T(0) + A_2(r_{inv}^2 - r^2) \ , \tag{4.21}$$

where $A_2 \sim U(0, a_2)$, i.e., A_2 is uniformly distributed on the interval $(0, a_2)$. Clearly, $2A_2$, which can be interpreted as the random curvature, has mean a_2 and standard deviation $a_2/\sqrt{3}$. (The average curvature a_2 has the interpretation of being the sawtooth amplitude in the center divided by r_{inv}^2.) Note that this sawtooth model assumes a (simple) spline for the temperature itself, instead of for the logarithm. Furthermore, the spline coefficient has a uniform instead of a normal distribution. The practical implications of these differences are likely to be small. It is noted that the sawtooth activity is confined to the inner region of the plasma. If this activity is the dominant source of plasma variation, then the matrix $\boldsymbol{\Sigma}$ in the above discussed models will be nearly singular. In model II this is modelled by a small contribution of $\sigma^2 \boldsymbol{W}_d$ and zero elements for $\boldsymbol{\Lambda}$, except for the row(s) and column(s) corresponding to the regression coefficient of $\mu'(r_1)$, where r_1 is near the sawtooth inversion radius.

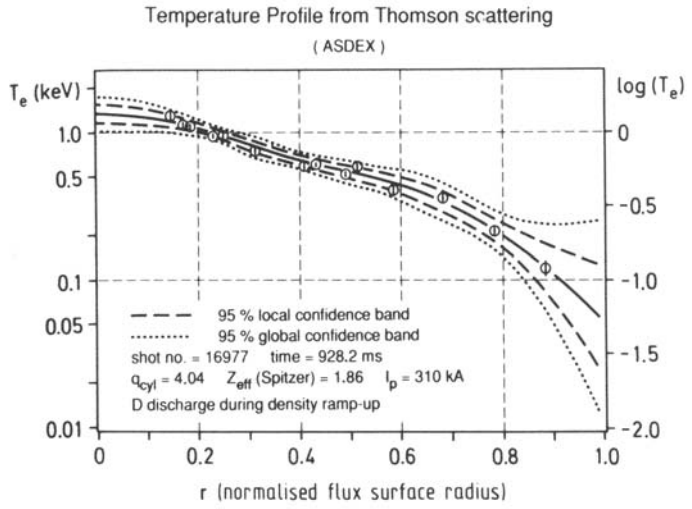

Fig. 4.3. Illustration of temperature profile from the ASDEX tokamak, fitted by a Hermitian spline model to measurements of the Nd:YAG–laser Thomson scattering diagnostic for an ohmic discharge. The average fitted profile as well as a 95% local and a 95% global confidence band are shown. The figure is reproduced from [344], by courtesy of IPP Garching.

4.4 Profile Dependence on Plasma Parameters

4.4.1 Mean Value and Error Structures

In principle, the profile shapes can be determined separately for each value of the plasma parameters $\boldsymbol{x}$. However, we are interested in parametric dependencies of the profile shapes in an entire domain of the plasma parameters. Initial statistical analyses should concentrate on relatively small domains of parameter space where one physical phenomenon is expected to be dominant. Such domains may be experimentally delineated. One can distinguish, for instance, between the 'standard' L–mode regime, the runaway electron regime, the density rollover regime, the ELM-free H–mode regime, type I, II, III ELMy H–mode regime, etc. Theoretical parameter domains may be characterised by local instability boundaries for the η_i and ballooning modes. After the scaling in these 'single phenomenon' regions has been established, the transition regions may be more easily explored. Two examples of scalings devised to describe the energy confinement of both the Ohmic and the auxilary heating regime are $\tau_E^{-2} = \tau_{E,ohm}^{-2} + \tau_{E,aux}^{-2}$ or $\tau_E = \frac{P_{ohm}^0}{P_L}\tau_{E,ohm} + \left(1 - \frac{P_{ohm}^0}{P_L}\right)\tau_{E,inc}$,

where P^0_{ohm} is the Ohmic heating power before the additional heating is turned on and P_L is the total heating power during the additional heating phase, see [233] and [616, 622]. In another example, physical models are described related to density roll-over associated with current-driven and collisional modes, respectively, applied to an early, versatile Japanese experiment [300], [107] containing a published dataset.

We now formulate approaches to determine the dependencies of the profiles on the plasma variables. Assuming that the coefficients α_j in (4.8) are functions of the plasma variables $\boldsymbol{x}$, we represent the profiles by $Y(r, \boldsymbol{x}) = \mu(r, \boldsymbol{x}) + E(r, \boldsymbol{x})$, where

$$\mu(r, \boldsymbol{x}) = \sum_{h=1}^{p'} \alpha_h(\boldsymbol{x}) f_h(r) = \sum_{h=1}^{p'} \sum_{k=0}^{w} \alpha_{h,k} g_k(\boldsymbol{x}) f_h(r) \ , \tag{4.22}$$

and $E(r, \boldsymbol{x})$ describes the measurement errors and plasma fluctuations. A simple, but natural choice for the plasma basis functions, at least in a single domain of parameter space where the sawtooth averaged profiles are expected to vary smoothly with the plasma parameters, is $g_o(\boldsymbol{x}) = 1$ and $g_k(\boldsymbol{x}) = \ln(x_k/x^*_k)$, where x^*_k is a typical value of x_k in the database of interest ($k = 1, \ldots, w$). In a more extensive investigation one can, just as for the radial dependence, consider polynomial or spline models in $\ln(x_k/x^*_k)$.

Given m profile measurements at p radial positions, the discrete analogue of (4.22) can be written as

$$\boldsymbol{Y} = \boldsymbol{X}_{rad} \boldsymbol{\alpha} \boldsymbol{X}_{cov} + \boldsymbol{E} \ , \tag{4.23}$$

where $\boldsymbol{Y}$ is a $p \times m$ data matrix, $\boldsymbol{E}$ is a $p \times m$ error matrix, and $\boldsymbol{\alpha}_{mat}$ is the $p' \times w$ matrix of regression coefficients. Note that $f_h(r)$ is represented by the hth column of the fixed radial design matrix $\boldsymbol{X}_{rad}$, and $g_k(\boldsymbol{x})$ by the kth row of the 'covariate design matrix' $\boldsymbol{X}_{cov}$. This model is sometimes called the Potthoff–Roy model [522], see also [405].

Alternatively, the regression model can be written as

$$\boldsymbol{Y}_j = \boldsymbol{X}_j \boldsymbol{\alpha} + \boldsymbol{E}_j \quad (j = 1, \ldots, m) \ . \tag{4.24}$$

Here, $\boldsymbol{Y}_j$ is the $p \times 1$ vector for the jth profile, $\boldsymbol{\alpha}$ is the vector of regression coefficients (obtained by vertical concatenation of the columns of $\boldsymbol{\alpha}_{mat}$), and $\boldsymbol{X}_j = \boldsymbol{X}^t_{cov,j} \otimes \boldsymbol{X}_{rad}$ is the tensor product of the transpose of the jth column of $\boldsymbol{X}_{cov}$ with $\boldsymbol{X}_{rad}$. ($\boldsymbol{X}^t_{cov,j}$ contains typically the logarithms of plasma variables for the jth profile. By definition, for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $(\boldsymbol{A} \otimes \boldsymbol{B})_{i,j} = A_{i,j}\boldsymbol{B}$.)

With respect to the errors, we assume that $\boldsymbol{E}_1, \ldots, \boldsymbol{E}_m$ are independent and $\boldsymbol{E}_j \sim N_p(0, \boldsymbol{\Sigma}_j)$. In the simplest case one has $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ satisfies model I, II, or III of Sect. 4.3.5. For models I and III, this assumption of 'homoskedasticity' can be tested using Bartlett's modification

of the likelihood ratio test (see Chap. 10 of [14]), which is a multivariate version of Bartlett's statistic for the homogeneity of variances, see Sect. 2.3. The reader will have noticed that (4.24) has the same structure as (4.20), which described the situation with fixed plasma parameters, but different radial design matrices. Hence, the same techniques of estimation, testing and confidence intervals can be applied.

The covariance structure may also depend on the plasma parameters. A natural analogue of model I^a is that, using continuous notation, $E(r, \boldsymbol{x})$ is a Gaussian random field with

$$\log \sigma(r, \boldsymbol{x}) = c_1 + c_2 \mu(r, \boldsymbol{x}) \; . \tag{4.25}$$

Quite generally, one could model $\log \sigma(r, \boldsymbol{x}, r', \boldsymbol{x}')$ by bilinear expressions of the basis functions in (4.22). Such general models are, of course, rather difficult to estimate in practice, and some simplifying assumptions have to be made to reduce the number of free parameters. In such a situation, it is useful to have a repeated measurement design (i.e., many profile measurements for fixed values of r and $\boldsymbol{x}$), which allows a separate estimation of the parameters of the mean–value structure and of the error structure.

4.4.2 Profile Invariance

The concept of profile invariance can be formulated as follows. A family of plasma profiles is invariant with respect to some plasma variable if the profile shape, i.e. $L^{-1}(r, \boldsymbol{x}) = \frac{\partial}{\partial r} \mu(r, \boldsymbol{x})$, is independent of that plasma variable. From (4.22), with $g_0(\boldsymbol{x}) = f_1(r) = 1$, it follows that

$$L^{-1}(r, \boldsymbol{x}) = \sum_{h>1} \alpha_{h,0} f_h'(r) + \sum_{h>1, k>0} \alpha_{h,k} f_h'(r) g_k(\boldsymbol{x}) \; . \tag{4.26}$$

Profile invariance holds for all plasma variables if the second term in this expression is 0, i.e. if all elements of $\boldsymbol{\alpha}_{mat}$ in (4.23), except those of the first row and the first column, vanish.

Similarly, one can consider the more general situation of profile invariance, with respect to some plasma variable (say, q_{eng}), on an interval $(r_{min}, r_{max}) \subseteq [0, 1]$. Now, the condition is

$$\sum_{h>1}^{p'} \alpha_{h,k} f_h'(r) = 0 \tag{4.27}$$

for all $r \in (r_{min}, r_{max})$ and for all k's that correspond to that plasma variable (e.g., $g_1(\boldsymbol{x}) = \log q_{eng}$, $g_2(\boldsymbol{x}) = (\log q_{eng})^2$, etc.). When a cubic spline model for $\mu(r, \boldsymbol{x})$ is used, this condition can be translated into the requirement that $i+2$ linear combinations of $\alpha_{2,k}, \ldots, \alpha_{p',k}$ are zero if i knot regions are needed to cover the interval (r_{min}, r_{max}).

Such a hypothesis can statistically be tested as soon as the distribution of the estimate of $\boldsymbol{\alpha}_{mat}$ has been derived. This route will be undertaken in Sects. 4.5 and 4.6. Alternatively, the hypothesis can conveniently be tested from a global confidence band for the derivative of $L^{-1}(r, \boldsymbol{x})$ with respect to q_{eng}, see Sect. 4.7. For a practical case study in which such confidence bands are used, the reader is referred to [451].

4.5 Estimation of Regression Coefficients

4.5.1 Least Squares and Maximum Likelihood

Initially, we consider only estimates for the model $\boldsymbol{Y}_j = \boldsymbol{X}_{rad}\boldsymbol{\alpha} + \boldsymbol{E}_j, j = 1, \ldots, m$, where $\boldsymbol{X}_{rad}$ is a *fixed* $p \times p'$ radial design matrix, and $\boldsymbol{E}_j \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ satisfies one of the error structures I, II, III, described in Sect. 4.3. Then, we make some remarks on estimates for the more general situation, described in Sect. 4.4, of different radial design matrices.

Let $\hat{\boldsymbol{\mu}} = m^{-1} \sum_j \boldsymbol{Y}_j$ denote the empirical average of the m profiles. It can easily be derived that for any $p \times p$ non-singular matrix, $\boldsymbol{G}$ such that $\boldsymbol{X}_{rad}^t \boldsymbol{G} \boldsymbol{X}_{rad}$ is also non-singular,

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{Q}\hat{\boldsymbol{\mu}} \,, \tag{4.28}$$

with

$$\boldsymbol{Q} = \left(\boldsymbol{X}_{rad}^t \boldsymbol{G}^{-1} \boldsymbol{X}_{rad}\right)^{-1} \boldsymbol{X}_{rad}^t \boldsymbol{G}^{-1} \tag{4.29}$$

is a normally distributed, linear, unbiased (i.e., $E(\hat{\boldsymbol{\alpha}}) = \boldsymbol{\alpha}$) estimator for $\boldsymbol{\alpha}$. The covariance matrix of $\hat{\boldsymbol{\alpha}}$ is

$$\boldsymbol{V}(\hat{\boldsymbol{\alpha}}) = \frac{1}{m} \boldsymbol{Q} \boldsymbol{\Sigma} \boldsymbol{Q}^t \,. \tag{4.30}$$

Such estimators for $\boldsymbol{\alpha}$ are called generalised least-squares estimators. Notice that the choice $\boldsymbol{G} = c\boldsymbol{I}$ corresponds to ordinary least squares regression, and $\boldsymbol{G}$ diagonal to weighted least squares regression, the weights being inversely proportional to the diagonal elements of $\boldsymbol{G}$. Note that by substitution of $\hat{\boldsymbol{\alpha}}$ into (4.8), a parametric estimate of the plasma profile is obtained. Equations (4.28) and (4.29) can, for symmetric and positive definite $\boldsymbol{G}$, be interpreted by the fact that the vector $\hat{\hat{\boldsymbol{\mu}}} = \boldsymbol{X}_{rad}\hat{\boldsymbol{\alpha}}$ is the projection of $\hat{\boldsymbol{\mu}}$ on the linear subspace generated by the columns of $\boldsymbol{X}_{rad}$, using the inner product defined by $\boldsymbol{G}$.

If $\boldsymbol{\Sigma}$ is known, then among all possible matrices $\boldsymbol{G}$, the optimal choice is $\boldsymbol{G} = \boldsymbol{\Sigma}$, since it gives the estimator that maximises the likelihood and has minimal covariance matrix among *all unbiased* estimators for $\boldsymbol{\alpha}$. The variance of this estimator reduces to

$$\boldsymbol{V}(\hat{\boldsymbol{\alpha}}) = \frac{1}{m} \left(\boldsymbol{X}_{rad}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_{rad}\right)^{-1} \,. \tag{4.31}$$

In practice, $\boldsymbol{\Sigma}$ is unknown and must be estimated as well. The 'best' estimator for $\boldsymbol{\Sigma}$, and the simplest way to calculate this estimator, depends on the assumed error structure.

Model I, corresponding to ordinary least squares regression, is the simplest case. Estimates for σ^2 are based on the residual sum of squares,

$$\hat{\sigma}^2 = \frac{1}{mp - p'} \sum_{j=1}^{m} (\boldsymbol{Y}_j - \boldsymbol{X}_{rad}\hat{\boldsymbol{\alpha}}_I)^t (\boldsymbol{Y}_j - \boldsymbol{X}_{rad}\hat{\boldsymbol{\alpha}}_I) \tag{4.32}$$

being the minimum variance unbiased estimator for σ^2, which is distributed as $\sigma^2/(mp - p')$ times $\chi^2_{mp-p'}$.

In model III an iterative procedure might seem to be needed, to solve simultaneously

$$\boldsymbol{\Sigma} = \frac{1}{m-1} \sum_{j=1}^{m} (\boldsymbol{Y}_j - \boldsymbol{X}_{rad}\boldsymbol{\alpha})(\boldsymbol{Y}_j - \boldsymbol{X}_{rad}\boldsymbol{\alpha})^t \tag{4.33}$$

and (4.28) – (4.29), starting for example with the unweighted least squares estimate for $\boldsymbol{\alpha}$. However, it can be proven (see, e.g., [381] [326]) that one gets directly the ML estimate $\hat{\boldsymbol{\alpha}}_{III}$ by inserting for $\boldsymbol{G}$ in (4.29) the estimated variance in the unsmoothed model, i.e., $\hat{\boldsymbol{\Sigma}} = m^{-1}\boldsymbol{S}$ as given by (4.6). The adjusted ML estimate of $\boldsymbol{\Sigma}$ in the smoothed model is then obtained by inserting $\hat{\boldsymbol{\alpha}}_{III}$ into (4.33). (Adjusted, because in the denominator $m - 1$ has been used instead of m.)

An elegant generalisation exists for the situation with covariates, as expressed by (4.23). As shown in [250, 381], among others,

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}_{rad}\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_{rad}^t)^{-1}\boldsymbol{X}_{rad}^t\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y}\boldsymbol{X}_{cov}^t(\boldsymbol{X}_{cov}\boldsymbol{X}_{cov}^t)^{-1} , \tag{4.34}$$

with

$$\widetilde{\boldsymbol{\Sigma}} = f^{-1}\boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{X}_{cov}^t\big(\boldsymbol{X}_{cov}\boldsymbol{X}_{cov}^t)^{-1}\boldsymbol{X}_{cov}\big)\boldsymbol{Y}^t , \tag{4.35}$$

and $f = m - w - 1 - (p - p')$, is in that case the maximum likelihood estimator (the constant f was chosen to simplify the next formula). An unbiased estimate of its covariance matrix is given by [250]

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}}) = (\boldsymbol{X}_{cov}\boldsymbol{X}_{cov}^t)^{-1} \otimes (\boldsymbol{X}_{rad}^t\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_{rad})^{-1}(m - w - 2)/(f - 1) , \tag{4.36}$$

where $\boldsymbol{\alpha}$ is the vertical concatenation of the columns of $\boldsymbol{\alpha}_{mat}$, and $\otimes$ denotes the tensor product.

The maximum likelihood estimates are asymptotically (i.e., as $m \to \infty$) normal and efficient. It is noted that $\hat{\boldsymbol{\Sigma}}$ may deviate considerably from $\boldsymbol{\Sigma}$, if the total number, mp, of observations is not much larger than the total number of free parameters, $\frac{1}{2}p(p+1) + p$ in the unsmoothed model. This leads, then, to an inefficient estimate for $\boldsymbol{\alpha}$ which is, in addition, not normally distributed. In such a case, it may be wise to use a model with fewer parameters.

We now consider model II in the simple case of a fixed design matrix with no covariates. If its special covariance structure is inserted for $\boldsymbol{G}$, one can derive [24, 536, 583, 660] that (4.28)–(4.30) reduce to

$$\hat{\boldsymbol{\alpha}}_{II} = \left(\boldsymbol{X}_{rad}^t \boldsymbol{W}_d^{-1} \boldsymbol{X}_{rad}\right)^{-1} \boldsymbol{X}_{rad}^t \boldsymbol{W}_d^{-1} \hat{\boldsymbol{\mu}} \tag{4.37}$$

and

$$\boldsymbol{V}(\hat{\boldsymbol{\alpha}}_{II}) = \frac{1}{m}\left(\sigma^2 \left(\boldsymbol{X}_{rad}^t \boldsymbol{W}_d^{-1} \boldsymbol{X}_{rad}\right)^{-1} + \boldsymbol{\Lambda}\right) . \tag{4.38}$$

Note that the estimator $\hat{\boldsymbol{\alpha}}_{II}$ does not depend at all on the parameters $\boldsymbol{\Lambda}$ and σ^2. Of course, the covariance matrix of $\hat{\boldsymbol{\alpha}}_{II}$ depends on these parameters, but it does so in a simple way. Equation (4.37) requires the inversion of a sometimes considerably smaller matrix than in the matrix $\boldsymbol{G}$ in (4.28) and (4.29).

In Model II, (4.37) was constructed to be the minimum–variance unbiased estimator for $\boldsymbol{\alpha}$. We now have to find an estimator for its variance $\boldsymbol{V}(\hat{\boldsymbol{\alpha}}_{II})$. It can be derived (see, e.g., [426, 519] that, if $\boldsymbol{W}_d$ is known,

$$\hat{\boldsymbol{\Sigma}}_{II} = \frac{1}{m-1} \sum_{j=1}^{m} (\boldsymbol{Y}_j - \boldsymbol{X}_{rad}\hat{\boldsymbol{\alpha}}_{II})(\boldsymbol{Y}_j - \boldsymbol{X}_{rad}\hat{\boldsymbol{\alpha}}_{II})^t \tag{4.39}$$

is the minimum variance unbiased estimator for $\boldsymbol{\Sigma}$. Inserting $\hat{\boldsymbol{\Sigma}}_{II}$ and

$$\boldsymbol{Q}_{II} = \left(\boldsymbol{X}_{rad}^t \boldsymbol{W}_d^{-1} \boldsymbol{X}_{rad}\right)^{-1} \boldsymbol{X}_{rad}^t \boldsymbol{W}_d^{-1} , \tag{4.40}$$

into (4.30), one gets the minimum variance unbiased estimator for $\boldsymbol{V}(\hat{\boldsymbol{\alpha}}_{II})$, which will be denoted by $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}}_{II})$.

A particularly nice feature is that $\hat{\boldsymbol{\alpha}}_{II}$ can be looked upon as the sample mean of the estimated regression coefficients of the individual profile fits, and that $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}}_{II})$ can be rewritten as the empirical covariance matrix of this sample mean, i.e.,

$$\hat{\boldsymbol{\alpha}}_{II} = \frac{1}{m} \sum_{j=1}^{m} \hat{\boldsymbol{\alpha}}_{jII} , \tag{4.41}$$

and

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}}_{II}) = \frac{1}{m(m-1)} \sum_{j=1}^{m} (\hat{\boldsymbol{\alpha}}_{jII} - \hat{\boldsymbol{\alpha}}_{II})(\hat{\boldsymbol{\alpha}}_{jII} - \hat{\boldsymbol{\alpha}}_{II})^t , \tag{4.42}$$

where $\hat{\boldsymbol{\alpha}}_{jII} = \boldsymbol{Q}_{II}\boldsymbol{Y}_j$.

Notice that one can construct separate estimators for σ^2 and $\boldsymbol{\Lambda}$. The usual estimator for σ^2 equals the weighted sample average of the squared residuals from the fitted individual profiles, with a correction for the fact that mp' parameters are estimated, i.e.,

$$\hat{\sigma}^2 = \frac{1}{m(p-p')} \sum_{j=1}^{m} (\boldsymbol{Y}_j - \boldsymbol{X}_{rad}\hat{\boldsymbol{\alpha}}_{jII})^t \boldsymbol{W}_d^{-1} (\boldsymbol{Y}_j - \boldsymbol{X}_{rad}\hat{\boldsymbol{\alpha}}_{jII}) . \tag{4.43}$$

Subsequently, $\boldsymbol{\Lambda}$ may be estimated by the relation

$$\hat{\sigma}^2 \left(\boldsymbol{X}_{rad}^t \boldsymbol{W}_d^{-1} \boldsymbol{X}_{rad}\right)^{-1} + \hat{\boldsymbol{\Lambda}} = m\hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}}_{II}) \;, \tag{4.44}$$

which follows directly from (4.38). This estimator for $\boldsymbol{\Lambda}$ has the disadvantage that, because of the subtraction, it may not be positive definite. If this occurs, one can consider (1) to reformulate the model, assuming that at least some elements of $\boldsymbol{A}$ are non-random, and (2) to estimate $\boldsymbol{\Lambda}$ and σ^2 numerically by maximum likelihood. It is stressed that in order for (4.37) and (4.42) to be sensible estimators, the weights associated with $\boldsymbol{W}_d$ must be known, or at least be independently estimable.

Since $\hat{\boldsymbol{\alpha}}_{II}$ is the UMVU estimator in a linear model, it has the property that for every $\boldsymbol{f}(r), \hat{\boldsymbol{\alpha}}_{II}^t \boldsymbol{f}(r)$ is the minimum–variance unbiased estimate for $\boldsymbol{\alpha}^t \boldsymbol{f}(r)$. This property is useful for predicting future profiles, see Sect. 4.7. If $\boldsymbol{Q}_w = \boldsymbol{X}_{rad}^t \boldsymbol{W}_d^{-1} \boldsymbol{X}_{rad}$ is rather close to a singular matrix, the restriction to unbiasedness leads to large variances, however, and one may be inclined to allow for some bias in order to get a substantial reduction in variance. Procedures to do this are ridge regression, see 3.4, which replaces $\boldsymbol{Q}_w$ by $\boldsymbol{Q}_w + \boldsymbol{R}$ for some positive definite $\boldsymbol{R}$ (frequently, $\boldsymbol{R} = k\boldsymbol{I}$), and latent root regression [257, 724] which discards small principal components of $\boldsymbol{Q}_w$ that are, in addition, uninformative for the regression. In the context of model II, empirical Bayes estimates have been developed [537], see also [651], that minimise the mean squared error of prediction (i.e., the variance plus the square of the bias), summed over the profiles in the data base. Here, we will restrict attention to ML and UMVU estimates, which is reasonable as long as the matrix $\boldsymbol{Q}_w$ is rather well conditioned.

Now we will discuss estimation procedures for the general case of unequal design matrices, where

$$\boldsymbol{Y}_j = \boldsymbol{X}_j \boldsymbol{\alpha} + \boldsymbol{E}_j \quad (j = 1, \ldots, m) \;, \tag{4.45}$$

and $\boldsymbol{E}_j \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$. For any set of non-singular matrices $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_m$, such that $\sum_j \boldsymbol{X}_j^t \boldsymbol{G}_j^{-1} \boldsymbol{X}_j$ is invertible, an unbiased estimator for $\boldsymbol{\alpha}$ is given by

$$\hat{\boldsymbol{\alpha}} = \Big(\sum_{j=1}^{m} \boldsymbol{X}_j^t \boldsymbol{G}_j^{-1} \boldsymbol{X}_j\Big)^{-1} \sum_j \boldsymbol{X}_j^t \boldsymbol{G}_j^{-1} \boldsymbol{Y}_j \;. \tag{4.46}$$

This property holds independent of the error structure of $\boldsymbol{E}_j$, the only condition being that $\boldsymbol{E}_j$ has expectation zero for $j = 1, \ldots, m$. Note that, although each $\boldsymbol{X}_j^t \boldsymbol{G}_j^{-1} \boldsymbol{X}_j$ may be singular, $\hat{\boldsymbol{\alpha}}$ can formally be regarded as a weighted average of the individual least-squares estimates $\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_m$, where each $\hat{\boldsymbol{\alpha}}_j$ is weighted according to $\boldsymbol{X}_j^t \boldsymbol{G}_j^{-1} \boldsymbol{X}_j$.

For known $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m$, the choice $\boldsymbol{G}_j = \boldsymbol{\Sigma}_j$ leads to the minimum variance unbiased estimator, which in this case coincides with the maximum likelihood (ML) estimator. Furthermore, in that case, $\boldsymbol{X}_j^t \boldsymbol{\Sigma}_j^{-1} \boldsymbol{X}_j$ equals the inverse of the covariance matrix of $\hat{\boldsymbol{\alpha}}_j$, and

$$\boldsymbol{V}(\hat{\boldsymbol{\alpha}}) = \big(\sum_j \boldsymbol{X}_j^t \boldsymbol{\Sigma}_j^{-1} \boldsymbol{X}_j\big)^{-1} . \tag{4.47}$$

In the random coefficient model, i.e., if $\boldsymbol{\Sigma}_j = \sigma^2 + \boldsymbol{X}_j^t \boldsymbol{\Lambda} \boldsymbol{X}_j$, insertion of $\boldsymbol{G}_j = \boldsymbol{\Sigma}_j$ in (4.46) leads to

$$\hat{\boldsymbol{\alpha}}_{II} = \big(\sum_j \boldsymbol{H}_j\big)^{-1} \sum_j \boldsymbol{H}_j \hat{\boldsymbol{\alpha}}_{(j)} , \tag{4.48}$$

where $\hat{\boldsymbol{\alpha}}_j = \big(\boldsymbol{X}_j^t \boldsymbol{W}_d^{-1} \boldsymbol{X}_j\big)^{-1} \boldsymbol{X}_j^t \boldsymbol{W}_d^{-1} \boldsymbol{Y}_j$ and $\boldsymbol{H}_j^{-1} = \boldsymbol{\Lambda} + \big(\boldsymbol{X}_j^t \boldsymbol{W}_d^{-1} \boldsymbol{X}_j\big)^{-1} \sigma^2$. Obviously, now, $\boldsymbol{V}(\hat{\boldsymbol{\alpha}}_{II}) = (\sum_j \boldsymbol{H}_j)^{-1}$. Note that for $\boldsymbol{\Lambda} = 0$, the individual estimates $\hat{\boldsymbol{\alpha}}_{(j)}$ are averaged with weights $\boldsymbol{X}_j^t \boldsymbol{W}_d^{-1} \boldsymbol{X}_j$, whereas for $\boldsymbol{\Lambda} \to \infty$ they are averaged with equal weights. For equal design matrices (4.48) reduces to (4.41) for any $\boldsymbol{\Lambda}$. In general, $\hat{\boldsymbol{\alpha}}_{II}$ depends on $\boldsymbol{\Lambda}$ and another equation is needed to estimate $\boldsymbol{\Lambda}$. One possibility is to consider

$$\hat{\sigma}^2 \sum_j \big(\boldsymbol{X}_j^t \boldsymbol{W}_d^{-1} \boldsymbol{X}_j\big)^{-1} + m\hat{\boldsymbol{\Lambda}} = \sum_j (\hat{\boldsymbol{\alpha}}_{jII} - \hat{\boldsymbol{\alpha}}_{II})(\hat{\boldsymbol{\alpha}}_{jII} - \hat{\boldsymbol{\alpha}}_{II})^t , \tag{4.49}$$

which is the analogue of (4.44). Obviously, σ^2 is estimated from the residual sum of squares from the individual regressions. Iterative solution of (4.48) and (4.49), starting, e.g., with $\boldsymbol{\Lambda} = 0$, gives consistent estimates of both $\boldsymbol{\alpha}$ and $\boldsymbol{\Lambda}$. These estimates can relatively easily be calculated, but they may not be 'efficient', i.e., they may not have asymptotically minimal variance, as $\boldsymbol{\Lambda}$ is not estimated by maximum likelihood.

We will discuss the method of maximum likelihood in a more general formulation, which contains the random coefficient model as special case. We assume that the covariance matrices are parameterised by a finite dimensional parameter $\boldsymbol{\theta}$, and for brevity we write $\underline{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$ for $\big(\boldsymbol{\Sigma}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{\Sigma}_m(\boldsymbol{\theta})\big)$. The log likelihood of the observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ is

$$l\big(\boldsymbol{\alpha}, \underline{\boldsymbol{\Sigma}}(\boldsymbol{\theta})\big) = C(m,p) - \frac{1}{2}\Big(\sum_{j=1}^m \log|\boldsymbol{\Sigma}_j| + \sum_{j=1}^m (\boldsymbol{Y}_j - \boldsymbol{X}_j\boldsymbol{\alpha})^t \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{Y}_j - \boldsymbol{X}_j\boldsymbol{\alpha})\Big). \tag{4.50}$$

The likelihood equations are

$$(\partial/\partial\boldsymbol{\alpha}, \partial/\partial\boldsymbol{\theta})\; l\big(\boldsymbol{\alpha}, \underline{\boldsymbol{\Sigma}}(\boldsymbol{\theta})\big) = (\boldsymbol{0}, \boldsymbol{0}) . \tag{4.51}$$

The maximum likelihood estimates are those values of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ that maximise (4.50). If the maximum does not occur at the boundary of the parameter region, then (4.51) constitutes a necessary condition. If the log-likelihood is a (strictly) concave function of $(\boldsymbol{\alpha}, \boldsymbol{\theta})$, then (4.51) is a sufficient condition for the (unique) maximum. Concavity is not always easy to demonstrate in practice, however. For a large number of parameters, it pays to look for an efficient computational algorithm [228, 324, 526].

As an example, we present the computational equations for the still rather general special case that each $\boldsymbol{\Sigma}_j = \sum_{i=1}^{f} \theta_i \boldsymbol{G}_{ij}$ for $j = 1, \ldots, m$. The derivative of $l\big(\boldsymbol{\alpha}, \underline{\boldsymbol{\Sigma}}(\boldsymbol{\theta})\big)$ with respect to $\boldsymbol{\alpha}$ yields

$$\sum_j \boldsymbol{X}_j^t \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{Y}_j - \boldsymbol{X}_j \boldsymbol{\alpha}) = \boldsymbol{0} \,, \tag{4.52}$$

whose solution corresponds precisely to (4.33) with $\boldsymbol{G}_j = \boldsymbol{\Sigma}_j$. Derivation with respect to θ_i gives [324, 519]

$$\sum_{j=1}^{m} tr \boldsymbol{\Sigma}_j^{-1} \boldsymbol{G}_{ij} (I - \boldsymbol{\Sigma}_j^{-1} \boldsymbol{e}_j \boldsymbol{e}_j^t) = 0 \quad (i = 1, \ldots, f) \,, \tag{4.53}$$

where $\boldsymbol{\Sigma}_j = \sum_{i=1}^{f} \theta_i \boldsymbol{G}_{ij}$ and $\boldsymbol{e}_j = \boldsymbol{Y}_j - \boldsymbol{X}_j \boldsymbol{\alpha}$. Equations (4.52) and (4.53) are to be solved iteratively. The covariance matrix of the ML estimator $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$ is estimated by the negative of the inverse of the matrix of second derivatives of the log likelihood at the maximum likelihood solution $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$.

The maximum likelihood estimator tends to be rather sensitive to the validity of the distributional assumption (multivariate normality) for the errors. Sometimes, it is more robust to iterate the likelihood equations only a few times using the ordinary least-squares estimates as starting values. In [324] three algorithms are discussed to solve (4.52) and (4.53) numerically. They are for instance implemented in the program BMDP5V, see [150]. Under certain conditions, asymptotic normality and consistency has been proven [519], even for estimators obtained by the first few iteration steps of (4.52) and (4.53).

4.5.2 Robust Estimation

The idea of robust statistics is to use estimators and test procedures that are insensitive to (moderately large) deviations from the probabilistic assumptions. They tend to give better fits to the *bulk* of the data than classical procedures do, and hence robust methods are also suitable for outlier detection. Here, we restrict attention to robust estimation of the mean–value structure (i.e., the parameter $\boldsymbol{\alpha}$), under the assumption $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is assumed either to be known, or independently estimable. We will discuss the multivariate situation of outlying profiles from the regression on the plasma parameters, rather than the somewhat simpler univariate case of single outlying measurement points (channels) from individual profile fits.

Generalised least-squares estimation minimises $-l(\boldsymbol{\alpha}, \boldsymbol{G})$ from (4.50) as a function of $\boldsymbol{\alpha}$ for some (symmetric, positive definite) matrix $\boldsymbol{G}$. In practice, one has to insert some sensible estimate of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ for G. A generalisation of this procedure (see, e.g., [292], [264]) is to estimate $\boldsymbol{\alpha}$ by minimising

$$\sum_{j=1}^{m} \rho\left((\boldsymbol{Y}_j - \boldsymbol{X}_j\boldsymbol{\alpha})^t \boldsymbol{G}_j^{-1} (\boldsymbol{Y}_j - \boldsymbol{X}_j\boldsymbol{\alpha})\right) \tag{4.54}$$

for some suitable, non-decreasing function ρ, and symmetric, positive definite matrices $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_m$. Such estimates are called M–estimates, see also Sect. 3.4.3.

Remarks.

1. Note that (4.54) permits arbitrarily weighted regression. Choosing $\boldsymbol{G}_j = \boldsymbol{W}_j \boldsymbol{\Sigma} \boldsymbol{W}_j^t$, with $\boldsymbol{W}_j$ diagonal and positive, corresponds to assigning the diagonal elements of $\boldsymbol{W}_j$ as weights for the jth profile, *in addition to* the weighting imposed by $\boldsymbol{\Sigma}$, $j = 1, \ldots, m$. These weights can be used to robustify the estimation procedure, at the cost of some loss of efficiency in case the multivariate normal error distribution happens to be a good approximation.
2. Equation (4.54) can be viewed as a log likelihood equation for a very specific probability distribution. More importantly, it has been derived (see [292, 563]), that for symmetric error distributions, in a number of situations, solution of (4.54) yields consistent and asymptotically normal parameter estimates.

The quantity $D_j = (\boldsymbol{Y}_j - \boldsymbol{X}_j\hat{\boldsymbol{\alpha}})^t \boldsymbol{G}_j^{-1} (\boldsymbol{Y}_j - \boldsymbol{X}_j\hat{\boldsymbol{\alpha}})$ can be interpreted as a squared residual, i.e., the squared distance between the observed and the fitted jth profile, in the metric defined by $\boldsymbol{G}_j$. A large value of D_j indicates an outlying profile.[11] An outlying profile is a candidate for a bad measurement, and should be checked on physical grounds. A weak point of least squares is that large *actual* residuals (i.e., deviations from the actual postulated model) that are, in addition, influential (i.e., have a large influence on the regression fit, because they are in plasma parameter space far away from the bulk of the dataset), can distort the regression fit in such a way that the corresponding *fitted* residuals are quite moderate. Hence, such points go undetected as outliers. One way to get around this difficulty is to calculate the jth residual from a dataset from which just the jth datapoint has been excluded. This can efficiently be done, see, e.g., [578, 581]. Outlier detection is an important part in practical regression. Justified deletion of outliers is an important step in robustifying a least-squares regression fit. Another expedient is to use robust (e.g., M–type) estimates that automatically downweight influential data points [264].

Least squares corresponds to $\rho(u) = u$, and is rather sensitive to outliers, as the residuals from the fit are squared. In robust regression, sensitivity to outliers is diminished by using functions for ρ that grow less strongly than lin-

[11] If m profiles are tested simultaneously, the cut-off point for outlier detection should be based on the distribution function of $\max_{j=1}^{n} D_j$, rather than on that of D_j, see [465].

early for large values of the residual. We present some common examples. The minimum absolute deviation (MAD) estimator corresponds to $\rho(u) = \sqrt{u}$. To avoid uniqueness problems, the absolute deviation norm may be replaced by $\rho(u) = \|u\|^r$ for some $0.5 < r < 0.7$, say. The 'Huber estimator' is defined by $\psi(u) = \rho'(u) = \sqrt{u}$ for $u < cst$ and $\psi(u) = \sqrt{cst}$ for $u > cst$. For cst between 0 and ∞, it ranges between the minimum absolute deviation and the least squares estimator. It turns out that this estimator is 'optimally robust' from two points of view, see Chaps. 2.6 and 2.7 of [264], under the assumption that only the response variable is contaminated with outliers. A problem is the choice of cst. The optimal choice depends on the (unkown) amount of contamination. For practical purposes, it seems sensible to choose cst near $\chi^2_{(p-p');0.05/m}$, which is the classical cut-off from normal least squares. Finally, one can apply iteratively reweighted least squares, downweighting the profiles with large values of D_j (by choosing, e.g., the weights $\boldsymbol{W}_j = w_j \boldsymbol{I}$, with w_j inversely proportional to D_j). It is easily shown that these are, in fact, M–estimates in a slightly disguised form [264].

The need for robust statistics depends on the quality of the dataset, and on the amount of effort one is willing to spend on residual analysis. Broadly speaking, for 'high quality' datasets, standard least squares estimation is a reasonably justified procedure. For 'medium quality' datasets, standard least squares is only justified if one spends considerable effort to detect and repair (wherever justified) any outlying data. Robust estimation may then be useful for better pinpointing these outliers. For 'low quality' datasets, robust estimation is indispensible for fitting the bulk of the data and for outlier identification. In general, solution of the estimating equations and determination of the covariance matrix of the regression estimates is more difficult for robust estimation than it is for least squares. The reader is referred to [264, 292, 566, 567] for a description of various robust methods.

4.6 Model Testing

We present some statistical tests for testing the null-hypothesis that a given model for the observations is correct. The general procedure may heuristically be described as follows. First, the parameters for the model to be tested, say M_2, are estimated, as are the parameters for a more comprehensive, competing, model M_1, which contains model M_2 as a special case. Then the distance between these two sets of parameters, suitably normed, is calculated. Under the null-hypothesis that model M_2 holds, this distance has some probability distribution. The characteristics of this distribution, or of an asymptotic approximation thereof, are determined. If the observed value of this distance is in the tail of the distribution, then the null-hypothesis is considered to be implausible. If one rejects at the 5% level, then the probability of rejecting the null-hypothesis incorrectly (when in fact it is true) is at most 5%. The probability of making an error of the second kind, i.e., of accepting the

null-hypothesis when in fact it is false, depends of course on how much the true underlying model differs from the hypothesised model M_2. Let us denote this probability by $\beta(\boldsymbol{\alpha}_{12})$, where $\boldsymbol{\alpha}_{12} \in V_{12}$ denote the parameters in model M_1 *in excess* of those of model M_1, i.e., the parameterisation is such that the origin of the space V_{12} corresponds to model M_2. Broadly speaking, the tests to be discussed have the property that in comparison with other tests $\beta(\boldsymbol{\alpha}_{12})$ is *relatively low* in *every direction* of the parameter space V_{12}. Statistically, a corresponding optimality criterion, which constitutes a practical compromise, is called 'asymptotically most stringent, against unrestricted alternatives' [5, 424, 625, 718].

In the following, M_1 and M_2 are identifiers that indicate in each situation the more comprehensive model and the more restricted model, respectively. Which specific models are chosen for them in practice changes with the context.

4.6.1 Discrete Versus Continuous Profile Representations

As in Sect. 4.3, we assume that, for fixed values of the plasma parameters, the available data consist of a single sample of m profile measurements $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$, each at p radial positions. For simplicity we assume that the radial positions correspond to constant flux–surface radii. As null-hypothesis we consider a continuous (e.g., spline) model with p' free, linear regression parameters and with a general $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ (i.e., (4.16) with covariance model III in Sect. 4.3). The more comprehensive model M_1 is the unsmoothed model $\boldsymbol{Y}_j = N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ denotes the unknown underlying profile vector. Let $\hat{\boldsymbol{\Sigma}}$ be the ML estimator of the covariance matrix under model M_1, see (4.6), and $\hat{\boldsymbol{\Sigma}}_{III}$ the ML estimator under model M_2, see (4.33). The statistical test is based on the following result (see, e.g., [326] and Chap. 5.3 of [440]): If the null-hypothesis holds, and $m > p - p'$, then

$$T_1 = \frac{\left|\hat{\boldsymbol{\Sigma}}_{III}\right| - \left|\hat{\boldsymbol{\Sigma}}\right|}{\left|\hat{\boldsymbol{\Sigma}}\right|} = \left(\hat{\boldsymbol{\mu}} - \boldsymbol{X}_{rad}\hat{\boldsymbol{\alpha}}_{III}\right)^t \hat{\boldsymbol{\Sigma}}^{-1} \left(\hat{\boldsymbol{\mu}} - \boldsymbol{X}_{rad}\hat{\boldsymbol{\alpha}}_{III}\right) \tag{4.55}$$

is distributed as

$$\frac{p - p'}{m - p + p'} F(p - p', m - p + p') \; . \tag{4.56}$$

For large values of T_1 the continuous model will be rejected. The test statistic T_1 can be interpreted in two ways:

(a) The right-hand side gives the distance, in p-dimensional space, between the estimated underlying profile under model M_1 and under model M_2, in the metric defined by the (estimated) residual covariance matrix under the more comprehensive model M_1.

(b) The determinant of a covariance matrix is a scalar measure of the total residual variation associated with a given model. (This determinant is sometimes called the generalised variance.) Hence, the statistic T estimates the fractional increase in generalised variance in going from M_1 to model M_2.

4.6.2 Different Covariance Structures

If a given smoothed profile model with general covariance matrix $\boldsymbol{\Sigma}$ is not rejected using the previous test, then we can proceed to test the more parsimonious covariance structure II within the smoothed model, i.e., $M_1 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{II}$ is tested against $M_2 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{III}$. One can do this by applying a large-sample likelihood ratio test. The likelihood ratio statistic is

$$T_2 = -2 \ln \frac{\max_{M1} L(\boldsymbol{\alpha}, \boldsymbol{\Sigma})}{\max_{M2} L(\boldsymbol{\alpha}, \boldsymbol{\Sigma})} , \tag{4.57}$$

where L(.,.) stands for the likelihood function, which has to be maximised under model M_1 and M_2, respectively. By writing down the multivariate normal densities, one derives that

$$-2 \ln L(\hat{\boldsymbol{\alpha}}_{III}, \hat{\boldsymbol{\Sigma}}_{III}) = m(\ln\left|2\pi \hat{\boldsymbol{\Sigma}}_{III}\right| + p) , \tag{4.58}$$

where $\hat{\boldsymbol{\alpha}}_{III}$ and $\hat{\boldsymbol{\Sigma}}_{III}$ are the ML estimates under Model 2, see (4.33). For the smoothed model M_1, the likelihood has to be maximised numerically. This can be done by using, for instance, a computer program such as BMDP5V. Assuming that some regularity conditions are satisfied, it follows from general theory developed by Wilks [736] and Wald [718] (see, e.g., Chap. 4 of [607], and Chap. 6e of [538]) that, asymptotically,

$$T_2 \sim \chi^2_{\frac{1}{2}p(p+1)-\frac{1}{2}p'(p'+1)-1} . \tag{4.59}$$

In practice, it is convenient to apply an asymptotically equivalent version of T_2, which is obtained by inserting the UMVU instead of the ML estimators for $\boldsymbol{\Sigma}$ into the likelihood ratio (4.57). In that case, we get, after some simplification,

$$\widetilde{T}_2 = m \ln \frac{|\boldsymbol{S}_{II}|}{|\boldsymbol{S}_{III}|} , \tag{4.60}$$

where $\boldsymbol{S}_{III} = m\hat{\boldsymbol{\Sigma}}_{III}$ and $\boldsymbol{S}_{II} = (m-1)\hat{\boldsymbol{\Sigma}}_{II}$, see (4.39). For large m, $\widetilde{T}_2$ has the same distribution as T_2. As in (4.55), a computationally more convenient form of the ratio of the generalised variances is

$$\frac{|\boldsymbol{S}_{II}|}{|\boldsymbol{S}_{III}|} = 1 + m(\hat{\boldsymbol{\alpha}}_{II} - \hat{\boldsymbol{\alpha}}_{III})^t \boldsymbol{X}^t_{rad} \boldsymbol{S}^{-1}_{III} \boldsymbol{X}_{rad} (\hat{\boldsymbol{\alpha}}_{II} - \hat{\boldsymbol{\alpha}}_{III}) . \tag{4.61}$$

It should be mentioned that test statistics based on other combinations of the eigenvalues of $\boldsymbol{S}_{II}$ and $\boldsymbol{S}_{III}$ than those implied by (4.60) are in use (see,

e.g., Chap. 5 of [480]). For reasonable sample sizes, the practical differences between these multivariate test statistics tend to be small. Special care is required if the (unconstrained) ML estimates of the parameters under Model 2 lead to an estimate of $\boldsymbol{\Sigma}$ which is not positive definite [24, 296, 583].

4.6.3 Different Continuous Profile Representations

Although the theory presented holds for general linear profile representations, we will, for concreteness, orient the discussion towards spline representations. Suppose, we have a Hermitian spline model with p' free parameters for the mean–value structure and with one of the errors structures I, II, III. We want to test the null-hypothesis that a spline sub-model holds. A spline sub-model is defined as a spline model where, with respect to the original model, the continuity index (see Sect. 4.3.4) of some of the knots has been increased. (Recall that imposing third order continuity for a third degree spline amounts to removing the knot.) To test such a null-hypothesis, which can be stated as linear restrictions on the parameters, there exist standard F-tests, especially in the case of the error structures I and III (see, e.g., [14, 440, 538]). Here, we will directly discuss the more relevant and interesting case of error structure II. We consider the general model M_1: $\boldsymbol{Y}_j = \boldsymbol{X}_{rad}\boldsymbol{A} + \boldsymbol{E}_j$, where $\boldsymbol{E}_j \sim N_p(0, \sigma^2\boldsymbol{W}_d)$ and $\boldsymbol{A} \sim N_{p'}(\boldsymbol{\alpha}, \boldsymbol{\Lambda}_1)$ have a multivariate normal distribution. Within this model, we test the sub-model $M_2 : A \sim N_{p''}(\boldsymbol{D\beta}, \boldsymbol{D\Lambda}_2\boldsymbol{D}^t)$ for some $p' \times p''$ matrix D.

We again use the likelihood ratio as an asymptotic test statistic. Consider the random variable

$$T_3 = m \ln \frac{|\boldsymbol{S}_{II,2}|}{|\boldsymbol{S}_{II,1}|} , \tag{4.62}$$

where $\boldsymbol{S}_{II,2}$ and $\boldsymbol{S}_{II,1}$ are the the residual SSCP matrices for the restricted model and for the general model, respectively. If the null-hypothesis holds, then asymptotically (i.e., as $m \to \infty$),

$$T_3 \sim \chi^2_{\frac{1}{2}p'(p'+1)+p'-\frac{1}{2}p''(p''+1)-p''} \cdot \tag{4.63}$$

If the observed value of T_3 is the tail of this χ^2 distribution, then the null-hypothesis has to be rejected. Similar to formula (4.55), the ratio of the generalised variances can be rewritten as

$$\frac{|\boldsymbol{S}_{II,2}|}{|\boldsymbol{S}_{II,1}|} = 1 + m(\hat{\boldsymbol{\alpha}}_{II} - \boldsymbol{D}\hat{\boldsymbol{\beta}}_{II})^t \boldsymbol{X}^t_{rad}\boldsymbol{S}^{-1}_{II,1}\boldsymbol{X}_{rad}(\hat{\boldsymbol{\alpha}}_{II} - \boldsymbol{D}\hat{\boldsymbol{\beta}}_{II}) . \tag{4.64}$$

As a concrete example, we consider testing the null-hypothesis that the underlying profiles can be described by Gaussian functions (with random coefficients) against the alternative that a general Hermitian spline model (also with random coefficients) holds, see (4.15). For the Gaussian model we have $\boldsymbol{\beta} = (\mu(0), b_0)^t$, the random Gaussian coefficients $\boldsymbol{B}$ satisfying $\boldsymbol{B} \sim$

$N(\boldsymbol{\beta}, \boldsymbol{\Lambda}_2)$, with $\boldsymbol{\Lambda}_2$ a general 2×2 covariance matrix. For the general spline model we have $\boldsymbol{\alpha} = (\mu(0), a_0, b_0, c_0, b_1, c_1, b_2, c_2, \ldots)^t$, and $\boldsymbol{A} \sim N(\boldsymbol{\alpha}, \boldsymbol{\Lambda}_1)$. Obviously,

$$\boldsymbol{D}^t = \begin{pmatrix} 1\,0\,0\,0\,\ldots\,0 \\ 0\,0\,1\,0\,\ldots\,0 \end{pmatrix} . \tag{4.65}$$

The Gaussian model has the simple design matrix

$$\boldsymbol{X}_{rad,g} = \boldsymbol{X}_{rad}\boldsymbol{D} = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ r_1^2 & r_2^2 & r_3^2 & \ldots & r_p^2 \end{pmatrix} . \tag{4.66}$$

Using the methods described in Sect. 4.5, one can estimate, within each of the random coefficient models, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (the estimates are denoted by $\hat{\boldsymbol{\alpha}}_{II}$ and $\hat{\boldsymbol{\beta}}_{II}$, respectively) as well as the residual SSCP matrices from both fits (denoted by $\boldsymbol{S}_{II,1}$ and $\boldsymbol{S}_{II,2}$). Hence, one can calculate the test statistic T_3 from either side of (4.64). (The left-hand side may computationally be somewhat more cumbersome for large matrices.) Note that for the Gaussian model, $\boldsymbol{D}^t\boldsymbol{X}_{rad}^t\boldsymbol{S}_{II,1}^{-1}\boldsymbol{X}_{rad}\boldsymbol{D}$ in (4.64) can be rewritten as $\boldsymbol{X}_{rad,g}^t\boldsymbol{S}_{II,1}^{-1}\boldsymbol{X}_{rad,g}$, which is readily calculated and can be interpreted as a constant times the estimated covariance matrix for the regression coefficients $\hat{\boldsymbol{\beta}}_{II}$, *provided* the covariance matrix of the p profile measurements is estimated from the residual SSCP matrix of the *full* spline model M_1.

4.6.4 Profile Invariance

We consider the model with covariates and error structure III, i.e.,

$$\boldsymbol{Y} = \boldsymbol{X}_{rad}\boldsymbol{\alpha}_{mat}\boldsymbol{X}_{cov} + \boldsymbol{E} , \tag{4.67}$$

with the columns of $\boldsymbol{E}$ independently $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ distributed, see (4.23). As argued in Sect. 4.2, the hypothesis of (full) profile invariance means that all elements of $\boldsymbol{\alpha}_{mat}$ except for the first row and the first column are zero. This can be expressed as

$$H_0{:}\boldsymbol{C}\boldsymbol{\alpha}_{mat}\boldsymbol{M} = 0 , \tag{4.68}$$

where $\boldsymbol{C}$ and $\boldsymbol{M}$ are diagonal matrices with 0 on the (1,1) position and unity on all other diagonal elements. The hypothesis is tested by a similar criterion as in the previous cases, namely

$$T_4 = \widetilde{m} \ln \frac{|\boldsymbol{S}_{II,2}|}{|\boldsymbol{S}_{II,1}|} , \tag{4.69}$$

where $\boldsymbol{S}_{II,2}$ denotes the SSCP matrix under H_0 and the $\boldsymbol{S}_{II,1}$ the SSCP matrix under the less restricted model. Now, we have

$$\boldsymbol{S}_{II,2} = \boldsymbol{C}(\boldsymbol{X}_{rad}^t(f\widetilde{\boldsymbol{\Sigma}})^{-1}\boldsymbol{X}_{rad})^{-1}\boldsymbol{C}^t , \tag{4.70}$$

with $f = m - w - 1 - (p - p')$. The expression for $\boldsymbol{S}_{II,1}$ is slightly more complicated. As derived in [250], it can be written as $\boldsymbol{S}_{II,2} + \boldsymbol{S}_{II,H}$, with

$$\boldsymbol{S}_{II,H} = (\boldsymbol{C}\boldsymbol{\alpha}_{mat}\boldsymbol{M})(\boldsymbol{M}^t\boldsymbol{R}\boldsymbol{M})^{-1}(\boldsymbol{C}\boldsymbol{\alpha}_{mat}\boldsymbol{M})^t \ , \tag{4.71}$$

where

$$\boldsymbol{R} = \boldsymbol{Q} + f^{-1}\boldsymbol{Q}\boldsymbol{X}_{cov}\boldsymbol{Y}^t\big(\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{\boldsymbol{\Sigma}}})\big)\boldsymbol{Y}\boldsymbol{X}_{cov}\boldsymbol{Q} \ , \tag{4.72}$$

with $\boldsymbol{Q} = (\boldsymbol{X}_{cov}\boldsymbol{X}_{cov}^t)^{-1}$ and

$$\boldsymbol{P}_{\widetilde{\boldsymbol{\Sigma}}} = \boldsymbol{X}_{rad}(\boldsymbol{X}_{rad}^t\widetilde{\boldsymbol{\Sigma}}\boldsymbol{X}_{rad})^{-1}\boldsymbol{X}_{rad}^t\widetilde{\boldsymbol{\Sigma}}^{-1} \ . \tag{4.73}$$

Note that $\boldsymbol{P}_{\widetilde{\boldsymbol{\Sigma}}}$ is the projection operator with respect to the inner product defined by $\widetilde{\boldsymbol{\Sigma}}$. For $\widetilde{m} = m-(w+1)-(p-p')-\frac{1}{2}(r_c-r_m+1)$, where r_c and r_m denote the rank of $\boldsymbol{C}$ and $\boldsymbol{M}$, respectively, T_4 is approximately distributed as a χ^2 variate with $r_c r_m$ degrees of freedom.

The hypothesis of partial profile invariance (for instance, with respect to q_{eng}, but only for $r > 0.6$), can also be expressed as H_0: $\boldsymbol{C}\boldsymbol{\alpha}_{mat}\boldsymbol{M} = 0$, for suitable matrices $\boldsymbol{C}$ and $\boldsymbol{M}$, and hence be tested by the just described procedure.

4.7 Confidence Bands and Regression

The presentation of only a single, smoothed profile estimate gives no information on which *other* profiles would *also* be plausible for the true underlying profile or for a new experimentally determined profile. Such information is given by confidence, and prediction bands, respectively. Although confidence bands can be considered as a limiting case of prediction bands, we shall discuss them consecutively, as the construction of prediction bands is more complicated.

It will be useful to make a distinction between 'local' and 'global' confidence bands. A local confidence band for an underlying unknown profile $\mu(r)$ is defined as the region between two random boundaries $\mu_{loc}^{(l)}(r)$, and $\mu_{loc}^{(h)}(r)$ such that, for any $r \in [0,1]$,

$$P\{\mu(r) \in \big(\mu_{loc}^{(l)}(r), \mu_{loc}^{(h)}(r)\big)\} = 1-\alpha \ , \tag{4.74}$$

which means that for any $r \in [0,1]$, the probability that the random interval $\big(\mu_{loc}^{(l)}(r), \mu_{loc}^{(h)}(r)\big)$ includes $\mu(r)$ equals $1-\alpha$. A global confidence band for $\mu(r)$ is defined by

$$P\{\mu(r) \in \big(\mu_{loc}^{(l)}(r), \mu_{loc}^{(h)}(r)\big) \quad \text{for all} \quad r \in [0,1]\} = 1-\alpha \ . \tag{4.75}$$

The quantity $1-\alpha$ is called the confidence coefficient of the band, which is frequently chosen to be 67% or 95%. The null-hypothesis that the 'true' plasma profile equals some predescribed profile $\mu_o(r)$ is not rejected at the 100α% level if and only if $\mu_o(r)$ is everywhere contained in a *global* confidence

band with confidence coefficient $1-\alpha$.
Remark. For each confidence band with confidence coefficient $1-\alpha$, there exists a statistical test rejecting at level α. Usually, one tries to select tests and confidence bands with some 'optimality criterion', for instance, to confidence bands with minimal area. Since, under our model assumptions, the estimates for $\mu(r)$ are symmetrically distributed and (usually) unbiased, we will consider only symmetrical confidence bands that can be written as $\hat{\mu}(r) \pm \delta_{loc}(r)$, and $\hat{\mu}(r) \pm \delta_{gl}(r)$, and that correspond to well-known and rather efficient (two-sided) statistical tests.

We will discuss now the construction of confidence bands for fixed plasma parameters and error structures I and II. The methods are easily adapted to error structure III, however. As in Sect. 4.3, we have the following model for the m stochastically independent profiles $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ at p radial positions: $\boldsymbol{Y}_j = \boldsymbol{X}_{rad}\boldsymbol{A} + \boldsymbol{E}_j$, with $\boldsymbol{A}$ either deterministically equal to $\boldsymbol{\alpha}$ (model I) or $\boldsymbol{A} \sim N_{p'}(\boldsymbol{\alpha}, \boldsymbol{\Lambda})$ (model II) and $\boldsymbol{E}_j \sim N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{W}_d)$, for $j = 1, \ldots, m$.

4.7.1 Local Confidence Bands

A local confidence band for $\mu(r) = \sum_{h=1}^{p'} \alpha_h f_h(r)$ is constructed by calculating for each $r \in [0,1]$ a confidence interval for the linear combination $\boldsymbol{\alpha}^t \boldsymbol{f}(r)$ of the regression parameters, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p'})^t$ and $\boldsymbol{f}(r) = \big(f_1(r), f_2(r), \ldots, f_{p'}(r)\big)^t$ From the fact that $\hat{\boldsymbol{\alpha}}^t$ has a multivariate normal distribution it follows that $\hat{\boldsymbol{\alpha}}^t \boldsymbol{f}(r)$ is normally distributed with expectation $\boldsymbol{\alpha}^t \boldsymbol{f}(r)$ and variance $\boldsymbol{f}^t(r) \boldsymbol{V}(\hat{\boldsymbol{\alpha}}) \boldsymbol{f}(r)$. By studentising (i.e., by taking into account the effect that the estimated instead of the true variance of $\hat{\boldsymbol{\alpha}}$ is inserted) we get

$$\hat{\boldsymbol{\alpha}}_I^t \boldsymbol{f}(r) \pm \big(\boldsymbol{f}^t(r) \hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}}_I) \boldsymbol{f}(r)\big)^{\frac{1}{2}} t_{mp-p';\alpha/2} \,, \tag{4.76}$$

where $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}}_I) = m^{-1} \left(\boldsymbol{X}_{rad}^t \boldsymbol{W}_d^{-1} \boldsymbol{X}_{rad}\right)^{-1} \hat{\sigma}$ with $\hat{\sigma}$ given by (4.32), as the required confidence interval under model I, and

$$\hat{\boldsymbol{\alpha}}_{II}^t \boldsymbol{f}(r) \pm \big(\boldsymbol{f}^t(r) \hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}}_{II}) \boldsymbol{f}(r)\big)^{\frac{1}{2}} t_{m-1;\alpha/2} \,, \tag{4.77}$$

where $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}}_{II})$ is given by (4.42), as the corresponding confidence interval under model II. As usual, t_f stands for the Student distribution with f degrees of freedom. Note that the unbiased estimation of the covariance structure in model II costs a considerable number of degrees of freedom. This does not always matter in practice, as for $f > 20$ and usual confidence levels, the Student distribution is very close to the standard normal.

4.7.2 Global Confidence Bands

A global confidence band for the unknown profile is derived from a k-dimensional confidence ellipse for $\boldsymbol{\alpha}$, which consists of all values $\boldsymbol{\alpha}_o$ such that

$$(\boldsymbol{\alpha}_o - \hat{\boldsymbol{\alpha}})^t \, \hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}})^{-1} \, (\boldsymbol{\alpha}_o - \hat{\boldsymbol{\alpha}}) \leq c(m, p, p'; \alpha) \; , \tag{4.78}$$

where $c(m, p, p'; \alpha)$ equals $F_{p', mp-p'; \alpha}$ for model I and $\frac{(m-1)p'}{m-p'} F_{p', m-p'; \alpha}$ for model II. (Asymptotically, both expressions tend to $\chi^2_{p';\alpha}$.) The extreme values of $\boldsymbol{\alpha}_o^t \boldsymbol{f}(r)$ under the restriction (4.78) are found to be (using, e.g., the method of Lagrange multipliers)

$$\hat{\boldsymbol{\alpha}}^t \boldsymbol{f}(r) \pm \big(\boldsymbol{f}^t(r) \hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}}) \boldsymbol{f}(r)\big)^{\frac{1}{2}} \big(c(m, p, p'; \alpha)\big)^{\frac{1}{2}} \; . \tag{4.79}$$

Evidently, a global band is wider than a local band with the same confidence coefficient. Notice from (4.76), (4.77) and (4.79) that in this case the bands are 'proportional', i.e., a global band with coefficient $1 - \alpha$ corresponds to a local band with coefficient $1 - \alpha' > 1 - \alpha$.

Under model I, similar formulae can be derived for local and global bands for unequal design matrices, and in particular for the regression model (4.22)–(4.24) that includes plasma variables as covariates. (In that case one may be particularly interested in a band that is global with respect to r, but local with respect to the plasma variables.) Under model II one must be satisfied with asymptotic confidence bands, based on some maximum likelihood estimate of $\boldsymbol{V}(\hat{\boldsymbol{\alpha}}_{II})$.

A practical application of confidence band construction is as follows. As indicated in Sect. 4.2, profile invariance is expressed by the requirement that the profile shape $L^{-1}(r, \boldsymbol{x}) = \frac{\partial}{\partial r} \mu(r, \boldsymbol{x})$ does not depend on the plasma parameters $\boldsymbol{x}$. For a graphical evaluation, it is useful to plot $\frac{\partial}{\partial g_k(\boldsymbol{x})} \hat{L}^{-1}(r, \boldsymbol{x})$, which is the mixed derivative of $\hat{\mu}(r, \boldsymbol{x})$, as a function of r, for $g_1(\boldsymbol{x}) = \ln I_p$, $g_2(\boldsymbol{x}) = \ln q_{eng}$, etc. The hypothesis of profile invariance is not rejected for those plasma variables, and for those radii, for which the plotted mixed derivative of $\hat{\mu}(r, \boldsymbol{x})$ does not 'significantly' differ from zero. For each given plasma variable, this can directly be tested once a global confidence band for this mixed derivative is constructed. Under the assumption that it is sufficient to the expand $\mu(r, \boldsymbol{x})$ only up to the first powers in the logarithms of the plasma parameters, $\frac{\partial}{\partial g_k(\boldsymbol{x})} \frac{\partial}{\partial r} \mu(r, \boldsymbol{x}) = \sum_{h=1}^{p'} \alpha_{h,k} f'_h(r)$ for $k = 1, \ldots, w$, see (4.22), and hence precisely the type of gobal confidence band as discussed above can be used. A practical elaboration of this for Ohmic ASDEX shots is found in [451].

4.7.3 Prediction Bands

We now turn to the somewhat more complicated case of prediction bands. The discussion will be tuned to error model II, where we have random measurement noise and intrinsic plasma variations. We start by defining a fairly general construct, which contains several useful special cases. Let $\langle Y \rangle_{m_0}(r, t)$ be the average of m_0 hypothetical, new measurements (on logarithmic scale) at time t and at some normalised flux–surface radius r.

Similarly, let $\langle Y\rangle_{m_0,m_1}(r,t) = m_1^{-1}\sum_{j=1}^{m_1}\langle Y\rangle_{m_0}(r,t_j)$, where $t_1,\ldots,t_{m_1}$ are equidistant timepoints over the time interval $(t-\frac{1}{2}\Delta t, t+\frac{1}{2}\Delta t)$. In other words, $\langle Y\rangle_{m_0,m_1}(r)$ 'estimates' a time-averaged profile over the just mentioned time interval on the basis of m_0 hypothetical, repeated measurements at each of m_1 equidistant timepoints. We assume Δt to be so large that, for $m_1\to\infty$, $\langle Y\rangle_{m_0,m_1}(r,t)\to\mu(r)$. (The assumed model implies that, for each fixed r, $\langle Y\rangle_{m_0}(r,t)$ is a stationary stochastic process. Hence, the right-hand side is independent of time.)

A *prediction* band for $\langle Y\rangle_{m_0,m_1}(r,t)$ is based on the the variance of the difference between $\langle Y\rangle_{m_0,m_1}(r,t)$ and its estimate $\hat{\boldsymbol{\alpha}}_{II}^t\boldsymbol{f}(r)$:

$$\begin{aligned} V\big(\langle Y\rangle_{m_0,m_1}(r,t) &- \hat{\boldsymbol{\alpha}}_{II}^t\boldsymbol{f}(r)\big) \\ &= V\big(\langle Y\rangle_{m_0,m_1}(r,t) - \boldsymbol{\alpha}^t\boldsymbol{f}(r)\big) + V\big((\hat{\boldsymbol{\alpha}}_{II}-\boldsymbol{\alpha})^t\boldsymbol{f}(r)\big) \\ &= \big(\frac{1}{m_1}+\frac{1}{m}\big)\big(\boldsymbol{f}^t(r)\boldsymbol{\Lambda}\boldsymbol{f}(r)\big) + \\ &+ \big(\frac{1}{m_1 m_0}+\frac{1}{m}\big)\boldsymbol{f}^t(r)\left(\boldsymbol{X}_{rad}^t\boldsymbol{W}_d^{-1}\boldsymbol{X}_{rad}\right)^{-1}\boldsymbol{f}(r)\sigma^2, \end{aligned} \tag{4.80}$$

see (4.38). (To understand this formula, it may be useful to consider first the various special cases: $m_1=1$, $m_0\to\infty$; $m_1=1$, m_0 arbitrary, etc.) Note that, because of the stationarity assumption, the right-hand side of (4.80) is independent of time.

For brevity, we denote the variance in (4.80) by V. In practice, the quantities $\boldsymbol{V}(\hat{\boldsymbol{\alpha}}_{II})$, σ^2, and $\boldsymbol{\Lambda}$ are unknown, and have to be estimated, for instance from (4.42)–(4.44), and inserted into 4.80. From the resulting estimate $\hat{V}$, asymptotic prediction bands for $\langle Y\rangle_{m_0,m_1}(r,t)$ are constructed (with some radial interpolation assumption for the first term if the variance is not assumed to be constant for the p measurement channels). Asymptotically (i.e., if m and p sufficiently large), a local m_0,m_1 prediction band with confidence coefficient $1-\alpha$ is given by $\hat{\boldsymbol{\alpha}}_{II}^t\boldsymbol{f}(r)\pm\hat{V}^{1/2}u_{\alpha/2}$. For $m_0\to\infty$, we have a confidence band for the underlying profile at a certain time t (if $m_1=1$), or of the time-averaged profile (if $m_1\to\infty$). An asymptotic global confidence band is obtained by replacing $u_{\alpha/2}$ by $(\chi^2_{p';\alpha})^{1/2}$, and letting $m_0\to\infty$. Studentised confidence bands, which take into account the error in estimating V, are constructed by replacing $u_{\alpha/2}$ by $t_{p-p';\alpha/2}$, and $\chi^2_{p';\alpha}$ by $p'F_{p',p-p';\alpha}$.

Although they are more complicated, prediction bands from the covariate model (4.23) can be derived along somewhat similar lines. Such an analysis is well worth the effort, as it enables one to make predictions for any value (within a reasonable range) of the plasma parameters $\boldsymbol{x}$, and because the accuracy in the prediction will be increased due to the extra information contained in the additional density and temperature profiles.

4.7.4 Confidence Intervals for Global Plasma Variables

There may be special interest in estimating volume-averaged profiles, instead of the profiles themselves. In particular, we consider

$$\begin{aligned}\langle n_e\rangle &= \int_0^1 n_e(r)h(r)dr \ ,\\ \langle p_e\rangle &= \int_0^1 n_e(r)T_e(r)h(r)dr \ ,\\ \langle T_e\rangle &= \langle p_e\rangle/\langle n_e\rangle \ ,\end{aligned} \tag{4.81}$$

where r denotes the flux–surface radius and $h(r)$ is some weight function arising from the transformation from geometrical radius to flux surface radius. For concreteness, we give the discussion for $\langle n_e\rangle$. For $\langle p_e\rangle$ the calculations are completely analogous, and as will be explained in the end, an approximate confidence interval for $\langle T_e\rangle$ can be constructed from those for $\langle p_e\rangle$ and $\langle n_e\rangle$. For simplicity of notation, we suppose, as in Sect. 4.3, that $\ln n_e(r) = \mu(r) = \sum_{h=1}^{p'} \alpha_h f_h(r) = \boldsymbol{\alpha}^t \boldsymbol{f}(r)$. The formulae are, however, analogous in the physically more general case $\mu(r) = \sum_h \sum_k f_h(r) g_k(\boldsymbol{x})$. We assume that we have unbiased and approximately normally distributed estimators for $\boldsymbol{\alpha}$ and $\mu(r)$, which are denoted by $\hat{\boldsymbol{\alpha}}$ and $\hat{\mu}(r)$, respectively. We introduce

$$\hat{n}(r) = e^{\hat{\mu}(r)} \ , \qquad \langle \hat{n}_e\rangle = \int_0^1 e^{\hat{\mu}(r)} h(r)dr \ . \tag{4.82}$$

It is assumed that $\langle n_e\rangle$ is accurately approximated (say, within 1 %) by a numerical integration routine, so that the difference between $\langle \hat{n}_e\rangle$ and its numerical estimate is negligible with respect to both the variance and the bias in $\langle \hat{n}_e\rangle$. It can be derived, see Sect. 4.9 (Appendix), that

$$\operatorname{var}\langle \hat{n}_e\rangle = \int_0^1 \int_0^1 e^{\boldsymbol{\alpha}^t(\boldsymbol{f}(r)+\boldsymbol{f}(r'))} \boldsymbol{f}^t(r) \boldsymbol{V}(\hat{\boldsymbol{\alpha}}) \boldsymbol{f}(r') h(r)h(r')drdr' \ . \tag{4.83}$$

To estimate this variance, estimators for $\boldsymbol{\alpha}$ and $\boldsymbol{V}(\hat{\boldsymbol{\alpha}})$ have to be inserted in this expression (see Sect. 4.5), and the integration has to be carried out either analytically (which may be rather complicated) or numerically.

As (4.82) is a non-linear transformation, an unbiased estimate $\hat{\mu}(r)$ will *not* lead to an unbiased estimate for $\langle n_e\rangle$. In fact, as derived in the Appendix, the bias is approximately

$$\operatorname{bias}\langle \hat{n}_e\rangle = \frac{1}{2} \int_0^1 e^{\boldsymbol{\alpha}^t \boldsymbol{f}(r)} \boldsymbol{f}^t(r) \boldsymbol{V}(\hat{\boldsymbol{\alpha}}) \boldsymbol{f}(r) h(r)dr \ . \tag{4.84}$$

To determine $\langle \hat{n}_e\rangle$, bias $\langle \hat{n}_e\rangle$, and $\operatorname{var}\langle \hat{n}_e\rangle$ numerically, the same type of integration routine is required. If $\mu(r)$ is approximated by splines, a low-order integration routine should be used to cope with the discontinuities of the third

derivative. Because the smoothed profiles behave very neatly otherwise, and only a relatively low accuracy is needed, the numerical integration should present no problem.

In practice, it may turn out that $\text{bias}^2\,\langle\hat{n}_e\rangle$ is negligible with respect to $\text{var}\,\langle\hat{n}_e\rangle$. In that case, an approximately 95% confidence interval for $\langle n_e\rangle$ is given by $\langle\hat{n}_e\rangle \pm 2\,\text{var}^{1/2}\langle n_e\rangle$.

If this is not the case, one can correct for the bias (which should be small anyhow), and assume that the variance is only negligibly influenced by this operation.

Finally, we discuss how to obtain a confidence interval for the density-weighted volume-averaged temperature. We assume that we fit the density and the temperature profiles on logarithmic scale, so that we have

$$\hat{n}_e(r) = e^{\hat{\boldsymbol{\alpha}}_n^t \boldsymbol{f}_n(r)}\,, \qquad \hat{T}_e(r) = e^{\hat{\boldsymbol{\alpha}}_T^t \boldsymbol{f}_T(r)}\,, \tag{4.85}$$

where $\hat{\boldsymbol{\alpha}}_n$ and $\hat{\boldsymbol{\alpha}}_T$ both have a multivariate normal distribution. For some reason, one might prefer in practice another set of basis functions for the density fit than for the temperature fit, so for generality we do not suppose that $\boldsymbol{f}_n(r) = \boldsymbol{f}_T(r)$.

Under the above model assumption,

$$\langle p_e\rangle = \int_0^1 e^{\hat{\boldsymbol{\alpha}}_n^t \boldsymbol{f}_n(r) + \hat{\boldsymbol{\alpha}}_T^t \boldsymbol{f}_T(r)} h(r)dr\,, \quad \langle n_e\rangle = \int_0^1 e^{\hat{\boldsymbol{\alpha}}_n^t \boldsymbol{f}_n(r)} h(r)dr\,, \tag{4.86}$$

and we consider the estimate $\langle\hat{T}_e\rangle = \langle\hat{p}_e\rangle/\langle\hat{n}_e\rangle$, which, for $\text{var}\left(\hat{\boldsymbol{\alpha}}_n^t \boldsymbol{f}_n(r)\right) \ll 1$ and $\text{var}\left(\hat{\boldsymbol{\alpha}}_T^t \boldsymbol{f}_T(r)\right) \ll 1$, has approximately zero bias and variance

$$\frac{1}{\langle n_e\rangle^2}\text{var}\,\langle\hat{p}_e\rangle + \frac{\langle p_e\rangle^2}{\langle n_e\rangle^4}\text{var}\,\langle\hat{n}_e\rangle - 2\frac{\langle p_e\rangle}{\langle n_e\rangle^2}\text{cov}\left(\langle\hat{n}_e\rangle, \langle\hat{p}_e\rangle\right)\,. \tag{4.87}$$

If necessary, the bias can be estimated. We will concentrate here on the variance. Each term in this expression, except for $\text{cov}\left(\langle\hat{n}_e\rangle, \langle\hat{p}_e\rangle\right)$, can be estimated according to the formulas derived above. For the cross term we have

$$\text{cov}\left(\langle\hat{n}_e\rangle, \langle\hat{p}_e\rangle\right) = \int_0^1 \int_0^1 \text{cov}\left(\hat{n}_e(r), \hat{p}_e(r)\right) h(r)h(r')drdr'\,, \tag{4.88}$$

with

$$\begin{aligned}\text{cov}\left(\hat{n}_e(r), \hat{p}_e(r')\right) = n_e(r)p_e(r')\big(&\boldsymbol{f}_n^t(r)\boldsymbol{V}(\hat{\boldsymbol{\alpha}}_n)\boldsymbol{f}_n^t(r') + \\ &+ \boldsymbol{f}_n^t(r)\boldsymbol{V}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\alpha}}_T)\boldsymbol{f}_T^t(r')\big)\,.\end{aligned} \tag{4.89}$$

Here, $\boldsymbol{V}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\alpha}}_T)$ denotes the matrix of covariances between $\hat{\boldsymbol{\alpha}}_n$ and $\hat{\boldsymbol{\alpha}}_T$. Obviously, this is zero if the temperature and the density measurements are independent, which is for instance not the case for the Nd:YAG laser experiments, where $\boldsymbol{V}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\alpha}}_T)$ has to be estimated from a simultaneous regression

of the temperature and density profiles. For volume-averaged profiles up to flux–surface radius r_0, all of the above formulae hold if only the upper integration limit is changed from 1 to r_0.

The important point to see is that although the expressions look somewhat complicated, their derivation and structure is quite straightforward. Given some tools for numerical integration, practical implementation is not a tremendous task.

It should be noted that the best (i.e., minimum variance) unbiased predictor for the local profile is not the best unbiased predictor for the volume averaged profile. In the latter case, because of the typical behaviour of $rn(r)$, the profile has to be known with a higher precision in the middle region (say $0.25 < r < 0.75$) than at both ends of the interval $[0, 1]$. On the other hand, an optimised profile fit with respect to MSE ($\langle \hat{n}_e \rangle$) may locally be bad, and for instance over- or underestimate the central density considerably. In a larger profile study, where the profiles are used for various purposes and where it is useful to have the estimates of the volume averages consistent with the fitted profiles, we prefer therefore local profile fits with good overall properties, such as those discussed in Sect. 4.5, and allow for the fact that the estimates for the volume averages do not have minimal variance.

4.8 Summary and Conclusions

In this chapter, we have presented a systematic approach to the parameterisation of temperature and density profile shapes, and provided a framework for analysing experimental profile-data in plasma physics obtained by active beam measurements, by employing various multivariate statistical regression techniques.The profiles are expanded in a double series of radial and plasma-parameter basis functions. Each product basis function is a covariate and the corresponding regression coefficients have to be estimated. The 'best' method of estimation depends on the postulated error structure. Partly motivated by the YAG–laser measurements performed at the ASDEX tokamak, the simplifying assumption was made that the measurement variations at a certain channel are uncorrelated in time. On the other hand, the deviations from an ideal profile fit may be radially correlated.

In Sects. 4.2 – 4.4 we have presented a variety of different representations for the plasma profiles and the error covariance matrix $\boldsymbol{\Sigma}$. These statistical models often form a hierarchy with the simple models embedded in the complex models as special cases. The more realistic models have more free parameters and therefore require more data. In a sense, the discrete point model of Sect. 4.2 constitutes the most accurate radial representation. However, a spline model with roughly equal numbers of measurement channels between radial knots yields a more effective representation for many datasets. The random coefficient model of $\boldsymbol{\Sigma}$, where the spline coefficients vary ran-

domly in addition to statistical noise, is expected to be a somewhat artificial but reasonably realistic model of the spatial correlations of the fluctuations.

In Sect. 4.5 three methods to estimate the spline coefficients are presented. The first method is by generalised least squares. Direct and elegant solutions are possible for relatively simple covariance structures, and even for the random coefficient model in the simple situation that one has a dataset of profiles for fixed values of the plasma parameters.

The second method is by maximum likelihood, which, in simple situations and for normally distributed errors, coincides with least squares. In more complex situations, differentiation of the log likelihood with respect to the mean value and the covariance parameters gives two coupled sets of equations, which have to be solved iteratively. In such a case, least squares can be considered as one such iteration step which maximises the likelihood with respect to the mean value parameters for a fixed (known or sensibly estimated) covariance matrix.

Finally, a discussion has been given of robust regression. The so-called M-estimators can be viewed as a generalisation of maximum likelihood estimators, in as far as they minimise some general function of the residuals, which may have nothing to do with the actual likelihood function. For relatively simple covariance structures, the asymptotic distribution of general M-estimators has been rigorously derived [264]. The purpose is to diminish the influence of erroneous, outlying datapoints on the regression estimates, by downweighting large residuals. Since a robust regression closely fits the bulk of the data, it is also suitable for clearly detecting outlying (i.e., suspicious) datapoints.

Sect. 4.6 presents a number of statistical tests to check if a simplification of the model, i.e., either of the profile representation or of the model for $\boldsymbol{\Sigma}$, produces significantly more than the expected increase in the residual error. As a special application, the empirical testing of profile invariance is considered.

Sect. 4.7 describes the construction of local and global confidence bands for the true profile shapes, and of prediction bands for observing new profiles. Furthermore, estimates and confidence intervals are given for corresponding volume-averaged quantities. A parameterised representation provides a compact summary of an experimental investigation: a table of spline coefficients is much more usable than a large database of all discharges. Knowledge of the dependence of profiles on the plasma parameters may lead to new physical insight. In view of the considerable experimental variation in individual profile measurements, it is however worthwhile to express by sound statistical analysis clearly the uncertainties associated with fitting a set of carefully measured profiles. Such parameterised profiles are then in a form suitable to be used by so-called interpretative plasma stability and transport codes.

It must be remarked that this chapter has concentrated on explaining the *basic theory* of statistical plasma *profile analysis*. This asks from the reader

some effort in algebraic abstraction and willingness to draw, where needed, associated 'pictures of principle' on a piece of paper. The rationale of this approach is that the reader's attention is directed to the essential mathematical aspects of the analysis. For a more graphicallly oriented analysis of profiles from the ASDEX tokamak (which has been rebuilt and is operating since 2003 as the HL-2A tokamak in Chengdu, China) the reader is referred to two case-studies on this topic, [343] and [451], which illustrate a substantial part, albeit not all, of the methods described above.

4.9 Appendix

4.9.1 Profile Representation by Perturbation Expansion: Moments

We consider the Hermite polynomial expansion given by (4.13). For any fixed value of c_0, the coefficients $a_1, a_2, \ldots$ are linearly related to the moments of the temperature distribution. To see this, we write (4.13) as

$$T(r) = T_0 \sum_{n=0}^{\infty} (-1)^n a_n \varphi_{c_0}^{(n)}(r) \qquad \text{with} \quad \varphi_{c_0}^{(n)}(r) = (d/dr)^n e^{-c_0 r^2} \ , \quad (4.90)$$

with $a_0 = 1$. From this it follows that the two-sided Laplace transform equals

$$\int T(r) e^{-sr} dr = T_0 \big(\frac{\pi}{c_0}\big)^{1/2} e^{\frac{s^2}{4c_0}} \sum_{n=0}^{\infty} (-1)^n a_n s^n \ . \quad (4.91)$$

Since the Laplace transform generates the moments of the temperature profile, we get, by expanding (4.15) as a power series in s,

$$T_0 (\frac{\pi}{c_0})^{1/2} \Big(a_{2k} + \frac{a_{2k-2}}{4c_0} + \cdots + \frac{a_0}{k!(4c_0)^k} \Big) = \frac{m_{2k}}{(2k)!} \ . \quad (4.92)$$

A similar expression holds for the odd moments. Hence, for each fixed value of c_0 one can, from the fitted regression coefficients $a_1, a_2, \ldots$ (and their estimated covariance matrix), easily calculate the moments $m_1, m_2, \ldots$ (and their covariance matrix). Reversely, by estimating the moments by some method, one could derive the corresponding estimates for the polynomial coefficients.
Remark. Note that for symmetric profiles, the Fourier transform of (4.13) is a Gaussian function times a power series with coefficients $a_1, a_2, \ldots$, i.e., assumes the form (4.12). In this sense, the representations given by (4.12) and (4.13) are dual.

4.9.2 Variance and Bias for Volume-Averaged Global Quantities

It is recalled that, by definition,

$$\mathrm{MSE}(\langle \hat{n}_e \rangle) = E(\langle \hat{n}_e \rangle - \langle n_e \rangle)^2 = \mathrm{var}\ \langle \hat{n}_e \rangle + \mathrm{bias}^2 \langle \hat{n}_e \rangle\ , \tag{4.93}$$

where var $\langle \hat{n}_e \rangle = E\big(\langle \hat{n}_e \rangle - E(\langle \hat{n}_e \rangle)\big)^2$, and bias $(\langle \hat{n}_e \rangle) = E(\langle \hat{n}_e \rangle) - \langle n_e \rangle$, and $E(X)$ denotes the mathematical expectation of a random variable X. The variance and the bias of $\langle \hat{n}_e \rangle$ depend in a simple manner on the covariance function

$$\mathrm{cov}\left(\hat{n}(r), \hat{n}(r')\right) = E\left(\hat{n}(r) - E\hat{n}(r)\right)\left(\hat{n}(r') - E\hat{n}(r')\right)\ . \tag{4.94}$$

By applying a Taylor series expansion, it is derived that, for var $\hat{\mu}(r) \ll 1$ and var $\hat{\mu}(r') \ll 1$,

$$\mathrm{cov}\left(\hat{n}(r), \hat{n}(r')\right) = n(r)n(r')\mathrm{cov}\left(\hat{\mu}(r), \hat{\mu}(r')\right)\ . \tag{4.95}$$

Because of the linear structure of our linear model for $\mu(r)$, we have

$$\mathrm{cov}\left(\hat{\mu}(r), \hat{\mu}(r')\right) = \boldsymbol{f}^t(r)\boldsymbol{V}(\hat{\boldsymbol{\alpha}})\boldsymbol{f}(r')\ . \tag{4.96}$$

As $\langle \hat{n}_e \rangle = \int \hat{n}(r)h(r)dr$, it is easily derived that the variance of $\langle \hat{n}_e \rangle$ is obtained by integrating $\mathrm{cov}\left(\hat{n}(r), \hat{n}(r')\right)$ over $h(r)dr$ and $h(r')dr'$. Hence, the full expression is given by (4.83).

The bias of $\langle \hat{n}_e \rangle$ can clearly be written as $\int_0^1 \mathrm{bias}\left(\hat{n}(r)\right)h(r)dr$. Again by Taylor expansion, it is derived that, for var $\hat{\mu}(r) \ll 1$,

$$\mathrm{bias}\ \hat{n}(r) \simeq \frac{1}{2}\mathrm{var}\ \hat{n}(r)\ , \tag{4.97}$$

where var $\hat{n}(r) = \mathrm{cov}\left(\hat{n}(r), \hat{n}(r)\right)$ can be estimated from (4.95). Hence, the full formula is given by (4.84).

Exercise 4.1. (*) Derive a formula and an estimate for the skewness of $\langle \hat{n}_e \rangle$ under the (asymptotic) approximation that $\hat{\mu}(r)$ is estimated with skewness γ_1 and excess of kurtosis γ_2 both equal to zero. Hint: A multi-dimensional analogue of Theorem 1.8 in Sect. 1.8 has to be utilised, while generalising the bivariate log-normal example explained in Chap. 5 of [33].

5 Discriminant Analysis

5.1 Introduction

The abbreviation ELM stands for edge localised mode. ELMs [447] have been observed from the beginning of H–mode research [373] and have aroused active interest ever since [117, 118, 251, 456, 671, 750, 752, 753]. For us, it suffices here to know that they are, usually regular, periodic perturbations of the plasma edge in which plasma particles and energy are expelled. In a very rough empirical classification, based solely on the H_α (or D_α) signal, one can distinguish between three types of H–mode: ELM-free H–mode, H–mode with giant ELMs, and H–mode with small ELMs. The latter two types are jointly called ELMy H–mode. The term H–mode, introduced in [714], stands for H(igh) confinement mode as opposed to L–mode, which stands for L(ow) confinement mode. In practice, the confinement time is some 50% to 100% higher in ELMy H–mode than in (normal) L–mode. While a plasma is sometimes called the 'fourth state of matter', the word mode is used here to denote some sub-state of the high-temperature plasma discharge, remotely akin to different states of superfluid helium (at very low temperatures). The discrimination of these three types of H–mode is an important concern for future machines, because the various H–mode discharges exhibit quite different behaviour. ELM-free H–mode discharges show the largest increase in confinement time with respect to L–mode, but are most vulnerable to the accumulation of impurities and the resultant radiation collapse [174, 668]. Long sustainment of ELM-free H–mode seems at present possible for approximately one or two confinement times only [94, 310, 439, 553]. The presence of giant ELMs during the H–mode moderates this problem, but the repetition of the large heat pulses is a serious constraint in designing the divertor plate and the first wall [274, 317, 513]. Large heat pulses cause a transient peak in the temperature of the divertor plate, so as to enhance the erosion rate and impurity generation. The H–mode with small and frequent ELMs is presently the best candidate for future plasmas aiming at sustained thermonuclear burning [512]. Hence, one needs a guiding principle to avoid the appearance of the ELM-free H–mode as well as the H–mode with giant ELMs, in order to enhance the probability of (quasi) steady-state operation in future experimental fusion reactors. Several experimental efforts have been made to

look for the most efficient control parameter to realise H–mode discharges with small frequent ELMs [160, 469, 635, 647, 713].

The ITER project induced the construction of an international global confinement H–mode database, which was assembled by a joint effort from several tokamak teams. Since 1992, the first version of this database (ITERH.DB1), see [110], contains global confinement data from the ASDEX[(1)], DIII–D[(2)], JET[(3)], JFT–2M[(4)], PBX–M[(5)], and PDX[(6)] tokamaks.[1] In the second version (ITERH.DB2), more attention than before was given to an accurate determination of the thermal confinement time, see [360,678]. In the third version (ITERH.DB3), the dataset was extended with additional data from the six original tokamaks as well as from a number of new machines (ALC C–MOD[(a)], ASDEX Upgrade[(b)], COMPASS[(c)], JT–60U[(d)], MAST[(e)], NSTX[(f)], START[(g)], TUMAN–3M[(h)], TCV[(i)], TdeV[(j)], TFTR[(k)], T–10[(l)])[2], see [362, 671][3]. The three types of H–mode that were discussed above have been routinely distinguished in this database. The scaling of the energy confinement time with respect to the plasma parameters is not precisely the same for the various types of H–mode [110, 348, 360]. Hence, also from this point of view, although secondary to divertor considerations, it is useful to identify the plasma parameter regimes where the various types of H–mode occur.

As a first step to analyse this question, we apply the method of discriminant analysis [207, 399, 404, 442, 480] to the H–mode database. We want to study the regions in plasma parameter space where two classes of H–mode, i.e., *class-1* (ELM-free or with Giant ELMs) and *class-2* (with Small ELMs) occur. The analysis is performed for each device separately. In the practical section of this chapter, attention is restricted to the analysis of data from ASDEX, JET, and JFT–2M. Attention is also restricted to ITERH.DB1, which is an established subset of ITERH.DB3, is generally available on a public internet server, see http://pc-sql-server.ipp.mpg.de/HmodePublic, and which has also been used for the global confinement-time analysis published in [110]. As explained in the preface, the datasets are analysed here primarily with a didactic purpose in mind.

The arrangement of this chapter is as follows. After an exposition of its theoretical background, discriminant analysis is applied to several subsets of the H–mode confinement dataset ITERH.DB1, in order to find the region in plasma parameter space in which H–mode with small ELMs (Edge Localised

[1] Affiliations: [(1)]IPP, Garching, Germany; [(2)]General Atomics, San Diego, USA; [(3)]Euratom/UKAEA, Culham, UK; [(4)]JAERI, Naka, Japan; [(5)]PPPL, Princeton, USA; [(6)]PPPL, Princeton, USA.

[2] Affiliations: [(a)]MIT, Cambridge, USA; [(b)]MPI, Garching, Germany; [(c)]UKAEA, Culham, UK; [(d)]JAERI, Naka, Japan; [(e)]UKAEA, Culham, UK; [(f)]PPPL, Princeton, USA; [(g)]UKAEA, Culham, UK; [(h)]Ioffe Institute, St. Petersburg, Russia; [(i)]CRPP, Lausanne, Switzerland; [(j)]INRS, Varennes, Canada; [(k)]PPPL, Princeton, USA; [(l)]Kurchatov Institute, Moscow, Russia.

[3] Still more machines have contributed to the L–mode database, see [362,370,747].

Modes) is likely to occur. The boundary of this region is determined by the condition that the probability for such a type of H–mode to appear, as a function of the plasma parameters, should be (1) larger than some lower bound, implying a sufficiently high 'typicality', and (2) larger than the corresponding probability for other types of H–mode (i.e., H–mode without ELMs or with giant ELMs).

In Sect. 5.2, we present a guided tour to those theoretical aspects of discriminant analysis which are needed to understand the background of the subsequent practical section. In particular, attention is paid to the relationship between discriminant analysis and regression analysis. In practice, the discrimination is performed for the ASDEX, JET and JFT–2M tokamaks (a) using four instantaneous plasma parameters (injected power P_{inj}, magnetic field B_t, plasma current I_p and line averaged electron density $\overline{n}_e$) and (b) taking also memory effects of the plasma and the distance between the plasma and the wall into account, while using variables that are normalised with respect to machine size.

In Sect. 5.3, we describe the datasets used, present a preliminary graphical analysis, and apply discriminant analysis in various settings to predict the occurrence of the various types of ELMs. In particular, we compare 'parametric discrimination' (using linear as well as quadratic discriminant functions) with 'non-parametric discrimination' (using kernel density estimates and the multinomial independence model, respectively). We discuss the explicit forms of the linear and quadratic boundaries, and compare the performance of the various methods. In Sect. 5.3.2 this is done using four or five instantaneous plasma parameters that are usually also applied in global confinement time analyses, i.e, plasma current, magnetic field, injected power, electron density, and isotope composition, all considered at the same time point as that of the occurrence of ELMs. In Sect. 5.3.3 the elapsed time since the L-H transition is taken into account, and the Ohmic target density replaces the instantaneous density as a discriminating variable. In other words, we examine an effect of the plasma's 'memory'. Furthermore, the plasma–wall distance is used as an additional variable, and all variables are normalised with respect to machine size. Condensed results of the analyses are presented in tables and physical interpretations are discussed in the main text whenever they are deemed to be appropriate. A summary and discussion is presented in Sect. 5.4.

Generally speaking, it is found that there is a substantial overlap between the region of H–mode with small ELMs and the region of the two other types of H–mode. However, both the ELM-free H–modes and the H–modes with giant ELMs appear relatively appear in the region, that, according to the analysis, is allocated to small ELMs. A reliable production of H–mode with only small ELMs seems well possible by choosing this regime in parameter space.

A more elaborate distinction between type I ELMs (large ELMs with good confinement), type III ELMs (small ELMs with somewhat reduced confine-

ment) and an intermediate type II ELMs (small ELMs with a relatively good confinement) is not made. The latter type of ELMs have a more sophisticated signature, and their identification [156] was not operational across the machines at the time the first version of the database (ITERH.DB1) was assembled. In the physical motivation of the fourth exercise dataset in Chap. 7 (Sect. 7.5) the reader can find some further background on the various types of ELMs. In this chapter we adhere to the practically easier distinction of *class-1* ('large ELMs or 'no ELMs') and *class-2* ('small ELMs') and we will use the abbreviation HSELM for H–mode with small ELMs and HGELM for H–mode with large ('giant') ELMs, the operational definition of the latter being that the ELM amplitude during H–mode exceeds the level of the D_α signal during L–mode.

Despite this restriction, the methods presented in Sect. 5.2 are obviously also applicable to more extended and somewhat differently oriented datasets, oriented toward the physical question of determining an appropriate operating window with suitable ELMy H–mode, compatible with an extended lifetime of the ITER divertor and at the same time maintaining an energy confinement time which is sufficiently long to establish a thermal isolation of the plasma producing prevalent alpha–particle heating in ITER [350]. In [226], discriminant analysis is used for automated regime identification in ASDEX Upgrade (L–mode, different variants of H–mode).

We will not attempt here to arrive at a unified discrimination across the machines. Projection from one machine to another is, in practice, a more complicated issue and is not treated in the present chapter of this book, where attention is restricted to the situation that a training sample from the device under consideration is available.

5.2 Theory

Wir machen uns Bilder der Tatsachen

Wittgenstein, Tractatus

5.2.1 Informal Introduction to Discriminant Analysis

In order to assess quantitatively in which regions of plasma parameter space the various types of ELM are likely to occur, we must formulate relevant parts of the mathematical theory of discrimination. Here we present some heuristic considerations. A more precise description of the theory is given in the next sub-section. An important aspect is to determine the mathematical form of the boundary of the regions of interest.

For simplicity, let us consider two classes of discharges with ELMs. We want to determine the boundary between the regions in plasma parameter

space where these two classes are expected to occur. One way to find this boundary is by estimating the probability distributions of the two classes of ELMy shots over plasma parameter space. We consider boundaries such that, locally, class-1 (e.g., non-HSELM) discharges are more probable than class-2 (e.g., HSELM) discharges on one side of the boundary, and vice versa on the other side. The probability density of finding class-j discharges is denoted by

$$f_j(\boldsymbol{x}) \qquad (j = 1, 2)\,, \tag{5.1}$$

where $\boldsymbol{x}$ is an n-dimensional vector of variables (i.e., plasma parameters), such as the injected power P_{inj}, plasma current I_p, magnetic field B_t, electron density $\overline{n}_e$, etc. These densities have to be estimated from the available data. The estimates depend on the assumed class of probability distributions, see Sect. 5.2.2. The boundary $\mathbf{B}$ is defined by $\mathbf{B} = \{\boldsymbol{x} | f_1(\boldsymbol{x}) = f_2(\boldsymbol{x})\}$. The region $\mathbf{R}_1$ satisfying the condition $\mathbf{R}_1 = \{\boldsymbol{x} | f_1(\boldsymbol{x}) > f_2(\boldsymbol{x})\}$ is the region where class-1 shots are expected, whereas class-2 shots are expected in the region $\mathbf{R}_2 = \{\boldsymbol{x} | f_1(\boldsymbol{x}) < f_2(\boldsymbol{x})\}$. The boundary $\mathbf{B}$ and the regions $\mathbf{R}_{1,2}$ can be calculated for each set of plasma variables $\boldsymbol{X} = (X_1, \ldots, X_p)$. The dimension p can be as large as the number of parameters which fully characterise the discharge. Among this large set of parameters, we look for key parameters which determine the boundary $\mathbf{B}$. In statistics, this is called the problem of variable selection. Obviously, this requires in practice a good interplay between physical considerations and statistical methods. Given a boundary in a higher dimensional space, one can identify which linear combinations of plasma parameters are locally important to discriminate between the classes: obviously, they consist of those linear combinations of which the hyperplanes of constant values are 'close' to the tangent plane of the boundary. This means that the gradient of the linear combination should have a large component in the direction of the gradient of the ratio $f_1(\boldsymbol{x})/f_2(\boldsymbol{x})$.

The allocation regions $\mathbf{R}_1$ and $\mathbf{R}_2$ are influenced by considerations that take the risks associated with the various types of misallocation into account. Usually, misclassifying a (new) observation as class-1 has different consequences than misclassifying a (i.e., another!) new observation as class-2. For instance, mispredicting ELM-free H–mode as ELMy H–mode is more harmful, from the viewpoint of impurity accumulation, than mispredicting ELMy H–mode as ELM-free. In that case, the boundary $\mathbf{B}$ between the regions $\mathbf{R}_1$ and $\mathbf{R}_2$ is no longer optimal, and improvement is possible by constructing boundaries so as to reduce the probability of dangerous errors, which amounts to minimising, in some sense, the expected losses.

Another extension is to consider an additional area in parameter space. For instance, one region is assigned to class-1, another to class-2, and a third region is the 'grey' area, in which the probabilities associated with the two classes are either not distinctively different or are both negligibly low. (In fact, distinguishing between these two reasons for non-allocation, one has two additional regions.) By using such 'non-allocation regions', the prediction

of 'class-1' (or 'class-2') can be made with less expected error. This kind of analysis is sometimes called 'discrimination in non-forced decision situations'.

Each of these cases require more precise formulations, which are discussed in the next section.

5.2.2 Statistical Aspects of Discriminant Analysis

Discriminant analysis is a well developed branch of statistics and also still a field of active research (see, e.g., [10, 215, 378, 404, 408, 442, 457, 538, 586, 587, 593, 637]). Part of the algorithms are implemented in specialised and/or general-purpose statistical software packages, such as [150,182,225,694]. Here, we will discuss only those parts of the theory that are needed to understand the background of the practical analyses in Sect. 5.3.

One can approach discriminant analysis from a purely data-descriptive point of view and from a probabilistic point of view. (Both approaches, but most easily the latter one, can be incorporated into a decision theoretical framework.) In the latter approach, a probabilistic model is used to describe the situation. The applicability of such a model in non-random situations may be questioned from a fundamental point of view. However, such a probabilistic framework is almost indispensible if one wants to estimate the performance of procedures in future situations, and to express uncertainties in various estimates. Moreover, it often leads to procedures that are also sensible from a data-descriptive point of view. Or, conversely: A specific procedure can often be viewed upon as a data-descriptive one, with little further interpretation, and as a probabilistic one, with considerably more interpretation, the validity of which is of course dependent on the adequacy of the framework. Sometimes a procedure developed in one probabilistic framework can also be interpretated in another probabilistic framework, which may be more relevant for the data at hand. We will return to this when discussing the relationship between discriminant analysis and regression analysis. The author holds the view that the use of such probabilistic models is justified as long as one does not take them 'too seriously', i.e., one should realise their approximative model character. Here we shall discuss methods of a probabilistic nature, but we take care to introduce the various probabilistic structures only step by step. The reader is referred to [358] for further reflections on the statistical and probabilistic background according to various schools of thought.

With the term (multivariate) observation one generally denotes a basic constituent (consisting of various components) that is liable to repetitive measurement or drawn from a population. In this context we use it for a vector of plasma parameters, say $(P_{inj},\ I_p,\ B_t,\ \overline{n}_e)$, measured at a certain time point (or even at several time points) during a discharge. The principal object of the excercise is to predict the occurrence of the type of ELMs as a function of the plasma parameters on the basis of historical 'training samples' of such observations for each class of ELMs.

In the case of two kinds of observations that are distributed according to two elliptical multivariate distributions, say in p dimensions, there is a geometrically evident family of boundaries that can be used for prediction: Assign a new observation to group 1 if $D_1 < cD_2$, where D_1 and D_2 are distances of the particular observation to the centers of gravity of group 1 and 2, respectively. Obviously, these distances have to be measured in the metric defined by the ellipses. Intuitively, the choice $c = 1$ seems best. (In a decision theoretic formulation, the constant c depends on the losses one wants to minimise. In a symmetric loss situation, the choice $c = 1$ is appropriate.) In this case the boundary becomes simple if two distributions differ only by their first moment: it is the $(p-1)$-dimensional plane conjugate to the line joining the expectation values of the two distributions. (We have assumed that the ellipses describing the contours of constant probability density are not degenerate.) 'Conjugate' means the same as 'perpendicular in the metric given by the ellipses'. Such a plane, obviously, passes through the (real or imaginary) generically $(p-1)$-dimensional intersection manifold ('ellipse') of any pair of p-dimensional ellipses with the same 'radius'.

A specialisation of the above elliptical situation is the case of two multivariate normal distributions, and a generalisation is the situation that we have two shifted multivariate distributions, with convex contours of equal probability density that define a seminorm on $\mathbb{R}^p$. For analogy with the normal case, it is convenient to 'scale' the seminorm according to the relation $f(\boldsymbol{x}) \propto \exp[-\frac{1}{2}D^2(\boldsymbol{x})]$, where $f(\boldsymbol{x})$ denotes the probability density of the multivariate distribution. Using an analogy from mechanics, we can call $-2\log f(\boldsymbol{x}) = D^2(\boldsymbol{x})$ the potential $V(\boldsymbol{x})$.

We shall now describe discriminant analysis while introducing some more probabilistic structure. Again, for simplicity, this is done for two groups. Generalisation to $k > 2$ groups is in principle straightforward, though in practice sometimes unwieldy because of the combinatorial complexity.

The basic method is: Estimate the probability density for the two groups, and allocate a new observation according to a rule based on (an estimate of) the above described family of boundaries. Actual allocation requires the determination of a single member of this family. This choice, other than the obvious one, $f_1(x) = f_2(x)$, may be influenced by two considerations: (a) the assignment of losses for the various types of error that can occur and (b) the assignment of prior probabilities with respect to the occurence of the two types of observations. Let $\rho_h, h = 1, 2$, denote these prior probabilities (with $\rho_1 + \rho_2 = 1$), and let $f_h(\boldsymbol{x})$ denote the probability density for group h. Then, according to Bayes' theorem, the conditional or 'posterior' probability $\rho_{h|\boldsymbol{x}}$ to belong to group h, given the vector of observations $\boldsymbol{x}$, can be written as

$$\rho_{h|\boldsymbol{x}} = \frac{\rho_h f_h(\boldsymbol{x})}{\sum_{h'=1}^{2} \rho_{h'} f_{h'}(\boldsymbol{x})} \,. \tag{5.2}$$

The above class of allocation rules considers regions of which the boundaries can, equivalently, be characterised by (i) the difference in metric or

potential, (*ii*) the ratio of probability densities, and (*iii*) the posterior probabilities. The relationship between these characterisations is given by the following equations, which can be easily derived:

$$V_2(\boldsymbol{x}) - V_1(\boldsymbol{x}) = D_2^2(\boldsymbol{x}) - D_1^2(\boldsymbol{x}) = 2d \;, \tag{5.3}$$

$$f_1(\boldsymbol{x})/f_2(\boldsymbol{x}) = e^d \;, \tag{5.4}$$

and

$$\log \frac{\rho_{1|\boldsymbol{x}}}{1-\rho_{1|\boldsymbol{x}}} = \log \frac{\rho_1}{1-\rho_1} + d \;. \tag{5.5}$$

The latter expression means that d equals the shift from the prior to the posterior probabilities on the logit scale. (The S-shaped logit function $\mathrm{logit}(p) = \log\big(p/(1-p)\big)$ transforms the interval $(0,1)$ into $(-\infty,+\infty)$.) Note that in going from (*i*) to (*iii*) more probabilistic structure is gradually introduced. A basic question is the choice of d. One choice is $d = 0$, another choice is to determine d by the requirement that the $\rho_{1|\boldsymbol{x}} = 50\%$ (in that case, d depends on ρ_1 and is obviously 0 for the 'Laplacian' choice to represent initial ignorance by $\rho_1 = 1/2$). A third choice is based on a loss-function formulation. Losses are associated with actions. In the two-group situation, there can be two or more relevant actions. We consider: a_1: assign to group 1, a_2: assign to group 2, and a_0: no assignment. The actions are to be based on the value of the experimental variable $\boldsymbol{x} \in \mathbb{R}^p$. The mapping $d : \mathbb{R}^p \to \{a_0, a_1, a_2\}$ is called a decision function. A decision procedure fixes a decision function, which in the above case can be characterised by three regions $\mathbf{R}_0$, $\mathbf{R}_1$, and $\mathbf{R}_2$, respectively, where

$$\mathbf{R}_0 = d^{-1}(a_0) = \{x \in \mathbb{R}^p \mid d(x) = a_0\} \;, \tag{5.6}$$

and similarly for $\mathbf{R}_1$ and $\mathbf{R}_2$. (If $\mathbf{R}_0$ is empty, we are in the two-decision situation described above.) The losses depend on the actions as well as on the true state of nature, and are conveniently summarised by a loss matrix. For instance:

	a_0	a_1	a_2
1	l_{01}	0	l_{21}
2	l_{02}	l_{12}	0

(The first index of ℓ denotes the action, and the second index of ℓ indicates the true state of nature.) Some losses have to be specified, be it $\ell_{01} = \ell_{02} = \infty$, and $\ell_{21} = \ell_{12} = 1$. Then the risk ('expected loss') can be expressed as a function of the true state of nature and the decision procedure:

$$\mathrm{R}(d,1) = \ell_{01} \int_{R_0} f_1(\boldsymbol{x}) d\boldsymbol{x} + \ell_{21} \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} \;, \tag{5.7}$$

$$\mathrm{R}(d,2) = \ell_{02} \int_{R_0} f_2(\boldsymbol{x}) d\boldsymbol{x} + \ell_{12} \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} \ . \tag{5.8}$$

One principle is to choose a decision rule d_{mm} such that the maximum expected loss is minimised, i.e.,

$$\max_h \mathrm{R}(d_{mm}, h) = \min_d \max_h \mathrm{R}(d,h) \qquad (h = 1,2) \ . \tag{5.9}$$

This is called the minimax rule. Another principle is to minimise some linear combination of the two risks. A natural candidate for the weights of this linear combination is the set of prior probabilities, i.e.,

$$\mathrm{R}(d, \boldsymbol{\rho}) = \rho_1 R(d,1) + \rho_2 R(d,2) \tag{5.10}$$

is minimised. This is called a Bayes decision rule with respect to the prior distribution $\boldsymbol{\rho} = (\rho_1, \rho_2)$. Its definition can be reformulated as: a Bayes decision rule with respect to $\boldsymbol{\rho}$ minimises, as a function of $\boldsymbol{x}$ the sum of the probability densities, weighted by both the losses and the prior probabilities, i.e., the sum of the 'risk densities'

$$\sum_{h=1}^{2} \ell\big(d(\boldsymbol{x}), h\big) \rho_h f_h(\boldsymbol{x}) \ . \tag{5.11}$$

One can easily see that, in the special case of 0–1 losses and no doubt region, the Bayes rule assigns an object with score $\boldsymbol{x}$ to population h according to the largest value of $\rho_h f_h(\boldsymbol{x})$, i.e., to the largest posterior probability. Unfortunately, prior probabilities are sometimes hard to assess, or even make little physical sense. Therefore it is useful to look at the Bayes rule with respect to a prior distribution $\boldsymbol{\rho}_{\mathrm{lf}}$, which is 'least favourable' in the sense that

$$\mathrm{R}(d, \boldsymbol{\rho}_{\mathrm{lf}}) = \max_{\boldsymbol{\rho}} \mathrm{R}(d, \boldsymbol{\rho}) \ . \tag{5.12}$$

It can be shown that under some regularity conditions, the Bayes rule with respect to the least favourable prior distribution coincides with the minimax rule, see [191]. For classical probabilities, in contrast to upper and lower probabilities (see [358]), $\rho_2 = 1 - \rho_1$.

Let us now consider some special cases. We consider the trivial (0–1) loss-function formulation. For multivariate normal densities with equal covariance matrices, the Bayes rule with respect to equal prior probabilities allocates an observation $\boldsymbol{x}$ to that group h for which the squared Mahalanobis distance (or 'potential')

$$D_h^2(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \boldsymbol{\Sigma}_{(h)}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}) \tag{5.13}$$

is minimal. For arbitrary prior probabilities, the 'effective' distance is changed into:

$$D_h'^2(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \boldsymbol{\Sigma}_{(h)}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}) - 2 \log \rho_h \ , \tag{5.14}$$

and for unequal covariance matrices into:

$$D_h''^2(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \boldsymbol{\Sigma}_{(h)}^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}) - 2\log \rho_h + \log\bigl(\det(\boldsymbol{\Sigma}_{(h)})\bigr) \,. \tag{5.15}$$

As usual, superscript t is used to denote vector transposition. The quantity $e^{-\frac{1}{2}D_h^2(\boldsymbol{x})}$, proportional to $f_h(\boldsymbol{x})$, and varying between 1 (for $D_h(\boldsymbol{x}) = 0$) and 0, is sometimes called the (classical) typicality of observation $\boldsymbol{x}$ to belong to group h.

In general, it requires rather elaborate numerical procedures to determine the minimax rule, and for simplicity one calculates a Bayes rule with respect to equal prior probabilities, or according to prior probabilities that are proportional to the observed relative frequencies of occurrence of the two groups in the sample.

The discrimination surfaces between group 1 and group 2 are given by

$$D_2^2(\boldsymbol{x}) - D_1^2(\boldsymbol{x}) = c \,, \tag{5.16}$$

where c depends on the ratio of the prior probabilities, the ratio of the determinants of the covariance matrices $\boldsymbol{\Sigma}_{(1)}$ and $\boldsymbol{\Sigma}_{(2)}$, and, in general, also on the ratio of the losses ℓ_{21} and ℓ_{12} (at least if $\ell_{01} = \ell_{02} = \infty$). In general, these surfaces are quadratic surfaces, in practice ellipses or hyperbolae. If and only if $\boldsymbol{\Sigma}_{(1)} = \boldsymbol{\Sigma}_{(2)} = \boldsymbol{\Sigma}$, they constitute parallel hyperplanes $\boldsymbol{w}^t\boldsymbol{x} = c'$, characterised by a vector of weights

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \,. \tag{5.17}$$

In discriminant analysis, these are often called the weights of Fisher's linear discriminant function. They share a number optimality properties and interpretations. Weight vectors belong mathematically to the dual of the data-space $\mathbb{R}^p$. Often they are only important up to a multiplicative factor, so that, mathematically speaking, they constitute a $(p-1)$-dimensional projective space. Plotting the weight vectors in the original data-space $\mathbb{R}^p$, they mark directions on which the data can be perpendicularly projected. For any random vector $\boldsymbol{X}_i$ in $\mathbb{R}^p$, $\boldsymbol{w}^t\boldsymbol{X}_i$ divided by the length of the vector $\boldsymbol{w}$ is such a projection. We give three interpretations, which hold if the origin of the coordinate system in which the vector $\boldsymbol{w}$, given by (5.17) with $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$ estimated by their sample analogues, is chosen somewhere on the line connecting the centers of gravity of the two groups.

(*i*) The above weights correspond to directions for which the ratio of the between-class variance and the sum of the within-class variances is maximal.

(*ii*) Considering data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ in the data space $\mathbb{R}^{(n_1+n_2)}$, where n_1 and n_2 denote the sample sizes, $w_1\boldsymbol{x}_1 + \cdots + w_p\boldsymbol{x}_p$ can be viewed as that linear combination (up to a proportionality factor) of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ which maximises the (sample) correlation coefficient with some vector

$\boldsymbol{y}$, where $\boldsymbol{y}$ denotes the class membership. Since the correlation coefficient is independent of both location and scale transformations of $\boldsymbol{y}$ as well as (uniformly) of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, the coding of the class membership and any proportionality factor in $\boldsymbol{w}$ are immaterial.

(*iii*) The weights in (5.17) can be obtained (up to a linear transformation) by linear regression of the vector $\boldsymbol{y}$ on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. If the vector $\boldsymbol{y}$ is coded as $c_1 = -\frac{n_2}{n_1+n_2}$ for group 1 and $c_2 = c_1 + 1$ for group 2, then precisely the weights in (5.17) are formally recovered by the linear regression.[4] This equivalence, which is useful in computational practice, goes back to R.A. Fisher [202], see Sect. 4.4 of [126] and also the book by Flury and Riedwyl [207], for a vivid practical account.

Discriminant analysis of $k > 2$ groups can be based on separate discriminant analyses between each of the $k(k-1)/2$ pairs of groups. However, the practical complexity of this approach increases rapidly with k. A more parsimoneous description is given by canonical correlation analysis between $k-1$ binary group membership codes and the explanatory data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. This is, among others, implemented in the procedure DISCRIM of the package SAS, see [578, 581]. (For two groups, it reduces to multiple linear regression analysis of a single group-membership vector.) In view of the general situation, the regression coeffcients are often called canonical coefficients. They depend on how the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are (separately) normalised. Often, one divides $\boldsymbol{x}_j$ by the square root of the jth diagonal element of some suitable covariance matrix, for instance, the 'total sample' covariance matrix or the 'pooled within-class' covariance matrix. The latter is estimated by $(n_1 + n_2 - 2)^{-1}(\boldsymbol{S}_{(1)} + \boldsymbol{S}_{(2)})$, where $\boldsymbol{S}_{(1)}$ and $\boldsymbol{S}_{(2)}$ are the sum of squares and cross-products (SSCP) matrices, corrected for the mean.

For simplicity, we shall consider in the practical sections only the 'parametric case', with multivariate normal densities, the 'non-parametric case', with arbitrary multivariate densities, and the 'multinomial independence' model, with a discretised data structure (and a global association factor).

In practice, the densities $f_h(\boldsymbol{x})$ have to be estimated. In the parametric case, this is done by substituting the standard estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In the non-parametric case, this can be done by so-called non-parametric density (e.g., kernel) estimators. For a discussion of such type of estimators, we refer to [6, 137, 266, 359] (which describes mainly the univariate situation) and to the description of PROC DISCRIM in the SAS manual [578, 581] and the references therein. The third approach starts with discretising the distribution by dividing each axis into k, say 5 or 10, intervals. In general, one then obtains a multinomial distribution. However, since the number of free parameters, $(k^p - 1)$ with p the number of discriminating variables, quickly exceeds the number of available data points, one has to make further simplifying as-

[4] Note that the usual probabilistic assumptions for linear regression models are different from those used in discriminant analysis.

sumptions. One possibility is to impose the multinomial independence model, which approximates the distribution by the product of the marginal multinomial distributions. In that case the number of free parameters is $p \times (k-1)$. It obviously only works well if the original continuous distribution can be thought to be generated by independent random variables, which will usually not be the case in practice. In the case of two groups, one could discretise the 'joint principal components' [207], which are at least uncorrelated, and then transform back to the original scale, in order to achieve approximate independence. The multinomial independence model is implemented in the program INDEP [225, 260]. A program with a number of user options for several situations, such as DIScrete, NORmal, MIXed models, and especially oriented toward expressing the uncertainty in both prior and posterior probabilities is POSCON [359, 694].

5.3 Practice

Bilder bergen die Gefahr, uns gefangenzuhalten

WITTGENSTEIN, Philosophische Untersuchungen

5.3.1 Description of the Dataset and Preliminary Analysis by Visual Inspection

We define two classes of H–mode plasmas, class–1 (without ELMs and with giant ELMs) and class–2 (with small ELMs). We look for the boundary in plasma parameter space which separates the regime of class–1 discharges from that of class–2 discharges.

The ITERH.DB1 dataset contains data from ASDEX, DIII–D, JET, JFT–2M, PBX–M, and PDX. It consists of about 2000 H–mode timeslices ('observations') of some 1000 discharges. For each timeslice, 76 global plasma parameters ('variables') have been included. Details of the dataset and its variables can be found in [110]. One of those variables is a label of the type of H–mode: ELM-free, with small ELMs, and with giant ELMs (denoted as H, HGELM, and HSELM, respectively). The characterisation of a timeslice as H, HGELM or HSELM has been made by each experimental group providing the data. The distinction between HSELM and HGELM was based on the height of the spikes in the H_α signal. If it was lower than or equal to the H_α level during the L–mode, the timeslice was labeled 'HSELM'. From the H–mode database, we dropped the pellet shots, and kept only the semi-stationary H–mode timeslices, for which $\dot{W}_{mhd}$ and, if available, $\dot{W}_{dia}$ are between -0.05 and $+0.35$ times the injected power. Otherwise we applied none of the restrictions used in the 'standard dataset' used in [110].

We do not intend to present here an exhaustive practical analysis. We selected subsets from the original dataset, each consisting of shots in a single magnetic configuration (either DN or SN)[5] from one of the machines ASDEX, JET, or JFT–2M. The reason for this is that according to a preliminary investigation, DN and SN discharges indeed seemed to be associated with somewhat different regions of ELM occurrence. In several of the six possible groups, the plasma parameters have been varied rather little and/or only a few datapoints were available. Hence, for convenience, we restrict attention to the ASDEX (DN), JET (SN), JFT2M (SN) data in the database. The total numbers of observations for each tokamak are given in the following summary table.

Device	*Number of Observations*	
ASDEX	351	(only DN: 206)
JET	522	(only SN: 466)
JFT–2M	384	(only SN: 373)
total	1257	(main subsets: 1045)

In addition, in Fig. 5.1, we made a graphical representation showing the distributions of the discharges in plasma parameter space. Attention is restricted to discharges in (a) the $P_{inj} - I_p$ plane, (b) the $P_{inj} - \overline{n}_e$ plane, and (c) the $\overline{n}_e - I_p$ plane, and (d) the $seplim/a - P_{inj}$ plane. As usual, P_{inj} is the injected power, I_p is the plasma current and $\overline{n}_e$ is the line averaged plasma density.

In Fig. 5.1, the symbols $\cdot$, $\blacklozenge$ and $\circ$ denote data from ASDEX (DN), JET (SN) and JFT–2M (DN) from the ITERH.DB1 dataset [110, 351]. In Figs 5.2, 5.3 and 5.4, the symbols $\cdot$, $\blacklozenge$ and $\circ$ denote ELM-free H–mode (i.e., H), H–mode with giant ELMs (i.e., HGELM) and H–mode with small ELMs (i.e., HSELM). The symbols $\cdot$ and $\blacklozenge$ correspond to class–1 shots, whereas the symbol $\circ$ to class–2 shots.

On first sight, it appears from these figures that the two classes of H–mode overlap considerably. Distinction, however, seems possible if we look more carefully. Figure 5.1 presents the overview plot for all three tokamaks. One can see, for instance, that the range in density is similar for the three tokamaks, and neither correlated with plasma current (I_p) nor with input power (P_{inj}). Plasma current and input power are correlated, where, obviously, the largest tokamak, JET, has the highest values. The fractional distance between the separatrix and the limiter (or, not equivalently, the wall), $seplim/a$, is nominally lowest for JET and highest for ASDEX, which was in its origin designed as a divertor tokamak.

Figure 5.2 gives the plots for ASDEX. At first sight, it appears from this figure that the two classes overlap considerably. Some distinction, however,

[5] In a DN configuration, the plasma separatrix has two X–points inside the vessel, and in a SN configuration only one X–point, see Sect. 7.10.

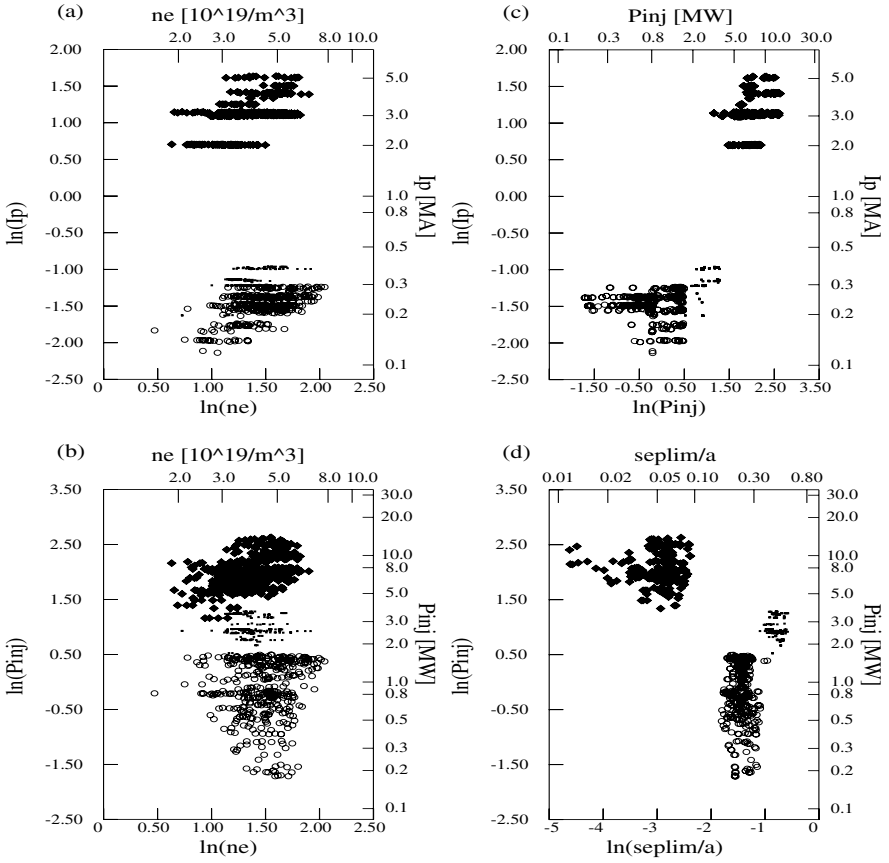

Fig. 5.1. Data from three tokamaks are displayed, corresponding to four basic variables used in discriminant analysis. Symbol · stands for DN (double-null) discharges from ASDEX, ♦ for SN (single-null) discharges from JET, and ∘ for DN discharges from JFT–2M. I_p is the plasma current, $\overline{n}_e$ is the line-average density, P_{inj} the injected heating power, and *seplim/a* the distance between the separatrix and either the limiter or the vessel.

seems possible if we look more carefully. From the plot of the distribution in the $\overline{n}_e - I_p$ plane, Fig. 5.2a, one can see that the plasma density is of discriminatory value in addition to the current. (The figure suggests that class–1 discharges occur at low current, at least if the density is not very low.) From Fig. 5.1b one can see that in the $\overline{n}_e - P_{inj}$ plane, the domain of class–1 discharges surrounds the region HSELM, except at low P_{inj}. From Fig. 5.2c, we see that H–mode with small ELMs (i.e., class–2) are predominant in the region of low P_{inj} and low I_p. For higher P_{inj} and I_p both classes

occur, but class–1 does so more frequently. These observations indicate that discrimination may only be possible by considering the joint influence of more than two plasma parameters.

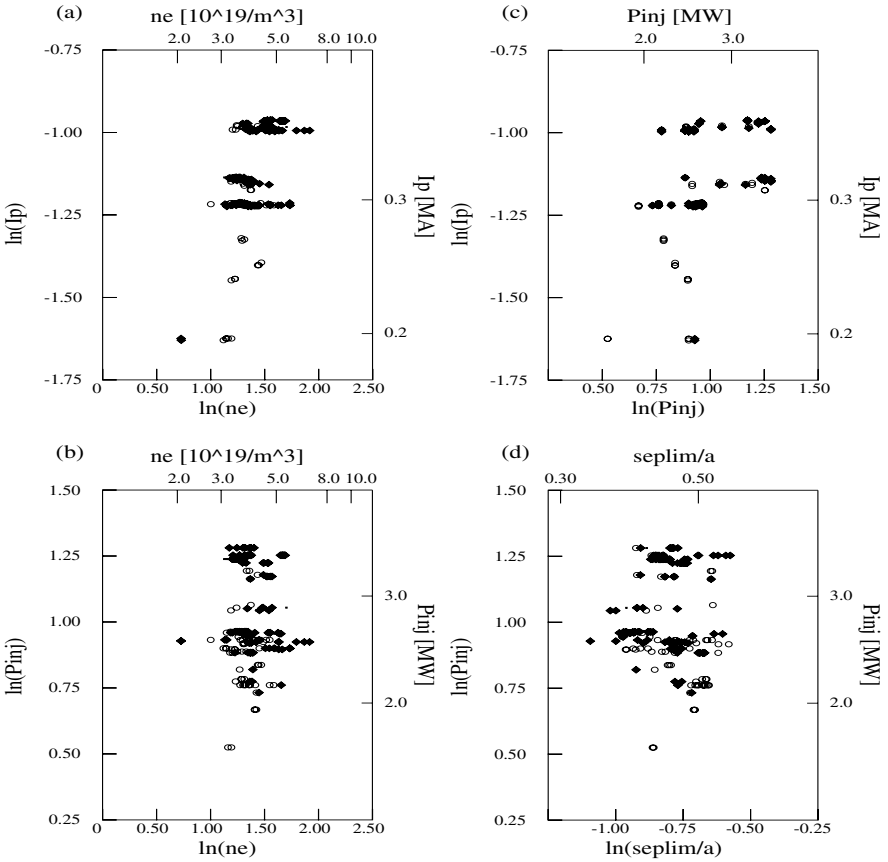

Fig. 5.2. ASDEX Double Null (DN) H $\rightarrow$ D^+ discharges. The symbols $\cdot$, $\blacklozenge$ and $\circ$ indicate ELM-free H–mode, H–mode with Giant ELMs, and H–mode with small ELMs, respectively. The first two types of H–mode belong to class–1 and the last type of H–mode to class–2. The data are projected onto the following planes: (a) $P_{inj} - I_p$, (b) $P_{inj} - \overline{n}_e$, (c) $\overline{n}_e - I_p$, and (d) *seplim/a* $- P_{inj}$. *Seplim/a* denotes the distance between the separatrix and the vessel, normalised by the minor radius.

Figure 5.3 gives the plots for JET. From Fig. 5.3a, it appears that also in the case of JET, the combination of current and density gives a better discrimination than each of the variables alone. The HSELM discharges do not occur at low current (2 MA). From Fig. 5.3b, one can see that HSELM and

HGELM, though they occur less often than ELM-free H–mode discharges, are well scattered over the density and power range. At low density and power, there tends to be a region where ELM-free discharges are less likely to occur. HGELM does not seem to occur for $\overline{n}_e < P_{inj}^{0.4}$. From Fig. 5.3c one can see that HSELM discharges tend to occur at high current (at any medium to high input power) and at medium current (3 MA) only for low input power.

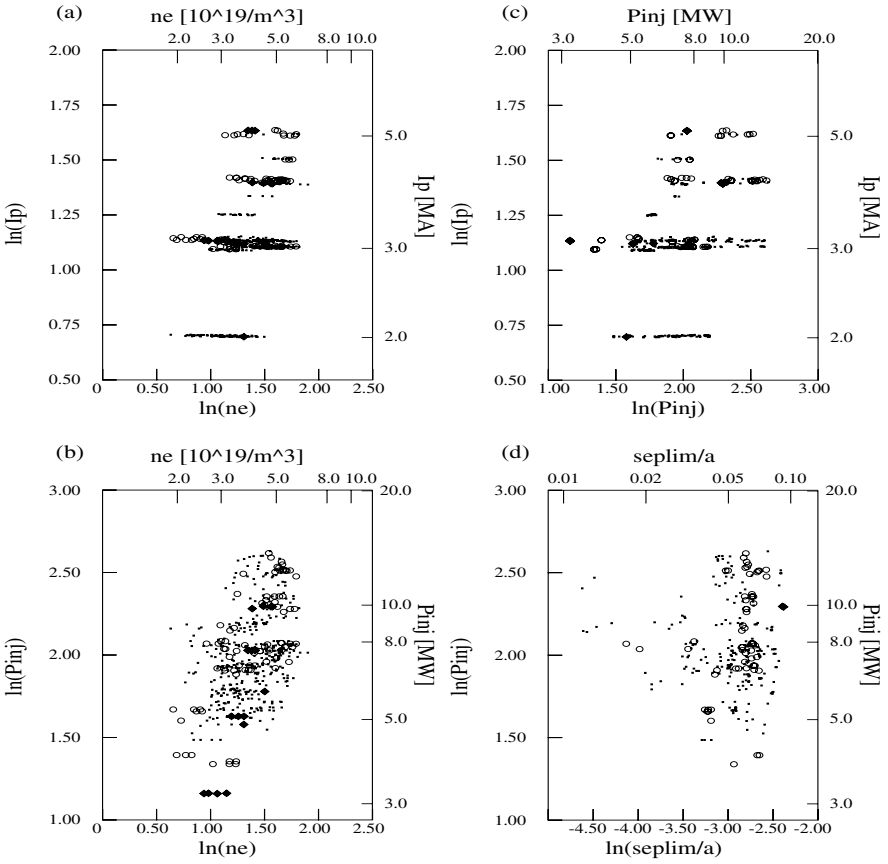

Fig. 5.3. JET Single Null (SN), $D \rightarrow D^+$ discharges. *seplim/a* denotes the distance between the separatrix and the limiter, normalised by the minor radius. For further explanation, see Fig. 5.2.

Figure 5.4 gives the plots for JFT–2M. In the present dataset, no discharge from JFT–2M showed large ELMs. One might ask whether this may be due to a (high) correlation between density and current. From Fig. 5.4a one can see that this is not the case, so that in summary, HSELM at JFT–2M seems

to avoid the region of low values for I_p, P_{inj} as well as $\overline{n}_e$. Both for hydrogen-injected deuterium H → D^+ and for hydrogen injected hydrogen H → H^+ plasmas (which are not distinguished in Fig. 5.4), it appears that HSELM does not occur at low I_p, nor at low $\overline{n}_e$. (It is noted that the latter type of discharges, H → H^+, turned out to have been performed with relatively high input power P_{inj}.)

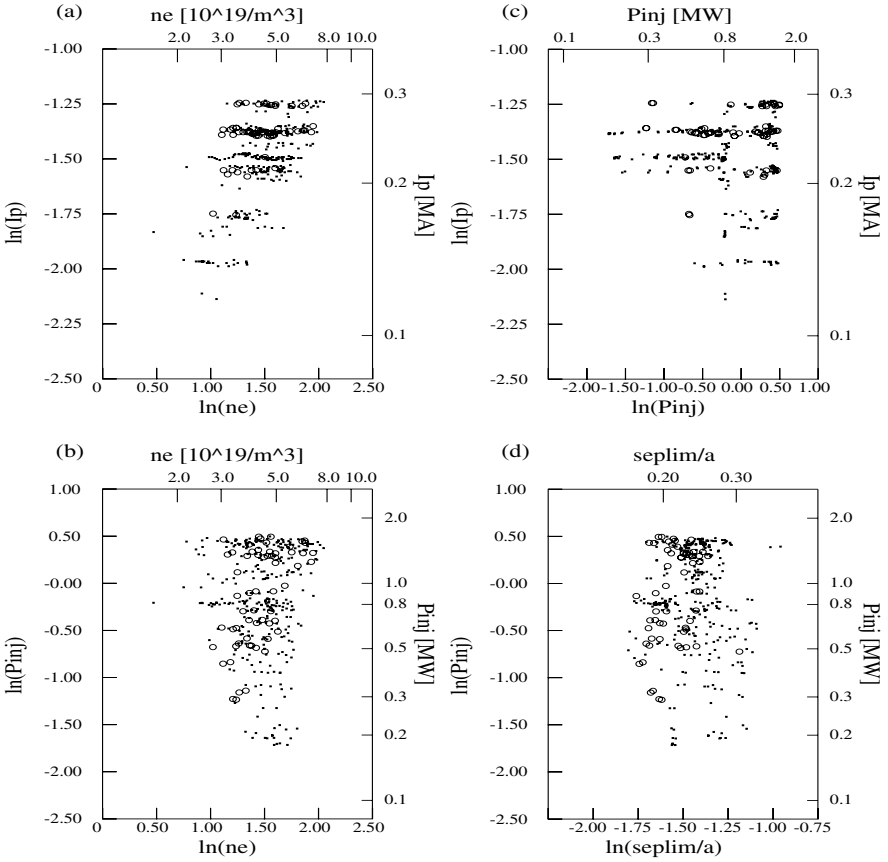

Fig. 5.4. JFT–2M Single Null (SN) discharges. *seplim/a* denotes the distance between the separatrix and the limiter, normalised by the minor radius. For further explanation, see Fig. 5.2.

The above interpretation of the graphs should be viewed as tentative, because of an incompleteness and an essential limitation.

The incompleteness stems from the fact that other plasma variables, such as the magnetic field, and, so far, the distance of the plasma to the wall have

not been considered. From other investigations, the last variable is expected to influence the occurrence of ELMs at least at ASDEX and JFT–2M. Here, as an example, it is plotted against $\ln P_{inj}$ in Figs. 5.2d, 5.3d and 5.4d. It can be seen that for JET and JFT–2M, there is an upper limit of this distance, above which not any HSELM is found (at $seplim/a \approx 0.075$ and 0.25, respectively). For ASDEX, the distance (which is measured to the wall and not to the limiter) seems, according to this 2–D projection, not to have a clear influence on the occurrence of HSELM. The reader is referred to Sect. 5.3.3, where the influence of this distance is discussed when q_{eng}, the plasma current and the time since the L–H transition are also (simultaneously) taken into account.

From this it is clear that interpreting all 10 graphs from all of the above 5 plasma variables for each tokamak in a similar way still has the limitation that only two-dimensional projections are considered, and in fact only a limited number of them: those corresponding to the coordinate planes of the chosen pairs of variables. The above graphical approach, however illustrative and supportive it may be, is limited by the fact that we cannot easily represent the data in more than 2 or 3 dimensions. Of course, an important objective is to try and find lower dimensional subspaces and surfaces in $\mathbb{R}^5$ that separate the data. In the following we will do this systematically by applying discriminant analysis.

5.3.2 Discriminant Analysis using Four Instantaneous Plasma Parameters

We first perform discriminant analysis taking instantaneous plasma parameters (on logarithmic scale) as discriminating variables. Choosing the engineering parameters P_{inj} (injected power of neutral beams), I_p (plasma current), B_t (toroidal magnetic field), and $\overline{n}_e$ (line averaged electron density), we define

$$X_P = \ln P_{inj} \ , \quad X_I = \ln I_p \ , \quad X_B = \ln B_t \ , \quad X_n = \ln \overline{n}_e \ . \tag{5.18}$$

The variables P_{inj}, I_p, B_t, and $\overline{n}_e$ are assumed to be measured in units of MW, MA, Tesla, and $10^{19}/\mathrm{m}^3$, respectively. We use the symbol $S_I = \{P, I, B, n\}$ to denote the set of indices corresponding to these instantaneous plasma parameters. The above choice of variables implies that we assume, for the time being, that the probability of appearance of each class of H–mode is not influenced by the history of the discharge. (The influence of the discharge history, or 'plasma memory', will be discussed in Sect. 5.3.3).

Two discrete variables in the dataset that may influence the occurrence of ELMs require special attention. One is the equilibrium configuration: there exist single-null (SN) and double-null (DN) discharges. In view of the possible qualitative differences associated with these configurations, we analyse SN and DN configurations separately. The other variable is the isotope composition, which we (drastically) simplify as follows. A distinction is made

between proton (H, GAS=1), deuteron (D, GAS=2), and mixed proton-deuteron (GAS is assumed to be 1.5) plasmas. Table 1 gives a break-down of the total number of discharges with respect to device, configuration and ion species. One can see that in this respect each tokamak tends to concentrate somewhat on its own 'favourite' type of discharges. In our analysis

Table 1 — H-mode Database (from ITERH.DB1)
Numbers of timeslices per tokamak, gas composition, and divertor configuration.

	Gas=1		Gas=1.5		Gas=2	All
	SN	DN	SN	DN	SN	
ASDEX	•	206	145	•	•	351
JET	•	•	•	56	466	522
JFT2M	104	11	269	•	•	384
Total	104	217	414	56	453	1257

Gas: 1=H, 2=D, 1.5 = mixture H and D

we consider only DN discharges from ASDEX and SN discharges from JET. The reason is that in the case of ASDEX SN, the class–2 (i.e., HSELM) discharges in the database occur only at a particular value of the magnetic field, whereas for JET all of the DN discharges in the database belong to class–1. The analysis with the same variables is difficult for ASDEX (SN) and JET (DN). We therefore choose the ASDEX (DN) and JET (SN) subsets.

We first apply parametric discriminant analysis using quadratic discriminant functions. Subsequently, non-parametric discriminant analysis is performed: (1) by SAS using uniform kernel estimates (with a certain radius that determines the amount of smoothing) for the probability densities, and (2) by the program INDEP which uses a product multinomial model for the discretised plasma variables (for each variable 10 classes have been taken, roughly according to the deciles of the distribution).

In each case, a (mis-) classification summary is given, which counts the numbers of correctly and of the various types of incorrectly classified observations in the present dataset. This gives information about the performance of the procedure. If the summary would be based on the discriminant function fitted to all data, it would yield an optimistic view (bias) of the performance of the procedure since the same data are used twice. Therefore, the leaving-one-out method (also called 'jackknife') has been used, which classifies the jth observation according to a discriminant function based on all observations except for observation j. This largely eliminates the bias (but increases the variance of the discriminant function estimates). If the performance of the parametric procedure is close to those of the non-parametric procedures, then, at least for the set of plasma variables used, the assumption of a quadratic form of the boundary seems to be appropriate, and we have good explicit formulas describing the region in which we can expect small ELMs.

ASDEX (DN)

The basic frequency table describing the occurrence of ELMs in the dataset is

	non-HSELM	HSELM	Total
Number of observations	134	72	206
Ratio (%)	65	35	100

The distributions of the two groups of data with respect to variables X_i ($i \in S_I$) have to be estimated. Therefore, we present the centers of mass, the standard deviations, and the correlations for both groups, see Table 2. (Together, they determine the first two moments of the distributions.) From

Table 2 — ASDEX (DN) data

(a) Mean values and standard deviations

	non-HSELM, N=134		HSELM, N=72			
log: ↓	Mean	SD	Mean	SD	T	F
P_{inj}	1.07	0.16	0.91	0.17	6.7	•
I_p	–1.10	0.12	–1.22	0.16	5.4	–
B_t	0.77	0.07	0.77	0.07	–0.8	•
$\overline{n}_e$	1.39	0.19	1.33	0.12	3.0	*

*: $0.01 < P < 0.05$, –: $0.05 < P < 0.2$, •: $0.3 < P$

(b) Correlation coefficients

		non-HSELM, (STD_0= 0.09)		HSELM, (STD_0= 0.12)	
log: ↓	log: →	P_{inj}	I_p	B_t	$\overline{n}_e$
P_{inj}		1	0.42	0.18	0.02
I_p		0.14	1	0.31	0.32
B_t		–0.22	0.16	1	0.13
$\overline{n}_e$		–0.23	0.60	0.19	1

The sample correlation coefficients of the HSELM class are displayed in the right upper corner and those of the non-HSELM class in the left lower corner.
STD_0 is one standard deviation of the sample correlation coefficient under the hypothesis of no actual correlation.

the table, one can see that HSELM occurs at lower current and lower injected power than non-HSELM. One can also see that the correlations are not very large, the largest being $r = 0.6$ between $\log I_p$ and $\log \overline{n}_e$ in the non-HSELM group. (However, many of them are more than two standard deviations different from zero, and not the same for the HSELM as for the non-HSELM group.) The other columns of Table 2a quantify how significant the differences of the mean values and the standard deviations are between the two groups. For each plasma variable X_i, the difference in mean value

between the two classes is to be compared with the standard deviation of this difference, which is estimated by $\hat{\sigma}_i = \sqrt{\mathrm{SD}_{i,1}^2/N_1 + \mathrm{SD}_{i,2}^2/N_2}$, where N_1 and N_2 denote the sample sizes. If this difference is significant, i.e. (for suitably large sample sizes) if $T_i = |\overline{X}_i^1 - \overline{X}_i^2|/\hat{\sigma}_i > 2$, then this variable is effective for (univariate) discrimination between the two classes. (The over-bar is used to denote the mean value and the second subscript j to denote the class: 1 = non-HSELM, 2 = HSELM.) The observed 'approximate t-values' T_i (see also [578]) are displayed in the fifth column (with header T) of Table 2a, and could be replaced by more accurate ones in the case that the null-hypothesis of equal variances in the two classes cannot be rejected. The inaccuracy due to the incurred loss of degrees of freedom is generally small in our case, however, and we are only interested in indicating the main effects. The sixth column (with header F) indicates whether the null-hypothesis of equal variances can be rejected according to the F-test at the (one-sided) testing levels shown below the table.

So, with some more statistical background than before, we can infer from the table that X_P and X_I are significantly larger in the non-HSELM class than in the HSELM class. This is not the case for X_B and X_n. However, the variation of the density is significantly larger for the non-HSELM class than for the HSELM class, in accordance with the visual inspection of Fig. 5.1.

Of course, such considerations are only univariate, taking the 'influence' of only one variable at a time into account. (Due to correlations, such 'influences' may be confounded with those of other variables.)

The simultaneous influence of P_{inj}, I_p, B_t, $\overline{n}_e$ is investigated by performing discriminant analysis for these variables (on a logarithmic scale). From the output of the program DISCRIM of SAS [578], we will discuss (*i*) the distance matrix between the (two) groups, (*ii*) the standardised canonical coefficients, and (*iii*) the (mis-) classification tables for quadratic and non-parametric discriminant analysis.

The (squared) distances between the two groups are calculated using as metrics (yardsticks) the covariance matrices of both groups of data. These distances are summarised in a matrix $\hat{\boldsymbol{D}}$, which allows an effective generalisation to k groups. Specifically,

$$\hat{D}_{ij} = (\overline{\boldsymbol{X}}_i - \overline{\boldsymbol{X}}_j)^t \hat{\boldsymbol{\Sigma}}_j^{-1} (\overline{\boldsymbol{X}}_i - \overline{\boldsymbol{X}}_j) \tag{5.19}$$

is used as an estimator of

$$D_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \; . \tag{5.20}$$

For ASDEX (DN), the matrix $\hat{\boldsymbol{D}}$ equals

$$\hat{\boldsymbol{D}} = \begin{pmatrix} 0 & 1.38 \\ 1.72 & 0 \end{pmatrix} \; . \tag{5.21}$$

The fact that $\hat{D}_{1,2}$ is smaller than $\hat{D}_{2,1}$ means that the estimated variance along the line connecting the means, which is (for group j) proportional to

$$(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)^t \hat{\boldsymbol{\Sigma}}_j (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \ ,$$

has a smaller value for group 1 than for group 2. This implies that $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$ are unequal. Of course, the difference between $\hat{D}_{1,2}$ and $\hat{D}_{2,1}$ is not tremendously large. The question is whether it is statistically significant, i.e., whether $D_{1,2} = D_{2,1}$. Assuming multivariate normal distributions for the two samples and neglecting the sampling variation of $\overline{\boldsymbol{X}}_1$ and $\overline{\boldsymbol{X}}_2$, the ratio between $D_{2,1}$ and $D_{1,2}$ is asymptotically (for $N \gg p$) distributed as an F-distribution with N_2 and N_1 degrees of freedom [442]. In our case, $N_1 = 134$, $N_2 = 72$, and $p = 4$. The critical value of the corresponding F-distribution at 5% is about 1.4. The ratio between $D_{1,2}$ and $D_{2,1}$ being $1.72/1.38 \simeq 1.25$, the effect is not statistically significant at the 5% level. Hence, from that point of view, one cannot conclude that $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are unequal. (Note that taking the sampling distribution of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ into account will lead to an even larger critical value and hence also not to statistical significance of the value 1.25.) Of course, this univariate consideration works only in one direction: $D_{1,2} = D_{2,1}$, does not imply at all that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. A multivariate test is required to determine whether or not the hypothesis $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ can be accepted. In fact, the modified likelihood ratio statistic (see, e.g, Chap. 7.4 of [480]) which compares the pooled covariance matrix with the individual ones, and programmed in SAS [578], gives a highly significant result: $45 > \chi^2_{10,.05} = 18$. Hence, the null-hypothesis $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ is indeed rejected in favour of $\det(\boldsymbol{\Sigma}_2) < \det(\boldsymbol{\Sigma}_1)$, though along the line connecting the two centers of mass, the two distributions do not have a significantly different spread.

The pooled within-class standardised canonical coefficients are given by

$$\hat{\boldsymbol{C}} = \begin{pmatrix} 0.72 \\ 0.45 \\ -0.19 \\ 0.25 \end{pmatrix} \ . \tag{5.22}$$

The vector $\hat{\boldsymbol{C}} = (\hat{C}_P, \hat{C}_I, \hat{C}_B, \hat{C}_n)^t = (0.72, 0.45, -0.19, 0.25)^t$ is proportional to

$$D_{ws}^{1/2} \hat{\boldsymbol{\Sigma}}_{ws}^{-1} (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \ , \tag{5.23}$$

where $\hat{\boldsymbol{\Sigma}}_{ws} = (n_1 + n_2 - 2)^{-1} \big((n_1 - 1) \hat{\boldsymbol{\Sigma}}_1 + (n_2 - 1) \hat{\boldsymbol{\Sigma}}_2 \big)$ is the pooled within-class covariance matrix and D_{ws} its diagonal part. The subscript denotes the class of discharge (1=non-HSELM, 2=HSELM). The word standardised means in this case that the coefficients pertain to the plasma variables $\ln P_{inj}$, $\ln I_p$, $\ln B_t$, $\ln \overline{n}_e$ normalised by their (pooled within sample) standard deviations. For that reason, the unstandardised (raw) coefficients are multiplied by $D_{ws}^{1/2}$, i.e., by the standard deviations of the corresponding plasma variables. As we have only two groups, the canonical coefficients are (up to a multiplicative constant) the same as the weights of Fisher's linear discriminant functions, and also the same as the coefficients obtained from linear

regression of the group membership vector (arbitrarily coded). The t-values (the ratio between the estimated coefficients and their estimated standard deviations), calculated from standard linear regression, are $(4.9, 2.6, -1.3, 1.5)$. This indicates that C_P and C_I are significantly (at least 2 standard deviations) different from zero, whereas C_B and C_n are not.

Linear discriminating boundaries are formed in terms of the standardised plasma variables by the hyperplanes perpendicular to the vector $\hat{\boldsymbol{C}}$, and in terms of the unstandardised plasma variables by the hyperplanes perpendicular to $\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2$ in the metric given by $\hat{\boldsymbol{\Sigma}}_{ws}$. The pooled within-class covariance matrix is used as a metric to measure distances and angles. This should be viewed as an approximate data-descriptive procedure, since we know that the hypothesis $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ is rejected by the data, in which case no single metric is satisfying from all points of view, and some improvement is expected by considering quadratic boundaries. At least, from the vector $\hat{\boldsymbol{C}}$ one can see that the (standardised) variables X_P and X_I are more important for discrimination than the (standardised) variables X_B and X_n, and that small ELMs ($j = 2$) are to be sought in the region of low current and injected power. (More precisely, for low values of $\hat{\boldsymbol{C}}^t \boldsymbol{X}$, where $\boldsymbol{X}$ is the (column-) vector of standardised variables X_p, X_I, X_B, X_n. However, the coefficients for X_B and X_n are not significantly different from zero.)

Table 3a shows the jackknifed ('cross-validated') resubstitution summary of the data in the dataset. Now, a quadratic discriminant function has been used. In the 2×2 performance table, the (1,2)-component is small (14.9%), but the (2,1)-component is large (47.2%). This result indicates that most of the non-HSELM discharges are correctly classified as 'non-HSELM', whereas a relatively large fraction of the HSELM discharges is misclassified as 'non-HSELM'. This is because, as is seen in Fig. 1a, with the exception of two almost identical outlying points at low current, the non-HSELM discharges are observed in a fairly narrow region, $\mathbf{R}_1$, in parameter space, where SELM discharges are also possible. Below this region is a region, $\mathbf{R}_2$, where only SELM discharges are possible. This result allows us to avoid the region where the non-HSELM discharges occur, which would be the more dangerous ones from the viewpoint of sustaining good confinement under stationary conditions.

The few outlying HGELM points (which are, according to the agreed definition, indeed HGELM, but somewhat borderline with HSELM), make the dataset from a statistical point of view an interesting example to try and compare various robust versions of discriminant analysis (based on concepts in [264, 292]), which downweight the points that do not conform to the majority. We shall refrain here from doing so. To ease our conscience, we furtively glanced at the additional ASDEX data from the ITERH.DB2 dataset [360,678]. There, an ELM-free point occurs very close to the outlying HGELM points, indicating that that spot does not belong to the region of (exclusively) HSELM. Hence, from a physical point of view it seems sensible not to downweight the outlying points in this case.

Table 3 — Classification performance for various models ASDEX (DN)

Prior probability: HSELM = 0.35

True class	Allocated class non-HSELM	HSELM	OTHER	TOTAL
a) Quadratic boundaries				
non-HSELM	114	20		134
Row %	85.1	14.9		
HSELM	34	38		72
Row %	47.2	52.8		
Total	148	58		206
Row %	71.8	28.2		
b1) Kernel density estimation (r=1)				
non-HSELM	130	4	0	134
Row %	97.0	3.0	0.0	
HSELM	43	29	0	72
Row %	59.7	40.3	0.0	
Total	173	33	0	206
Row %	84.0	16.0	0.0	
b2) Kernel density estimation (r=0.5)				
non-HSELM	116	9	9	134
Row %	86.6	6.7	6.7	
HSELM	34	32	6	72
Row %	47.2	44.4	8.3	
Total	150	41	15	206
Row %	72.8	19.9	7.3	
c) Multinomial independence model				
non-HSELM	111	23		134
Row %	82.8	17.2		
HSELM	27	45		72
Row %	37.5	62.5		
Total	138	68		206
Row %	67.0	33.0		

Table 3 (b1) shows the jackknifed resubstitution summary of the discrimination based on kernel estimates of the probability densities (with a uniform kernel with radius 1 on natural logarithmic scale). In comparison with Table 3 (a), the (2,1) component of the matrix (i.e., the probability to misclassify HSELM discharges as non-HSELM) changes from 47.2 to 59.7%. The (1,2) component of the matrix (i.e., the probability to misclassify non-HSELM discharges as HSELM) reduces from 14.9%, which is the value for the parametric discrimination, to 3.0%. This indicates that even (multi-dimensional) elliptical contours do not demarcate very well the region where only HSELM

discharges occur, and that some further improvement of prediction is possible, at the cost of a more complicated boundary, if one wants to avoid entirely the non-HSELM discharges. This more complicated boundary does not classify all HSELM shots correctly, possibly because the HSELM shots are rather scattered throughout 'non-HSELM region'. By adjusting the kernel radius, one can get some trade-off between the numbers of (1,2)-type and (2,1)-type misclassifications. This is illustrated in Table 3 (b2), where the same type of discrimination is performed, but now with kernel radius 0.5. One sees a reduction of the number of (2,1)-type and an increase of the number of (1,2)-type of misclassification. Note also that 9 non-HSELM and 6 HSELM observations are not allocated to either group, and hence fall into the category 'other'. This is because those observations are outside a radius 0.5 of all other observations in the dataset, whence, by this method, both estimated probability densities are zero. Such a situation does not occur if the radius of the kernel is increased to 1.0. Which kernel radius will be 'optimal', depends on the relative losses associated with the two types of misclassification.

In Table 3 (c), the results from applying the multinomial independence model, obtained by using the program INDEP [225], are shown. For simplicity, the discretisation was done on the original variables, not on the (joint) principal components. (As one can see from Table 2 (b), the four discriminating variables are not very highly correlated.) Ten groups per variable were used, as far as sensible roughly according to the deciles. Zero cells were avoided by using a 'flattening constant' [192], which has some Bayesian justification. One can see that the performance of the multinomial independence model is comparable with that of the quadratic boundary model.

A very simple way to compare the three approaches is to look at the crude error rates (CER), i.e, the total fraction of misclassifications. From this point of view, the estimated performance is similar (CER = 26%, 23%, and 24%, respectively). (In Table 3 (b2), the CER would be 24.5% if not classifying an observation is considered half as serious as making a misclassification.) The two types of misclassification occur in a different ratio for the kernel density estimate approach than for the other two approaches, however. Clearly, the CER is a sensible criterion only if the two types of misclassification are considered to be about equally serious, which is not the case here. To cope with this situation, one has to either associate losses to the misclassifications, or analyse both types of misclassifications jointly. Some theory and analysis using the last approach, albeit in another physical context, can be found in [342].

JET (SN)

The basic frequency table is

	non-HSELM	HSELM	Total
Number of observations	383	83	466
Ratio (%)	82	18	100

Table 4 — JET (SN) data

(a) Mean values and standard deviations

	non-HSELM, N=383		HSELM, N=83			
log: ↓	Mean	SD	Mean	SD	T	F
P_{inj}	1.97	0.26	2.07	0.34	–2.5	*
I_p	1.08	0.22	1.30	0.19	–8.4	–
B_t	0.90	0.15	1.03	0.11	–8.7	*
$\overline{n}_e$	1.36	0.24	1.38	0.30	–0.4	–

*: $0.01 < P < 0.05$, –: $0.05 < P < 0.2$

(b) Correlation coefficients

		non-HSELM, (STD_0= 0.05)		HSELM, (STD_0= 0.11)	
log: ↓	log: →	P_{inj}	I_p	B_t	$\overline{n}_e$
P_{inj}		1	0.55	0.60	0.70
I_p		0.18	1	0.30	0.45
B_t		0.45	0.51	1	0.52
$\overline{n}_e$		0.35	0.58	0.45	1

The sample correlation coefficients of the HSELM class are displayed in the right upper corner and those of the non-HSELM class in the left lower corner.
STD_0 is one standard deviation of the sample correlation coefficient under the hypothesis of no actual correlation.

The centers of mass and the standard deviations of the distributions are shown in Table 4. The column T of this table gives, for each of the four variables, the difference in mean value divided by its estimated standard deviation, and the column F roughly indicates the significance of the difference between the standard deviations. In contrast to the case of ASDEX, X_P and X_I are larger for HSELM than for non-HSELM discharges. Also, the average value of X_B is larger for HSELM than for non-HSELM (for ASDEX there was no significant difference). The question whether the density distributions of the HSELM and non-HSELM shots can be considered to be the same is addressed in the same way as for ASDEX. The distance between the two groups, in the two corresponding metrics, is given by

$$\hat{\boldsymbol{D}} = \begin{pmatrix} 0 & 3.33 \\ 1.75 & 0 \end{pmatrix} . \tag{5.24}$$

Now $D_{1,2}$ is larger than $D_{2,1}$, and significantly so, since the ratio $\simeq 1.9 > F_{379,79;.05} \simeq 1.4$, which indicates that, for JET, group 1 (non-HSELM) does

not have the same covariance matrix as group 2 (HSELM). This is confirmed by the modified likelihood ratio test [480,578] which gives a highly significant result: $72 \gg \chi^2_{10,0.05} = 18$, $\det(\boldsymbol{\Sigma}_2)$ being smaller than $\det(\boldsymbol{\Sigma}_1)$. Hence, the discrimination boundary cannot accurately be expressed by linear combinations of X_j.

The pooled within-class standardised canonical coefficients are

$$\hat{\boldsymbol{C}} = \begin{pmatrix} 0.11 \\ 0.87 \\ 0.51 \\ -0.72 \end{pmatrix} , \tag{5.25}$$

where $\hat{\boldsymbol{C}} = (C_P, C_I, C_B, C_n)^t$. Using this simple yardstick, we see that the injected power is, in comparison with ASDEX (DN), less important and the current is more important for discrimination between HSELM and non-HSELM at JET. Also, the canonical coefficients for the magnetic field and the density are somewhat larger in absolute value than in the case of ASDEX (DN). The t-values of these coefficients, calculated from linear regression analysis, are $(1.0, 7.1, 4.1, -5.6)$. Hence, except for C_P, all coefficients are clearly significant. Table 5 shows the jackknifed (mis-) classification summary for (a) the quadratic discriminant function and (b) non-parametric discriminant functions based on density estimates with uniform kernels and radii $r = 1$ and $r = 0.7$, respectively. The estimated probability of (1,2) misclassification is 3.4% in case (a), 5% in case (b1), and 5.5% in case (b2). (The latter applies if making no classification is considered half as serious as making a wrong classification.) This means that the quadratic fit is better suited to exclude the non-HSELM shots than in the ASDEX (DN) case. Therefore, we will describe this boundary more explicitly. The squared Mahalanobis distance, or potential, of an observation at point $\boldsymbol{x}$ to the center of gravity, $\boldsymbol{\mu}_j$, of group j can be written as

$$D_j^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_j)^t \boldsymbol{A}_j (\boldsymbol{x} - \boldsymbol{\mu}_j) + \boldsymbol{b}_j^t (\boldsymbol{x} - \boldsymbol{\mu}_j) + c_j \ , \tag{5.26}$$

where the index $j = 1, 2$ indicates the class, and $\boldsymbol{A}_j$ and $\boldsymbol{b}_j$ are a matrix and a vector of coefficients, respectively. The posterior probability density for an observation at point $\boldsymbol{x}$ to belong to group j is given by

$$f_j(\boldsymbol{x}) = e^{-\frac{1}{2} D_j^2(\boldsymbol{x})} / \Sigma_{k=1}^2 e^{-\frac{1}{2} D_k^2(\boldsymbol{x})} \ . \tag{5.27}$$

An observation is allocated to group j if the posterior probability density to belong to group j is larger than 50% ($j = 1, 2$).

The boundary $\mathbf{B}$ separating the two allocation regions is given by

$$\mathbf{B} = \{\boldsymbol{x} | f_1(\boldsymbol{x}) = f_2(\boldsymbol{x})\} \ . \tag{5.28}$$

The coefficients $\{\boldsymbol{A}_j, \boldsymbol{b}_j^t, c_j\}$, multiplied by $-\frac{1}{2}$, where $\boldsymbol{A}_j = (A_{jik})_{i,k \in S_I}$ and $\boldsymbol{b}_j^t = (b_{jk})_{k \in S_I}$, for $j = 1, 2$, are tabulated in Table 6. In both cases, among the quadratic coefficients $A_{i,k}$, the diagonal ones, $A_{k,k}$, are dominant.

Table 5 — Classification performance for various models JET (SN)

Prior probability: HSELM = 0.18

True class	Allocated class non-HSELM	HSELM	OTHER	TOTAL
a) Quadratic boundaries				
non-HSELM	370	13		383
Row %	96.6	3.4		
HSELM	44	39		83
Row %	53.0	47.0		
Total	414	52		466
Row %	82.2	11.2		
b1) Kernel density estimation (r=1.0)				
non-HSELM	362	19	0	383
Row %	94.5	5.0	0.0	
HSELM	24	59	0	83
Row %	28.9	71.1	0.0	
Total	386	78	0	466
Row %	82.2	17.8	0.0	
b2) Kernel density estimation (r=0.7)				
non-HSELM	358	17	8	383
Row %	93.5	4.4	2.1	
HSELM	14	69	0	83
Row %	16.9	83.1	0.0	
Total	327	86	8	466
Row %	79.8	18.5	1.7	
c) Multinomial independence model				
non-HSELM	325	58		383
Row %	84.9	15.1		
HSELM	35	48		83
Row %	42.2	57.8		
Total	360	106		466
Row %	77.3	22.7		

JFT–2M (SN)

The original ITERH.DB1 dataset contains only a few JFT–2M observations with small ELMs. After having made some analyses with them, the author turned back to the original time traces plotted on the review sheets [110]. Looking at the H_α signal, it was noticed that in a number of cases, small ELMs did occur just after the last time point that had been selected for inclusion in the database. Since for the present investigation, in contrast to the confinement analysis in [110], it is important to know whether or not

Table 6 — Quadratic discrimination

JET (SN)

Coefficients of the quadratic form representing –0.5 times the squared Mahalanobis distance

To: non-HSELM

log: ↓	log: →	P_{inj}	I_p	B_t	$\overline{n}_e$
P_{inj}		–9.7	–2.3	7.6	2.8
I_p		–2.3	–18.4	10.2	7.9
B_t		7.6	10.2	–38.8	2.5
$\overline{n}_e$		2.8	7.9	2.5	–14.9
Linear term		21.8	9.2	11.0	7.9
Constant		–30.2			

To: HSELM

log: ↓	log: →	P_{inj}	I_p	B_t	$\overline{n}_e$
P_{inj}		–11.5	5.7	10.2	5.4
I_p		5.7	–20.1	–2.2	1.7
B_t		10.2	–2.2	–62.4	5.0
$\overline{n}_e$		5.5	1.7	5.0	–11.3
Linear term		–3.1	28.8	78.0	–5.9
Constant		–46.4			

small ELMs were produced during a substantial period of time during the shot, which may be outside the two time points selected for the confinement analysis, some of the shots were reclassified as being 'HSELM'. A description of the criteria used and a list of shots and time intervals that were classified as 'HSELM' can be found in the Appendix of [351].

The reclassification was based solely on the H_α signal. It should be noted that sometimes the H_α spikes are provoked by sawtooth oscillations. Hence, the label HSELM is, more precisely speaking, used to denote 'shots with a spiky H_α signal indicating either (small) ELMs or sawteeth'. A further study in which the H_α signals are compared with those from interferometry in order to separate the ELMy from the sawtooth cases, would therefore be valuable.

For the ASDEX and JET data, the reclassification problem is less pressing. For those two tokamaks, usually three, sometimes four, time-points per shot have been included in the dataset. This covers somewhat better than in the case of JFT–2M the interval between the beginning of the H–mode and the time that W_{dia} is maximal. Nevertheless, in a future investigation, one could consider to investigate as response variable the total duration of the periods (standardised in some way) with and without small ELMs during the discharge. At present, no such information is available in the database.

The basic frequency table for JFT–2M (SN) is: The centers of mass and the standard deviations of the distributions are shown in Table 7. Again, column F indicates which SD's, and column T indicates which mean values are to be considered different between HSELM and non-HSELM. From the table

Table 7 — JFT-2M (SN) data

(a) Mean values and standard deviations

	non-HSELM, N=312		HSELM, N=61			
log: ↓	Mean	SD	Mean	SD	T	F
P_{inj}	–0.18	0.57	–0.13	0.51	–0.7	•
I_p	–1.52	0.19	–1.39	0.11	–7.0	**
B_t	0.22	0.07	0.24	0.01	–4.5	***
$\overline{n}_e$	1.48	0.26	1.46	0.22	0.7	•
Gas	0.32	0.17	0.16	0.20	6.6	–

***: P < 0.001, **: 0.001 < P < 0.01, –: 0.05 < P <0.2, •: 0.2 < P < 0.5

(b) Correlation coefficients

		non-HSELM, (STD_0= 0.06)			HSELM, (STD_0= 0.13)	
log: ↓	log: →	P_{inj}	I_p	B_t	$\overline{n}_e$	Gas
P_{inj}		1	0.16	0.15	0.51	–0.77
I_p		–0.13	1	0.02	0.40	–0.22
B_t		–0.09	0.18	1	0.11	–0.25
$\overline{n}_e$		–0.02	0.62	0.17	1	–0.26
Gas		–0.41	0.15	–0.11	0.21	1

The sample correlation coefficients of the HSELM class are displayed in the right upper corner and those of the non-HSELM class in the left lower corner.
STD_0 is one standard deviation of the sample correlation coefficient under the hypothesis of no actual correlation.

one can see that the HSELM discharges generally occur at higher values of I_p, B_t, and at a lower value of GAS than the non-HSELM discharges do. The last fact can also be seen from Table 8: For the H into H^+ (GAS=1) discharges, a larger fraction HSELM observations occurs than for the H into D^+ discharges (GAS=1.5). Of course, such considerations are only univariate, taking the 'influence' of only one variable at a time into account. Due to correlations, such 'influences' may be confounded with those of other variables. The simultaneous influence of P_{inj}, I_p, B_t, $\overline{n}_e$ and GAS is investigated by

Table 8 — JFT-2M (SN) data
Numbers of timeslices per ELM type and gas composition

Gas	H → H	H → D	Total
non-HSELM	67	245	312
HSELM	37	24	61
Total	104	269	373

performing discriminant analysis for these variables on logarithmic scale. We summarise the main results.

The estimated pairwise distances between two groups are given by the matrix

$$\hat{\boldsymbol{D}} = \begin{pmatrix} 0 & 4.77 \\ 1.82 & 0 \end{pmatrix} . \tag{5.29}$$

The fact that $\boldsymbol{D}_{1,2}$ is significantly larger than $\boldsymbol{D}_{2,1}$ indicates that, like at JET and unlike at ASDEX, see (5.21) and (5.24), the HSELM shots occur in a smaller region of plasma parameter space than the non-HSELM shots. The imbalance between $\boldsymbol{D}_{1,2}$ and $\boldsymbol{D}_{2,1}$ implies that the discrimination boundary is expected not to be efficiently representable by a linear combination of the above five, logarithmically transformed, variables.

The (pooled within-class) standardised canonical coefficients are estimated by

$$\hat{\boldsymbol{C}} = \begin{pmatrix} 0.19 \\ -0.87 \\ -0.06 \\ 0.47 \\ 0.83 \end{pmatrix} . \tag{5.30}$$

The t-values of the estimates

$$\hat{\boldsymbol{C}}^t = (C_P, C_I, C_B, C_n, C_G)$$

are

$$(1.4, -6.0, -0.5, 3.2, 6.1) .$$

As in the univariate analysis, we have a significant dependence on plasma current and on gas composition. However, in contrast to the univariate analysis, the effect of the plasma density turns out to be significant. This (significant) density dependence is to be interpreted when the four other plasma parameters are kept constant. In the univariate analysis, this fact was masked by the correlation between density and current, which is about 0.6 for non-HSELM, and 0.4 for HSELM, see Table 7b. Compared with the univariate analysis, the injected power remains insignificant, and the magnetic field has become insignificant in the presence of the other variables. To understand the latter, we note that two magnetic field scans exist in the present dataset. As can be seen from Fig. 1, no HSELM shots occurred in these scans, which led to the suspicion of B_t being significant from univariate considerations. However, the two magnetic field scans were made only with H into D^+, i.e., GAS=1.5, and are somewhat negatively correlated ($r = -0.3$) with I_p. The coefficients of the simultaneous analysis express the fact that the non-occurrence of HSELM can be ascribed to the high value of GAS and the relatively low value of I_p for the 16 observations from the 2 scans. In addition to these two effects, which are estimated from all data, the magnetic field does not exhibit additional discriminatory value. Obviously this holds within the simple power–law type model used so far. Table 9 shows the jackknifed (mis-) classification summary of the discriminant analysis using: (a) a quadratic discriminant function, (b) non-parametric density estimates with uniform kernels and radii $r = 1$ and $r = 0.5$, respectively, and (c) the multinomial independence model with a global association factor. It is noted that the discrimination with the uniform kernel density estimates improves the number of (dangerous) misclassifications of non-HSELM shots. Nevertheless, a fair amount of misclassifications

Table 9 — Classification performance for various models JFT-2M (SN)

Prior probability: HSELM = 0.16

True class	Allocated class: non-HSELM	HSELM	OTHER	TOTAL
a) Quadratic discriminant analysis				
non-HSELM	239	73		312
Row %	76.6	23.4		
HSELM	11	50		61
Row %	18.0	82.0		
Total	250	123		373
Row %	67.0	33.0		
b1) Kernel density estimation (r=1.0)				
non-HSELM	263	42	7	312
Row %	84.3	13.5	2.2	
HSELM	17	42	2	61
Row %	27.9	68.9	3.3	
Total	280	84	9	373
Row %	75.1	22.5	2.4	
b2) Kernel density estimation (r=0.5)				
non-HSELM	238	12	62	312
Row %	76.3	3.9	19.9	
HSELM	22	21	18	61
Row %	36.1	34.4	29.5	
Total	260	33	80	373
Row %	69.7	8.9	21.5	
c) Multinomial independence model				
non-HSELM	296	16		312
Row %	94.9	5.1		
HSELM	45	16		61
Row %	73.8	26.2		
Total	341	32		373
Row %	91.4	8.6		

remains for $r = 1$, whereas for $r = 0.5$ there is a large number of unclassified observations. The multinomial independence model reduces the number of non-HSELM misclassifications, albeit at the cost of a larger number of the (less dangerous) HSELM misclassifications.

5.3.3 Discriminant Analysis taking Plasma Memory and Plasma-Wall Distance into Account

We now want to investigate some possible influence of 'plasma memory' on the class of plasma discharges that will occur. For convenience of exposition, we use expressions such as 'influence', with a possible causal connotation. However, we keep well in mind that statistical relationships alone, which is our main investigation, do not yield information about the direction of any causal relationship, that time ordering is a necessary but not sufficient aspect of causality, and that in fact there may be underlying 'hidden' causal factors which we simply ignore here for simplicity, see [127, 128, 281]. First, we will only replace the instantaneous plasma density by the target plasma density, i.e., we use as the fourth discrimination variable the logarithm of the line-average density in the Ohmic phase preceding the H-mode phase

$$X_{\overline{n}_{e,ohm}} = \ln(\overline{n}_{e,ohm}) \;, \tag{5.31}$$

and keep the other three components the same as in Sect. 5.3.2. The effect of this replacement will be investigated for ASDEX (DN) and JET (SN) plasmas.

The 'influence' of the plasma density during the Ohmic phase of the discharge may still be felt during the subsequent H–mode phase, which we call for convenience, somewhat elliptically, a 'plasma memory effect', even though we concentrate on the aspect of temporal order only, and, as just remarked, the physical effect (if at all) may be quite indirect.

The estimated canonical coefficients are

$$\hat{\boldsymbol{C}} = \begin{pmatrix} 0.53 \\ 0.99 \\ -0.12 \\ 0.64 \end{pmatrix} \;. \tag{5.32}$$

Comparing these values with the canonical coefficients in Sect. 5.3.2, we see the importance of P_{inj} and I_p and the unimportance of B_t for linear discrimination confirmed. However, the density dependence stands out more clearly now. This means that the appearance of HSELM discharges depends on the target plasma density rather than on the instantaneous density.

The difference between the target density and the instantaneous density in their effect on the discrimination suggests the importance of the elapsed time after the transition to the H–mode. Furthermore, the distance between the separatrix and the wall has been known by the experimentalists to influence the occurrence of ELMs [110].

To see these effects quantitatively, we choose as new set of variables

$$X_t = \ln\{(t - t_{LH})\} \;, \quad X_{PV} = \ln(P_{inj}/V_p) \;, \quad X_{\overline{n}_{e,ohm}} = \ln(\overline{n}_{e,ohm}) \;, \tag{5.33}$$

$$X_d = d_{SW} \text{ or } d_{SL} \;, \quad X_q = \ln(q_{eng}) \;, \quad X_{BIV} = \ln(B_t I_p/V_p) \;, \tag{5.34}$$

where $t - t_{LH}$ is the elapsed time after the L-H transition (expressed in units of 50 ms for ASDEX and JFT–2M, and of 500 ms for JET), d_{SL} is the distance between the separatrix and 'limiter' and d_{SW} the distance between the separatrix and the outer wall, both normalised by the minor radius and defined more precisely below, $q_{eng} = 5\frac{B_t(T)}{I_p(MA)}\frac{ab}{R}$ is an 'elliptical cross-section' approximation to the engineering q-value, and $V_p = 2\pi^2 Rab$ to the plasma volume (in m^3). The impact of the parameter d_{SL} has been extensively studied experimentally (see, e.g., [713]). As some values of X_d were close to zero, it was decided not to use the logarithm of this variable. Note that for constant V_p, the transformation from (X_B, X_I) to (X_q, X_{BIV}) can be described by a simple rotation of the coordinate frame. With these variables, we perform discriminant analysis for ASDEX (DN) and JET (SN).

ASDEX (DN)

Table 10 describes and compares the distributions of the non-HSELM and HSELM observations in plasma parameter space. We can see, for instance, that the difference in X_t is 0.25, i.e., on average, the non-HSELM mode appears $e^{0.25} \times 50 \simeq 65$ ms later during the shot than the HSELM mode. The difference in d_{SW} means that the distance between the separatrix and the outer wall for the HSELM observations is on average $0.02 \times 40 = 0.8$ cm larger than for the non-HSELM observations. Discriminant analysis applied to the $N = 206$ DN observations from ASDEX gives the following estimated (pooled within-class standardised) canonical coefficients.

$$\hat{\boldsymbol{C}} = \begin{pmatrix} 0.53 \\ 0.54 \\ -0.52 \\ -0.49 \\ -0.66 \\ 0.29 \end{pmatrix} . \tag{5.35}$$

The t-values of $\hat{\boldsymbol{C}}^t = (\hat{C}_t, \hat{C}_{PV}, \hat{C}_{\overline{n}_{e,ohm}}, \hat{C}_d, \hat{C}_q, \hat{C}_{BIV})$, obtained from standard linear regression, are $(-4.5, -4.5, 3.2, 3.9, 4.2, -1.6)$. These results indicate that the variables $t - t_{LH}$, $\overline{n}_{e,ohm}$, and d_{SW}) are about as important for discrimination as the heating power.

The interpretation of the last two coefficients in (5.35) is: for fixed values of the four other plasma parameters, the direction of highest discrimination is given by the logarithm of $q^{-.66/0.14}(B_t I_p/V_p)^{0.29/0.18}$, where 0.14 and 0.18 are the pooled within-class standard deviations of X_q and X_{BIV}, estimated from Table 10a. By comparing (5.35) with (5.22) and (5.32) one can see that the q-value is more important for discrimination than the absolute value of the toroidal magnetic field. Table 11 shows the result of the jackknifed (mis-) classification summary. The performance is better than the performance of the discriminant analysis based on the four instantaneous variables. The (2,1)

Table 10 — ASDEX (DN) data

(a) Mean values and standard deviations

	non-HSELM, N=134		HSELM, N=72			
log: ↓	Mean	S.D.	Mean	S.D.	T	F
$t - t_{LH}$	0.10	0.54	–0.15	0.53	3.1	•
P_{inj} / V_p	–0.61	0.16	–0.76	0.17	6.5	•
$\overline{n}_{e,\,ohm}$	1.25	0.17	1.23	0.24	–0.6	*
sepwall / a	0.44	0.04	0.46	0.05	–3.0	–
q_{cyl}	1.15	0.13	1.28	0.17	–5.6	–
$B_t I_p / V_p$	–2.02	0.16	–2.12	0.21	3.7	–

*: $0.01 < P < 0.05$, –: $0.05 < P < 0.3$, •: $0.3 < P$

(b) Correlation coefficients

		non-HSELM, (STD_0= 0.09)				HSELM, (STD_0= 0.12)	
log: ↓	log: →	$t - t_{LH}$	P_{inj} / V_p	$\overline{n}_{e,\,ohm}$	sepwall / a	q_{cyl}	$B_t I_p / V_p$
$t - t_{LH}$		1	0.00	0.20	0.19	–0.02	0.10
P_{inj} / V_p		–0.44	1	0.17	–0.15	–0.39	0.45
$\overline{n}_{e,\,ohm}$		0.25	0.14	1	0.31	–0.49	0.69
sepwall / a		0.03	0.16	0.55	1	–0.13	0.33
q_{cyl}		0.00	–0.27	–0.69	–0.58	1	–0.73
$B_t I_p / V_p$		0.26	0.08	0.69	0.60	–0.59	1

The sample correlation coefficients of the HSELM class are displayed in the right upper corner and those of the non-HSELM class in the left lower corner. STD_0 is one standard deviation of the sample correlation coefficient under the hypothesis of no actual correlation.

component of the (mis-)classification table is 9.0% for the quadratic discriminant analysis, and 7.8% for the analysis using uniform kernel density estimation with $r = 1.5$. The assumption of the quadratic fitting is better than in Sect. 5.3.2.

It is noted that for X_d the variable SEPLIM as available in ITERH.DB1 was used, which denotes for ASDEX an estimate of the distance between the plasma boundary and the outside torus wall (in the horizontal plane), not taking the position of the ICRH antenna into account.

We also did the discriminant analysis while taking the position of the ICRH antenna into account, as well as with an estimate of the closest distance between the plasma and the wall (not necessarily in the horizontal plane). These data are not available in ITERH.DB1.In both cases, the performance of the discrimination turned out to be less than in the above case. This suggests that, for instance, the magnetic field ripple may be of more importance for the occurrence of ELMs than the closest distance between the plasma and the wall. Such a suggestion would not be in contradiction with [298], where ballooning-type instabilities with toroidal modes numbers $n = 8$ to 15 were considered to be likely candidates for ELM precursors. This topic can be investigated more precisely in the situation that more accurate estimates of the minor plasma radius than presently available in ITERH.DB1 are available.

Table 11 — Classification performance for various models ASDEX (DN)

Priors HSELM = 0.35

True class	Allocated class non-HSELM	HSELM	OTHER	TOTAL
a) Quadratic discriminant analysis				
non-HSELM	122	12		134
Row %	91.0	9.0		
HSELM	23	49		72
Row %	31.9	68.1		
Total	145	61		206
Row %	70.4	29.6		
b1) Kernel density estimation (r=1, threshold=0.5)				
non-HSELM	108	10	16	134
Row %	80.6	7.5	11.9	
HSELM	16	36	20	72
Row %	22.2	50.0	27.8	
Total	124	46	36	206
Row %	60.2	22.3	17.5	
b2) Kernel density estimation (r=1.5, threshold=0.5)				
non-HSELM	122	9	3	134
Row %	91.0	6.7	2.2	
HSELM	27	43	2	72
Row %	37.5	59.7	2.8	
Total	149	52	5	206
Row %	72.3	25.2	2.4	

The coefficients for the Mahalanobis distances to the two centers of gravity are given in Table 12. The discrimination surfaces are surfaces of constant difference between the two Mahalanobis distances. Note that they can be multidimensional ellipses and that they can also have an hyperbolic character. The reader is referred to [352] for a graphical two-dimensional section of the discriminant surface for these data.

JET (SN)

For JET, we used for d_{SL} the variable SEPLIM as available in ITERH.DB1, which is an estimate of the minimum distance between the separatrix and the 'limiter', i.e., any part of the wall. In Table 13, the univariate summary statistics and the correlation matrices are given. One can see that, of the six variables in Table 13, the Ohmic density and the variable $B_t I_p / V_p$ are clearly

Table 12 — Quadratic discrimination

ASDEX (DN)

Coefficients of the quadratic form representing –0.5 times the squared Mahalanobis distance

To: non-HSELM

Quadratic term	log: →	$t - t_{LH}$	P_{inj} / V_p	$\overline{n}_{e,\,ohm}$	sepwall / a	q_{cyl}	$B_t\, I_p$ / V_p
log: ↓							
$t - t_{LH}$		–2.61	–3.80	2.65	–4.01	1.86	2.16
P_{inj} / V_p		–3.80	–26.84	2.91	–1.33	–6.21	0.40
$\overline{n}_{e,\,ohm}$		2.65	2.91	–44.18	14.28	–24.50	16.60
sepwall / a		–4.01	–1.33	14.28	–477.34	–42.96	50.68
q_{cyl}		1.86	–6.21	–24.50	–42.96	–65.69	–8.05
$B_t\, I_p$ / V_p		2.16	0.40	16.60	50.68	–8.05	–47.38
Linear term		–2.75	–22.05	224.10	689.55	209.63	–259.17
Constant		–669.29					

To: HSELM

Quadratic term	log: →	$t - t_{LH}$	P_{inj} / V_p	$\overline{n}_{e,\,ohm}$	sepwall / a	q_{cyl}	$B_t\, I_p$ / V_p
log: ↓							
$t - t_{LH}$		–1.91	0.33	1.06	3.51	0.46	–0.49
P_{inj} / V_p		0.33	–26.86	–4.40	–30.16	–1.05	15.08
$\overline{n}_{e,\,ohm}$		1.06	–4.40	–18.96	1.23	–0.00	16.36
sepwall / a		3.51	–30.16	1.23	–290.90	15.80	41.59
q_{cyl}		0.46	–1.05	–0.00	15.80	–39.62	–24.61
$B_t\, I_p$ / V_p		–0.49	15.08	16.36	41.59	–24.61	–48.28
Linear term		–9.16	64.56	108.64	357.55	–19.43	–198.11
Constant		–313.37					

the most important ones for discrimination. Also, the HSELM shots seem to occur at somewhat higher values of P_{inj}/V_p and at lower values of q_{eng} than the non-HSELM discharges do. The time since the onset of the H–mode does not exhibit a significant effect. However, these univariate considerations do not necessarily give the correct estimates for the simultaneous influence of these variables. In fact, the estimated standardised canonical coefficients from discriminant analysis, which are, as we have seen before, the multiple regression coefficients after standardising all variables by the pooled within-class variances, are ($N = 277$)

$$\mathbf{C} = \begin{pmatrix} -0.004 \\ 0.038 \\ 0.217 \\ 0.0026 \\ -0.037 \\ 0.785 \end{pmatrix} . \tag{5.36}$$

Table 13 — JET (SN) data

(a) Mean values and standard deviations

	non-HSELM, N=218		HSELM, N=59			
log: ↓	Mean	S.D.	Mean	S.D.	T	F
$t - t_{LH}$	0.54	0.58	0.65	0.62	–1.2	•
P_{inj} / V_p	–2.68	0.26	–2.55	0.30	–3.2	–
$\overline{n}_{e,\,ohm}$	0.39	0.32	0.72	0.21	–9.5	*
seplim / a	0.055	0.017	0.059	0.012	–1.9	+
q_{cyl}	1.10	0.17	1.01	0.19	3.6	–
$B_t I_p / V_p$	–2.72	0.34	–2.33	0.23	–10.4	*

*: 0.01 < P < 0.05, +: 0.05 < P < 0.10, –: 0.10 < P < 0.3, •: 0.3 < P

(b) Correlation coefficients

		non-HSELM, (STD_0= 0.05)				HSELM, (STD_0= 0.11)	
log: ↓	log: →	$t - t_{LH}$	P_{inj} / V_p	$\overline{n}_{e,\,ohm}$	seplim / a	q_{cyl}	$B_t I_p / V_p$
$t - t_{LH}$		1	0.29	0.22	–0.04	0.05	–0.01
P_{inj} / V_p		–0.07	1	0.59	0.24	–0.29	0.62
$\overline{n}_{e,\,ohm}$		0.18	0.16	1	0.28	–0.24	0.68
seplim / a		0.16	–0.03	0.21	1	–0.13	0.52
q_{cyl}		–0.25	0.08	–0.48	0.01	1	–0.58
$B_t I_p / V_p$		0.16	0.33	0.83	0.13	–0.34	1

The sample correlation coefficients of the HSELM class are displayed in the right upper corner and those of the non-HSELM class in the left lower corner. STD_0 is one standard deviation of the sample correlation coefficient under the hypothesis of no actual correlation.

The t-values are $(-0.04, 0.26, 0.89, 0.02, -0.25, 3.2)$. This means that, except for $B_t I_p/V_p$, none of the coefficients are statistically significant, nor large in absolute value! This illustrates the effect of confounding, due to correlations between the discrimination variables. Indeed, HSELMs are associated with a larger Ohmic density than non-HSELMs, but a higher Ohmic density is also correlated with a higher value $B_t I_p/V_p$ ($r = 0.85$). At a constant value of $B_t I_p/V_p$, the Ohmic density does not have a statistically significant predictive value for the occurrence of small ELMs. The correlations between $B_t I_p/V_p$ and P_{inj}/V_p and q_{eng} are smaller ($r = 0.4$ and $r = -0.4$, respectively), but apparently sufficient, as one can see by comparing Table 13 (a) with (5.36), to provoke a smaller amount of confounding.

The question may arise whether the correlations between the variables are sufficiently high to make the dataset ill-conditioned for simultaneous regression or discriminant analysis. This is investigated by principal component analysis. The square root of the smallest eigenvalue of the correlation matrix equals 0.36, and the associated eigenvector is mainly associated with $B_t I_p/V_p$ and $\overline{n}_{e,ohm}$. As the measurement accuracy of these 2 quantities is better than, say, 0.2 times 0.36 (i.e., than about 7%), the estimated bias in the estimates of the canonical coefficients induced by neglecting such (random) measurement errors is not more than a few percent of those estimates. The square root of

second smallest eigenvalue is considerably larger (0.74). From these considerations, the condition is sufficiently good for a well-behaved simultaneous linear discriminant analysis.

The canonical coefficients suggest that the minimum distance between the separatrix and the limiter (or any part of the wall, in absense of a limiter) is not important for (linear) discrimination. This should, however, not be misinterpreted. About 150 observations with missing values for d_{SL} have not been used in the analysis. They are, in overwhelming majority, shots for which at least one of the X-points is outside or nearly outside the vessel, so that the missing values correspond to $d_{SL} \leq 0$. Using an indicator variable for those missing values (0=missing, 1=non-missing), instead of the variable d_{SL}, gives a significant discrimination coefficient ($t = 2.6$), while the other coefficients remain roughly the same! Also, coding the missing values as $d_{SL} = 0$, and retaining the positive values of d_{SL} as they are in the database, gives a significant coefficient ($t = 2.45$). So the negative values of d_{SL} seem to be far more interesting for discrimination than the positive ones. HSELMs are associated (in the simultaneous regression) with positive d_{SL} (i.e., with X points within the vessel), but the precise value of d_{SL} is rather unimportant. A fortunate coincidence is that restricting attention to the ob-

Table 14 — Quadratic discrimination
JET (SN)
Coefficients of the quadratic form representing –0.5 times the squared Mahalanobis distance

To: non-HSELM

Quadratic term	log: →	$\overline{n}_{e,\,ohm}$	$B_t\, I_p\, /\, V_p$
log: ↓			
$\overline{n}_{e,\,ohm}$		–16.0	12.6
$B_t\, I_p\, /\, V_p$		12.6	–14.3
Linear term		81.0	–87.6
Constant		–132.3	

To: HSELM

Quadratic term	log: →	$\overline{n}_{e,\,ohm}$	$B_t\, I_p\, /\, V_p$
log: ↓			
$\overline{n}_{e,\,ohm}$		–21.0	12.8
$B_t\, I_p\, /\, V_p$		12.8	–17.1
Linear term		89.8	–98.2
Constant		–144.7	

servations with both X-points inside the vessel leads to a region where the desirable HSELM discharges are expected, and at the same time simplifies the discriminant analysis by fading out the influence of all plasma variables, except for $B_t I_p/V_p$. Hence, as a compact representation of the case that both X-points are inside the vessel, we present in Table 14 the quadratic discriminant function using the variables $B_t I_p/V_p$ and $\overline{n}_{e,ohm}$ only. The table is based on the $N = 277$ observations with $d_{SL} > 0$. In two dimensions one can easily make plots of the discriminant curves such as in [352]. In our situation, one can see that the difference between the two Mahalanobis distances is positive definite. Hence, the discriminant curves are ellipses.

Finally, we check that, using only $\overline{n}_{e,ohm}$ and $B_t I_p/V_p$, the canonical coefficients are $\boldsymbol{C}^t = (0.22, 0.81)$, with t-values $(0.97, 3.6)$. This is not very different from the corresponding coefficients in the linear analysis with the 6 variables discussed above. Table 15 (a) shows the result of the jackknifed (mis-) classification summary of the quadratic analysis with the 6 and the 2 variables described above. One can see that the quadratic discrimination with the six variables performs very well with respect to the non-parametric discrimination. With the two variables $\overline{n}_{e,ohm}$ and $B_t I_p/V_p$ only, the performance is somewhat worse, but still very reasonable for practical purposes.

5.4 Summary and Discussion

In this chapter, we looked at various methods for determining the plasma parameter regime where H–mode discharges with small ELMs can be expected. After a description of its theoretical aspects, discriminant analysis was applied to the ASDEX, JET, and JFT–2M data of the ITERH.DB1 dataset. We divided the H–mode discharges into two classes: class-1 (ELM-free or with large ELMs) and class-2 (with small ELMs). This distinction was motivated by the fact that H–mode with small ELMs is favourable from the viewpoint of a long sustenance of improved confinement while at the same time avoiding a heavy heat load on the divertor plates. The distributions of the two classes of discharges in plasma parameter space overlap each other. The general statistical methodology for describing these distributions and for discriminating between the two classes of H–mode has been discussed in Sect. 5.2.

In Sect. 5.3, linear and quadratic discriminant analysis on a logarithmic scale was used to find explicit expressions for combinations of variables that are efficient to predict the region where small ELMs will occur. A general aspect based on the analysis of ASDEX (DN), JET (SN) and JFT–2M (SN) data is that linear discrimination (on logarithmic scale) is not very accurate since the covariance matrices for the two classes are significantly different. Hence, the boundary **B** cannot effectively be expressed by a simple power law in terms of the engineering variables, and hence also not as a simple power law in terms of the dimensionless plasma parameters.

Table 15 — Classification performance for various models JET (SN)

Priors HSELM = 0.21

True class	Allocated class non-HSELM	HSELM	OTHER	TOTAL
a) Quadratic discriminant analysis				
a1) with the 6 variables from table 13				
non-HSELM	199	19		218
Row %	91.3	8.7		
HSELM	24	35		59
Row %	40.7	59.3		
Total	223	54		277
Row %	80.5	19.5		
a2) with $\bar{n}_{e,\,ohm}$ and $B_t\, I_p / V_p$ only				
non-HSELM	200	18		218
Row %	91.7	8.3		
HSELM	26	33		59
Row %	44.1	55.9		
Total	226	51		277
Row %	81.6	18.4		
b1) Kernel density estimation (r=1, threshold=0.5)				
non-HSELM	191	4	23	218
Row %	87.6	1.8	10.6	
HSELM	10	36	13	59
Row %	17.0	61.0	22.0	
Total	201	40	36	277
Row %	78.7	14.4	13.0	
b2) Kernel density estimation (r=1.5, threshold=0.5)				
non-HSELM	203	10	5	218
Row %	93.1	4.6	2.3	
HSELM	12	43	4	59
Row %	20.3	72.9	6.8	
Total	215	53	9	277
Row %	77.6	19.1	3.3	

Instead, the discriminant surfaces, i.e., the surfaces on which the difference in Mahalanobis distance to the centers of gravity of the two classes of H–mode discharges is constant, are on logarithmic scale explicitly described by quadratic equations.

The performance of the quadratic discriminant analysis was estimated by using the jackknife method on the available datasets of ASDEX, JET and JFT–2M, and expressed in (mis-) classification tables. A comparison was made with the performance of discrimination based on non-parametric density estimates and of discrimination using a multinomial independence

model. These more flexible non-parametric methods showed a better performance than the quadratic discriminant analysis, however considerably less so when the 'plasma memory' and the plasma-wall distance were taken into account. The non-parametric methods do not permit a simple representation of the discrimination surfaces.

It was found that for quadratic discriminant functions a larger fraction of non-SELM discharges is (correctly) classified as non-SELM, than SELM discharges are (correctly) classified as SELM. The last feature is seen by inspecting the (mis-) classification tables and can be explained by assuming, as is partly seen from the scatter plots, that there is a mixed region where both classes of discharges (HSELM and non-HSELM) occur, and a region where predominantly HSELM discharges occur. This feature of the data is of importance for operating future machines since non-HSELM discharges have undesirable properties for long burning plasmas. (The ELM-free H–mode is vulnerable to impurity accumulation, and giant ELMs may produce damage from strong repetitive heat loads on the divertor plates.) By using discriminant analysis we have presented a method for avoiding these unfavourable modes to a considerable extent.

In the analysis of the ASDEX (DN) H into D^+ discharges, the injected power P_{inj} and the plasma current (or q_{eng}) were shown to be important for discriminating between HSELM and non-HSELM. For JFT–2M (SN) discharges, plasma current and gas composition were the most important discriminating variables (with some additional effect of density), and for JET (SN) D into D^+ discharges, it was plasma current, magnetic field and density. Quantitative estimates of the linear discriminant functions have been given in the main text. They can be used as simple approximations to the discriminant surfaces, though it was shown that quadratic surfaces give a more accurate description. Explicit quadratic coefficients have been presented for ASDEX (DN) H into D^+ discharges in Table 12. A comparison shows that the discriminant surfaces are not the same for the three devices. For instance, low P_{inj} is favourable for getting small ELMs in ASDEX, but in JET and JFT–2M, the variable P_{inj} does not play an important role. This sets a limit to the ability to predict, from the present analysis, the presence of small ELMs in future devices. In a further analysis, it is of interest to investigate in more detail the change of the discriminant surfaces with machine size.

An important role of the 'plasma memory' during the discharge was found. The target plasma density is more important for the prediction of the class of H–mode than the instantaneous density. For ASDEX, the elapsed time is as important for discrimination as is the injected power. In JET, the elapsed time does not seem important. These contrasting results may be due to the fact that the time slices for JET have been chosen differently, usually more at the end of the H–mode phase, than the time slices for ASDEX. They also suggest an improvement of the database and the ensuing analysis. Ideally, all the time points of transition between H–mode with SELM and the other

types of H–mode should be recorded. From this information one can estimate, using methods of survival analysis (as a function of the plasma parameters and the elapsed time since the L–H transition) the 'hazard rate' of transition of HSELM to another type of H–mode, or the fraction of time the discharge dwells in H–mode with small ELMs.

Two results from the analysis of the JET data are the following. (1) The distance between the separatrix and the wall d_{SL} is not important for predicting the type of H–mode, perchance unless d_{SL} is negative, which corresponds to the X-point being (nearly) outside the vessel. The precise negative values of d_{SL} are not available in the present database. Compared to $d_{SL} > 0$, discharges with $d_{SL} < 0$ tend to be less frequently associated with HSELM. (2) For discharges with the X-point inside the vessel,[6] the discrimination is much simplified by the fact that only the target density and $B_t I_p / V_p$ are by far the most important variables.

This result from JET may be in agreement with the fact that for ASDEX a somewhat better discrimination was found using d_{SW} (the distance between the separatrix and the wall, not taking the ICRH antenna into account) instead of using d_{SL} (the distance between the separatrix and the limiter, taking the ICRH antenna into account). From a physical point of view, it appears interesting to investigate empirically, with an improved dataset, whether the magnetic field ripple, rather than the closest distance between the plasma and the wall, is the factor that influences the occurrence of small ELMs.

It has not been the objective to present here an exhaustive study, but rather to provide a background of discriminant analysis as well as to illustrate its practical usage by identifying regions in plasma parameter space where various types of H–mode can be produced. The approach can also be used for identifying, e.g., the regions of H–mode and L–mode, see [226, 569],which is one of the objectives of the power threshold database investigations [360, 569, 626, 664].

In this chapter, the usefulness of discriminant analysis was illustrated by taking a relevant, and interesting, problem in fusion-oriented plasma-physics. Similar methods can be applied to a wide variety of other subjects, such as the classification of wines [478], Swiss bank–notes (Chaps. 1 and 8 of [207]) or human skulls [585, 699].

It is hoped that the reader could perceive from the present case study that the specific ties to concrete applications provide a motivating stimulus for a subject-matter oriented investigation, and that thereby methods of analysis useful for practical purposes can be developed. Occasionally also some theoretical aspect can be elucidated or a substantiated indication given for further research.

[6] see Fig. 7.10 for a drawing

6 Statistical Software

Praeteriti saecli supremo mense creatus
Prisci memor hodierna sequor et crastina curo[†]
B.L. Kroon, 1899–1982

The times of Nightingale (1858), Galton (1889), Fisher (1925) and even Tukey[1] (1962), see [327, 393, 396], during which most of the practical statistical analysis was done by hand, or while using at most a desk-calculator, and where the computational labour, in addition to the conceptual difficulties, stimulated in many cases the consultation of an, if arm-chaired, experienced statistician [129], have changed since the advent and the proliferation of powerful, versatile *statistical computer packages* with extensive user interfaces. The variation in depth, scope and orientation of these packages is considerable. Each of them offers practical capabilities allowing us to perform analyses in a fraction of the time that was required in the past.

6.1 Overview

Around the early seventies of the previous century, several general-purpose statistical programs such as BMDP, GENSTAT, MINITAB, SAS and SPSS, were developed on mainframe. These packages have evolved during the course of the years while maintaining to a considerable degree backward compatibility and ensuring a reasonable longevity of existing programs. They are now well adapted to workstation (UNIX) and PC environment (MS Windows/LINUX). Around the middle of the eighties, PC-oriented packages such as Stata, Statistica, Statistix, and Statview were started, which presently also incorporate many features. Commercial packages tend to be broader in scope and are oriented towards a more extensive user support than is normally provided by the more specialised software stemming from academic research, such as, for instance, (originally) GLIM, which stands for generalised linear interactive modeling, LIMDEP for econometric regression models, Xtremes for extreme value analysis, XLispStat and XploRe for exploratory analysis with dynamic graphics, PEP for regression analysis with errors–in–variables, PROGRESS for robust multiple regression, and POSCON and CONCERN,

[†] Born during the last month of the previous century, through remembering the past, I follow the present and care for the future.

[1] Commemorative articles can be found in The Annals of Statistics (Dec. 2002) and in Statistical Science (Aug. 2003).

which incorporate confidence regions in discriminant analysis and survival analysis, respectively. These academic computer programs are, of course, exposed to software erosion due to computer hardware developments, and the influence of some of these programs has been stimulation of further research (due to their originally published use) rather than frequent practical application.

In [186] an interesting classificatory graph is presented of 18 statistical software packages. The appendix in [181] provides some overview information of several of these packages, with an emphasis on generalised linear models. An electronically available journal on statistical software, maintained at UCLA is [691]. A substantial amount of (usually open source) statistical software and a number of practical data sets have been made electronically available through StatLib, maintained at Carnegie Mellon University, see http://lib.stat.cmu.edu. In the following description, we somewhat liberally refer to internet addresses for further information on statistical packages. One should be aware of the fact that the exact locations are liable to change and may even disappear from the web. In such cases, the reader can use any general search machine, to obtain some further information from a proxy internet address.

Several software systems which were originally primarily focussed on somewhat different technological areas, such as signal processing, matrix language, symbolic mathematics or graphics, have added modules for statistical analysis. We will attempt here a brief survey of some of them. MATLAB by MathWorks Inc. (see http://www.mathworks.com) is a comprehensive, matrix oriented, scientific numerical computing environment based on C/C++. It integrates symbolic mathematics and extensive graphics and contains toolboxes for data acquisition, signal analysis, (financial) time series, image processing, neural networks, optimisation, wavelets, system identification, splines, curve fitting and statistics. Reference [523] provides a tutorial introduction. The associated package SIMULINK uses block–diagram representation for modeling, analysing and simulating dynamical systems. Similarly, IDL by Research Systems Inc. (see http://www.rsinc.com) started as an interactive data–analysis language and evolved into a fairly comprehensive system for signal processing and graphics, with an integrated widget environment and with a number of mathematical as well as statistical features. A comparable software product is PV–WAVE (see http://www.vni.com). ORIGIN (see http://www.originlab.com) is a statistical graphing software package working in an MS Windows environment, in which the user can employ either interpreter scripts (LabTalk) or a compiled language, Origin-C, similar to C++. Some special features are annotated graphics, non-linear curve fitting and peak analysis, useful for spectroscopy, and a link facility to the NAG-C library. Mathematica (see http://www.wolfram.com) and Maple (see http://www.maplesoft.com) are originally languages for symbolic calculations with integrated graphics. They have developed a notebook environ-

ment and user–application packages for a number of areas. Mathstatica is a statistical application package based on Mathematica, see [564]. Gauss (see http://www.Aptech.com) is a matrix programming language with graphical and statistical functions. Mathcad (see http://www.mathcad.co.uk/) provides symbolic, numerical and graphical analysis in an interactive spread–sheet type environment. California based Systat Software Inc. (see http://www.systat.com) markets SYSTAT, which it has acquired from the Science Unit of SPSS Inc. (see http://www.spss.com), and offers a range of complementary products for graphically oriented algorithmic analysis under MS Windows (TableCurve 2D/3D, PeakFit, Autosignal). A precious non-commercial, interactive, multi-dimensional data–visualisation tool is Xgobi [659] developed at AT&T. Dynamic graphics are combined with statistical methods in XLispStat [673] based on XLisp, a LISP language variant for X–Windows, and in XploRe (see http://www.xplore-stat.de), which is based on C++. On a yearly basis, Scientific Computing and Instrumentation (http://www.scimag.com) publishes reader's choice information on scientific software, hardware and instrumentation, from which one can observe a somewhat rotating popularity among the various commercial statistical packages. Evidently, a considerable amount of overlap exists between the functionality of these packages. An article by Gort in [186] presents an interesting examination of the usage of a few dozen statistical packages at an agricultural university in the Netherlands with distributed departments and a variety of users. While realising that it involves a considerable amount of oversimplification, we tend to group the packages with similar features, somewhat akin to the conclusions in [186], as follows:
(a) SAS/SPSS (+ BMDP) (general purpose);
(b) STATA/STATISTICA/STATISTIX/STATVIEW/SYSTAT/MINITAB - (PC-based);
(c) MAPLE/MATHEMATICA/REDUCE/MACSYMA (symbolic mathematics kernel);
(d) MATLAB/IDL/PV–WAVE/ORIGIN;
(e) S–PLUS/R/GENSTAT;
(f) Gauss/XlipStat/XploRe.
Many of these packages provide some form of guided analysis, directed towards professional scientists which have little statistical background. Areas that appear to be relatively weakly covered to date and that seem to be of increasing interest for the future are dedicated *learnware* (to assist students learning statistics), *teachware* (to assist professors preparing, computerised, statistics courses), and *siftware* (to assist researchers sifting correct information from redundant and partly erroneous data in large datasets).

In the sequel, we will restrict attention to some introductory aspects of the systems SAS and S–PLUS. The fact that these two packages have been selected for a further examination is not to be considered as a comparative valuation with regard to the other packages. SAS can be characterised as an

all-round packaged data analysis system and S–PLUS, at least in its original form, as a flexible and extendable, interpretative, graphical-statistical language. Both systems have found wide application in industry and at universities. In recent years, a considerable amount of attention was directed towards 'data warehousing' (SAS) and 'interactive graphics' (S–PLUS), which have their use in data management and in exploratory data analysis, respectively, and are not to be viewed as a substitute for theory-based subject–matter oriented statistical analysis.

Numerous textbooks exist for SAS (see http://www.sas.com/service/publications.html) and S–PLUS (see http://www.insightful.com/support/splusbooks.asp). We mention [142], [398] and [599]. More comprehensive monographs are, among others, [182] for SAS and [700] for S–PLUS.

The packages SAS (version 8) and S–PLUS (version 6) are in many respects complementary. The terminology describing the underlying data structures and data-transformation procedures differs, somewhat unfortunately, substantially and it takes some adaptive effort to become thoroughly familiar with them. For full menu-driven interactive analyses, SAS/Enterprise Guide has been created for X–Windows platforms. While having a more specific scope, the module SAS/INSIGHT can be used under UNIX, as well as MS Windows, and SAS/JMP is available on MacIntosh. The package S–PLUS is based on the S language. (The GNU public license version of S is named R, see http://www.r-project.org.) S–PLUS provides a facility to read directly SAS datasets, and it also provides interface tools to FORTRAN, C, and Mathematica. Evidently, mixed language programming has its subtle complications, but is sometimes needed for computer-intensive applications which are faster in a compiled language than in an interpreted language. Version 6 of S–PLUS supports simple data-types of FORTRAN–77 as a subset of FORTRAN–90/95. S–PLUS version 6.2 includes rather extensive logging capabilities for batch processing and XML-based reporting facilities with output in HTML, PDF and RTF format.

Precisely written user documentation, well organised on-line help in HTML format and a sufficient number (S–PLUS), or even ample collection (SAS), of example programs are provided by the proprietary software. The world-wide headquarters of SAS is located in Cary, North Carolina, USA, and the international headquarters in Heidelberg, Germany. On the world-wide web, an active user–group community exists, see for instance www.sas.de/academic. The global headquarters of the Insightful Corporation ('S–PLUS'), is located in Seattle, Washington, USA. European offices are located in France, Germany, Switzerland, and the United Kingdom. In addition, as mentioned above, several didactical monographs explain the use of the various procedures against a statistical background. Therefore, the major remaining obstacle for a novice user is how to start. To facilitate this facet of working with these packages, we will give here, after a brief overview

of their scope, some introductory information about how to start SAS and S–PLUS on UNIX systems.

Although the user interface has been considerably extended, basic SAS programming has remained essentially the same over many years. Practical examples of the usage of SAS more than a decade ago are found in [579] and, illustrated by a plasma-physical dataset from the JFT–2M tokamak in Naka, Japan, in [356]. Large scale User Group Conferences, which create annual proceedings, are being held since 1974 in the USA, and since 1981 in Europe and Japan. In addition, more specialised events, for instance 'DISK' and 'KSFE' in Germany, are organised country-wise.

It must be mentioned that, somewhat unfortunately, examples in the SAS and S–PLUS user documentation quite often only illustrate the simplest correct use of a particular syntax and the user has to find out by trial and error what goes wrong in more complicated situations, while being just somewhat comforted, in some cases, by information on the web site http://www.sashelp.com. Nevertheless, program debugging in SAS (and S–PLUS) is usually markedly less time consuming than for FORTRAN–77 and even FORTRAN–90/95 programs, because of the linear program structure (at least when the SAS macro language is not utilised) and since virtually all errors in SAS are captured by the compiler which provides them with informative error messages.

NAG (by The Numerical Algorithms Group, see http://www.nag.co.uk) and IMSL (by Visual Numerics, see http://www.vni.com/products/imsl) offer a range of numerical as well as statistical software routines, which can be integrated in a FORTRAN–90 programming environment. The range of statistical algorithms is similar to that of SAS and S–PLUS. On the other hand, despite the definite improvement in structured programming of FORTRAN–90/95 with respect to FORTRAN–77 and further developments of the FORTRAN language being planned (see http://www.j3-fortran.org), many dataset manipulations, as well as (semi-) interactive data analysis and integration of graphical user interfaces, require considerably less programming work in SAS and S–PLUS than in a bare FORTRAN environment unadorned by utility programs such as X/Winteracter (see http://www.winteracter.com).

The pictures in this book, except for those in Sect. 4.4, have been made using either the special–purpose graphics program KOMPLOT made by J. Kraak, see http://rc60.service.rug.nl/∼oldhpcv/hpc/vc/kom8/komplot.html, or S–PLUS (version 3.4) or SAS/GRAPH (version 8.2).

In Sects. 6.2 and 6.3 we describe how to start using SAS and S–PLUS under UNIX. Under MS Windows, while the functionality is basically the same as under UNIX, both systems are, even more so than under UNIX, encapsulated in a self-demonstrating windowing environment.

6.2 SAS

A SAS program consists of a sequence of *procedure calls* which are interpreted and executed by the SAS system. Because of clever default settings, these procedures require for routine usage only a very few specifications from the user.

The SAS system itself consists of several *modules*, which have been developed for several areas of application. Each module consists of a set of procedures, which can be called according to a *uniform syntax*. These procedures can be used interchangeably, without any need to specify the module to which they belong. We mention a subset of the SAS modules that are available in version 8.2 (and higher) on UNIX:

SAS/BASE – Data access, SQL, elementary analysis and data presentation

SAS/STAT – Comprehensive package for statistical analysis

SAS/GRAPH – High-resolution graphical package

SAS/ETS – Time series analysis

SAS/IML – Integrated matrix language

SAS/QC – Quality control and design of experiments

SAS/AF – Application facility for object-oriented graphical user interfaces

SAS/FSP – Product for interactive viewing and processing of SAS datasets

SAS/ANALYST – Integrated elementary menu-driven system (graphics, statistics, reporting)

SAS/ASSIST – Menu driven system which accesses most of the SAS procedures

SAS/SHARE – Concurrent access, update and management of SAS data across platforms

SAS/CONNECT – Parallel data processing by SAS jobs in a distributed environment

SAS/ACCESS – Interface to relational databases (Oracle, ODBC, MS–SQL, among others)

SAS/MDDB – Facility to extract and represent multi-dimensional data structures

SAS/EIS – Facility to manage and visualise MDDB's, SAS datasets and metadata structures

SAS/INSIGHT – Module for investigative interactive data analysis under X–Window systems

SAS/SPECTRAVIEW – Module for visualising 2-D to 4-D datasets

SAS/Warehouse Administrator – Environment for managing data warehouses

SAS/Enterprise Miner – Module for data mining and predictive analysis, which includes a wide range of different algorithms and offers capabilities for both web mining and text mining

SAS/Enterprise Guide – Project-oriented thin-client application module on X–Windows, designed to access a large part of the analytic facilties of SAS via user-friendly dialogs.

The reader is referred to the web-site http://support.sas.com/documentation/onlinedoc for online documentation of the current SAS version.

How to invoke SAS.

There are basically two ways to use SAS on UNIX.

a) Interactively:

Type 'sas' on the command line. Three windows appear ('SAS: Explorer', 'Log', 'Editor'). Below the three windows is a small 'Toolbox' window. Underneath the three windows, two more windows exist ('Results' and 'Output'). Sometimes all six windows appear at different locations. From a seventh window, called session management, one can, if needed, externally interrupt or terminate the entire SAS session. In the window called Program Editor, one can type one's own SAS program (or include an existing SAS program stored in a UNIX directory). The menu items in the three windows can be activated by using the mouse. The small, horizontally elongated ToolBox window contains illustrated buttons to activate useful commands, tailored to any of the other activated windows ('Program–Editor', 'Output', 'Log', 'Results', and 'Explorer'). The name of an executable command associated with a ToolBox button is automatically displayed when the mouse cursor is located at that button. Most of these commands are also available in the main menu of these other windows. For instance, a program in the Program–Editor window is submitted by activating `Submit` under the main menu `Run`, or by pushing the ToolBox button picturing a running figure.
Results of the calculations are written in the Output window, while comments from the SAS–interpreter are written in the Log window. The SAS editor is similar, with a few important exceptions, to the XEDIT editor, an orthodox but time-honoured command-line editor from IBM, available on a number of different platforms.
To use SAS/INSIGHT, one has to select, one after the other: `Solutions, Analysis, Interactive Data Analysis`. In SAS/INSIGHT, hard copies of data tables, graphs and analysis summaries are conveniently produced by saving a postscript file (e.g., by menu options `File, Save, Graphics File`) and sending this to a printer, for instance by using the UNIX command 'lpr'. Alternatively, the graphs can be stored in a graphical catalog, a scrollable picture book, which can be processed later

by PROC GREPLAY and in which each graph can be edited further with a graphical editor, invoked by selecting `View, Graph`.
It is stressed that any programs one may have written during an interactive session are automatically cleared after the session, unless one took the precaution to save them as a file under UNIX (use the option `Save` under the main menu item `File`).

b) Non-Interactively:

Using a UNIX editor of one's choice, one can write a SAS program and store that in a file with extension sas, e.g. *mp1.sas*. This file is submitted by typing 'sas mp1'. The output appears in the file *mp1.lst*, and messages from the SAS–interpreter appear in the file *mp1.log*.
The somewhat traditional non-interactive mode of operation is still a viable alternative to develop, on a routine basis, programs in SAS that rely on long-term reuse and lasting documentation. In this case, one can use one's familiar editor, window and file system under the UNIX operating system.

For contact information on SAS see http://www.sas.com/offices. Other internet addresses are http://www.sas.com and http://www.sas.de. An address for e-mail correspondence is academic.club@ger.sas.com.

6.3 S–Plus

While being based on a built-in *object-oriented language*, and having interfaces to SAS, FORTRAN–77, C and Mathematica, the computer package S–PLUS Version 6 covers the following areas of data analysis and statistics:

Analysis of proportions
Cluster analysis
Descriptive statistics
Graphics (2D and 3D)
Hypothesis testing
Interactive graphics
Missing data imputation
Multivariate analysis
Multiple-comparison analysis
Quality control charts
Power-function and sample-size estimation
Regression analysis (Linear/Non-linear/Non-parametric/Mixed effects,Robust)

Resampling techniques

Survival analysis

Time–series analysis

Tree-based analysis (regression/classification)

Variance analysis

As of the time of writing this book, the following additional modules or toolkits exist for S–PLUS:

S–PLUS Analytic Server (multi-user, distributed applications)

S–PLUS for ArcView GIS, a Geographical Information System

S+ArrayAnalyzer (applications in genomics and bioinformatics)

S+DOX (design of experiments)

S+EnvironmentalStats (environmental statistics)

S+FinMetrics (econometric and financial timeseries)

S+GARCH (Generalised Auto-Regressive Conditional Heteroskedastic) models

S–PLUS Graphlets (web-based interactive graphics)

S+NuOpt (large scale optimisation and non-linear regression)

S+SDK (a S–PLUS Software Development Kit, designed for C/C++ programmers)

S+SeqTrial (sequential analysis)

S+SpatialStats (spatial statistics)

S–PLUS StatServer (an intranet analysis environment)

S+Wavelets (wavelet analysis)

Insightful Miner is designed for data mining applications and InFact for automated textual analysis. Both products allow investigations of large datasets. CONNECT/C++ and CONNECT/JAVA are special interfaces to C++ and Java programs, respectively. S–PLUS Version 6 on UNIX/LINUX and on MS–Windows are cast in a Graphical User Interface and utilize Java-based Web Graphlets.

Because of its interface facilities, the kernel of the S–PLUS package could even be viewed as an extendable *module* of SAS, or alternatively as a front–end arthropod in the web of statistical software products. This is an important feature, because also for statistical software the maxim holds *non omnia possumus omnes* (*eodem exiguo tempore*)[2].

[2] Not all of us can do everything (in the same limited amount of time).

How to invoke S–PLUS.

The user–interface of S–PLUS can be easily adapted to create a familiar environment. The traditional S–PLUS invocation on UNIX from Version 3.3 onwards runs as follows:

a) Interactively:
Type 'Splus' on the UNIX command line. The program responds with few information messages followed by a prompt 'SPL>'. S–PLUS (like Mathematica) is an interpreter that evaluates the expressions one types in at the command line during the session. Results of the calculations and the interpreter messages appear in the S–PLUS window.

b) Non-Interactively:
Using a UNIX editor of one's choice, one can write an S–PLUS program and store this program in a file with extension spl, e.g. *mp1.spl*. This file is submitted by typing 'spl mp1'. Results of the calculations and the interpreter messages appear in the file *mp1.lst*.

One can combine both approaches by typing 'source("mp1.spl")' at the SPL> prompt. In this way, the file *mp1.spl* is read and executed during an interactive session.

7 Annotated Plasma Physical Datasets

γηράσκω δ' αἰεὶ πολλὰ διδασκόμενος[†]
Solon, 634-560 BC

7.1 Introduction

In this chapter, we describe some experimental datasets intended for further study. The level is intermediate between the simplified exercises in the first two chapters, which serve as a quick check for the reader whether he has understood a specific part of the theory, and actual practical case studies, which involve a vivid interplay between statistics and subject–matter considerations. To pose the right questions and to construct adequate solutions in such situations require communicative and creative skills that are more effectively learned by guided interaction during statistical consultation than from textbooks. Nevertheless, it is hoped that some of the flavour of practical statistics will transpire from the following 'stylised case studies'. In each of the case descriptions below, we provide a short introduction of the physics and some references to the literature that contain information about how the measurements were actually performed. The introductory sections, which at times may be somewhat condensed for the general reader, are intended in particular to provide a motivation for the subject of the exercise as well as sufficient background information for those who actually perform the exercises. The monographs [636] and [594] are devoted to the area of plasma edge physics. A case study on plasma edge profiles is presented in [155].

Further explanation of the physical background of several items and its implications for the basis of a Next Step tokamak experiment can be found in [671]. Most of the datasets discussed below stem from ASDEX Upgrade [254, 366, 394]. In a single instance, case 1D, the exercise dataset consists of a subset (a selection of data records as well as of variables) of a public release of the international global confinement database, available at http://efdasql.ipp.mpg.de/HmodePublic/. While it is understood that the results are based, in each case, on collaborative work of an entire tokamak team, for each of the datasets, name(s) of individual persons have been explicitly mentioned. They are usually the physicists who are or have been responsible for a pertinent diagnostic or a certain area of activity. All of the named persons were involved in the final stage of the data processing chain.

[†] I am growing old in the process of always learning many things.

The datasets are available under UNIX in the public domain area */afs/ipp /fusdat/course* and on the data carrier attached to this book. The variable names and the data are provided in standard ASCII format. In the text below, at times a more generic mathematical font is used. We trust that the small typographical differences between the electronic files and the corresponding description in this chapter will not hamper the reader's comprehension.

7.2 Case I: Scalings of the Energy Confinement Time

Physical Motivation: The scaling of the stored plasma energy is of practical interest for estimating the size of additionally heated future reactor–scale devices, where the power generated by the fusion process should exceed the external input power by a certain factor, somewhere between 10 and 50. Empirical multi-tokamak confinement–time scalings of a simple ('log-linear' or 'power–law') type have been derived, among others, in [107,233,369,370,617,653,747] for L–mode and in [123,360,591,671] for H–mode. They serve as a yardstick for (semi-) empirical local transport scalings, see for instance [40,41], and constitute the basis of at least one of three complementary approaches to plasma performance predictions for future devices, see [350, 485].

Progress in tokamak physics and international collaboration has led to the conceptual and technical design of an international thermonuclear experimental reactor ('ITER'), with a major radius of 8.14 m which was more recently replaced by the design of ITER FEAT with a major radius of 6.2 m, see [484, 671, 672, 684]. ELMy H–mode is foreseen as a reference scenario in the ITER FEAT tokamak [23, 484, 671]. Due to the improvement in energy confinement time τ_E of the plasma in H–mode, see, e.g., [362, 372, 373, 634, 668, 671], with respect to what one expects from L–mode scalings, see, e.g., [368, 370, 747], the machine size needed for achieving the mission of sustained operation beyond break-even between auxiliary heating and alpha-particle heating could be substantially reduced. The improvement in confinement time over L–mode has appreciably alleviated the necessary engineering R&D, see [175, 177, 307, 684].

While combining H–mode data from a number of machines, the dependence of τ_E on plasma parameters and on machine size has been extensively investigated [110, 123, 349, 360, 362, 470, 591, 663, 671, 678] and several empirical scalings, under varying assumptions, have been derived. In contrast to a scaling suggested in an early investigation [253], for additionally heated plasmas, the global energy confinement time in ELMy H–mode markedly decreases with heating power, at constant current approximately as $P^{-2/3}$, and a sufficient amount of heating [672, 734] is required to achieve driven burn in a next step device [350]. The empirical scaling ITERH-98(y,2), based on the ITERH.DB3v5 dataset, was used as reference scaling during the design phase of the ITER–FEAT tokamak, see [350, 484, 671, 672].

Invariance principles have been invoked to constrain the form of log-linear confinement–time scalings, see [116]. At least for simple power–law scalings, even if the heating power is used instead of the temperature as a regression variable, various such models can be expressed as linear constraints on linear combinations of the exponents, and hence conveniently be tested to empirical data, see [108, 346]. Some of the interaction between the development of power–law confinement scalings for H–mode plasmas, with expressed margins of uncertainty, and the design of the next-step ITER experiment can be retraced from [22, 110, 349, 350, 360, 483–485, 591, 671, 672].

It must be mentioned that the true regression surface is expected to be more accurately described by non-linear scalings, see, e.g., [110, 353, 361, 616, 662, 747], which are more intricate to fit, but which can be connected to, for instance, core–pedestal plasma models [122, 269, 287, 362, 692]. Specific areas are the confinement scaling of Ohmic discharges [85, 624] and a unification of L–mode and H–mode confinement [353, 355] related to bifurcation models [305, 306].

The stored energy of the plasma is related in a simple way to the confinement time $\tau_E = W_{th}/P_{L'}$ and to the effective heat diffusivity[1] $\langle\chi\rangle = \kappa/n \simeq a^2/\tau_E$. (Here, n stands for the plasma density, κ for the effective heat conductivity, and a for the plasma minor radius.) Elucidating the detailed mechanism behind plasma heat transport is still an unresolved research problem in theoretical plasma physics [303]. Sophisticated models have been developed which describe several aspects of plasma heat transport through micro-turbulence, see [306, 308, 335, 336, 603]: drift waves, influenced by, e.g., magnetic shear and curvature [589, 701], among which trapped electron modes (TEM) [667]; ion and electron temperature–gradient (ITG, ETG) modes, [565, 726]; pressure driven modes, such as ballooning modes (with high mode number) [38, 119]. Besides ELMs (see case IV^A) and sawteeth [729], also magnetic-island formation, magneto-hydrodynamic tearing modes [61, 62, 258, 259, 380], and, in case of heating by neutral beams or alpha particles, toroidal Alfvén eigenmodes due to velocity space anisotropy (see [46, 185] and Chap. 5 of [671]) can play a role. A special area is (anomalous) transport barrier physics [261, 503]. Merely classical heat transport, as described in [86], and neoclassical heat transport (which takes toroidal

[1] The word effective means here that an average is taken of ion and electron diffusivity, while both of them are averaged over flux-surfaces and these averages are again, now harmonically, averaged in radial direction with considerable weighting near the last closed flux-surface, see Appendix C of [622]. For simplicity, a circular cross-section has been assumed and an approximately constant factor is ignored. More generally (see, e.g., Sect. 4.3 of [667]), the approximations $\langle\chi\rangle \simeq Area/(14\tau_E) \simeq S/(450\tau_E)$ hold, where $Area$ is the plasma cross-sectional area and S the plasma surface area. Normalised by these constants, $\langle\chi\rangle$ is about 1 m^2/s for ELMy H-mode and varies, except for devices with a very high density and magnetic field, in practice between between 0.5 and 2.0 m^2/s, which are approximately the 5% and 95% quantiles in the ITERH.DB3 database.

effects into account), see [216, 279] and Chap. 6 of [632], lead to effective heat–diffusivity predictions that are typically about two orders of magnitude lower than those experimentally observed in magnetically confined toroidal plasmas.

A fundamental article describing the discovery of the H–mode at the ASDEX tokamak is [714]. Some historic records describing the state of the art concerning the subject of this exercise before the engineering design phase of the ITER project, are [173, 253, 295, 411, 413]. For a general introduction to fusion-oriented plasma physics, the reader is referred to [20, 82, 84, 105, 234, 335, 365, 471, 646, 729], and for its fifty years history to [84, 89, 273, 482, 613].

Case Ia: Scalings of the Energy Confinement Time, Part I

Dataset conf1.dat (prepared by O. Kardaun)

Data Description: This artificial dataset describes a 'hypothetical dependence' of the thermal plasma energy on some basic characteristics of plasma discharges in ASDEX Upgrade. These characteristics are called 'variables' in statistical and 'plasma parameters' in physical jargon. A particular type of (compound) power–law model has been used to generate the 'true values' of the stored thermal energy. The basic plasma parameters have been varied according to a special, 'central composite', statistical design. For apposite background, see [113, 205, 278] and [125, 487, 527]. One of the first references describing actual usage of statistical experimental design in plasma physics is [176]. The errors satisfy the standard assumptions of ordinary least squares regression. (The thermal energy data have been generated using an artificial model and do not contain any real measurements from ASDEX Upgrade.)

Table 7.1. The first three and last two observations in the file conf1.dat are:

wth	ip	bt	nel	pl
0.0804	0.8	2	4	2
0.1132	0.8	2	4	4
0.1343	0.8	2	4	6
...				
0.1348	1	2.5	2.6	6
0.1659	1	2.5	2.6	8

The thermal energy confinement time is defined as the ratio between W_{th} and P_L.

The variables in the dataset are:
W_{th}: Thermal plasma energy [MJ]
I_p: Plasma current [MA]

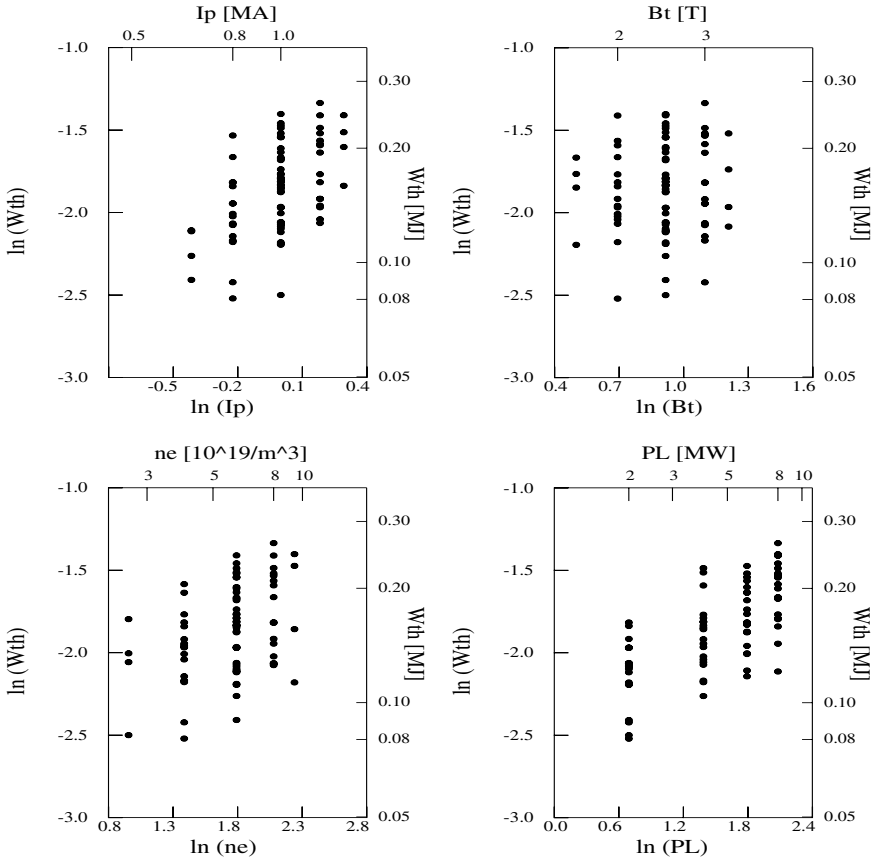

Fig. 7.1. Dataset conf1.dat: Thermal plasma energy (W_{th}) against plasma current (I_p), magnetic field (B_t), electron density (n_e) and heating power (P_L) – artificial data, for didactic purpose only –

B_t: Magnetic field[2] [T]
$\overline{n}_e$ (= nel): Line-averaged electron density [10^{19}/m^3]
$P_{L'}$: Absorbed input heating power [MW]

Exercise 7.1. (*) Analyse this dataset and find the correct scaling for W_{th} using for instance SAS, PROC/INSIGHT.

[2] More accurately, B_t is the toroidal magnetic field at the centre of the plasma, i.e., the geometrical centre of the last closed flux-surface, see [176].

Case Ib: Scalings of the Energy Confinement Time, Part II

Datasets conf2.dat and conf2b.dat (prepared by O. Kardaun)

Also the dataset conf2.sas7bdat contains *simulated* AUG-like data (intended for tutorial purposes only). The variables W_{th}, $W_{th,e}$, $W_{th,cf}$, and $W_{th,cf2}$ all contain the 'measured' thermal energy, each generated with a different deviation from the standard assumptions, i.e., independent and normally distributed errors with constant variance, of ordinary least-squares regression.

Table 7.2. The first three and last two observations in the file conf2.dat are:

wth	wthe	wthcf	wthcf2	ip	bt	nel	pl
0.0741	0.0911	0.0801	0.0785	0.8	2	4	2
0.1232	0.1153	0.0987	0.0987	0.8	2	4	4
0.1470	0.1233	0.1018	0.1219	0.8	2	4	6
...							
0.1348	0.1442	0.1144	0.1369	1	2.5	2.6	6
0.1636	0.1693	0.1247	0.1493	1	2.5	2.6	8

The physical units are the same as those in case Ia.

Exercise 7.2. Estimate W_{th} as a function of I_p, B_t, $\overline{n}_e$, P_L.
What can you say about the error distribution?
What kind of systematic deviations do you detect from a simple power law?

Dataset conf2b.sas7bdat is of the same type as conf2.sas7bdat. However, the plasma energy is now stored in the variable *wthvs* and the dataset contains some gross outliers (imputed data).

Exercise 7.3. Try to detect the outliers using robust regression methods (or otherwise). What are your best regression estimates?

Case Ic: Scalings of the Energy Confinement Time, Part III

Dataset conf3.dat (prepared by F. Ryter, with special contributions from C. Fuchs, L. Horton, A. Kallenbach, O. Kardaun, B. Kurzan, A. Stäbler, J. Stober)

Data Description: This is quite a complicated dataset containing actual measurements from ASDEX Upgrade under a variety of conditions. The experiments have not been performed according to a specific statistical design. The variable W_{th} is the response variable which contains the measured thermal plasma energy. The first two columns represent the discharge ('shot') numbers and the times at which the measurements were taken.

Regression variables in the dataset, in addition to those described in Case Ia, are:
$H \to H^+$ (=hh): indicator for plasma isotope and beam isotope (hh=1 for hydrogen $H \to H^+$, hh=0 for deuterium $D \to D^+$; only discharges with the same isotope for the beam particles as well as for the working gas are included in the data set)
$n_e(0)$ (= ne0): Central electron density $[10^{19}/\mathrm{m}^3]$
$\langle n_e \rangle$ (= nev): Volume-averaged electron density $[10^{19}/\mathrm{m}^3]$
δ (=delta): Triangularity of the last closed magnetic surface, related to the plasma shape; for its (geometric) definition, see [176].
ind: Indicator variable for outliers, which has been set to -1 when one of the two electron density peaking factors, $\gamma_0 = n_e(0)/\overline{n}_e$ or $\gamma_v = \overline{n}_e/\langle n_e \rangle$, is outside its usual range and substantial systematic measurement inaccuracies are expected; the effect of omitting these observations, in order to achieve more robust and physically more plausible regression results, can be investigated.

Obviously, all kinds of functional relationships and deviations from the 'standard error assumptions' (independent and normally distributed errors with constant variance on logarithmic scale) are possible. Approximate measurement accuracies are 10% (W_{th}), 1% (I_p), 1% (B_t), 3.5% ($\overline{n}_e$), 7.5% ($P_{L'}$), 12% ($n_e(0)$), 12% ($\langle n_e \rangle$), 8%(F_q) and 8%$(1+\delta)$. The quantities δ and F_q are dimensionless. The units of the other quantities are the same as in the previous two cases. The variable F_q is equal to q_{95}/q_{eng}, where q_{95} is the ratio between the number of toroidal and polodial turns of a magnetic field line at a normalised flux–surface radius 0.95, and $q_{eng} = 5(B_t/I_p)(Area/\pi R)$ with $Area$ the plasma cross-sectional area inside the separatrix and R the major radius of the plasma.[3] The quantity F_q is correlated with the geometric shape of the plasma.

Table 7.3. The first three and last two observations in the file conf3.dat are:

shot	time	wth	hh	ip	bt	nel	ne0	nev	pl	delta	Fq	ind
6483	4.000	0.422	0	1.002	2.090	9.923	.	.	4.553	0.121	1.313	0
7439	3.000	0.328	0	0.997	2.499	4.502	5.340	.	2.575	0.094	1.284	0
7440	3.000	0.326	0	0.997	2.499	4.773	5.630	.	2.605	0.094	1.283	0
...												
15840	6.650	0.759	0	1.000	2.092	5.777	.	.	5.241	0.155	1.371	0
15930	4.900	0.398	0	0.719	1.847	6.881	.	.	2.332	0.420	1.432	0

[3] To a practical extent, similar quantities are $q_{cyl} = 5\frac{B_t}{I_p}\frac{ab}{R}$ and $q_{cyl,*} = 5\frac{B_t}{I_p}\frac{(L/2\pi)^2}{R}$, where b is the vertical minor radius and L the contour length of the separatrix. For stellarators, the rotational transform $\iota/2\pi = 1/q_{cyl,*}$ is routinely used.

Experimental Description: For a fixed device, the thermal plasma energy is largely (but not entirely) determined by four groups of plasma parameters: Plasma current, heating power, shape of the last-closed magnetic flux surface and plasma density. For all discharges, deuterium was used both as working gas, introduced by gas valves ('gas–puff'), and, in as far as applied, as the gas injected by the neutral-beam additional heating system. Part of the discharges were additionally heated by Ion Cyclotron Resonance Heating (ICRH) or by combined ICRH and neutral-beam injection (NBI). The measurement techniques for each of these four groups of plasma parameters is briefly described below.

(a) The plasma current is measured by Rogowskij coils, and the toroidal magnetic field by magnetic flux loops, see [729].

(b) For plasmas heated by neutral beam injection (NBI), the heating power $P_{L'}$ (=pl) is equal to $P_{\mathrm{OHM}} + P_{\mathrm{NBI}} - P_{\mathrm{LOSS}}$, where the Ohmic power P_{OHM}, equals the 'loop voltage' times the plasma current, P_{NBI} is the power of the neutral beams injected into the torus, and P_{LOSS} is the power lost by shine through, unconfined orbits and charge exchange of the beam particles in the plasma. The physics behind these processes, described in [501], is implemented in the FAFNER Monte Carlo code, see [431]. For ASDEX Upgrade, a parametric fit of these losses, the total amount of which varies between some 10% to 30% of P_{NBI}, as a function of density and average temperature to a designed database of modeled deuterium discharges has been used to estimate P_{LOSS} for the discharges in the present dataset, see [706]. For ion cyclotron heating, it was assumed that the power coupled to the plasma is 0.8 times the power delivered from the grid.

(c) δ and $F_q = q_{95}/q_{eng}$, which was introduced as q_{sh} in [349], are based on transforming the external measurements of the poloidal flux and of the poloidal magnetic field into plasma shape parameters by applying function parameterisation to an extensive database of plasma equilibria calculated by solving the Grad–Shafranov equation, described in Case VII below, see [83, 449].

(d) the line-averaged density $\overline{n}_e$ is based on DCN interferometry[4] (see Appendix 12 in [594]), and $\langle n_e \rangle$ is equal to the total number of particles, estimated by a profile fit, see [343,451], of YAG–laser Thomson scattering measurements, see [562] and Appendix 13 in [594], divided by the

[4] The name (DCN) stems from the fact that the laser used for this type of interferometry [221, 222, 499] utilises deuterium cyanide, D−C≡N, rather than hydrogen cyanide as laser-active molecule. A classic textbook on laser operation is [620]. The wavelength of the DCN–laser, $\lambda \simeq 1950$ Å, is about 60% of that of the HCN–laser. This shorter wavelength is more suitable for measuring larger line–integrals of the plasma electron density and is less sensitive to laser–light deflection in the plasma boundary layer.

plasma volume, which is determined by function parameterisation. The YAG–laser diagnostic at ASDEX Upgrade consists of a vertical system with six laser units, each with 4 spectral channels and 20 Hz repetition rate, as well as a horizontal system. (For all discharges between 6483 and 12249, the YAG–laser profiles were fitted, during the Ohmic phase, to the profiles from a combination of the DCN line density and the lithium–beam edge diagnostic. After discharge 12249, an improved absolute Raman calibration was substituted for this procedure.)

(e) The thermal energy of the plasma is determined by MHD equilibrium[5] fits based on function parameterisation and corrected for the (density and temperature dependent) contributions from fast particles by application of a parametric fit, see [706], to FREYA / FAFNER Monte–Carlo code calculations, see [431]. It is noted that two other ways to determine the thermal energy are (1) by measuring the plasma diamagnetism associated with the plasma pressure, see [171, 729], and (2) 'kinetically', i.e., by measuring the electron and ion temperature profiles as well as the electron density profiles, while using measurements of $Z_{\rm eff}$ (see [639]) and specific assumptions on the main impurity species to estimate the ratio between ion and electron density profiles, see [176, 678]. As described for instance in [349], the relative consistency generally achieved experimentally between these three types of energy measurements varies between some 5% and 20%.

Exercise 7.4. a) Estimate the coefficients of a log–linear scaling, while using the four regression variables from Part I under standard assumptions for least-squares regression;
b) idem, while assuming that the errors in the regression variables are as given above;
c) compare your result with the ITERH–98P(y,2) scaling (see ITER Physics Basis, 1999, p. 2208), which is based on data from 9 tokamaks.

Exercise 7.5. (*) Analyse the residual dependence of

$$\log H_{98y2} = \log(\mathrm{wth}/\mathrm{wth}_{\mathrm{H98y2}}), \tag{7.1}$$

with respect to the density peaking *ne0/nel* and to δ where

$$W_{th,H98y2} \propto I_p^{0.93} B_t^{0.15} P_{L'}^{+0.31} \overline{n}_e^{0.41} \tag{7.2}$$

stands for the prediction of W_{th} according to the ITERH–98P(y,2) scaling (apart from a proportionality factor).
a) Verify by univariate regression that $\log \mathrm{H}_{98\mathrm{y}2}$ decreases with nel and increases with $\log(\mathrm{F_q}), \log(1+\delta)$, and $\log(\mathrm{ne0/nel})$;
b) eliminate a small number of points indicated by ind $= -1$, which happen to

[5] MHD is an abbreviation for magneto-hydrodynamics or its adjective (see, e.g., [729]).

have abnormal confinement degradation, and omit the hydrogen discharges, which are marked by hh=1; compare a log–linear regression of $\log H_{98y2}$ with respect to log (ne) with a log–cubic regression;
c) apply a regression of $\log H_{98y2}$ against log(nel/nev) and compare this with the results from inverse regression; derive the regression equation based on principal component analysis assuming independent (random) errors of magnitude 10% and 10% in wth and nel/nev, respectively;
d) perform, on a logarithmic scale, simultaneous regression of $\log(H_{98y2})$ against F_q, ne0/nel, $1+\delta$, nel, and a log–linear interaction term $(1+\delta)^{\log(nel)}$;
e) check the residuals of the last regression for the presence of a significant curvature with respect to the density by fitting a log–quadratic regression model. Notice that the presence of a (negative) curvature, related to two 'empty triangles' (one at low and another at high density), and which is usually called 'the density roll-over effect', reduces the magnitude of the interaction (on a logarithmic scale) between nel and $1+\delta$.

Case Id: Scalings Auxiliary to those of the Energy Confinement Time

Dataset conf4.dat (prepared by O. Kardaun, A. Kus and the ASDEX team, F. Ryter and the AUG team, J. DeBoo and the DIII–D team, K. Thomsen, J. Cordey, D. McDonald and the JET team, Y. Miura, T. Takizuka, H. Urano and the JT–60U team, and S. Kaye, M. Bell, C. Bush and the TFTR team, to the public version of the international ITERH.DB3 global confinement dataset, see http://efdasql.ipp.mpg.de/HmodePublic/)[6]

Data Description: This dataset is a subset of the ELMy H–mode data of the international H–mode confinement database ITERH.DB3 [362,679] and comprises actual measurements from six tokamaks. Only neutral-beam injected discharges with deuterium beam and deuterium plasma working gas ($D \rightarrow D^+$) have been included.In this case the response variables are $\gamma_0 = n_e(0)/\overline{n}_e$ and $\gamma_v = \overline{n}_e/\langle n_e \rangle$, which are two different characterisations of the peakedness of the electron density profile. The regression variables are divided into three groups:

(a) four plasma shape variables: inverse aspect ratio $\varepsilon = a/R_{geo}$, 'area' elongation[7] $\kappa_a = Area/(\pi a^2)$, triangularity δ, and shape parameter $F_q = q_{95}/q_{eng}$;

[6] The abbreviations stand for Axially Symmetric Divertor Experiment (ASDEX), ASDEX Upgrade (AUG), Doublet-III D-shaped (DIII–D), Joint European Torus (JET), Japanese Torus – 60 m^3 Upgrade (JT–60U), and Tokamak Fusion Test Reactor (TFTR), respectively, see also Sect. 4.1.
[7] To a good approximation, $Area = V/(2\pi R_{geo})$, where V is the plasma volume.

(b) three plasma–physical variables:[8] the normalised (ion) poloidal Larmor radius $\rho_* = \rho/(L/2\pi) = c\frac{\sqrt{M}}{Z}\frac{\sqrt{T}}{I_p} \sim \frac{c}{5}\frac{\sqrt{M}}{Z}\frac{\sqrt{T}}{\overline{B}_p(L/2\pi)}$, and the collisionality $\nu_* = c' n R_{geo}/T^2$ (for some constants[9] c, and c'), as well as the Greenwald parameter $\overline{n}_e/n_G = \overline{n}_e\kappa_a/(10 j_p)$;

(c) three variables related to beam penetration: the absorbed heating power $P_{L'}$, the density times the plasma minor radius $\overline{n}_e a$ and the beam particle energy $E_{\rm NBI}$.

The objective of this exercise is to perform, in a non-automatic way, stepwise multiple regression analysis of γ_0 and γ_v, while using a 'bottom-up' approach (i.e., going from simple models with a very few regression variables to more complicated ones), and applying errors-in-variables techniques if this is deemed needed in view of multi-collinearity between the regression variables.

Regression variables in the dataset, in addition to those described in Cases Ia–Ic are:
tok: Acronym of tokamak device
R_{geo} (= rgeo): Plasma major radius [m]
a_{min} (= amin): Plasma minor radius [m]
κ_a (= kappaa): Plasma ('area') elongation, i.e., the ratio between the cross-sectional area and πa^2
q_{eng} (=eng): ('engineering') safety factor $5\frac{B_t}{I_p}\frac{Area}{\pi R_{geo}}$
$E_{\rm NBI}$ (= enbi): Injected beam-particle energy [MeV]
ρ_* (= rhostar): Normalised, average (ion) Larmor radius, $c\frac{\sqrt{M}}{Z}\frac{\sqrt{T}}{I_p}$, expressed in percent (%)
ν_* (= nustar): Normalised, average (ion) collisionality, $c' R_{geo}\langle n_e\rangle/T^2$, expressed in percent (%)

[8] In the formulas, we use I_p for the plasma current and L for the contour length of the separatrix, $\overline{B}_p$ is the poloidal magnetic field averaged over the separatrix contour, such that $\overline{B}_p L = \mu_0 I_p$ with $\mu_0 = 4\pi \times 10^{-7}$, M and Z are the isotope mass and charge numbers, which are for D → D^+ discharges equal to 2 and 1, respectively, $T \sim W_{th}/nV$, with V the plasma volume, is the average temperature, $n = \langle n_e\rangle$ is the volume-averaged electron density, R is the plasma major radius and $j_p = I_p/Area$ the average current density of the plasma.

[9] Specifically, we used $c \simeq 2.26\%$ and $c' \simeq 0.226\%$ for I_p in MA, n in $10^{19} m^{-3}$ and T in keV. Following [109], ν_* used here is the global quantity corresponding to $\nu_* = \nu R_{geo}/v_{th,i}$, where ν is the Spitzer collision frequency and $v_{th,i} = v_\perp = \sqrt{2kT_i/m_i}$ is the thermal ion velocity perpendicular to the magnetic field. The factor $q(r)/\varepsilon(r)^{3/2}$ in the usual definition of the (local) collisionality parameter (see, e.g., [667]) has not been included. In the same units, assuming $T_i \simeq T_e$, for three degrees of freedom of the particle motion, the thermal plasma energy is $W_{th}(MJ) = 4.8 \times 10^{-3} nTV$. Also the normalised (ion) poloidal Larmor radius, related to the normalised width of the ion banana–orbit [365], $\rho_* = \frac{m}{q}\frac{v_\perp}{\overline{B}_p(L/2\pi)}$, with m and q the ion mass and charge, respectively, is based on two degrees of freedom, for which $\frac{1}{2}mv_\perp^2 = kT$ (see, e.g., [105, 108, 471, 671]).

$\overline{n}_e/n_G$ (=ngr): Greenwald density parameter, equal to $(\overline{n}_e/10)/(I_p/\pi a^2)$
ind: Indicator variable for outliers, which has been set to -1 for outliers, $+1$ for 'improved H–mode' discharges at AUG, and to 0 otherwise.
Approximate measurement accuracies are 1% (R_{geo}), 2% ($\overline{n}_e a_{min}$),3% (a_{min}), 4% (κ_a), 3.5% (a_{min}/R_{geo}), 4% ($\overline{n}_e/n_G$), 5% (q_{eng}), 5% ($E_{\rm NBI}$), 10% (T), 10% (ρ_*), 20% (ν_*), 10% (γ_0) and 10% (γ_v). There are some correlations between the errors, for instance between those of $\overline{n}_e$ and a_{min} (approximately $-5/6$), ρ_* and ν_* (approximately $-3/4$), as well as between those of q_{eng} and (R_{geo},a_{min},κ_a), which can be directly estimated from the definitions above.

Table 7.4. The first two observations and the last observation in the file conf4.dat (sometimes rounded to two significant digits) are:

tok	shot	time	date	nel	ne0	nev	rgeo	amin	kappaa	delta
tok	shot	time	hh qeng	Fq	pl	eenbi	rhostar	nustar	ngr	ind
ASDEX	31138	1.263	19900220	2.91	3.88	2.67	1.69	0.38	1.01	0.00
ASDEX	31138	1.263	0 3.351	1.14	1.27	0.054	8.37	3.42	0.39	0
ASDEX	31140	1.270	19900220	2.95	4.15	2.56	1.69	0.38	1.01	0.00
ASDEX	31140	1.270	0 3.589	1.15	1.55	0.054	8.59	4.06	0.43	0
...										
TFTR	78174	3.700	19940721	4.28	7.37	3.30	2.52	0.87	1.00	0.00
TFTR	78174	3.700	0 5.275	1.38	20.28	0.10	3.83	0.53	0.70	0

Experimental Description: For the experimental background of the method of measurement the reader is referred to Case Ic for ASDEX Upgrade and to [678, 679] for further specific details on the data from the other tokamaks (ASDEX, DIII–D, JET, JT–60U). In the following exercise it is in particular interesting to look at the correlation between plasma-physical variables, such as ν_* and ρ_* and measured characteristics of the density profile, at various levels of sophistication, by: (1) univariate regression; (2) multivariate regression, without and with a proxy variable, while using ordinary least squares; (3) multivariate regression using errors-in-variable techniques.

Exercise 7.6. a) Plot $\gamma_0 = n_e(0)/\overline{n}_e$ against $\gamma_v = \overline{n}_e/\langle n_e\rangle$ and investigate graphically the dependence of γ_0 and γ_v on (separately) ρ_* and ν_*, while using a different colour for each of the six tokamaks. Investigate possible systematic trends of the residuals from the two univariate log-linear regressions against basic plasma variables, such as $\overline{n}_e$, $P_{L'}$, $\varepsilon = a_{min}/R_{geo}$, q_{eng} and q_{95}. Examine the influence of omitting the timeslices with $ind = +1$. (*) Do the same after fitting a two-term power–law model of γ_0 as a function of ν_*.
b) (*) Perform multiple regression analysis of γ_0 and γ_v with respect to (simultaneously) (a) ε, κ_a and either F_q or δ; (b) ρ_* and ν_*; (c) $P_{L'}$, $\overline{n}_e a$ and $E_{\rm NBI}$. Check the absence of a possible systematic trend of the residuals with

respect to the basic plasma variables. Omit the timeslices with $ind = -1$, and examine the influence of omitting, in addition, the timeslices with $ind = +1$. Investigate the influence of restricting attention to observations not close to the Greenwald limit, i.e., to $\overline{n}_e/n_G < 0.85$. Calculate the partial correlation matrix between (ρ_*, ν_*) and (γ_0, γ_v), where the contribution of all other variables is 'partialled out'. (*) Perform robust regression using either the FORTRAN program PROGRESS[10] or its implementation in SAS/IML [580].

c) Perform multiple regression as under *b*) while replacing ν_* by $\overline{n}_e/n_G$, and compare the goodness-of-fit in these two cases. Standardise the regression variables (such that they have unit standard deviation) before applying them in another multiple regression analysis.[11]

d) Plot ν_* against $\overline{n}_e/n_G$ and note that a rather strong correlation between these two variables exists. At least when $\overline{n}_e/n_G$ is added to the regression variables under b), regression with an errors-in-variables technique is thought to be more adequate than OLS.

e) Investigate, more formally, the condition of the dataset for the regression variables under b) by principal component analysis. Determine the eigenvectors corresponding to the two or three smallest principal components. Do the same when $\overline{n}_e/n_G$ is added as a regression variable.

f) Perform errors-in-variables regression of γ_0 and γ_v against the variables under b) and in addition $\overline{n}_e/n_G$, using the error estimates as given above (i) while neglecting the correlations between the errors in the variables, and (ii) assuming, on a logarithmic scale, correlations (0.1,0.3) for $\big((\nu_*,\overline{n}_e/n_G),\overline{n}_e a\big)$, (−0.2,−0.2) for $\big((\rho_*,\overline{n}_e/n_G),\nu_*\big)$, (0.5,0.45,−0.2,−0.5) for $\big((\overline{n}_e/n_G,q_{eng},F_q,\kappa_a),a/R\big)$, −0.7 for (F_q,q_{eng}) and (−0.2,−0.1) for $\big((\overline{n}_e a),\gamma_0\big)$, respectively. Compare the results with those from ordinary least squares regression, which is in this case somewhat impaired by the multi-collinearity between the regression variables in the dataset.

g) For some of your best fitting models, predict, with a rough estimate of the prediction accuracy, γ_0 and γ_v for a machine like FIRE*, see [459], with parameters (nel, rgeo, amin, kappaa, delta, qeng) = (50, 2.14, 0.595, 1.85, 0.7, 2.0) and (Fq, pl, eenbi, rhostar, nustar, ngr) = (1.58, 34.0, 1.0, 1.2, 0.64, 0.7), the same units having been used as in the table above, under the (somewhat artificial) approximation that the heating profile is similar to that of the discharges in the dataset.[12]

---

[10] Program available from http://win-www.uia.ac.be/u/statis/Robustn.htm.

[11] Standardised in this way, the regression coefficients reflect reasonably well the relative importance of the various regression variables, see for instance Chap. 14, Sect. 3.1.2 of [347].

[12] In fact, in view of the contribution of the alpha–particle heating, more or less akin to Ohmic heating, the heating deposition profile is expected to be more centrally located in such a device than in present-day machines.

7.3 Case II: Halo Currents at ASDEX Upgrade (AUG)

Dataset halo1.dat (prepared by G. Pautasso)

Physical Motivation: Large mechanical forces of electro-magnetic nature arise during the fast current decay ('current quench phase') of a plasma disruption. Poloidal currents, called halo currents, flow from the plasma halo region into the vessel structures and vice versa. The plasma halo region corresponds to the region of open magnetic flux surfaces outside the plasma, which intersect with the vessel structures. Halo currents are produced by several effects: (a) the time variation of the toroidal magnetic flux across the shrinking poloidal cross-section of the plasma, (b) by electro-magnetic induction and (c) 'anomalous' diffusion of the plasma current ('skin effect'). The magnitude of halo currents is usually up to some 50% of the plasma current, and hence they are a major contribution to the mechanical forces on the machine during a disruption. While the halo current can be easily measured with different methods, further physical characterisation of the halo region through measurement of its extension, of the densities and temperatures and of the voltage drop at the plasma–wall interface, or even in-situ measurements of forces and stresses in vessel components, are often not available. The prediction of the magnitude of the halo currents for discharges in an already explored operational space, and its extrapolation to future devices, still relies on empirical scalings rather than on fundamental theory or simulations. Some theoretical models of disruptions have been described in [729] and statistics of a disruption database from the ASDEX tokamak can be found in [751]. For an description of practical understanding of disruptions and possible stratagems for their avoidance or mitigation, see [505, 506, 597], and Chap. 3.4 of [671]. Online prediction and disruption mitigation by pellet injection at ASDEX Upgrade is described in [507]. Disruption heat loads on the divertor at JET are analysed in [551].

Data Description: The halo currents in ASDEX Upgrade are measured by resistive rods ('shunts'), which are mounted between the tiles and the support structure of the divertor. The data have been assembled from the early phase of DIV I of ASDEX Upgrade, during which two poloidal measurement arrays existed, one on the inner and one on the outer divertor plate, at one toroidal position. The halo current is poloidally asymmetric. The halo current and its degree of asymmetry depends on I_p and B_t/I_p. For an overview of divertor physics investigations during this phase of ASDEX operation, see [412].

The variables in the dataset are:
shot: plasma shot number
maxI_o: maximum value of the halo current flowing poloidally between plasma and the outer divertor plate [A]
maxI_i: as above, for the inner divertor plate [A]

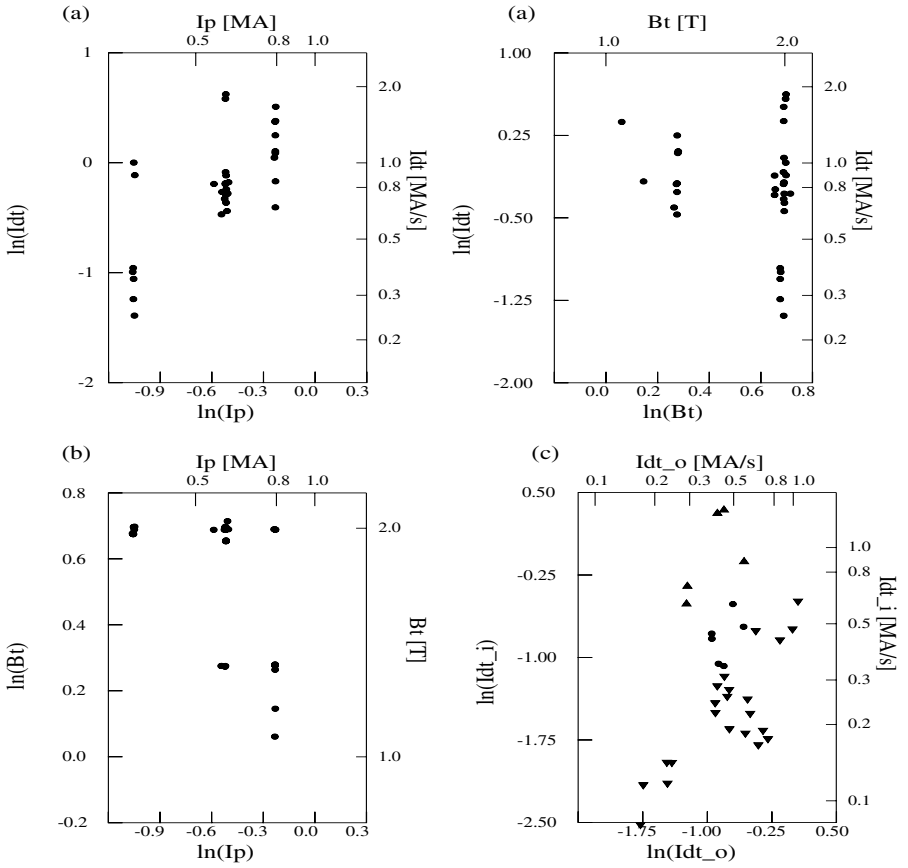

Fig. 7.2. Dataset halo1.dat: (*a*) time-integrated halo–current (Idt) against plasma current (I_p) and magnetic field (B_t); (*b*) magnetic field (B_t) against plasma current (I_p); (*c*) integrated halo current at the inner divertor against the halo current at the outer divertor, observations with poloidal asymmetry larger than 0.15 being marked by triangles

Idt_o: time integral of the halo current flowing onto the outer divertor plate [A s]
Idt_i: as above, for the inner divertor plate [A s]
polasy: (Idt_i−Idt_o)/(Idt_i+Idt_o) degree of in-out asymmetry of the measurements
I_p: plasma current [A]
B_t: magnetic field [T]

Exercise 7.7. Investigate by regression analysis whether or not the available data suggest a statistically significantly stronger than linear increase of the

Table 7.5. The first three and the last two observations in the file halo1.dat are:

shot	maxI_o	maxI_i	Idt_o	Idt_i	polasy	ip	bt
1936	87230	213700	292.0	709.4	0.2216	348700	2.007
1943	86810	137900	289.1	604.0	0.1903	350800	2.009
1944	86730	240100	413.6	1376	0.2802	593800	2.006
...							
5025	135700	98700	475.8	190.5	0.2574	793800	1.302
5036	66700	132000	387.4	457.1	0.01442	794500	1.157

(time-integrated) halo current as a function of plasma current
a) when all data are put together;
b) when the data with high poloidal asymmetry (polasy > 0.15) are removed.

Can you detect influential or outlying data points?

7.4 Case III: Density Limit (AUG)

Dataset dlim1.dat (prepared by R. Neu)

Physical Motivation: The attainable performance of future tokamak reactors is limited by several operational boundaries, see, e.g., [335], Chap. 3 of [671], and [96]. One of them is the so-called density limit. Several competing causes may determine the maximum attainable density in the high confinement ('H–mode') and low confinement ('L–mode') regime, respectively. The maximum attainable density in L–mode and Ohmic discharges is primarily determined by the balance between the radiated power, proportional to $\overline{n}_e^2$, and the heating power, for Ohmic discharges proportional to the square of the current density and to the effective charge number Z_{eff}. Above the critical density, expanding MARFEs occur,[13] which are usually the precursor for a pending 'density limit disruption'. When the plasma is operated in H–mode, a rather dramatic relapse into L–mode occurs, depending somewhat on profile effects, at densities which are in practice, to a reasonable approximation ($\pm 30\%$) described by simple expressions such as $\overline{n}_e \leq I/(\pi a^2)$, where I is the plasma current and a the plasma minor radius [246], see [72] for an adjustment. The physical mechanisms of the underlying processes are still largely unsolved. In [740], while using the basic heat-conduction equation, the maximum attainable density has been related to a bifurcation of the temperature as a function of the ratio between radiated power and heating power. For a

[13] The acronym MARFE stands for Multifaceted Asymmetric Radiation From the Edge.

further description of the physical background and the phenomenology of operational boundaries of the plasma density, the reader is referred to Chap. 3 of [671] and [71, 72, 246, 335, 461, 462, 486, 631], as well as the review article [245]. The lecture [170] provides background information on the role of impurities in tokamaks, especially JET.

Data Description: The objective of this exercise is to investigate the dependence of the maximum attainable density on the concentration of carbon and oxygen impurities, which were the main impurities during the operating period considered (March 1993 until March 1995), during which the machine was equipped with an open, so-called 'type-I' divertor, see [492]. Attention is restricted to Ohmic discharges at a constant plasma current (0.6 MA) and constant magnetic field (2T).

The (important) variables in the dataset are:
shot: shot number identifying the plasma discharge
gr: 1: original ('training') dataset, 2: incremental addition to the original dataset
shotbo: shot number of the first discharge after the most recent boronisation procedure
nmax: maximum attained density
dco: spectroscopically measured oxygen concentration
dcc: spectroscopically measured carbon concentration
(The other variables are not used in this exercise. To retain a practical flavour, they have not been removed from the dataset.)

Table 7.6. The first three and last two observations in the file dlim1.dat are:

shot	gr	shotbo	nmax	tdl	n12	io	co	ic	cc	dco	dcc
2788	1	2533	5.47	1.84	2.65	540	2.5	100	2.4	0.41	0.69
2832	1	2533	5.40	1.82	2.70	740	3.0	120	2.7	0.50	0.80
2903	1	2533	5.43	1.85	2.88	650	0.5	110	0.5	0.42	0.53
...											
5898	2	5704	5.16	1.68	2.65	440	0	95	0	0.27	0.80
5997	2	5704	5.20	1.69	2.65	400	0	65	0	0.26	0.50

Experimental Description: A good reference describing the C–O monitor system at ASDEX Upgrade is [491]. The carbon and oxygen concentrations are measured spectroscopically by two independent soft X–ray Bragg spectrometers. The vertical line–of–sight passes approximately at a normalised flux–surface radius (r/a) = 0.3 through the mid-plane. The O VIII (λ = 18.97 Å) and C VI (λ = 33.74 Å) Lyman–α line radiation is selected by first order Bragg reflection. For the oxygen line a potassium acid phthalate crystal

(d=13.29 Å) with a spectral resolution $\Delta\lambda/\lambda = 350$ is used and for the carbon line a PbSt crystal ($d = 50.2$ Å) with a spectral resolution of about 50. The radiation is detected by gas–flow proportional counters of a multistrip gaseous chamber type. The measured intensity is proportional to the product of four quantities: the line–integral of the emitted radiation times the impurity density in the H-like ionisation stage times the electron density (measured by DCN Interferometer) times an excitation rate coefficient which depends on the electron temperature (measured by ECE).[14] The corona equilibrium model (see, e.g., Chap. 3.5.1 of [636], [87, 598] and for futher discussion [169,458]), was used to estimate the concentration of C and O impurities. Because of the sensitive dependence on temperature and density, the line integral effectively reduces to a measurement near the location of maximum emitted intensity, which is $\rho_{pol} = 0.7$ for O and $\rho_{pol} = 0.8$ for C. An absolute calibration of the oxygen monitor was performed off-line using an X–ray tube, which led to an estimated accuracy of ±20%. The carbon monitor was 'indirectly calibrated' by using the C concentration obtained by charge–exchange spectroscopy and bremsstrahlung measurements as a 'silver standard'. This yielded an estimated accuracy of ±40%. From the data, it is easy to see that the maximally attainable density increases with decreasing plasma impurity content (C as well as O). However,it is not so easy to estimate in a correct way from the available data the separate influences from C and O on the density limit. From Fig. 7.3 (b) one can see that C and O are quite strongly correlated, the slope corresponding to approximately CO at medium concentrations. For low as well as for high concentrations, additional sources of carbon play a role.

Exercise 7.8. Estimate the density limit *nmax* as a function of the measured oxygen and carbon concentrations *dco* and *dcc*:
a) by applying simple OLS regression (each of the three variables on the other two);
b) by using principal components, taking the measurement errors into account, which are estimated to be approximately 2% for *nmax*, 20% for *dco* and 40% for *dcc*;
c) determine numerically a 95% confidence interval for α_2 as the interval where the null–hypothesis H_0: $a_2 = a_{2,0}$ is not rejected;
Hint: the null–hypothesis $\alpha_2 = 0$ can be tested by using a 'non-standard' F–statistic: $F = (SS_2 - SS_1)/SS_1$, where SS_2 is here the sum of squared *perpendicular* deviations to the two-dimensional regression plane and SS_1

[14] An earlier reference on calculated and measured rate coefficients of (helium-like) oxygen lines as a function of plasma temperature and density is [172]. The abbreviation ECE stands for electron cyclotron emission, an area of interest for high-temperature fusion reactors, see [350], as well as for plasma diagnostics, see [97] for an application to ASDEX. The subject has a thorough theoretical underpinning [45, 70] depending on the Larmor formula for the radiation of gyrating electrons [312].

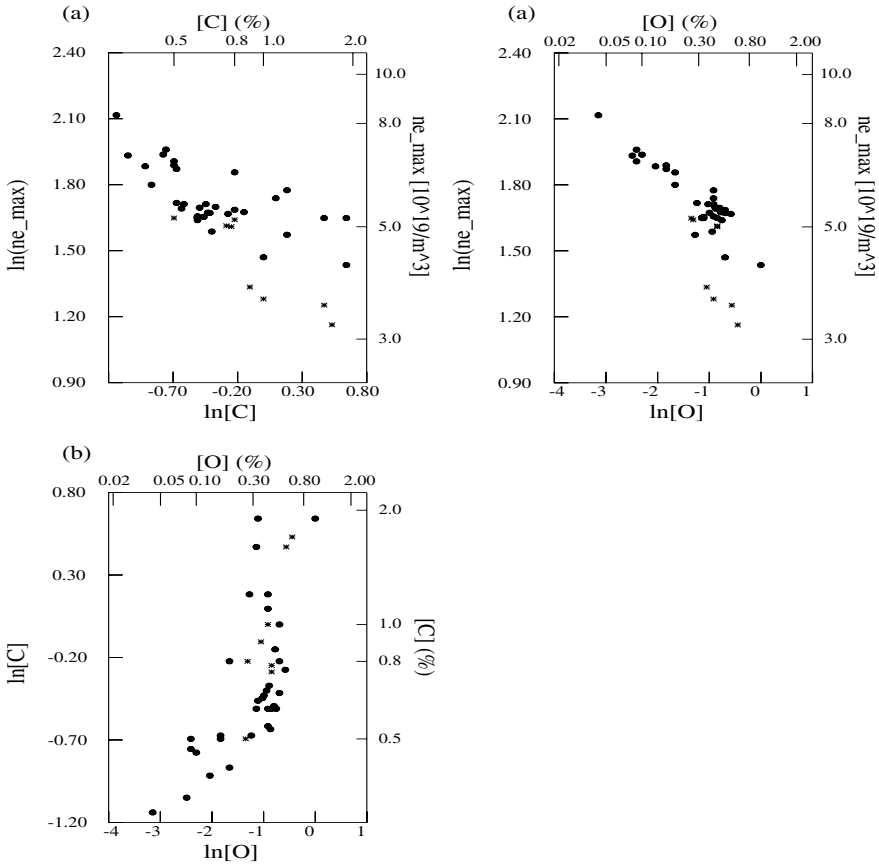

Fig. 7.3. Dataset dlim1: Attained maximum line-averaged electron density against carbon and oxygen concentration at ASDEX Upgrade. The dots correspond to the first group, and the stars to the second group of data.

the sum of deviations to the one-dimensional line $\alpha_0 + \alpha_1 \ln(dco)$. In [223] it is shown by simulation that also in this situation the distribution of F is well approximated by $n^{-1}F_{1,n}$, with n the number of data points.
d) multiply the available ranges of $\ln(dco)$ and $\ln(dcc)$ with the point estimates found at b) to obtain an estimate of the relative importance of C and O on the density limit, see Chap. 14.3 of [347], and [223];
e) check your findings qualitatively by performing visual inspection by using SAS Insight or X-gobi;
f) predict the density limit at AUG for carbon and oxygen concentrations equal to 0.1%.

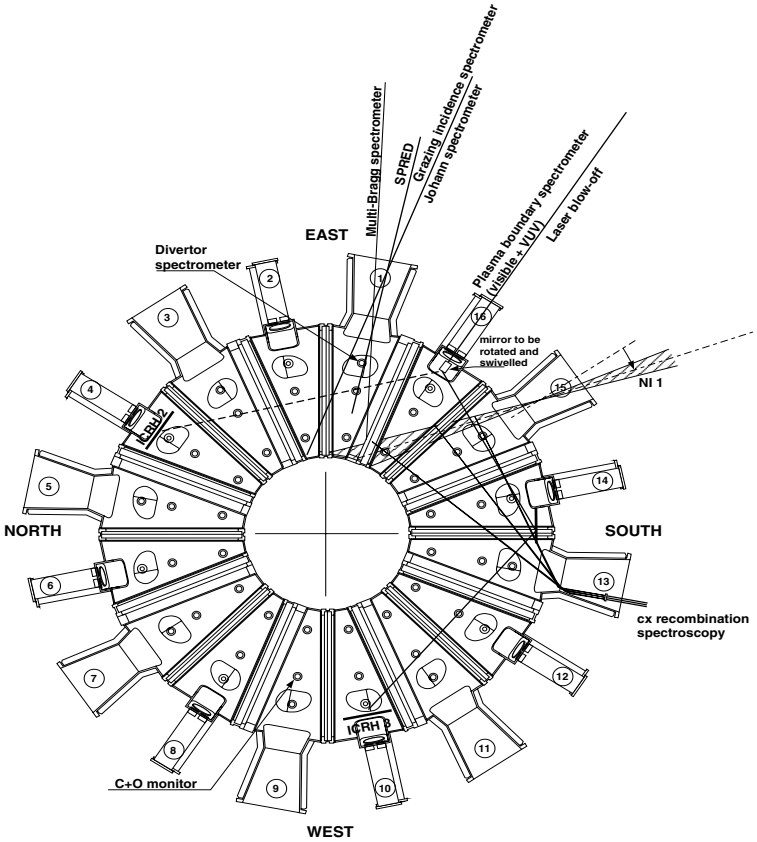

Fig. 7.4. Toroidal view of ASDEX Upgrade which shows the vacuum vessel containing the plasma and the lines–of–sight of several diagnostic systems used for plasma spectroscopy. Also the injection cone of one neutral beam system and the toroidal locations of two antennas for ICRH heating are displayed. Drawing reproduced by courtesy of IPP Garching.

7.5 Case IV: High-frequency MHD Oscillations (AUG)

Case IVa: ELM Data from Mirnov–Coils

Dataset elm_3703.dat (prepared by M. Maraschek)

Physical Motivation: Simply speaking, ELMs ('edge-localised modes') are relaxation oscillations of the plasma, localised near the plasma boundary and observed during high confinement ('H–mode') discharge phases, where the edge pressure profiles are steeper than during L–mode. They have some resemblance to sawtooth oscillations (which occur in the plasma centre) and in

some respects to minor disruptions. ELMy H–mode discharges are foreseen as a standard scenario for reactor-grade plasmas in next-step tokamak experiments such as ITER, see [350, 484, 615, 671]. From a technical viewpoint, their interesting effects are

(a) expulsion of particles and impurities from the plasma, thus preventing density accumulation and radiation collapse eminent in ELM-free H–mode;

(b) reduction of the energy confinement time, depending on the circumstances and the type of ELMs, by some 5% to 25 %;

(c) production of an intermittently peaked heat–load on the divertor.

The technical requirements of the divertor design to withstand this heat load increases with increasing size of the tokamak. A large variety of ELMs exists, which phenomenologically have been broadly classified into three groups: Type I, occurring at input power well above the threshold power for the transition from L–mode to H–mode. Typically, they have large amplitudes and their frequency increases with input power. Type III ELMs occur close to the L–H threshold power, usually have lower amplitude and their frequency decreases with increasing input power. While they produce less severe heat-load peaks to the divertor than do type I ELMs, type III ELMs are sometimes associated with a somewhat larger reduction in confinement time. In high density and highly shaped plasmas, type II ELMs can occur, sometimes in combination with type I ELMs. They are possibly related to a 'second regime' of MHD stability against ballooning modes, see [464]. Type II ELMs [254] possibly combine, for next step fusion experiments such as ITER, the attractive, complementary features of type I and type III ELMs, but have presently not been attained at low collisionalities foreseen in ITER. Models concerning ELMs are described in [117, 118, 309, 311, 657, 750]. According to conventional theory, they relate to the proximity of ideal MHD (notably type I ELMs) or resistive MHD (notably type III ELMs) limits of the pressure gradient near the plasma boundary, and are sometimes coupled with limit cycles of transitions from H–mode to L–mode or to (highly diffusive) M–mode [304, 309], where the abbreviation M stands for magnetic braiding. ELMs can be detected by measurement of the D_α light emitted by the plasma, by soft X–ray measurements and, as illustrated in this exercise, also magnetically by so-called Mirnov coils, which measure the time variation of the poloidal and radial magnetic field at various locations inside the vessel and outside the main plasma. Magnetic pick-up coils (also called field probes) have already been used in an early stage of plasma physics research, for instance in the theta pinch experiment ('Zeta') at Harwell, see [265]. They received their (now generic) name since, by signal correlation analysis, ('Mirnov type') instabilities were detected during the current ramp-up phase in the T–3 tokamak at Kurchatov Institute, see [468]. Type III ELMs exhibit clear magnetic precursors with toroidal mode numbers in the range 8–15. This feature can

be used to discriminate them from type I and type II ELMs, based on magnetic measurements only. Further background information about ELMs can be found in [117, 118, 306, 311, 657, 668, 671, 750].

Data Description: The dataset contains measured Mirnov–coil data from the neutral-beam heated discharge #3703 during the time interval t= 2.805–2.811 s, sampled with 500 kHz (N=3000 observations). It contains 28 time traces from C09_01 to C09_32, where C09_08 and C09_25 are omitted, because measurements were not available. The coils C09_13 and C09_19 are not installed. The locations of the coils are shown in Fig. 7.10. The positions of the coils were partly determined on the basis of the available space in the vessel and do not follow a precise regular pattern.

In the given time interval, one can see the occurrence of an ELM event and afterwards a strong $m = 1$ mode from a sawtooth precursor. The high-frequency fluctuations from the ELMs can be seen in all temporal eigenvectors from a singular-value decomposition analysis, whereas the low frequency modulation from the $m = 1$ mode appears only in the first two eigenvectors with the largest eigenvalues. In the subsequent eigenvectors one can see also the plasma movement caused by the ELM.

Exercise 7.9. Analyse the data with SAS/INSIGHT or with S–PLUS using multivariate regression methods and try to separate the different plasma activities: ELM, plasma movement, and $m = 1$ mode.

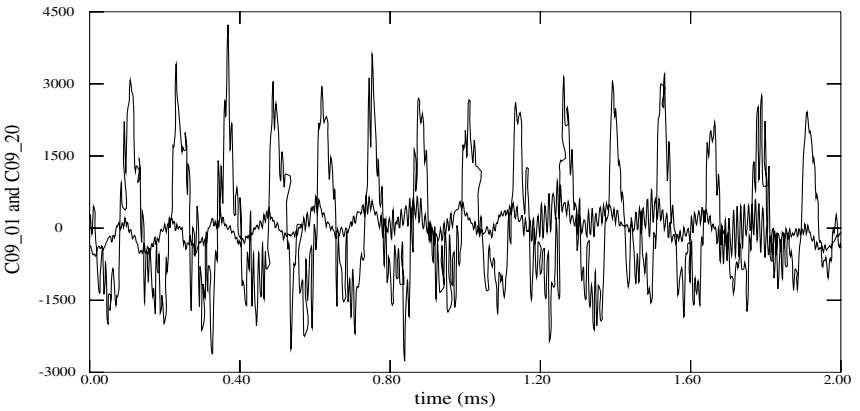

Fig. 7.5. Dataset elm_3703: Time–traces from two out of 28 Mirnov–coil measurements

Case IVb: Appearance of Toroidal Alfvén Eigenmodes (TAE) and an m=1 Mode during Neutral Beam Injection

Dataset tae_3703.dat (prepared by M. Maraschek)

Physical Motivation: In future fusion devices, fast α-particles are produced by fusion reactions. In addition, the various heating and current drive methods [388,729] engender fast plasma ions. Depending on the plasma current and density profiles, these fast particles can interact with resonant MHD modes within the plasma. One type of MHD mode with wave–particle interaction is the so-called fishbone, manifesting itself by low frequency (some 10 kHz) Mirnov oscillations, the graphical record of which resembles the bony framework of a fish, see Chap. 5.4 of [335] and Chap. 7.5 of [671]. Another class of such modes is the Alfvén type mode, a special species of which is the Toroidal Alfvén Eigenmode ('TAE'). TAE modes are of technical interest because they re-distribute and expel to a certain extent fast particles which are required to heat fusion-grade plasmas, see Chap. 4.2 of [671]. Toroidicity induces a 'gap' in the continuous Alfvén mode spectrum, somewhat similar to the Brillouin gap for electrons in solid state physics. Since discrete frequency TAEs are formed and undamped whenever this gap opened over the entire minor plasma radius, they are sometimes also called 'gap-modes'. The dispersion relation of the Alfvén continuum depends on the current and density profiles. The 'Brillouin gap' is located at $\omega_{TAE} = \mathrm{v}_A(q)/2qR$, where $\mathrm{v}_A \sim B/\sqrt{n_e}$ is the Alfvén velocity. For typical profiles, this leads to TAE frequencies in the range $\nu_{TAE} = 50 - 300$ kHz, where, for instance, also ion-acoustic waves can occur, see [105,320] and, for solitary solutions, [3,437,629,687]. TAE modes have been observed in various tokamaks in regimes with a large fraction of fast particles, see [185,745], but also in other regimes, where the excitation mechanisms are not yet fully understood, see [184,438].

Data Description: The dataset contains measured Mirnov–coil data from the neutral-beam heated discharge #03703 during the time interval t= 2.824–2.826 s, sampled with 500 kHz (N=1000 observations). For a description of the variables and the measurement locations, see Case IVa.

During the time interval presented in Fig. 7.6 one can see a strong $m = 1$ mode from a sawtooth precursor. During this $m = 1$ mode some bursts of TAE–mode activity appear. In the temporal eigenvectors one can see a separation between the high frequency fluctuations from the TAE–mode (approx. 100 kHz) and the low frequency $m = 1$ mode (10 kHz). The first two eigenvectors contain the low $m = 1$ mode, the next two eigenvectors represent noise from the divertor region and the next three eigenvectors contain the high frequency fluctuations from the TAE modes. An FFT analysis of the TAE mode part shows the typical structure of fingers in the frequency domain.

Exercise 7.10. (*) Try to separate the above mentioned three types of plasma activity using singular-value decomposition and/or Fourier analysis.

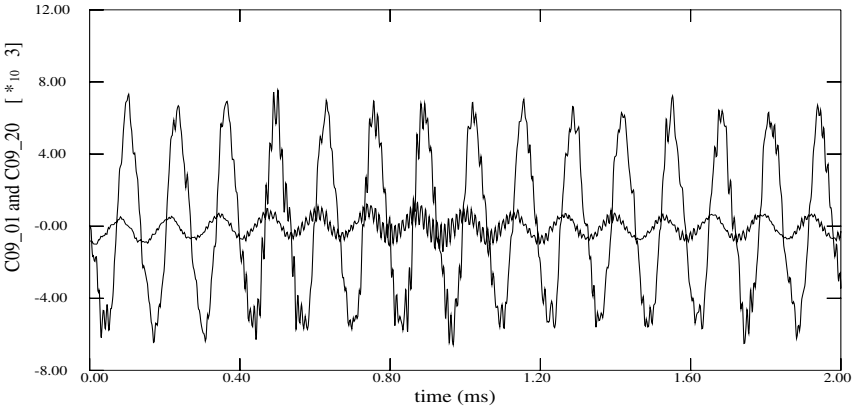

Fig. 7.6. Dataset tae_3703: Time–traces from two out of 28 Mirnov–coil measurements

7.6 Case V: Heat Flux Profiles at the Outer Divertor Plate (AUG)

Dataset: divheat2.dat (prepared by A. Herrmann)

Physical Motivation: Under stationary discharge conditions, the total energy per unit time produced by the plasma, through Ohmic and alpha–particle heating and introduced by the auxiliary heating systems, has to be dissipated at the same rate. Basically, this heat dissipation occurs (a) by electromagnetic radiation, and (b) by absorption of (neutral as well as ionised) hot particles to the vessel and to the divertor plates. Quantitative experimental investigation of the heat load to the divertor is interesting (i) from a technical perspective: to estimate, by appropriate scaling from different devices, whether the divertor material in large next-step devices, such as ITER, can stand the anticipated (stationary and transient) heat load, in standard operation and under disruptive events, and (ii) from a scientific viewpoint: to test empirically various models describing heat transport across the separatrix and the scrape-off layer, where the magnetic field lines are not closed and normally intersect with the divertor plates, see Fig. 7.10. Naturally, the two objectives are somewhat intertwined. For scrape-off layer physics, we refer the reader to [636], for aspects of the ITER–FEAT divertor, and plasma facing materials, to [44, 187, 316, 406, 445] and for the W7–X divertor concept to [549]. The first proposal of a divertor to reduce the plasma impurity content is attributed to L. Spitzer, see [92]. A practical overview of divertor heat-load physics is given in Chap. 4 of [671] and [274].

Data Description: The dataset is an update of divheat1.dat, released previously, and contains two characteristic parameters (λ and $q_{m,plate}$) of the heat–flux profiles at the outer divertor plate of ASDEX Upgrade, which were measured by the thermography diagnostic, as response variables and a few basic discharge parameters (the line-averaged density and the safety factor) as regression variables. The plasma current and magnetic field are also provided. The measured total power on the divertor plate, P_{plate}, is either a regression variable or a response variable.

The variables in the dataset are:
shot: shot number
$\overline{n}_e$: line averaged density measured by the DCN–interferometer [m^{-3}]
I_p: total plasma current [A]
B_t: toroidal magnetic field on axis [T]
q_{95}: safety factor at 95% of the normalised poloidal magnetic flux
λ (= *lambda:*) e-folding length of the heat–flux profile at the divertor target plate [mm]
qm_plate: maximum heat flux at the outer target plate [W/m^2]
P_plate: total power to the outer target plate [W]

Table 7.7. The first three and last two observations in the file divheat2.dat are:

shot	n_e	Ip	Bt	q95	lambda	qm_plate	P_plate
5975	7.96E+19	1200000	-2.00	2.71	28.6	1860000	2360000
5976	7.49E+19	1200000	-2.00	2.71	27.8	1910000	2470000
6011	8.86E+19	1010000	-2.00	3.24	47.2	2340000	2010000
...							
6074	5.69E+19	1000000	2.49	4.03	49.6	1950000	1270000
6075	5.86E+19	1010000	2.49	3.98	62.9	3270000	2700000

The heat flux profile on the divertor plates is estimated by measuring the infra-red radiation emitted by the divertor surface using a high-resolution infrared line camera. The spatially resolved heat flux to the target is calculated from the temporal evolution of the measured surface temperature by numerically solving the 2-D non-linear heat conduction equation with the so-called THEODOR code. The maximum of the heat flux is stored as $q_{m,plate}$ and the spatially integrated heat flux as P_{plate}. The spatial distribution is highly asymmetric and can be roughly approximated by an exponential shape. The corresponding spatial decay length is stored as λ. The total heat flux is compared with calorimetric measurements of the cooling water of the divertor plates. Further details can be found in [274, 275], and in Appendix 4 of [594].

Exercise 7.11. a) Find a scaling of the maximum heat flux, $q_{m,plate}$, and of the decay length, λ, as a function of $\overline{n}_e$, q_{95} and P_{plate} using a simple

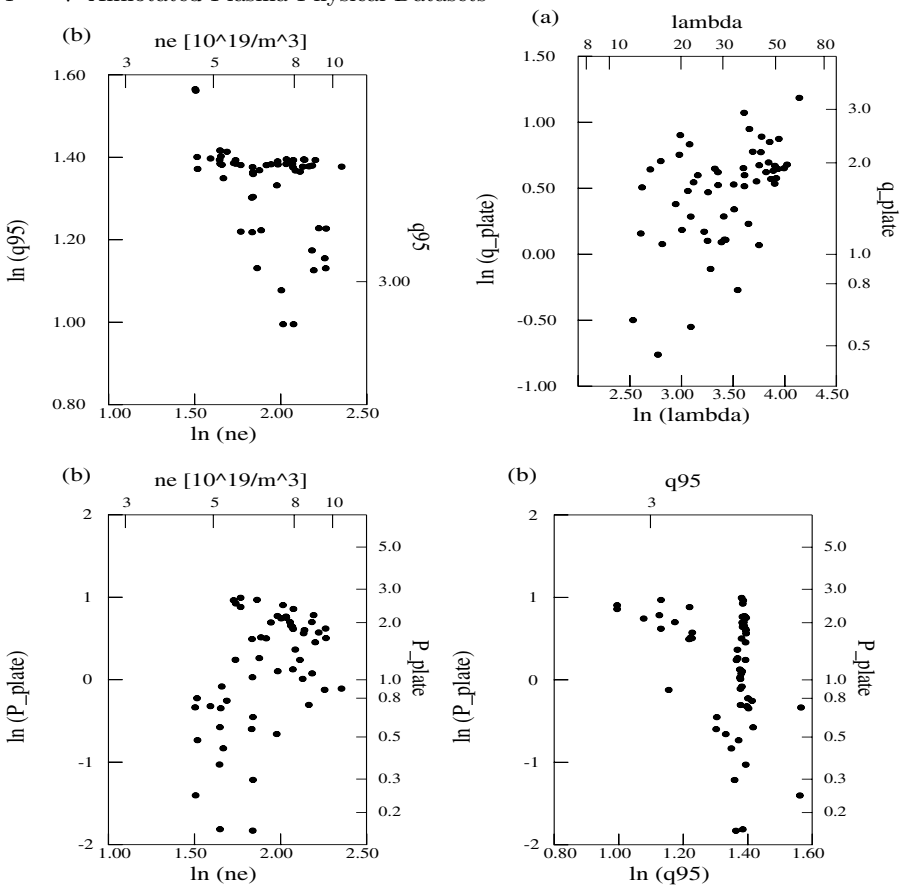

Fig. 7.7. Dataset divheat1.dat: (*a*) response variables: maximum heat flux at the outer divertor plate and the spatial decay length of the heat flux profile; (*b*) regression variables: edge safety factor (q_{95}), electron density ($\overline{n}_e$), and total power (P_{plate}) to the outer divertor plate

power–law ansatz; Investigate whether there is sufficient evidence that the current, or the (absolute value of the) magnetic field, is playing an additional role.

b) the measurement errors in P_{plate} and $q_{m,plate}$ are independent and both about 15%. How does this affect the power–law scaling?

c) the sign of the magnetic field in the last three discharges is different from those of the other discharges. This indicates that plasma has been operated with a 'reversed magnetic field'. Investigate whether these three discharges appear to scale somewhat differently from the rest. How is the power–law scaling affected when these three discharges are omitted from the analysis?

d) are there indications of a significant systematic deviation from a simple power–law dependence?
e) show that canonical correlation analysis gives a physically well-interpretable result: to a conspicuously good approximation, the first canonical vector of the two response variables is proportional to $\ln(\lambda \times q_{m,plate})$ and the second vector to $\ln(\lambda/q_{m,plate})$, which can be interpreted as the total heat load and the broadness/peakedness of the heat–deposition profile, respectively;
f) show quantitatively that both a high electron density and a high value of q_{95} tend to broaden the heat deposition profile, albeit in a different manner: the density mainly by reducing the maximum heat load and q_{95} by enlarging the spatial decay length;
g) to first order, $\ln(\lambda \times q_{m,plate})$ should depend linearly on P_{plate}. Test whether the deviation from linearity is statistically significant. Remark. Deviation from a linear dependence indicates that the heat deposition profile is, over the full heat-load range, not accurately described by exponential profiles with $q(x, y) = q_{m,plate} f(y) e^{-x/\lambda}$, where $f(y)$ is an arbitrary function describing the toroidal dependence.

7.7 Case VI: Density Profiles of Neutral Particles (AUG)

Data set: neutral1.dat (prepared by H.-U. Fahrbach and J. Stober).

Physical Motivation: Separate determination of the electron and ion temperature profiles in fusion plasmas is interesting from both a theoretical and practical outlook. Over a number of years, a variety of theory-based and semi-empirical models, with varying complexity, have been developed that try to capture quantitatively the main mechanisms of plasma heat transport ('confinement') and of the processes determining the boundaries of the plasma parameter regions that are accessible to operate the machine ('operational limits'). A few single key elements in such theories are ion temperature gradients, electron or ion drift waves, see [387] for an experimental set-up, as well as pressure driven ballooning modes near the plasma edge, and magnetic island formation near resonant surfaces, where the local safety factor, for a circular plasma $q(r) \sim \frac{B_t}{R} \frac{r}{B_p}$ $(0 < r < a)$, assumes 'bantam' fractional-integer values (1, 3/2, 4/3, 2, 5/2, 3, 4, ...). As usual, B_t denotes the toroidal and B_p the poloidal magnetic field, R the major radius and a the minor radius of the plasma. The ion–temperature profile prediction of these models varies substantially, as is illustrated in Fig. 23 of Chap. 2 in [671]. Hence, accurate measurement of the ion temperature profile is useful for discriminating between these models. A distinction is usually made between the central plasma

('sawtooth') region, an intermediate ('conductive heat transport') region and the plasma edge region, which makes some sense from both the perspective of the physical processes occurring in the plasma and that of the experimental accuracy attainable by a given diagnostic. From a more technical point of view, by applying PRETOR code transport calculations, based on a semi-empirical 'RLW' model, see [544], it has been estimated that the fusion output power in ITER–FEAT may vary between some 400 MW and 600 MW, if the ratio between the ion and electron diffusivity varies between 0.5 and 2, all other aspects being equal. In large tokamaks, such as ITER–FEAT, measurement of the alpha–particle temperature profile is of direct theoretical and practical importance. The neutral-particle analysis system can be used to measure not only the ion temperature, but also the relative concentration of various ions (hydrogen, deuterium, tritium, helium), and even, with some effort, the neutral density profile along some part of the line–of–sight. The latter plays a role in several theories about H–mode sustainment and suppression. The present exercise concentrates on fitting parametric profiles to the neutral density along the line–of–sight. We do not treat here the more extensive physical problem of how these profiles, which depend in fact on three spatial coordinates since the neutrals are not confined by the magnetic field, are affected by various plasma parameters and wall conditions.

Data Description: The dataset contains neutral density profiles as a function of the poloidal radius ρ for a particular line–of–sight near the horizontal mid-plane of the plasma. The neutral density profile (as a function of $1-\rho$) can be viewed as a survival function. The problem is here to find a parametric class of survival functions (in terms of $1-\rho$) that approximately fit these five cases. The curves presented in Fig. 7.8 were obtained by fitting the edge value $\ln n_{n,w}(1)$ and the inverse fall-off length k_n in

$$\ln n_{n,w}(\rho) = \ln n_{n,w}(1)\big(1 + k_n(1-\rho)\big)^{-1/3} \tag{7.3}$$

to the density profiles calculated by EIRENE simulations [548], while using regression weights proportional to $\big(\ln n_{n,w}(\rho)\big)^3$.

The variables in the dataset are:
shot: shot number
phase: plasma discharge phase (Ohmic, H–mode, CDH–mode)
ρ (= *rho*): normalised poloidal radius
pda: neutral density (particles per m^3)

Experimental Description: The ion temperature can be measured either by spectroscopy ('CXRS'), which yields the temperature of selected impurities from Doppler broadening, by measuring the neutron yield from fusion reactions,

Table 7.8. The first three and last two observations in the file neutral1.dat are:

shot	phase	rho	pda
7540	ohmic	0.0475	19490
7540	ohmic	0.1445	47160
7540	ohmic	0.2436	84540
...			
8002	ohmic	1.042	2079000000
8002	ohmic	1.047	2170000000

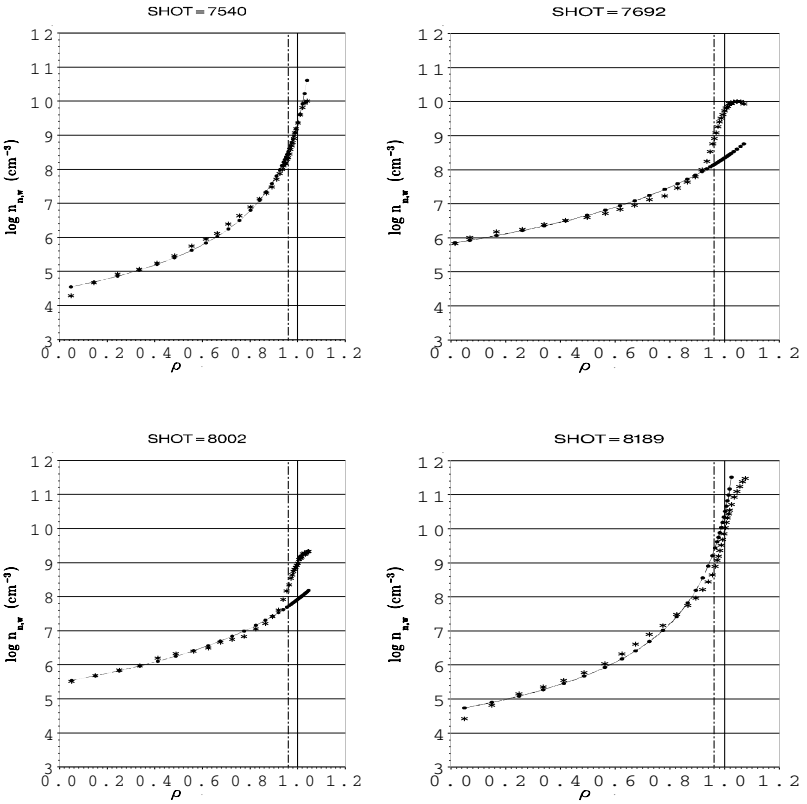

Fig. 7.8. Dataset neutral1.dat: Density of neutrals as a function of the normalised radius in the midplane

cape from the plasma ('NPA'). Experimental data from the latter type of diagnostic are presented here. Fast neutrals from the plasma center, born by

charge–exchange or 'volume recombination', are not confined by the magnetic field. However, they risk becoming ionised again by various processes, one of them being charge exchange ('CX'), which produces other neutrals and so on. Eventually, some neutrals escape from the plasma. Hence, the neutral particle flux is a rather complicated convolution of the temperature and density of neutrals, ions and electrons as well as of the cross–sections for charge–exchange and ionisation, see [151, Ch. 4]. In a simple approximation, the slope of the logarithmic neutral flux plotted against the energy describes, at sufficiently high energies, the temperature near the plasma center. For a more accurate evaluation, especially needed for high density plasmas in medium size and large devices, a rather sophisticated code ('CENS', analysing charge–exchange neutral spectra) was developed during a Volkswagen grant contract [383]. The standard option of the CENS code uses as input variables the neutral spectra from neutral particle analysis (NPA), the measured profiles of electron density and temperature, e.g., from DCN or YAG–laser and from electron cyclotron emission (ECE) or YAG laser, respectively. The CENS code can provide simultaneous estimates, with a suitable error propagation, of the ion temperature profile and of the neutral density profile, the latter along the line of sight. Under the hypothesis of a known ion temperature, e.g., as estimated from another ion temperature diagnostic, or, in the case of high density, where T_i is close to T_e, from an electron temperature diagnostic, the CX measurements can be used to estimate the neutral density profile, along the line of sight, with higher accuracy. In the present version of the code, the validity of the neutral profile is restricted to $\rho < 0.95$. It could be extended by using measurements of another diagnostic (LENA), which is especially adapted to low-energy neutrals, see [702]. This method is based on measuring the particle velocity by time–of–flight instead of by deflection in a magnetic field, and can hence not distinguish between the various isotopes. The 'observed' neutral densities in this exercise have been determined by a 3-D EIRENE Monte Carlo Code, which makes use of NPA and LENA measurements and, in addition, of H_α and neutral pressure measurements, see [648]. While a more flexible model would be required to describe the neutral density near the edge, the class of neutral density profile fits in the exercise, expected to be suitable in the region $0 < \rho < 0.95$, is presently used as an option for (parametric) neutral density and ion temperature estimation in CENS. For further information, see [180] and Appendix 9 of [594].

Exercise 7.12. Fit $\ln(pda)$ as a function of ρ for each profile separately, using
a) a simple exponential model;
b) the sum of two exponentials;
c) a generalised linear model which considers the expectation value of $\log(pda)$ to the power -3 as a linear function of ρ;
d) your own favourite (lower dimensional) parametric model.
(*) Compare the goodness-of-fit in these three cases.

7.8 Case VII: Plasma Parameter Recovery from External Magnetic Measurements (AUG)

Data set: magn1.dat (prepared by P.J. McCarthy).

Physical Motivation: An array of magnetic flux–loops and field probes to measure the magnetic field outside plasma is routinely installed in many tokamaks. They enable, on the basis of classical electrodynamics, one to estimate various plasma parameters such as the elongation κ, $\beta_p = W_{th}/\langle B_p \rangle^2$, and the plasma inductivity l_i. The latter parameter can be viewed as a normalised moment of the poloidal magnetic field or, by Ampère's law, the current density profile (see [729], and [83]), a peaked current profile corresponding to a high value of l_i. For circular plasmas, the Shafranov shift $l_i/2 + \beta_p$ can be recovered from the external magnetic measurements, but not l_i and β_p separately. In practice, recovery of l_i becomes increasingly difficult for low values of the elongation κ. In statistical terminology, this phenomenon is called an identifiability problem.

Data description: From a dataset of $N = 1200$ lower X–point equilibria (where the separatrix contains a lower X–point inside the vessel, see Fig. 7.10) three basic plasma parameters and 20 principal components of the correlation matrix of the 58 (calculated) magnetic measurements from (32 tangential and 8 normal) magnetic field probes and 16 flux differences from 17 poloidal flux loops, as well as 2 saddle loop fluxes are given, see Fig. 7.10. The principal components have been taken from the measurements as calculated from an equilibrium code, which should rather well reflect the actual plasma behaviour.

The variables in the dataset are:
r_{in}: absolute (horizontal) position of the innermost point on the separatrix [m]
r_{out}: absolute (horizontal) position of the outermost point on the separatrix [m]
rz_{min}: absolute (horizontal) position of the lowest point on the separatrix [m]
z_{min}: absolute (vertical) position of the lowest point on the separatrix [m]
rz_{max}: absolute (horizontal) position of the highest point on the separatrix [m]
z_{max}: absolute (vertical) position of the highest point on the separatrix [m]
β_p: ratio between the thermal energy and the energy stored in the poloidal magnetic field
l_i: (normalised) internal plasma inductance (ratio of the volume-averaged to the surface-averaged density of the poloidal magnetic field energy)
κ: plasma elongation (the ratio between vertical and horizontal plasma minor radius)
pc_1–pc_{20}: the first 20 principal components of the calculated magnetic measurements
Remark. The absolute horizontal ('R') positions are measured with respect

to the axis of symmetry of the torus, whereas the vertical positions ('z') are measured with respect to the mid-plane of the torus.

Table 7.9. The first two and the last observation in the file magn1.dat are:

rin	rout	rzmin	zmin	rzmax	zmax	bp	li	k	
pc1	pc2	pc3	pc4	pc5	pc6	pc7	pc8	pc9	pc10
pc11	pc12	pc13	pc14	pc15	pc16	pc17	pc18	pc19	pc20
1.0456	2.1532	1.4400	-0.6979	1.5310	0.5863	0.7139	0.5760	1.1531	
-219	-1202	-5581	1180	-1941	641	637	404	-1113	304
26	-288	-86	-23	-7	-449	-93	-14	135	-52
1.0431	2.0015	1.4857	-0.5333	1.5130	0.5588	0.5994	0.8486	1.1384	
2757	-962	-3255	-4173	1583	-2612	-1357	896	-134	839
294	-459	-433	446	-165	194	122	-65	-64	-58
...									
1.0495	2.1528	1.3800	-0.8828	1.5450	0.5769	0.6932	1.5003	1.2957	
-343	3057	-2884	2266	-1061	1625	31	174	-379	295
-1075	-607	375	-235	349	-310	321	24	-147	-94

Experimental description: Actual plasma parameter recovery is based on solving an inverse problem: The tangential and normal components of the magnetic field, as well as the differences in poloidal flux, at any location outside the plasma are calculated for a set of given plasma equilibria, which are obtained by solving the (weakly non-linear) elliptic Grad–Shafranov equation for a specified class of plasma current and pressure profiles.
In practice, to calculate the equilibria the Garching equilibrium code has been used, described in [409] and extended by McCarthy et al. In agreement with the properties of the Grad–Shafranov equation, the current profile is decomposed into two source terms (one, related to the plasma pressure gradient, proportional to R, and the other one, related to the poloidal current, proportional to $1/R$), with six free shaping parameters each. The second part of this approach consists in recovering, on a fast time scale, from the actual measurements of the magnetic field and the poloidal flux (at discrete positions outside the plasma) the basic 'geometric' plasma parameters related to the current and pressure profile. The method, sometimes called 'function parameterisation', has been introduced to plasma physics through B. Braams's stay at CERN summer school (1973/1974), where H. Wind et al. had developed such an approach for analysing spark chamber data in elementary particle physics, see [423, 737, 738]. As described in [83, 449], this is done by performing multivariate 'inverse' regression analysis on a database of several thousand calculated equilibria. The regression response surface is used for a 'fast on-line equilibrium reconstruction' based on the magnetic field and poloidal flux measurements. The accuracy of the actual measurements is of the order of 1% of the 'standard deviation', as determined from the range

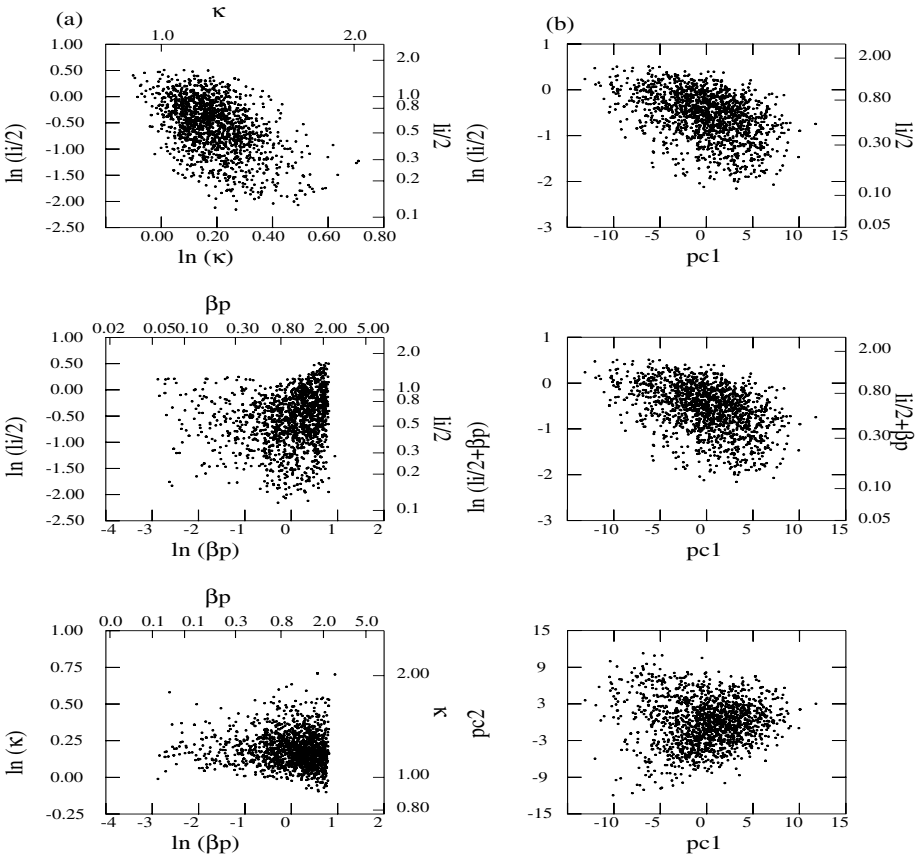

Fig. 7.9. Dataset magn1.dat: (*a*) physical plasma parameters: elongation, κ, inductance, l_i, and poloidal beta, $\beta_p = W_{th}/\int B_p^2 dV$; (*b*) the first two principal components of the magnetic measurements outside the plasma

over the simulated database. In practice, plasma equilibrium reconstruction is often complemented by additional empirical input from (electron) pressure profile measurements (YAG–laser diagnostic), local poloidal magnetic field measurements (MSE diagnostic) and global plasma energy measurements (diamagnetic loop, kinetic energy determination).[15] For further details of this area of integrated plasma equilibrium recovery, the reader is referred to the description of the interpretative code CLISTE [450].

[15] As described in [103, 213, 460], MSE measurements permitted the establishment of a central region with nearly zero toroidal current in certain internal barrier discharges, notwithstanding any contribution of the poloidal current.

Exercise 7.13. a) Perform a simple regression of $l_i/2$ against pc_1 to pc_{20}, to the entire dataset and to the two subsets which satisfy $\kappa < 1.4$ and $\kappa > 1.4$, respectively. Notice that for these two cases, a marked difference in RMSE, and in R^2 exists. (As usual, R^2 stands for the squared multiple correlation coefficient between the values of l_i in the database and its predictions from the measurements.)
b) Perform quadratic regression analysis, e.g., by SAS proc RSREG, and assess approximately whether the quadratic components are important for the regression.
c) Applying stepwise regression by ordinary least squares on $pc_1, \ldots, pc_{20}$, using $\alpha_{enter} = 0.15$ and $\alpha_{remove} = 0.001$, show that $R^2 = 0.85$ is reached by some 9 linear and 28 quadratic and cross-terms, each of which has a value $SL = \hat{\alpha}_j^2 / s_{\hat{\alpha}_j}^2$ of at least 10. Using only the linear terms $pc_1, \ldots, pc_{20}$ a value $R^2 = 0.7$ is obtained, which is markedly lower. Repeat the regression by applying weighted least squares with weights κ^{-p} $(p = 1, 2)$.
d) Establish by canonical correlation analysis of (β_p, l_i) against either $pc_1, \ldots, pc_{20}$ or against a selection of linear and higher order terms you make that (i) the first canonical combination of β_p and l_i which can very well be recovered by linear regression is proportional to $\beta_p + l_i/2$, and (ii) that the second linear combination of β_p and l_i, orthogonal to $\beta_p + l_i/2$, can be relatively well recovered for 'medium and high kappa' $(\kappa > 1.4)$, and quite poorly for 'low kappa' $(\kappa < 1.225)$.

Remark. In this situation, the response variable as calculated without any random error, and the regression variables contain some calibration error. The reader is referred to [449] for an approximate way to take this particular aspect into account.

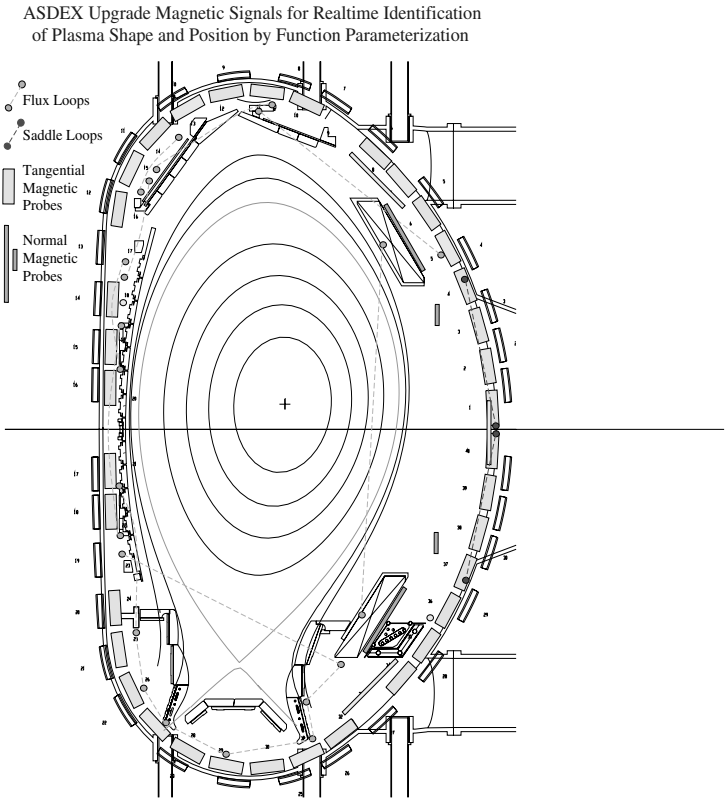

Fig. 7.10. Cross-section of ASDEX Upgrade with magnetic flux–surface contours and the locations of the magnetic measurements (magnetic coils and flux–loops). Inside the separatrix, the magnetic flux–surfaces are closed. The X–point of the separatrix is located close to the divertor near the bottom of the vessel. There is not an X–point inside the vessel near the top of the machine. Hence, a cross-section is shown for a single null (SN) discharge. The divertor at the bottom, named type-II or 'lyra' divertor, is of a closed type since the neutral gas is contained in a divertor chamber and cannot easily flow back into the main plasma. The divertor at the top is open, which permits a larger plasma volume in case of a bottom SN discharge, and has the same structure as in the type-I divertor operation period, during which also the bottom divertor was of an open type, see [492]. Drawing reproduced by courtesy of IPP Garching.

Some Annotated Statistical Literature

The list of books below is a very limited and subjective selection of statistical literature stemming from the last half of the twentieth century, made by the author on occasion of a course in elementary probability theory and statistics at the Max Planck Institute for Plasma Physics in Garching, and for that purpose provided with some brief motivating remarks. The reader should attach no meaning to the fact that many other important and valuable books in statistics are not included. Except for a classical reference and a popular paperback on Bayesian statistics, as well as the book with statistical biographies from Johnson and Kotz, the list below has not been expanded since, and the (more or less) random order has been maintained. Entries have been updated solely in as far as a new edition of (essentially) the same book has appeared. No reference is made to collective works, such as the Encyclopedia of Statistical Sciences (Vols. 1–9, Wiley, 1982, update Vols. 1–3, and supplement, 1992–2003), edited by Kotz, Johnson et al., and the Handbook of Statistics (Vols. 1–22, Elsevier, from 1980 onwards), edited by Rao, Krishnaiah et al.

---

(1) Rozanov, Y.A., *Probability Theory, a concise Course.*, Dover Publ. Inc. New York (1977).
Translated from Russian by R.A. Silverman.
Brief and clear. With exercises. Undergraduate level. (148 p.)

(2) Van Kampen, N.G., *Stochastic Processes in Physics and Chemistry,* North-Holland, Amsterdam (1985).
The first chapter is on probability theory, the second one on stochastic processes. There are 14 chapters in total, mainly on applications in various branches of physics. Very well written, but 'on graduate level'. (419 p.)

(3) Girault, M., *Calcul des Probabilités, en vue des Applications,* Dunod, Paris (1972).
Also clearly written; less concise than (1) and (2), undergraduate level 'pour les grandes Ecoles'. Emphasis on basic notions, and 'gentle' applications. (223 p.)

(4) Feller, W., *An Introduction to Probability Theory and its Applications,* Vol. I, II, Wiley, New York (1970).
A classic on probability theory, including random walks, diffusion problems and Markov chains. ($\simeq$ 1000 p.)

(5) Popper, K.R., *The Logic of Scientific Discovery,* Hutchinson, London (1972).
The first edition, *Logik der Forschung*, appeared in 1935, the tenth

German edition in 1994. A standard philosophical work on the possibility of justification of theories in the light of the data. Based on probability theory and degrees of corroboration, while not (yet) statistically inclined. The appendices contain an interesting alternative to the Kolmogorov axioms for conditional probabilities. (480 p.)

(6) Stegmüller, W., *Jenseits von Popper und Carnap: Die logischen Grundlagen des statistischen Schließens*, Springer–Verlag, New York, Berlin, Heidelberg (1973).
Philosophical recovery of part of the statistical gap left by Popper. (275 p.)

(7) Ledermann, W. & Lloyd, E. (editors), *Handbook of Applicable Mathematics*, Vol. II *Probability*, Wiley, New York (1980).
Extensive reference, undergraduate level. (440 p.)

(8) Letac, G., *Problèmes de Probabilité*, Collection SUP, Vol. 6, Presses universitaires de France, Paris (1970).
Contains 74 exercises - with solutions. Small, but advanced. ('2^{me} - 3^{me} cycle'). (118 p.)

(9) Barra, J.R. & Baille, A., *Problèmes de Statistique Mathématique*, Dunod, Paris (1969).
A collection of extensively worked exercises in (elementary) probability theory and traditional mathematical statistics (estimation theory, hypothesis testing), including examination problems from French and Swiss Universities. ('2^{me} cycle'). (321 p.)

(10) Székely, G.J., *Paradoxa, klassische und neue Überraschungen aus Wahrscheinlichkeitsrechnung und Mathematischer Statistik*, Harri Deutsch (1990).
Paradoxes (used in a broad sense: 'unexpected results') with their solution. With nice historical remarks and anecdotes about the origin of the problems. (240 p.)

(11) Burington, R.S. & D.C., *Handbook of Probability and Statistics*, with tables. McGraw-Hill, New York (1970).
Clear compendium, not too large, with precisely the amount of additional information needed to be intelligible. (462 p.)

(12) Graf, U., Henning, H.J., Stange, K. & Wilrich, P.-Th., *Formeln und Tabellen der angewandten mathematischen Statistik*, Springer–Verlag, New York, Berlin, Heidelberg (1987).
A thoroughly written standard work (1st edition 1954).
More extensive than Burington, with emphasis on statistics. (529 p.)

(13) Ledermann, W. & Lloyd E. (editors), *Handbook of applicable Mathematics*, Vol. VI A & VI B, *Statistics*, Wiley, New York (1984).
A good overview of statistics. 'English style' (clear, practice oriented). Very readable, despite its size. ($\simeq$ 1000 p.)

(14) Ferguson, T.S., *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, San Diego (1967).
A classical and well-written textbook on mathematical statistics, including the Bayesian contribution to statistical optimality theory. Graduate level. Especially suitable as a 'second course' in fundamental statistics. (396 p.)

(15) Kaufmann, A., *The Science of Decision-making*, World University Library, London (1968).
A broad overview of the (classical) area of decision theory. Written in a colloquial style, not using more than (modern) high-school mathematics. (253 p.)

(16) Fuller, W.A., *Measurement Error Models*, Wiley, New York (1987).
A clear, useful, and advanced book on a complex subject-matter: How to perform (also non-linear) regression if the regression variables are contaminated with random errors. (440 p.)

(17) Hillegers, L.T.M.E., *The Estimation of Parameters in Functional Relationship Models*, Ph.D. Thesis, University of Eindhoven (1986).
Treating the same area as Fuller, but presenting different numerical algorithms, and oriented towards non-linear models. Complemented with a worked example on phase separation in polymers. High level but very clearly written. (174 p.)

(18) Vincze, I., *Mathematische Statistik mit industriellen Anwendungen*, Band 1, 2, B.I. Wissenschaftsverlag, Mannheim (1984).
For the practitioner. Few geometrical interpretations. (502 p.)

(19) Cox, D.R. & Hinkley D.V., *Theoretical Statistics*, Chapmann & Hall, London (1974).
Complicated concepts are explained in a didactical and precise colloquial style without emphasis on mathematical detail, somewhat in the vein of van Kampen, but now on statistics. Graduate level (with undergraduate mathematics). With many examples and exercises, and an accompanying solutions manual. (511 p. and 139 p.)

(20) Box, G.E.P. & Tiao, G.C. *Bayesian Inference in Statistical Analysis*, Wiley Classics Library (1973, 1992).
An insightful book explaining carefully the Bayesian approach to statistics. It concentrates on non-informative priors and the relation with classical statistics based on frequentist (physical) probabilities. Hierarchical models are treated in detail, before their flowering in the computer age. Advanced undergraduate/graduate level. (608 p.)

(21) Kendall, M., Stuart, A. & Ord, K., *The Advanced Theory of Statistics*, Vol. I, II, III, Griffin and Co, London (1987).
An extensive and classic work. A number of editions since 1946. Large parts written in M. Kendall's didactical style. A valuable source of reference, even in this computerised age. ($\simeq$ 2000 p.) The new edition,

published by Edward Arnold 1994-1998, is called Kendall's Advanced Theory of Statistics and includes contributions by S. Arnold and A. O'Hagan.

(22) Rasch, D., *Mathematische Statistik*, Ambrosius Barth, Leipzig (1995). Quite extensive in depth and scope. Probability theory, statistical estimation theory, (non-)linear regression and design of experiments, all written in a thorough German biometrical tradition, which does not leave the details to statistical software packages. (851 p.)

(23) Bates, D.M. & Watts, D.G., *Nonlinear Regression Analysis and its Applications*, Wiley, New York (1988). A clear book about the geometry, the ideas behind some common algorithms, and the computational and graphical aspects of least squares methods for non-linear regression and compartment models (the last ones represent systems of coupled differential equations) with worked examples from chemical datasets. (365 p.)

(24) Stanford, J.L. & Vardeman, S.B. (editors), *Statistical Methods for Physical Science*, Academic Press, San Diego (1994).
A tutorial overview on probability, classical statistics, random as well as spatial processes, and time series analysis in 17 chapters, each written by a different author or group of authors, and often illustrated by practical data. Written in a uniform and colloquial style, well accessible to experimental physicists. (542 p.)

(25) Mardia, K.V., Kent, J.T. & Bibby, J.M., *Multivariate Analysis*, Academic Press, London (1979).
A very good book on multivariate analysis, with motivations, examples, geometrical interpretations. Not verbose, but a pleasant mathematical style. Advanced undergraduate - beginning graduate level. (518 p.)

(26) Lehmann, E.L., *Nonparametrics: Statistical Methods based on Ranks*, McGraw Hill, New York (1973), revised edition Prentice-Hall (1998).
Simple mathematics. Nevertheless intermediate level, between, e.g., Siegel's 'Nonparametric statistics for the social sciences' (1960,1980), and Hajek & Sidak's 'Theory of Rank Tests' (1967,1999). (457 p.)

(27) Lehmann, E.L., *Elements of Large Sample Theory*, Springer–Verlag, New York (1999). (630 p.);
Lehmann, E.L., *Theory of Point Estimation*, Wiley, New York (1983), (506 p.), second edition jointly with G. Casella, Springer–Verlag, New York (1998). (589 p.);
Lehmann, E.L., *Testing Statistical Hypotheses*, Wiley, New York (1959), second edition: Wiley, New York (1986), and Springer–Verlag, New York (1997). (600 p.)
Three classics in the area, the last book since its first edition in 1959. Didactically written, advanced undergraduate – graduate level.

Worked solutions by Kallenberg et al. (1987), CWI syllabus 3, Amsterdam. Statistical Science, Vol. 12 (1997) No.1, contains a reflection by the author.

(28) Rao, C.R., *Linear Statistical Inference and Its Application*, Wiley, New York (1973, 2002).
A standard textbook of (application oriented) mathematical statistics at the graduate level, covering the full classical area of linear statistical inference, with a didactical inclination to mathematical and conceptual detail. (656 p.)

(29) Hinkley, D.V., Reid, N. & Snell, E.J. (editors), *Statistical Theory and Modelling: in Honour of Sir David Cox, FRS*, Chapman and Hall (1991).
A varied and didactical overview explaining the status of many areas in statistics at the beginning of previous centuries' last decade. (349 p.)

(30) Serfling, R.J., *Approximation Theorems of Mathematical Statistics*, Wiley, New York (1980).
A very clear account of basic asymptotic statistical theory (χ^2, ML, von Mises functionals, the classes of M, L, R, U statistics, asymptotic relative efficiency), preceded by a chapter on the relevant convergence theorems in probability theory. Advanced undergraduate - beginning graduate level. (371 p.)

(31) Kotz, S. & Johnson, N.L. (editors), *Breakthroughs in Statistics*, Vols I, II, III Springer–Verlag, New York, Berlin, Heidelberg (1993).
Reprints of some 40 fundamental articles that appeared between 1900-1980 and that marked, often in retrospect, a breakthrough in statistical science. Each article is preceded by a 4 page introductory summary and historical annotation by well-known modern authors. (931 p.)

(32) Johnson, N.L & Kotz, S. (editors), *Leading Personalities in Statistical Sciences, from the Seventeenth Century to the Present*, Wiley, New York (1997).
Very readable, brief biographical sketches on the work, life and environment of some 110 leading statistical scientists who have passed away, written by 75 living statistical experts. The biographies are grouped into several areas. The book is about equally divided between fundamentals (sections: Statistical Inference, Statistical Theory, Probability Theory) and applications (sections: Government and Economic Statistics, Applications in Medicine and Agriculture, Applications in Science and Engineering). The statistical biographies of several physicists have been included: Bernoulli, Boltzmann, Gauss, Laplace, Maxwell, Newton, and Poisson. (421 p.)

(33) Kalbfleisch, J.G., *Probability and Statistical Reference*, Vols I, II, Springer–Verlag, New York, Berlin, Heidelberg (1986).

The standard areas of basic probability and statistics are explained in a highly didactical way, with emphasis on practical insight, and using undergraduate mathematics. No decision theory, nor computerintensive methods. (703 p.)

(34) Miller, R.G., *Beyond ANOVA*, Wiley, New York (1986).
Clearly and enthusiastically written overview of what is (was) known and what can be done if various standard assumptions of analysis of variance are violated. (317 p.)

(35) Dijkstra, J.B., *Analysis of Means in some non-standard Situations*, Ph.D. Thesis, Technical University of Eindhoven (1987), CWI Tract 47 (1988).
In the same vein as the book by Miller (but with a certain trade-off between deepness and scope). Restricted towards testing and multiple comparison theory in the k-sample situation. Clear evaluation of the efficiency of various procedures by computer simulations. (138 p.)

(36) Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A., *Robust Statistics, The Approach based on Influence Functions*, Wiley, New York (1986).
Motivating the flourishing of this relatively new and important area. At times provocatively written (the 'break-down' of classical approaches), yet refreshing reading for those interested in the theory of sytematically damping out regression pathologies induced by contaminated and influential data points. The third author's Least Median of Squares Regression paper appeared in Kotz and Johnson, Vol. III. (302 p.)

(37) Björck, A., *Least Squares Methods, in: Handbook of Numerical Analysis*, Vol. I, Ciarlet, P.G., and Lions, J.L. (editors), Elsevier Science Publishers, North Holland (1990).
High-level mathematical background on the algorithms for constrained linear and non-linear least-squares problems. (168 p.)

(38) Barth, F. & Haller, R., *Stochastik Leistungskurs*, Ehrenwirth, München (1994).
An interestingly written and very pedagogical book on basic probability and statistics, with a lot of diagrams and historical annotations. Advanced high school level. (432 p.)

(39) Huff, D., *How to Lie with Statistics*, Pinguin Books, Middlesex, England (1986). (First ed. 1954).
Vintage wine from the present author's birth-year. Pre high-school level, but very vividly and humorously written, pin-pointing many misuses of and fallacies in statistics. (124 p.)

(40) Beck-Bornholdt, H.-P. & Dubben H.-H., *Der Schein der Weisen: Irrtümer und Fehlurteile im täglichen Denken*, Hoffmann und Campe Verlag, Hamburg (2001); Rowolt Taschenbuch Verlag (2003).
Written in the humorous and sceptical vein of Huff, a plethora of in-

teresting examples from actual practice (mostly from medicine and law) which can be moulded into one or several two times two tables, are discussed in a dialogue style. They are analysed by applying the (classical) Bayesian approach. High school level. (270 p.)

Keywords

Probability Theory

1. event
2. random variable
3. conditional probability
4. probability distribution
5. probability density
6. distribution function
7. expectation value
8. standard deviation
9. variance
10. skewness
11. kurtosis / excess of kurtosis
12. variation coefficient
13. correlation
14. central limit theorem
15. characteristic function
16. moment generating function
17. cumulant generating function
18. asymptotic error propagation
19. fuzzy measure
20. belief measure
21. plausibility measure
22. possibility measure
23. necessity measure
24. confidence measure
25. lattice

Statistics

1. observation
2. sample size
3. frequency distribution
4. empirical distribution
5. hypothesis testing
6. test statistic
7. estimate
8. estimator
9. mean
10. bias
11. degrees of freedom
12. unbiased (estimator)
13. unbiased (test statistic)
14. residual sum-of-squares
15. least squares estimator
16. maximum likelihood estimator
17. p-value
18. power function
19. significance level
20. point estimation
21. interval estimation
22. set estimation
23. confidence level
24. confidence region
25. confidence interval
26. acceptance region
27. rejection region
28. confidence band
29. UMVU estimator
30. uniformly most powerful test
31. sufficient statistic
32. linear parameter restriction
33. estimability condition
34. principal component

Applied Regression

1. response variable
2. (in)dependent variable
3. generalised linear model
4. predicted residual
5. weighted least squares
6. outlier
7. robust estimator

8. contaminated data
9. model selection
10. forward selection
11. backward elimination
12. selection bias
13. goodness-of-fit test
14. cross validation
15. biased regression
16. collinearity
26. stability
27. low frequency oscillations
28. high frequency waves
29. Mirnov coils
30. grid
31. lattice

Plasma Physics

1. tokamak
2. thermal energy
3. nuclear fusion
4. confinement time scaling
5. plasma current
6. equilibrium
7. toroidal magnetic field
8. poloidal magnetic flux
9. magnetic configuration
10. electron density
11. electron temperature
12. absorbed heating power
13. ohmic discharge
14. neutral injection
15. electron (ion) cyclotron heating
16. boundary layer
17. neutral density
18. penetration length
19. heat flux
20. divertor plate
21. halo current
22. disruption
23. isotope composition
24. impurity concentration
25. density limit

Stichworte

Wahrscheinlichkeitstheorie

1. Ereignis
2. Zufallsvariable / -größe
3. bedingte Wahrscheinlichkeit
4. Wahrscheinlichkeitsverteilung
5. Wahrscheinlichkeitsdichte
6. Verteilungsfunktion
7. Erwartungswert
8. Standardabweichung
9. Varianz
10. Schiefe
11. Kurtosis / Exzeß
12. Variationskoeffizient
13. Korrelation
14. zentraler Grenzwertsatz
15. charakteristische Funktion
16. momenterzeugende Funktion
17. Kumulanten erzeugende Funktion
18. asymptotische Fehlerfortpflanzung
19. Fuzzy-maß
20. Trauensmaß
21. Glaubhaftigkeitsmaß
22. Möglichkeitsmaß
23. Notwendigkeitsmaß
24. Vertrauensmaß
25. Verband

Statistik

1. Beobachtung
2. Stichprobenumfang
3. Häufigkeitsverteilung
4. empirische Verteilung
5. Hypothesenprüfung
6. Prüfgröße
7. Schätzwert
8. Schätzfunktion
9. Mittelwert
10. Verzerrung
11. Freiheitsgrade
12. erwartungstreue/unverzerrte (Schätzfunktion)
13. unverfälschte (Prüfgröße)
14. Summe der Abweichungsquadrate
15. Methode der kleinsten Quadrate
16. Maximum-Likelihood Schätzfunktion
17. Überschreitungswahrscheinlichkeit
18. Trennschärfefunktion / Gütefunktion
19. Signifikanzniveau
20. Punktschätzfunktion
21. Intervallschätzfunktion
22. Bereichsschätzfunktion
23. Konfidenzniveau
24. Konfidenzbereich/Mutungsbereich
25. Konfidenzintervall/Mutungsintervall
26. Akzeptanzbereich/Annahmebereich
27. Verwerfungsbereich/Ablehnungsbereich
28. Konfidenzband
29. erwartungstreue Schätzfunktion mit gleichmäßig kleinster Varianz
30. gleichmäßig trennschärfste Testfunktion
31. vollständige / erschöpfende / suffiziente Schätzfunktion
32. lineare Parameterrestriktion
33. Schätzbarkeitsbedingung
34. Hauptkomponente

Angewandte Regression

1. Regressand
2. (un-)abhängige Variable
3. verallgemeinertes lineares Modell
4. vorhergesagtes Residuum

5. Methode der gewichteten / gewogenen kleinsten Quadrate
6. Ausreißer
7. robuste Schätzfunktion
8. kontaminierte Daten
9. Modellauswahl
10. Vorwärtsauswahl
11. Rückwärtseliminierung
12. Verzerrung durch Auswahl
13. Anpassungstest
14. Kreuzvalidierung
15. nichterwartungstreue Regression
16. Kollinearität

Plasmaphysik

1. Tokamak
2. thermische Energie
3. Kernfusion
4. Einschlußzeitskalierung
5. Plasmastrom
6. Gleichgewicht
7. toroidales Magnetfeld
8. poloidaler magnetischer Fluß
9. magnetische Konfiguration
10. Elektronendichte
11. Elektronentemperatur
12. absorbierte Heizleistung
13. Ohmsche Entladung
14. Neutralinjektion
15. Elektronen- (Ionen-) Zyklotronheizung
16. Randschicht
17. Neutraldichte
18. Eindringtiefe
19. Wärmefluß
20. Divertorplatte
21. Halostrom
22. Disruption
23. Isotopenmischverhältnis
24. Verunreinigungskonzentration
25. Dichtelimit
26. Stabilität
27. Niedrigfrequenzoszillationen
28. Hochfrequenzwellen
29. Mirnov-Spulen
30. Gitter
31. Gitter / Raster

Mots Clés

Théorie de Probabilité

1. événement
2. variable aléatoire
3. probabilité conditionelle
4. loi de probabilité
5. densité de probabilité
6. fonction de répartition
7. espérance mathématique
8. déviation standard
9. variance
10. asymmétrie
11. aplatissement / excès
12. variabilité
13. corrélation
14. théorème central limite
15. fonction caractéristique
16. fonction génératrice des moments
17. fonction génératrice des cumulants
18. propagation asymptotique des erreurs
19. mesure floue
20. mesure de croyance
21. mesure de plausibilité
22. mesure de possibilité
23. mesure de nécessité
24. mesure de confiance
25. treillis

Statistique

1. observation
2. taille de l'échantillon
3. distribution de fréquence
4. distribution empirique
5. éprouvement des hypothèses
6. statistique de test
7. valeur estimée
8. estimateur
9. moyenne
10. biais
11. degrés de liberté
12. (estimateur) sans biais / non biaisé
13. (statistique d'épreuve) sans distorsion
14. somme des résidues à moindres carrés
15. estimateur par les moindres carrés
16. estimateur de maximum de vraisemblance
17. valeur p
18. fonction de puissance
19. niveau du test
20. estimation ponctuelle
21. estimation par intervalle
22. estimation ensembliste
23. niveau de confiance
24. région de confiance
25. intervalle de confiance
26. région d'acceptation
27. région de rejet
28. bande de confiance
29. estimateur sans biais uniformément de variance minimale
30. épreuve uniformément la plus puissante
31. résumé exhaustif / statistique exhaustif
32. restriction linéaire des paramètres
33. condition d'estimabilité
34. composante principale

Régression Appliquée

1. variable de réponse

2. variable (in)dépendante
3. modèle lineaire généralisé
4. résidu prédicté / prévisé
5. moindres carrés pondérés
6. valeur aberrante
7. estimateur robuste
8. données contaminées
9. sélection du modèle
10. sélection en avant
11. élimination en arrière
12. biais de sélection
13. test d'adéquation du modèle
14. validation de croisement
15. régression biaisée
16. colinéarité

Physique du Plasma

1. tokamak
2. énergie thermique
3. fusion nucléaire
4. loi de temps de confinement
5. courant du plasma
6. équilibre
7. champ magnétique toroïdal
8. flux magnétique poloïdal
9. configuration magnétique
10. densité des électrons
11. température des électrons
12. puissance d'échauffage absorbée
13. décharge ohmique
14. injection de neutres
15. chauffage cyclotronique des électrons (ions)
16. couche limite
17. densité des neutres
18. profondeur de pénétration
19. flux de chaleur
20. plaque de divertor
21. courant halo
22. disruption
23. composition isotopique
24. concentration des impuretés
25. limite de densité
26. stabilité
27. oscillations à frequence basse
28. ondes à haute fréquence
29. bobines de Mirnov
30. grille
31. réseau

Palabras Claves

Teoría de Probabilidad

1. evento
2. variable aleatoria
3. probabilidad condicional
4. distribución de probabilidad
5. densidad de probabilidad
6. función de distribución
7. valor de esperanza
8. desviación estándar
9. varianza
10. asimetría
11. curtosis / exceso
12. coefficiente de variación
13. correlación
14. teorema del limite central
15. función caracteristica
16. función generatriz de los momentos
17. función generatriz de cumulantes
18. error estándar asintótico
19. medida fuzzy
20. medida de creencia
21. medida de plausibilidad
22. medida de possibilidad
23. medida de necesidad
24. medida de confianza
25. látice

Estadística

1. observación
2. tamaño de la muestra
3. distribución de frecuencias
4. distribución empírica
5. pruebando de hipótesis
6. estadístico de una prueba
7. estimación
8. estimador
9. media
10. sesgo
11. grados de libertad
12. (estimador) insesgado
13. (estadístico de una prueba) insesgado
14. suma de los cuadrados resíduos
15. estimador por mínimos cuadrados
16. método del máximo de verosimilitud
17. valor p
18. función de potencia
19. nivel de significación
20. estimación puntual
21. estimación por intervalo
22. estimación por conjunto
23. nivel de confianza
24. región de confianza
25. intervalo de confianza
26. región de aceptación
27. región de rechazo
28. banda de confianza
29. estimador insesgado con varianza uniformemente mínima
30. estadístico de una prueba uniformemente más poderoso
31. estadístico de una prueba suficiente
32. restrición lineal de parámetros
33. condición de estimabilidad
34. componente principal

Regresión aplicada

1. variable de ripuesta

2. variable (in)dependiente
3. modelo lineal generalizado
4. residuo predicado
5. mínimos cuadrados ponderados
6. valor atípico
7. estimador robusto
8. datos contaminados
9. selección de modelo
10. selección hacia adelante
11. eliminación hacia atrás
12. selección sesgada
13. prueba de la bondad del ajustamiento
14. validación cruzada
15. regresión sesgada
16. colinealidad

Física del plasma

1. tokamak
2. energía térmica
3. fusion nuclear
4. escalamiento del tiempo de confinamiento
5. corriente del plasma
6. equilibrio
7. campo magnético toroidal
8. flujo magnético poloidal
9. configuración magnética
10. densidad de los electrones
11. temperatura de los electrones
12. potencia de calefacción absorbida
13. descarga ohmica
14. inyección neutral
15. calefacción ciclotrónica de electrones (de iones)
16. capa limitativa
17. densidad neutral
18. profundidad de penetración
19. flujo de calor
20. plato de divertor
21. corriente de halo
22. disrupción
23. composición de isótopos
24. concentración de impureza
25. limite de densidad
26. estabilidad
27. oscilaciones de frecuencia baja
28. ondas de frecuencia alta
29. bobinas de Mirnov
30. rejilla
31. red

Стрежневые Слова

Теория Вероятности

1. событие
2. случайная переменная / величина
3. условная вероятность
4. распределение вероятности
5. плотность вероятности
6. функция распределения
7. математическое ожидание
8. стандартное отклонение
9. дисперсия
10. показатель асимметрии
11. избыток / эксцесс
12. коэффициент изменчивости
13. корреляция
14. центральная предельная теорема
15. характеристическая функция
16. производящая funktsiya моментов
17. производящая funktsiya кумулянтов
18. асимптотическое распространение ошибок
19. нечёткая мера
20. мера убеждения
21. мера правдоподобия
22. мера возможности
23. мера необходимости
24. мера доверия
25. структура / связка

Статистика

1. наблюдение
2. объём / размер выборки
3. частотные распределение
4. выборочное / эмпирическое распределение
5. проверка гипотез
6. тест / критерий
7. оценка
8. оценочная функция
9. среднее
10. смещение
11. степени свободы
12. несмещенная (оценка)
13. несмещенный (тест)
14. остаточная сумма квадратов
15. оценка метода наименьших квадратов
16. принцип максимального правдоподобия
17. значимость
18. функция мощности
19. уровень значимости
20. точечное оценивание
21. интервальное оценивание
22. множественное оценивание
23. доверительный уровень
24. доверительная область
25. доверительный интервал
26. область принятия
27. область отклонения
28. доверительная полоса
29. несмещенная оценка равномерно наименьшей дисперсии
30. равномерно наиболее мощный критерий
31. достаточная / исчерпывающая статистика
32. линейное ограничение параметров
33. условие оцениваемости

34. главная компонента

Прикладная Регрессия

1. переменная отклика
2. (не-)зависимая переменная
3. обобщенная линейная модель
4. предсказанный остаток
5. метод взвешенных минимальных квадратов
6. выпадающее наблюдение
7. устойчивая оценка
8. загрязненные данные
9. выбор модели
10. выбор прямого направления
11. исключение обратного направления
12. смещение из-за выбора
13. критерий согласия
14. перекрестная проверка
15. смещенная регрессия
16. коллинеарность

Физика Плазмы

1. токамак (тороидальная камера с магнитными катушками)
2. тепловая энергия
3. ядерный синтез
4. закон подобия времени удержания
5. ток плазмы
6. равновесие
7. тороидальное магнитное поле
8. полоидальный магнитный поток
9. магнитная конфигурация
10. плотность электронов
11. электронная температура
12. поглощенная тепловая мощность
13. омический разряд
14. инжекция нейтралов
15. электронно-(ионно-) циклотронный нагрев
16. пограничный слой
17. плотность нейтралов
18. длина проникновения
19. поток тепла
20. пластина дивертора
21. ток гало
22. срыв
23. изотопный состав
24. концентрация примесей
25. предел плотности
26. устойчивость
27. низкочастотные колебания
28. высокочастотные волны
29. Мирнова катушки
30. сетка
31. решётка

KANJI	Hiragana	Romanji	English	Unicode
確	かく	KAKU	evident	78BA
率	りつ	RITSU	proportion	7387
論	ろん	RON	theory	8AD6

PROBABILITY THEORY

事	じ	JI	affair	4E8B
象	しょう	SHOU	image	8C61
event				1.
偶	ぐう	GUU	by chance	5076
然	ぜん	ZEN	as it is	7136
変	へん	HEN	variation	5909
数	すう	SU	number	6570
random variable				2.
条	じょう	JOU	clause	6761
件	けん	KEN	matter	4EF6
付き	つき	TSUKI	attached	4ED8
分	ぶん	BUN	part	5206
布	ぷ	PU	spread	5E03
conditional distribution				3.
確	かく	KAKU	evident	78BA
率	りつ	RITSU	proportion	7387
分	ぶん	BUN	part	5206
布	ぷ	PU	spread	5E03
probability distribution				4.
確	かく	KAKU	evident	78BA
率	りつ	RITSU	proportion	7387
密	みつ	MITSU	dense	5BC6
度	ど	DO	degree	5EA6
probability density				5.

分	ぶん	BUN	part	5206
布	ぷ	PU	spread	5E03
関	かん	KAN	barrier	95A2
数	すう	SUU	number	6570
distribution function				6.
期	き	KI	expect	671F
待	たい	TAI	wait	5F85
値	ち	CHI	value	5024
expectation value				7.
標	ひょう	HYOU	mark	6A19
準	じゅん	JUN	norm	6E96
偏	へん	HEN	deviating	504F
差	さ	SA	difference	5DEE
standard deviation				8.
分	ぶん	BUN	part	5206
散	さん	SAN	scatter	6563
variance				9.
歪	わい	WAI	distort	6B6A
度	ど	DO	degree	5EA6
skewness				10.
尖	せん	SEN	pointed	5C16
度	ど	DO	degree	5EA6
kurtosis				11.
変	へん	HEN	variation	5909
動	どう	DOU	movement	52D5
係	けい	KEI	connect	4FC2
数	すう	SUU	number	6570
variation coefficient				12.

相	そう	SOU	mutual	76F8
関	かん	KAN	barrier	95A2
correlation				13.

中	ちゅう	CHUU	middle	4E2D
心	しん	SHIN	heart	5FC3
極	きょく	KYOKU	extreme	6975
限	げん	GEN	limit	9650
定	てい	TEI	fix	5B9A
理	り	RI	reason	7406
central limit theorem				14.

特	とく	TOKU	special	7279
性	せい	SEI	nature	6027
関	かん	KAN	barrier	95A2
数	すう	SUU	nunber	6570
characteristic function				15.

積	せき	SEKI	accumulate	7A4D
率	りつ	RITSU	proportion	7387
母	ぼ	BO	mother	6BCD
関	かん	KAN	barrier	95A2
数	すう	SUU	number	6570
moment generating function				16.

キュムラント	きゅむらんと	kyumuranto	cumulant	
母	ぼ	BO	mother	6BCD
関	かん	KAN	barrier	95A2
数	すう	SUU	number	6570
cumulant generating function				17.

漸	ぜん	ZEN	gradually	6F38
近	きん	KIN	near	8FD1
誤	ご	GO	mistake	8AA4
差	さ	SA	difference	5DEE
伝	でん	DEN	transmit	4F1D
播	ぱん	BAN	ordering	756A

asymptotic error propagation 18.

ファジイ	ふぁじい	fazii	fuzzy	
測	そく	SOKU	measure	6E2C
度	ど	DO	degree	5EA6

fuzzy measure 19.

所	しょ	SHO	one´s	6240
信	しん	SHIN	opinion	4FE1
性	せい	SEI	nature	6027
測	そく	SOKU	measure	6E2C
度	ど	DO	degree	5EA6

belief measure 20.

確	かく	KAKU	certain	78BA
実	じつ	JITSU	real	5B9F
性	せい	SEI	nature	6027
測	そく	SOKU	measure	6E2C
度	ど	DO	degree	5EA6

plausibility measure 21.

可	か	KA	potential	53EF
能	のう	NOU	-ity	80FD
性	せい	SEI	nature	6027
測	そく	SOKU	measure	6E2C
度	ど	DO	degree	5EA6

possibility measure 22.

必	ひつ	HITSU	inevitable	5FC5
然	ぜん	ZEN	as it is	7136
性	せい	SEI	nature	6027
測	そく	SOKU	measure	6E2C
度	ど	DO	degree	5EA6
necessity measure				23.

信	しん	SHIN	belief	4FE1
頼	らい	RAI	trust	1075
性	せい	SEI	nature	6027
測	そく	SOKU	measure	6E2C
度	ど	DO	degree	5EA6
confidence measure				24.

格	こう	KOU	rank	683C
子	し	SHI	child	5B50
lattice				25.

統	とう	TOU	unite	7D71
計	けい	KEI	compute	8A08
学	がく	GAKU	study	5B66
STATISTICS				

観	かん	KAN	view	89B3
測	そく	SOKU	measure	6E2C
observation				1.

標	ひょう	HYOU	mark, sign	6A19
本	ほん	HON	basis	672C
の	の	no	´s	
大きさ	おおきさ	ookisa	size	5927
sample size				2.

度	ど	DO	degree	6570
数	すう	SUU	number	E5A6
分	ぶん	BUN	part	5206
布	ぷ	PU	cloth	5E03
frequency distribution				3.

経	けい	KEI	pass through	7D4C
験	けん	KEN	test	9A0E
分	ぶん	BUN	part	5206
布	ぷ	PU	cloth	5E03
empirical distribution				4.

仮	か	KA	pseudo	4EEE
説	せつ	SETSU	theory	8AAC
検	けん	KEN	examine	63A8
定	てい	TEI	fix	5B9A
hypothesis testing				5.

検	けん	KEN	examine	63A8
定	てい	TEI	fix	5B9A
統	とう	TOU	unite	7D71
計	けい	KEI	compute	8A08
test statistic				6.

推	すい	SUI	infer	63A8
定	てい	TEI	fix	5B9A
値	ち	CHI	value	5024
estimate				7.

推	すい	SUI	infer	63A8
定	てい	TEI	fix	5B9A
関	かん	KAN	barrier	5909
数	すう	SUU	number	6570
estimator				8.

平	へい	HEI	equal	5E73
均	きん	KIN	balanced	5747
mean				9.
偏	かたよ	katayo	one-sided	5D4F
り	り	ri	-ness	
bias				10.
自	じ	JI	self	81EA
由	ゆう	YUU	use	7531
度	ど	DO	degree	5EA6
degrees of freedom				11.
不	ふ	FU	not	4E0D
偏	へん	HEN	biased	5D4F
(推	(すい	(SUI	infer	(63A8
定	てい	TEI	decide	5B9A
量)	りょう)	RYOU)	quantity	91CF)
unbiased (estimator)				12.
不	ふ	FU	not	4E0D
偏	へん	HEN	biased	5D4F
(検	(けん	(KEN	(examine	(63A8
定	てい	TEI	fix	5B9A
統	とう	TOU	unite	7D71
計)	けい)	KEI)	compute)	8A08)
unbiased (test statistic)				13.
残	ざん	ZAN	remain	6B8B
差	さ	SA	difference	5DEE
二	じ	JI	second	4E57
乗	じょう	JOU	power	548C
和	わ	WA	sum	548C
residual least squares				14.

最	さい	SAI	most	6700
小	しょう	SHOU	small	5C0F
二	じ	JI	second	4E57
乗	じょう	JOU	power	548C
推	すい	SUI	infer	63A8
定	てい	TEI	fix	5B9A
関	かん	KAN	barrier	5909
数	すう	SUU	number	6570

least squares estimator 15.

最	さい	SAI	most	6700
尤	ゆう	YUU	plausible	5C24
推	すい	SUI	infer	63A8
定	てい	TEI	fix	5B9A
関	かん	KAN	barrier	5909
数	すう	SUU	number	6570

maximum likelihood estimator 16.

P—	P—	P-	P-	
値	ち	CHI	value	5024

P-value 17.

分	ぶん	BUN	part	5206
割	かつ	KATSU	divide	5272
可	か	KA	ability	53EF
能	のう	NOU		80FD
関	かん	KAN	barrier	5909
数	すう	SUU	number	6570

power function 18.

有	ゆう	YUU	have	6709
意	い	I	meaning	610F
水	すい	SUI	water	6C34
準	じゅん	JUN	level	6E96

significance level 19.

地	ち	CHI	place	5730
点	てん	TEN	point	70B8
推	すい	SUI	infer	63A8
定	てい	TEI	fix	5B9A
point estimation				20.
区	く	KU	area	533A
間	かん	KAN	interval	9593
推	すい	SUI	infer	63A8
定	てい	TEI	fix	5B9A
interval estimation				21.
集	しゅう	SHUU	collect	96C6
合	ごう	GOU	interval	5408
推	すい	SUI	infer	63A8
定	てい	TEI	fix	5B9A
set estimation				22.
信	しん	SHIN	belief	4FE1
頼	らい	RAI	trust	1075
水	すい	SUI	water	6C34
準	じゅん	JUN	level	6E96
confidence level				23.
信	しん	SHIN	belief	4FE1
頼	らい	RAI	trust	1075
領	りょう	RYOU	territory	9818
域	いき	IKI	region	57DF
confidence region				24.
信	しん	SHIN	belief	4FE1
頼	らい	RAI	trust	1075
区	く	KU	area	533A
間	かん	KAN	interval	9593
confidence interval				25.

採	さい	SAI	pick	63A1
択	たく	TAKU	select	629E
領	りょう	RYOU	territory	9818
域	いき	IKI	region	57DF

acceptance region 26.

棄	き	KI	abandon	68C4
却	きゃく	KYAKU	remove	5374
領	りょう	RYOU	territory	9818
域	いき	IKI	region	57DF

rejection region 27.

信	しん	SHIN	belief	4FE1
頼	らい	RAI	trust	1075
帯	たい	TAI	belt	5E2F
域	いき	IKI	region	57DF

confidence band 28.

均	きん	KIN	uniform	5747
一	いつ	ITSU	-ly	4E00
最	さい	SAI	most	6700
少	しょう	SHOU	small	5C11
分	ぶん	BUN	part	5206
散	さん	SAN	scatter	6563
不	ふ	FU	not	4E0D
偏	へん	HEN	biased	5D4F
推	すい	SUI	infer	63A8
定	てい	TEI	fix	5B9A
関	かん	KAN	barrier	5909
数	すう	SUU	number	6570

uniformly minimum variance unbiased estimator 29.

均	きん	KIN	uniform	5747
一	いつ	ITSU	-ly	4E00
最	さい	SAI	most	6700
大	だい	DAI	large	5927
分	ぶん	BUN	part	5206
割	かつ	KATSU	divide	5272
可	か	KA	ability	53EF
能	のう	NOU		80FD
検	けん	KEN	examine	691C
定	てい	TEI	fix	5B9A

uniformly most powerful test 30.

十	じゅう	JUU	ten	5341
分	ぶん	BUN	part	5206
統	とう	TOU	unite	7D71
計	けい	KEI	compute	8A08
関	かん	KAN	barrier	5909
数	すう	SUU	number	6570

sufficient statistic 31.

線	せん	SEN	line	7DDA
形	けい	KEI	shape	5F62
パラメタ	ぱらめた	parameta	parameter	
制	せい	SEI	restrict	5236
限	げん	GEN	limit	9650

linear parameter restriction 32.

推	すい	SUI	infer	63A8
定	てい	TEI	fix	5B9A
可	か	KA	ability	53EF
能	のう	NOU		80FD
条	じょう	JOU	article, clause	6761
件	けん	KEN	matter	4EF6

estimability condition 33.

主	しゅ	SHU	main	4E38
要	よう	YOU	important	8981
成	せい	SEI	constitute	6210
分	ぶん	BUN	part	5206

principal component 34.

応	おう	OU	appropriate	5FDC
用	よう	YOU	use	7528
回	かい	KAI	turn back	56DE
帰	き	KI	return	5E30

APPLIED REGRESSION

反	ほん	HON	react	53CD
応	のう	NOU	respond	5FDC
変	へん	HEN	variation	5909
数	すう	SUU	number	6570

response variable 1.

独	どく	DOKU	alone	72EC
立	りつ	RITSU	stand	7ACB
変	へん	HEN	variation	5909
数	すう	SU	number	6570

independent variable 2a.

従	じゅう	JUU	follow	5F93
属	ぞく	ZOKU	subordinate	5C5E
変	へん	HEN	variation	5909
数	すう	SU	number	6570

dependent variable 2b.

一般	いっぱん	IPPAN	general-	822C
化	か	KA	ised	5316
線	せん	SEN	line	7DDA
形	けい	KEI	shape	5F62
モデル	もでる	moderu	model	
generalised linear model				3.
予	よ	YO	in advance	4E88
測	そく	SOKU	conjecture	6E2C
潜	せん	SEN	hide, dive	6F5C
在	ざい	ZAI	be, reside	5728
predicted residual				4.
加	か	KA	add	52A0
重	じゅう	JUU	weight	91CD
最	さい	SAI	most	6700
小	しょう	SHOU	small	5C0F
二	じ	JI	second	4E57
乗	じょう	JOU	power	548C
weighted least squares				5.
異	い	I	different	7570
常	じょう	JOU	regular	5E38
値	ち	CHI	value	5024
outlier				6.
強	きょう	KYOU	strong	5F37
推	すい	SUI	infer	63A8
定	てい	TEI	decide	5B9A
量	りょ	RYOU	quantity	91CF
robust estimator				7.
汚	お	O	dirty	6C5A
染	せん	SEN	infect	1640
デタ	でた	deta	data	
contaminated data				8.

型	かた	KATA	model	5F62
の	の	no	´s	
選	せん	SEN	choose	9078
別	べつ	BETSU	classify	5225
model selection				9.

前	ぜん	ZEN	ahead	524D
進	しん	SHIN	proceed	9032
選	せん	SEN	choose	9078
別	べつ	BETSU	classify	5225
forward selection				10.

後	こう	KOU	behind	9001
退	たい	TAI	retreat	5F8C
消	しょう	SHOU	exstinguish	6D88
去	きょ	KYO	remove	53BB
backward elimination				11.

抜	ばっ	BAT	extract	629C
擢	てき	TEKI	select	64E2
偏	かたよ	katayo	one-sided	5D4F
り	り	ri	-ness	
selection bias				12.

当てはまり	あてはまり	Atewamari	fitting	5F53
の	の		´s	
良さ	よさ	YOsa	goodness	826F
検	けん	KEN	examine	63A8
定	てい	TEI	fix	5B9A
goodness of fit test				13.

相	そう	SOU	mutual	76F8
対	たい	TAI	oppose	5BFE
妥	だ	DA	come to terms	59A5
当	とう	TOU	hit the mark	5F53
化	か	KA	turn into	5316
cross validation				14.
偏	かたよ	katayo	one-sided	5D4F
り	り	ri	-ness	
回	かい	KAI	turn back	56DE
帰	き	KI	return	5E30
biased regression				15.
共	きょう	KYOU	common	5171
線	せん	SEN	line	7DDA
性	せい	SEI	nature	6027
collinearity				16.

プラズマ	ぷらずま	purazuma	plasma	
物	ぶつ	BUTSU	thing	7269
理	り	RI	reason	7406

PLASMA PHYSICS

トカマク	とかまく	tokamaku	tokamak	
tokamak				1.
熱	ねつ	netsu	heat	71B1
エネルギ	えねるぎ	enerugi	energy	
thermal energy				2.

核	かく	KAKU	nucleus	6838
融	ゆう	YUU	fuse	878D
合	ごう	GOU	combine	5408
nuclear fusion				3.
閉じ	とじ	TOji	close	9589
込め	こめ	KOme	put in	8FBC
時	じ	JI	time	6642
間	かん	KAN	interval	9593
比	ひ	HI	ratio	6BD4
例	れい	REI	example	4F8B
則	そく	SOKU	rule	5247
confinement time scaling				4.
プラズマ	ぷらずま	purazuma	plasma	
電	でん	DEN	electricity	96FB
流	りゅう	RYUU	current	6D41
plasma current				5.
平	へい	HEI	even	5E73
衡	こう	KOU	balance	8861
equilibrium				6.
トロイダル	とろいだる	toroidaru	toroidal	
磁	じ	JI	magnet	78C1
場	ば	BA	field	5834
toroidal magnetic field				7.
ポロイダル	ぽろいだる	poroidaru	poloidal	
磁	じ	JI	magnet	78C1
束	そく	SOKU	bundle	675F
toroidal magnetic flux				8.

磁	じ	JI	magnet	78C1
場	ば	BA	field	5834
配	はい	HAI	distribute	914D
位	い	I	position	4F4D
magnetic configuration				9.
電	でん	DEN	electricity	96FB
子	し	SHI	child	5B50
密	みつ	MITSU	dense	5BC6
度	ど	DO	degree	5EA6
electron density				10.
電	でん	DEN	electricity	96FB
子	し	SHI	child	5B50
温	おん	ON	hot	6E29
度	ど	DO	degree	5EA6
electron temperature				11.
吸	きゅう	KYUU	absorb	5438
収	しゅう	SYUU	take in	53CE
加	か	KA	add	52A0
熱	ねつ	NETSU	heat	71B1
パワア	ぱわあ	PAWAA	power	
absorbed heating power				12.
オオミック	おおみっく	oumikku	ohmic	
放	ほう	HOU	release	653E
電	でん	DEN	electrcity	96FB
ohmic discharge				13.
中	ちゅう	CHUU	middle	4E2D
性	せい	SEI	nature	6027
粒	りゅう	RYUU	grain	7C92
子	し	SHI	child	5B50
入	にゅう	NYUU	enter	5165
射	しゃ	SHA	shoot	5C04
neutral particle injection				14.

電	でん	DEN	electricity	96FB
子	し	SHI	child	5B50
イオン	いおん	ion	ion	
サイクロトロン	さいくろとろん	saikurotoron	cyclotron	
加	か	KA	add	52A0
熱	ねつ	NETSU	heat	71B1
electron (ion) cyclotron heating				15.

境	きょう	KYOU	boundary	5883
界	かい	KAI	border	754C
層	そう	SOU	stratum	5C64
boundary layer				16.

中	ちゅう	CHUU	middle	4E2D
性	せい	SEI	nature	6027
粒	りゅう	RYUU	grain	7C92
子	し	SHI	child	5B50
密	みつ	MITSU	dense	5BC6
度	ど	DO	degree	5EA6
neutral particle density				17.

透	とう	TOU	penetrate	900F
過	か	KA	pass through	904E
長	ちょう	CHOU	length	9577
penetration length				18.

熱	ねつ	NETSU	heat	71B1
流	りゅう	RYUU	current	6D41
束	そく	SOKU	bundle	675F
heat flux				19.

ダイバアタ	だいばあた	daibaata	divertor	
板	ばん	BAN	plate	677F
divertor plate				20.

ハロオ	はろお	haroo	halo	
電	でん	DEN	electricity	96FB
流	りゅう	RYUU	current	6D41
halo current				21.
ディスラプション	でぃすらぷしょん	disurapushion	disruption	
disruption				22.
同	どう	DOU	same	540C
位	い	I	position	4F4D
体	たい	TAI	body	4F53
要	よう	JOU	essential	8981
素	そ	SO	element	7D20
isotope composition				23.
不	ふ	FU	not	4E0D
純	じゅん	JUN	pure	7D14
物	ぶつ	BUTSU	matter	7269
濃	のう	NOU	concentrated	6FC3
縮	しゅく	SHUKU	shrink	7E2E
impurity concentration				24.
密	みつ	MITSU	dense	5BC6
度	ど	DO	degree	5EA6
限	げん	GEN	limit	9650
界	かい	KAI	area	754C
density limit				25.
安	あん	AN	peace	5B89
定	てい	TEI	fixed	5B9A
性	せい	SEI	nature	6027
stability				26.

低	てい	TEI	low	4F4E
周	しゅう	SHUU	cycle	5468
波	は	HA	wave	6CE2
振	しん	SHIN	swing	632F
動	どう	DOU	movement	52D5
low frequency oscillation				27.

高	こう	KOU	high	9AD8
周	しゅう	SHUU	cycle	5468
波	は	HA	wave	6CE2
振	しん	SHIN	swing	632F
動	どう	DOU	movement	52D5
high frequency oscillation				28.

ミルノフコイル	みるのふこいる	mirunofu coiru	Mirnov coil	
Mirnov coil				29.

格	こう	KOU	rank	683C
子	し	SHI	child	5B50
grid				30.

格	こう	KOU	rank	683C
子	し	SHI	child	5B50
lattice				31.

Legend: The first column of each entry in the Japanese keyword list contains the Kanji (Chinese characters) of the component meanings. The second column gives the Japanese pronunciation written in Hiragana and the third column the corresponding transcription in Latin characters. The third column uses capitals for ON (i.e., originally Chinese) reading and lower case letters for KUN (Japanese) reading of the kanji. For foreign words, which are marked by Katakana in Japanese, lower case Latin characters are used also.The fourth column contains an English equivalent and the fifth column exhibits the Unicode number for each single Kanji. The meaning of the full keyword in English (which is read top-down in Japanese) and the keyword number are located on the last line of each entry.

References

1. Mokhtar Bin Abdullah. On a robust correlation coefficient. *The Statistician*, pages 455–460, 1990.
2. S.A. Adelfio and C.F. Nolan. *Principles and Applications of Boolean Algebras*. Hayden, Rochelle Park, first edition, 1964.
3. M. Agop, V. Melnig, D. Ruscanu, and G. Popa. Cnoidal modes oscillations as a generalization of ion acoustic waves and soliton. In *Controlled Fusion and Plasma Physics, (Proc. 25th Eur. Conf., Prague, 1998)*, volume 22 C, pages 272–275, Geneva, 1998. European Physical Society. Available on internet, URL=http://epsppd.epfl.ch/Praha/WEB/AUTHOR_P.HTM.
4. H. Akaike. Information theory and an extension of the maximum likelihood principle. In P. Petrov and F. Csàki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Reprinted in: Johnson, N.L. and Kotz, S. (editors), *Breakthroughs in Statistics*, Vol I, Springer–Verlag (1992) 610–624, with an introduction by J. de Leeuw.
5. J.C. Akkerboom. *Testing Problems with Linear or Angular Inequality Constraints*. PhD thesis, Groningen University, Groningen, 1989. Also: Springer Lecture Notes in Statistics 62 (1990).
6. C.J. Albers. *Distributional Inference: the Limits of Reason*. PhD thesis, Groningen University, Groningen, 2003.
7. C.J. Albers and W. Schaafsma. How to assign probabilities if you must. *Statistica Neerlandica*, 55(3):346–357, 2001.
8. S. Albeverio, J.E. Fenstad, R. Høegh-Krohn, and T. Lindstrøm. *Nonstandard Methods in Stochastic Analysis and Mathematical Physics*. Pure and Applied Mathematics, Vol. 122. Academic Press, 1986.
9. H.-W. Alten, A. Djafari-Naini, M. Folkerts, H. Schlosser, K.-H. Schlote, and H. Wußing. *4000 Jahre Algebra – Geschichten, Kulturen, Menschen*. Vom Zählstein zum Computer. Springer–Verlag, Heidelberg, 2003.
10. A.W. Ambergen. *Statistical Uncertainties in Posterior Probabilities*. PhD thesis, Groningen University, Groningen, 1989. Also: CWI tract 93, Mathematical Centre, Amsterdam, 1993.
11. T.W. Anderson. Asymptotically efficient estimation of covariance matrices with linear structure. *Annals of Statistics*, 1:135–141, 1973.
12. T.W. Anderson. Estimating linear statistical relationships (The 1984 Wald memorial lectures). *Annals of Statistics*, 12:1–45, 1984.
13. T.W. Anderson. R.A. Fisher and multivariate analysis. *Statistical Science*, 11:20–34, 1996.
14. T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, third edition, 2003. First edition: 1958.

15. T.W. Anderson and Y. Amemiya. The asymptotic normal distribution of estimators in factor analysis under general conditions. *Annals of Statistics*, 16:759–771, 1988.
16. A. Araujo and E. Giné. *The Central Limit Theorem for Real and Banach Space Valued Random Variables*. Wiley, New York, 1980.
17. J. Arbuthnot. An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society*, 27:186–190, 1710. Reprinted in: M.G. Kendall and R.L. Plackett (editors). Studies in the History of Statistics and Probability, Vol. 2, Griffin, London, 1977.
18. V.I. Arnold. *Bifurcation Theory and Catastrophe Theory*, chapter Catastrophe Theory, pages 207–271. Springer–Verlag, Heidelberg, 1999. Also: Dynamical Systems V, Encyclopedia of Mathematical Sciences 5. Translated by N.D. Kazarinov. Original Russian edition: Итоги науки и техники, Современные проблемы математики, Фундаменталъные направления, V, Динамические Системы 5, VINITI, Moscow, 1986.
19. M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.
20. L.A. Artsimovich and R.S. Sagdeev. *Plasmaphysik für Physiker*. Teubner–Verlag, Stuttgart, 1983. Translated into German by H.–P. Zehrfeld.
21. R. Aymar. ITER: an integrated approach to ignited plasmas. *Philosophical Transactions of the Royal Society of London, Series A*, 357:471–487, 1999.
22. R. Aymar, V.A. Chuyanov, M. Huguet, R. Parker, and Y. Shimomura. The ITER project: A physics and technology experiment. *Proceedings of the 16th Conference on Fusion Energy, Montreal 1996*, 1:3–17, 1997. IAEA–CN–64/O1–1.
23. R. Aymar, V.A. Chuyanov, M. Huguet, Y. Shimomura, ITER Joint Central Team, and ITER Home Teams. Overview of ITER–FEAT – the future international burning plasma experiment. *Nuclear Fusion*, 41:1301–1310, 2001.
24. A. Azzalini. Growth curve analysis for patterned covariance matrices. Technical report, Department of Statistical Sciences, University of Padua, 1985.
25. L. Bachelier. Théorie de la spéculation. *Annales scientifiques de l'Ecole Normale Supérieure*, 17:21–86, 1900. Thèse, Reprinted: Editions Jacques Gabay, 1995, Paris.
26. P. Bailache. *Essai de logique déontique*. Vrin, Paris, 1991.
27. S. Banach and A. Tarski. Sur la décomposition des ensembles de points en parties respectivement congruentes. *Fundamenta Mathematicae*, pages 244–277, 1924.
28. E.W. Barankin and A.P. Maitra. Generalization of the Fisher–Darmois–Koopman–Pitman theorem on sufficient statistics. *Sankhyā*, 25:217–244, 1963.
29. R.E. Barlow. Introduction to de Finetti (1937) Foresight: Its logical laws, its subjective sources. In S. Kotz and N.L. Johnson, editors, *Breakthroughs in Statistics*, Springer Series in Statistics, pages 127–133. Springer–Verlag, 1992.
30. O.E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, New York, 1978.
31. O.E. Barndorff-Nielsen. Processes of normal inverse Gaussian type. *Finance and Stochastics*, 2:41–68, 1998.
32. O.E. Barndorff-Nielsen, P. Blaesild, and C. Halgreen. First hitting-time models for the generalised Gaussian distribution. *Stochastic Processes and Applications*, pages 49–54, 1978.

33. O.E. Barndorff-Nielsen and D.R. Cox. *Asymptotic Techniques for Use in Statistics*. Number 31 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1989.
34. O.E. Barndorff-Nielsen and D.R. Cox. *Inference and Asymptotics*. Number 52 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1994.
35. O.E. Barndorff-Nielsen, J.L. Jensen, and W.S. Kendall. *Networks and Chaos – Statistical and Probabilistic Aspects*. Number 50 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1993.
36. M.S. Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London*, 160:268–282, 1937.
37. M.S. Bartlett and D.G. Kendall. The statistical analysis of variance–heterogeneity and the logarithmic transformation. *Journal of the Royal Statistical Society* (*Supplement*), 8:128–138, 1946.
38. G. Bateman and D.B. Nelson. Resistive ballooning-mode equations. *Physical Review Letters*, pages 1804–1807, 1978.
39. T. Bayes. Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:370–418, 1763. Reprinted in: Biometrika 45 (1958) 293–315 (with an introduction by G.A. Barnard).
40. G. Becker. Empirical scaling law for the effective heat diffusivity in ELMy H-mode plasmas. *Nuclear Fusion*, 36:527–530, 1996.
41. G. Becker. Transport simulations of ITER with empirical heat diffusivity scaling. *Nuclear Fusion*, 38:293–312, 1998.
42. G. Becker. Study of anomalous inward drift in tokamaks by transport analysis and simulations. *Nuclear Fusion*, 44:933–944, 2004.
43. R. Behrisch. *Sputtering by Particle Bombardment II*, chapter Sputtering of Solids with Neutrons, pages 179–229. Topics in Applied Physics. Springer–Verlag, Heidelberg, 1983.
44. R. Behrisch, M. Mayer, and C. García-Rosales. Composition of the plasma facing material tokamakium. *Journal of Nuclear Materials*, 233–237:673–680, 1996.
45. G. Bekefi. *Radiation Processes in Plasmas*. Wiley, New York, 1966.
46. V.S. Belikov and O.A. Silvira. Excitation of TAE instability by velocity anisotropy of fast ions. *Nuclear Fusion*, 34:1522–1526, 1994.
47. D.A. Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Wiley, New York, 1991.
48. A. Ben-Israel. The Moore of the Moore–Penrose inverse. *The Electronic Journal of Linear Algebra*, 9:150–157, 2002.
49. A. Ben-Israel and T.N.E. Greville. *Generalised Inverses: Theory and Applications*. Canadian Mathematical Society Books in Mathematics, Vol. 15. Springer–Verlag, Heidelberg, second edition, 2003. First edition: Wiley, 1974.
50. J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer–Verlag, Heidelberg, 1985.
51. M. Berger. *Geometry I, II*. Springer–Verlag, Heidelberg, 1987. Corrected second printing: 1994 and 1996.
52. J. Berkson. Minimum discrimination information, the 'no interaction' problem, and the logistic function. *Biometrics*, 28:443–468, 1972.
53. J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, New York, 1994.
54. Ph.D. Besse. Models for multivariate analysis. In *Proceedings in Computational Statistics XI*, pages 271–285, Heidelberg, 1994. Physica–Verlag.

55. R.N. Bhattacharya and R.R. Rao. *Normal Approximation and Asymptotic Expansions*. Wiley, New York, 1976.
56. P.J. Bickel and K.A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice Hall, Englewood Cliffs, NJ, second edition, 2000. First edition: Holden Day, San Francisco, 1977.
57. P. Billingsley. *Probability and Measure*. Wiley, New York, third edition, 1995.
58. C. Bingham. *R.A. Fisher: An Appreciation*, chapter Distribution on the sphere, pages 171–180. Number 1 in Lecture Notes in Statistics. Springer–Verlag, Heidelberg, 1980. S.E. Fienberg and D.V. Hinkley (editors).
59. G. Birkhoff and T.C. Bartee. *Modern Applied Algebra*. McGraw Hill, New York, 1970.
60. G. Birkhoff and S. MacLane. *A Survey of Modern Algebra*. Peters, Wellesley, Mass., 1997. First edition: Macmillan, New York, 1950.
61. D. Biskamp. *Nonlinear Magnetohydrodynamics*. Cambridge University Press, 1997.
62. D. Biskamp. *Magnetohydrodynamic Turbulence*. Cambridge University Press, 2003.
63. G.S. Bisnovatyi-Kogan. *Stellar Physics*, volume 1. Springer–Verlag, New York, 2001.
64. P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, 2001.
65. D. Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 18:105–110, 1947.
66. K.A. Bollen. *Structural Equations with Latent Variables*. Wiley, New York, 1989.
67. G. Böme. *Fuzzy–Logik*. Springer–Verlag, Heidelberg, 1993.
68. G. Boole. *An Investigation of the Laws of Thought, on which are founded the mathematical theories of logic and probabilities*. MacMillan, Cambridge, MA, 1854. Reprinted: Dover, New York, 1958.
69. A. Boomsma, M.A.J. Van Duijn, and T.A.B. Snijders, editors. *Essays on Item Response Theory*, volume 157 of *Lecture Notes in Statistics*. Springer–Verlag, Heidelberg, 2001. Volume dedicated to Ivo W. Molenaar on occasion of his 65th birthday.
70. M. Bornatici, R. Cano, O. De Barbieri, and F. Engelmann. Electron cyclotron emission and absorption in fusion plasmas (review paper). *Nuclear Fusion*, 23:1153–1257, 1983.
71. K. Borrass et al. Recent H–mode density limit studies at JET. *Nuclear Fusion*, 44:752–760, 2004.
72. K. Borrass, J. Lingertat, and R. Schneider. A scrape-off layer based density limit for JET ELMy H–modes. *Contributions to Plasma Physics*, 38:130–135, 1998.
73. N. Bourbaki. *Algèbre, Ch. 1–3, 4–7, 8–10*. Eléments de Mathématique. Hermann, Paris, 1970, 1981, 1973. Actualités Scientifiques et Industrielles.
74. N. Bourbaki. *Intégration, Ch. 1–4,5–9*. Eléments de Mathématique. Hermann, Paris, deuxième edition, 1973, 1959. Actualités Scientifiques et Industrielles.
75. N. Bourbaki. *Eléments d'Histoire des Mathématiques*. Hermann, Paris, 1974. Histoire de la Pensée IV. English translation: Springer–Verlag, Heidelberg, 1999.
76. N. Bourbaki. *Topologie Générale, Ch. 1–4, 5–10*. Eléments de Mathématique. Hermann, Paris, 1974. Actualités Scientifiques et Industrielles.

77. N. Bourbaki. *Théorie des Ensembles, Ch. 1–4*. Eléments de Mathématique. Hermann, Paris, deuxième edition, 1998. Actualités Scientifiques et Industrielles.
78. K.O. Bowman and L.R. Shenton. *Properties of Estimators for the Gamma Distribution*. Marcel Dekker, New York, 1988.
79. G.E.P. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26:211–252, 1964.
80. G.E.P. Box and G.M. Jenkins. *Time Series Analysis: Forecasting and Control*. Wiley Classics Library. Holden Day, 1976.
81. G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley Classics Library. Wiley, 1973.
82. T.J.M. Boyd and J.J. Sanderson. *The Physics of Plasmas*. Cambridge University Press, Cambridge, 2003.
83. B.J. Braams. *Computational Studies in Tokamak Equilibrium and Transport*. PhD thesis, Utrecht University, Utrecht, 1986.
84. C.M. Braams and P. Stott. *Nuclear Fusion: Half a Century of Magnetic Confinement Fusion Research*. Institute of Physics Publishing, 2002.
85. G. Bracco and K. Thomsen. Analysis of a global energy confinement database for JET Ohmic plasmas. *Nuclear Fusion*, 37:759–770, 1997.
86. S.I. Braginskii. *Reviews of Plasma Physics*, volume 1, chapter Transport Properties in a Plasma, pages 205–311. Consultants Bureau, New York, 1965. M.A. Leontovich, editor.
87. C. Breton, C.D. De Michelis, and M. Mattioli. Ionization equilibrium and radiative cooling of a high temperature plasma. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 19:367–379, 1978.
88. H.W. Broer, B. Krauskopf, and G. Vegter, editors. *Global Analysis of Dynamical Systems*. Institute of Physics Publishing, 2001. Festschrift dedicated to Floris Takens for his 60th birthday.
89. J.L. Bromberg. *Fusion: Science, Politics and the Invention of a new Energy Source*. MIT Press, Cambridge, MA, 1983.
90. M.B. Brown and A.B. Forsythe. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69:363–367, 1974. Corrigendum: JASA 69 (1974) 840.
91. R.J. Buehler. Some ancillary statistics and their properties. *Journal of the American Statistical Association*, 77:581–589, 1982.
92. C.R. Burnett, D.J. Grove, R.W. Palladino, T.H. Stix, and K.E. Wakefield. The divertor, a device for reducing the impurity level in a stellarator. *Physics of Fluids*, 1:438–445, 1958.
93. R. Burton and E. Waymire. Scaling limits for associated random measures. *Annals of Probability*, 13:1267–1278, 1985.
94. C.E. Bush, R. Maingi, M.G. Bell, et al. Evolution and termination of H-modes in NSTX. *Plasma Physics and Controlled Fusion*, 44:A323–A332, 2002.
95. D.J. Campbell. Physics and goals of RTO/RC–ITER. *Plasma Physics and Controlled Fusion*, 41:B381–B393, 1999.
96. D.J. Campbell, D. Borba, J. Bucalossi, D. Moreau, O. Sauter, J. Stober, and G. Vayakis. Report on the 10th European fusion physics workshop (Vaals, The Netherlands, 9–11 December 2002). *Plasma Physics and Controlled Fusion*, 45:1051–1067, 2003.

97. D.J. Campbell and A. Eberhagen. Studies of electron cyclotron emission from high density discharges in the ASDEX tokamak. *Plasma Physics and Controlled Fusion*, 26:689–702, 1984.
98. C. Carathéodory. Über das lineare Maß von Punktmengen – eine Verallgemeinerung des Längenbegriffs. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen*, pages 404–426, 1914.
99. H. Cardano. *Opera omnia*, chapter Liber de Ludo Aleae. Hugetan et Ravaud, Lyon, 1663. Manuscript 1525, facsimile reprint 1966 (Frommann–Holzboog, Stuttgart–Bad Cannstatt).
100. R.J. Carroll, D. Ruppert, and L.A. Stefanski. *Measurement Error in Nonlinear Models*. Number 63 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1995.
101. H. Caussinus. *Multidimensional Data Analysis*, chapter Models and Uses of Principal Component Analysis. DSWO Press, Leiden, 1986. J. De Leeuw et al (editors).
102. J.T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51:287–317, 1997.
103. A.V. Chankin, V.S. Mukhovatov, T. Fujita, and Y. Miura. Modelling of the current hole in JT–60U. *Plasma Physics and Controlled Fusion*, 45:323–336, 2003.
104. B.F. Chellas. *Modal Logic. An Introduction*. Cambridge University Press, 1980.
105. F.F. Chen. *Introduction to Plasma Physics and Controlled Fusion*. Plenum Press, second edition, 1984. First edition: 1974.
106. G. Chen, R.A. Lockart, and M.A. Stephens. Box–Cox transformations in linear models: Large sample theory and tests of normality (with discussion). *The Canadian Journal of Statistics*, 30:177–234, 2002.
107. R. Chin and S. Li. L–mode global energy confinement scaling for ion cyclotron heated tokamak plasmas. *Nuclear Fusion*, 32:951–965, 1992.
108. J.P. Christiansen, J.G. Cordey, O.J.W.F. Kardaun, and K. Thomsen. Application of plasma physics constraints to confinement data. *Nuclear Fusion*, 31:2117–2129, 1991.
109. J.P Christiansen, J.G. Cordey, and K. Thomsen. A unified physical scaling law for tokamak energy confinement. *Nuclear Fusion*, 30:1183–1196, 1990.
110. J.P. Christiansen, J. DeBoo, O.J.W.F. Kardaun, S.M. Kaye, Y. Miura, et al. Global energy confinement database for ITER (special topic). *Nuclear Fusion*, 32:291–338, 1992. Corrigendum 32:1281.
111. J.P. Cleave. *A Study of Logics*. Oxford Logic Guides. Clarendon Press, Oxford, 1991.
112. L. Cobb and S. Zacks. Applications of catastrophe theory for statistical modeling in the biosciences. *Journal of the American Statistical Association*, 80(392):793–802, 1985.
113. W.G. Cochran and G.M. Cox. *Experimental Designs*. Wiley, New York, second edition, 1992.
114. P. Cohen. *Set Theory and the Continuum Hypothesis*. W.A. Benjamin, 1966.
115. D.L. Cohn. *Measure Theory*. Birkhäuser, Boston, 1980.
116. J.W. Connor. Invariance principles and plasma confinement. *Plasma Physics and Controlled Fusion*, 30:619, 1988.
117. J.W. Connor. Edge-localised modes – physics and theory. *Plasma Physics and Controlled Fusion*, 40:531–542, 1998.

118. J.W. Connor. A review of models for ELMs. *Plasma Physics and Controlled Fusion*, 40:191–213, 1998.
119. J.W. Connor, R.J. Hastie, and J.B. Taylor. Shear, periodicity and plasma ballooning modes. *Physical Review Letters*, 40:396–399, 1978.
120. R.D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Number 18 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1982.
121. T.W. Copenhaver and P.W. Mielke. Quantit analysis: A quantal assay refinement. *Biometrics*, 33:175–186, 1977.
122. J.G. Cordey for the ITPA H-mode Database Working Group and the ITPA Pedestal Database Working Group. A two-term model of the confinement in ELMy H-modes using the global confinement and pedestal databases. *Nuclear Fusion*, 43:670–674, 2003.
123. J.G. Cordey for the ITER Confinement Database Working Group. Energy confinement scaling and the extrapolation to ITER. *Plasma Physics Controlled Fusion*, 39:B115–B127, 1997.
124. D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.
125. D.R. Cox and N. Reid. *Theory of the Design of Experiments*. Number 86 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 2000.
126. D.R. Cox and J.E. Snell. *Analysis of Binary Data*. Number 32 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, second edition, 1988. First edition: Methuen, London, 1970.
127. D.R. Cox and N. Wermuth. *Multivariate Dependencies: Models, Analysis and Interpretation*. Chapman and Hall, London, 1996.
128. D.R. Cox and N. Wermuth. Some statistical aspects of causality. *European Sociological Review*, 17:65–74, 2001.
129. G. Cox. Statistical frontiers. *Journal of the American Statistical Association*, 52:1–10, 1957. Reprinted in: Johnson, N.L. and Kotz, S. (editors), *Breakthroughs in Statistics*, Vol I, Springer–Verlag (1992) xvi–xli, with an editorial note and comments by G.A. Barnard, I.J Good, D.V. Lindley, F. Mosteller and P.K. Sen.
130. R.T. Cox. *The Algebra of Probable Inference*. John Hopkins University Press, Baltimore, 1961.
131. H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, 1946. Reprinted: 1999 (Princeton Landmarks in Mathematics and Physics).
132. R. D'Agostino and E.S. Pearson. Tests for departure from normality, empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika*, 60:613–622, 1973.
133. G. Darmois. Sur les lois de probabilité à estimation exhaustive. *Contes Rendues de l'Académie des Sciences à Paris*, pages 1265–1266, 1935.
134. O. Darrigol. *Electrodynamics from Ampère to Einstein*. Oxford University Press, first edition, 2000. First paperback edition, 2003.
135. S. Das Gupta. *R.A. Fisher: An Appreciation*, chapter Distribution of the Correlation Coefficient, pages 9–16. Number 1 in Lecture Notes in Statistics. Springer–Verlag, Heidelberg, 1980. S.E. Fienberg and D.V. Hinkley (editors).
136. P.L. Davies. Data features. *Statistica Neerlandica*, 49:185–245, 1995.

137. R. de Bruin, D. Salomé, and W. Schaafsma. A semi-Bayesian method for non-parametric density estimation. *Computational Statistics and Data Analysis*, 30:19–30, 1999.
138. B. de Finetti. *Teoria delle Probabilità*. Einaudi, 1970. English translation: Theory of Probability, Vols I & II, Wiley and Sons (1990), New York.
139. M.H. De Groot. *Optimal Statistical Decisions*. McGraw–Hill, New York, 1970.
140. A. de Moivre. *The Doctrine of Chances: or A Method of Calculating the Probability of Events in Play*. W. Pearson, London, 1718. Third edition: 1756, reprinted by Irvington publishers, 1967 and (with solutions) by the AMS, 2000.
141. C. Delacherie and P.-A. Meyer. *Probabilités et Potentiel*, volume I, II. Hermann, Paris, 1975, 1980.
142. L.D. Delwiche and S.J. Slaughter. *The Little SAS book: A Primer*. SAS Institute Inc., Cary, NC, 2nd edition, 1999.
143. A.W. DeSilva. The evolution of light scattering as a plasma diagnostic. *Contributions to Plasma Physics*, 40:23–35, 2000.
144. P.J. Dhrymes and S. Schwartz. On the invariance of estimators for singular systems of equations. *Greek Economic Review*, 9:231–250, 1987.
145. B. Diekmann and K. Heinloth. *Energie*. Teubner–Verlag, second edition, 1997.
146. F. Diener and M. Diener, editors. *Nonstandard Analysis in Practice*. Springer–Verlag, Heidelberg, 1995.
147. Dieudonné. *Abrégé d'Histoire des Mathématiques, 1700-1900*. Hermann, Paris, 1978. Translated into German by L. Boll et al., Vieweg, Wiesbaden, 1985.
148. J.B. Dijkstra. *Analysis of Means in some Non-Standard Situations*. PhD thesis, Technical University of Eindhoven, 1987. Also: CWI Tract 47, Mathematisch Centrum, Amsterdam, 1988.
149. J.B. Dijkstra. Trade regression. In Y. Dodge and Whittaker J., editors, *Computational Statistics X* (*Proc. 10th Symposium on Computational Statistics, Neuchâtel 1992*), volume I, pages 453–458, Heidelberg, 1992. Physica–Verlag.
150. W.J. Dixon, M.B. Brown, L. Engelman, M.-A. Hill, and R.I. Jennrich. *BMDP Statistical Software Manual*. University of California Press, Berkeley, CA, 1990. First edition (BMD): 1965.
151. Yu.N. Dnestrovskij and D.P. Kostomarov. Математическое Моделирование Плазмы. Физико-математическая литература, Nauka, Moscow, 2nd edition, 1993. First edition: *Numerical Simulation of Plasmas*, Springer–Verlag, Heidelberg, 1987.
152. G.G. Dolgov-Saveliev, V.S. Mukhovatov, V.S. Strelkov, M.N. Shepelev, and N.A. Yavlinksi. Investigation of a toroidal discharge in a strong magnetic field. In N.R. Nilsson, editor, *Proceedings of the Fourth International Conference on Ionization Phenomena in Gases, August 17–21, 1959*, volume II, Uppsala, 1960. North–Holland Publishing Company, Amsterdam.
153. E. Domany, J.L. van Hemmen, and K. Schulten. *Models of Neural Networks*, volume I, II, III, and IV. Springer–Verlag, Heidelberg, 1995–2001.
154. T. Dorlas. *Statistical Mechanics*. Institute of Physics Publishing, Bristol, 1999.
155. V. Dose, J. Neuhauser, B. Kurzan, H. Murmann, H. Salzmann, and ASDEX Upgrade Team. Tokamak edge profile analysis employing Bayesian statistics. *Nuclear Fusion*, 41:1671–1685, 2001.
156. E.J. Doyle, R.J. Groebner, K.H. Burell, et al. Modifications in turbulence and edge electric fields at the L–H transition in the DIII–D tokamak. *Physics of Fluids B*, 3:2300–2323, 1991.

157. N.R. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, third edition, 1998. First edition: 1966.
158. D. Dubois and H. Prade, editors. *Fundamentals of Fuzzy Sets*, volume 7 of *Handbooks of Fuzzy Sets*. Kluwer, Boston, 2000.
159. W.E. Dusenberry and K.O. Bowman. The moment estimator for the shape parameter of the gamma distribution. *Communications in Statistics – Simulation and Computation*, 6:1–19, 1977.
160. M.J. Dutch, F. Hofmann, B.P. Duval, et al. ELM control during double-null ohmic H-modes in TCV. *Nuclear Fusion*, 35:650–656, 1995.
161. M. Eaton. *Multivariate Statistics: A Vector Space Approach*. Wiley, New York, 1983.
162. H.-D. Ebbinghaus, H. Hermes, F. Firzebruch, M. Köcher, K. Mainzer, J. Neukirch, A. Prestel, and R. Remmert. *Zahlen*. Springer–Verlag, Heidelberg, third edition, 1992. First edition: 1983.
163. G. Edgar. *Integral, Probability and Fractal Measures*. Springer–Verlag, Heidelberg, 1998.
164. B. Efron. The geometry of exponential families. *Annals of Statistics*, 6:362–376, 1978.
165. B. Efron. Fisher in the 21st century (with discussion). *Statistical Science*, 13:95–122, 1998.
166. J. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, New York, 1975.
167. J. Elstrodt. *Maß– und Integrationstheorie*. Springer–Verlag, Heidelberg, 1998. First edition: 1996.
168. P. Embrechts, C. Kippelberg, and T. Mikosh. *Modelling Extremal Events*. Springer–Verlag, 1997.
169. W. Engelhardt. *Spectroscopy in Fusion Plasmas*. Commission of the European Communities, Directorate General XII – Fusion Programme, Brussels, 1982. EUR 8351–I EN.
170. W. Engelhardt. Wall effects and impurities in JET. *Philosophical Transactions of the Royal Society of London, Series A*, 322:79–94, 1987.
171. W. Engelhardt. Is a plasma diamagnetic? Unpublished manuscript, preprint available on request, Garching, 2000.
172. W. Engelhardt, W. Köppendörfer, and J. Sommer. Measurement of the depopulation of the 2 $^3P_{0,1,2}$ levels of heliumlike ions by electron collisions. *Physical Review A*, 6(2):1908–1914, 1972.
173. F. Engelmann. Extrapolation of tokamak energy confinement to next step devices. *Plasma Physics and Controlled Fusion*, pages 1101–1113, 1990.
174. F. Engelmann, N. Fujisawa, J. Luxon, V. Mukhovatov, et al. ITER: Physics R and D programme. *Proceedings of the 13th Conference on Plasma Physics and Controlled Nuclear Fusion Research, Washington 1990*, 3:435–442, 1991. IAEA–CN–53/F–III–18.
175. F. Engelmann, M.F.A. Harrison, R. Albanese, K. Borrass, O. De Barbieri, E.S. Hotston, A. Nocentini, J.-G. Wégrowe, and G. Zambotti. NET physics basis. In *NET Status Report*. Commission of the European Communities, Directorate General XII – Fusion Programme, Brussels, 1987. EUR–FU/XII-80/88-84.
176. F. Engelmann, O.J.W.F. Kardaun (coordinators), L. Pieroni, D.F. Düchs, N. Suzuki, D.R. Mikkelsen, Yu.V. Esipchuck, et al., FT Group, JET Team, JFT-2M Group, TFTR Team, and T-10 Team. Tokamak global confinement data (special topic). *Nuclear Fusion*, 30:1951–1978, 1990.

177. F. Engelmann for the NET Team. Concept and parameters for NET. *Proceedings of the 11th Conference on Plasma Physics and Controlled Nuclear Fusion Research, Kyoto 1986*, 3:249–257, 1987. IAEA–CN–47/H–1–1.
178. V. Erckman for the W7–AS Team et al. H–mode like transitions in the W7–AS stellarator with high power 140 GHz ECRH. *Proceedings of the 14th Conference on Plasma Physics and Controlled Nuclear Fusion Research, Würzburg 1992*, 2:469–481, 1993. IAEA–CN–56/C–1–8.
179. A. Erdélyi, W. Magnus, F. Oberhettinger, and F.G. Tricomi. *Higher Transcendental Functions*, volume I-III. McGraw Hill, 1953. (Bateman Manuscript Project).
180. H.-U. Fahrbach, O.J.W.F. Kardaun, J. Stober, Yu.N. Dnestrovskij, W. Herrmann, and A.V. Khutoretsky. Fast determination of T_i profiles from analysis of neutral flux measurements. In *Controlled Fusion and Plasma Heating,* (*Proc. 24th European Conference, Berchtesgaden, 1997*), volume 21A, pages 1501–1504, Geneva, 1997. European Physical Society.
181. L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer–Verlag, Heidelberg, second edition, 2001. First edition: 1994.
182. M. Falk, R. Becker, and F. Marohn. *Angewandte Statistik mit SAS.* Springer–Verlag, Heidelberg, 1995.
183. K.T. Fang, S. Kotz, and K.-W. Ng. *Symmetric Multivariate and Related Distributions.* Number 36 in Monographs on Statistics and Applied Probability. Chapman and Hall, New York, 1990.
184. A. Fasoli, J.A. Dobbing, C. Gormezano, J. Jacquinot, J.B. Lister, S. Sharapov, and A. Sibley. Alfvén eigenmode excitation by ICRH heat–waves. *Nuclear Fusion*, 36:258–263, 1996.
185. A. Fasoli, J.B. Lister, S. Sharapov, et al. Overview of Alfvén eigenmode experiments in JET. *Nuclear Fusion*, 35:1485–1495, 1995.
186. F. Faulbaum and W. Bandilla. *SoftStat–95: Advances in Statistical Software 5.* Lucius and Lucius, Stuttgart, 1996.
187. G. Federici et al. Key ITER plasma edge and plasma-material interaction issues. *Journal of Nuclear Materials*, 313–316:11–22, 2003.
188. G. Federici, C.H. Skinner, J.N. Brooks, et al. Plasma–material interactions in current tokamaks and their implications for next step fusion reactors. *Nuclear Fusion*, 41:1967–2137, 2001.
189. W. Feller. *An Introduction to Probability Theory and its Applications*, volume I. Wiley, third edition, 1968. First edition: 1950.
190. W. Feller. *An Introduction to Probability Theory and its Applications*, volume II. Wiley, second edition, 1991. First edition: 1966.
191. T.S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press, New York, 1967.
192. S.E. Fienberg and P.W. Holland. On the choice of flattening constants for estimating multinomial probabilities. *Journal of Multivariate Analysis*, 2:127–134, 1972.
193. T.L. Fine. *Theories of Probability: An Examination of Foundations.* Academic Press, New York, 1973.
194. D. Firth. *Statistical Theory and Modelling, In Honour of Sir David Cox, FRS*, chapter Generalized Linear Models. Chapman and Hall, London, 1991. Hinkley, D.V., Reid, N. and Snell, E.J. (editors).

195. G.H. Fischer and I.W. Molenaar, editors. *Rasch Models – Foundations, recent Developments and Applications*. Springer–Verlag, 1995.
196. N.I. Fisher, T. Lewis, and B.J.J. Embleton. *Statistical Analysis of Spherical Data*. Cambridge University Press, Cambridge, 1987.
197. R.A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10:507–521, 1915.
198. R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 222:309–368, 1922.
199. R.A. Fisher. On a distribution yielding the error functions of several well known statistics. In *Proceedings of the International Congress of Mathematics*, volume 2, pages 805–815, Toronto, 1924.
200. R.A. Fisher. Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 25:528–535, 1930.
201. R.A. Fisher. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London, Series A*, 144:285–307, 1934.
202. R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.
203. R.A. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, second edition, 1967. First edition: 1956.
204. R.A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, fourteenth edition, 1970. First edition: 1925.
205. R.A. Fisher. *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford University Press, Oxford, 1990. A re-issue of Statistical Methods for Research Workers, The Design of Experiments, Statistical Methods and Scientific Inference, edited by J.H. Bennett, with a foreword by F. Yates.
206. B.D. Flury. Estimation of principal points. *Applied Statistics*, 42:139–151, 1993.
207. B.D. Flury and H. Riedwyl. *Multivariate Statistics: A Practical Approach*. Chapman and Hall, London, 1988.
208. B. Francis, M. Green, and C. Payne, editors. *The GLIM System: Release 4 Manual*. Oxford University Press, Oxford, 1993.
209. D.A.S. Fraser. *The Structure of Inference*. Wiley, New York, 1968.
210. J.P. Friedberg. *Ideal Magnetohydrodynamics*. Plenum Press, 1987.
211. U. Frisch. *Turbulence*. Cambridge University Press, Cambridge, 1995.
212. A. Fujisawa. Experimental studies of structural bifurcation in stellarator plasmas. *Plasma Physics and Controlled Fusion*, 45:R1–R88, 2003.
213. T. Fujita, T. Oikawa, T. Suzuki, et al. Plasma equilibrium and confinement in a tokamak with nearly zero central current density in JT–60U. *Physical Review Letters*, 87:245001, 2001.
214. W.A. Fuller. *Measurement Error Models*. Wiley, New York, 1987.
215. W.K. Fung. Diagnosing influential observations in quadratic discriminant analysis. *Biometrics*, 52:1235–1241, 1996.
216. A.A. Galeev and R.S. Sagdeev. *Reviews of Plasma Physics*, volume 7, chapter Theory of Neoclassical Diffusion, pages 257–343. Consultants Bureau, New York, 1979. M.A. Leontovich, editor; translated from Russian by H. Lashinsky.
217. A.R. Gallant. *Nonlinear Statistical Models*. Wiley, New York, 1987.
218. F. Galton. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15:246–263, 1886.

219. F. Galton. *Natural Inheritance*. Macmillan, London, 1889.
220. J.W. Garson. *The Stanford Encyclopedia of Philosophy* (*Winter 2003 Edition*), *E.N. Zalta* (*editor*), chapter Modal Logic. The Metaphysics Research Laboratory at the Center for the Study of Language and Information, Stanford, CA, 2003. Available on internet, URL = http://plato.stanford.edu/archives/win2003/entries/logic-modal.
221. O. Gehre. The HCN–laser interferometer of the divertor tokamak ASDEX. *International Journal of Infrared and Millimeter Waves*, 5:369–379, 1984.
222. O. Gehre. Measurement of electron density profiles on ASDEX by HCN–laser interferometry. In *Basic and Advanced Techniques for Fusion Plasmas*, volume II, pages 399–407, Varenna, 1986. Commission of the European Communities, Directorate General XII – Fusion Programme. EUR 10797 EN.
223. A. Geier. *Aspekte des Verhaltens von Wolfram im Fusionsexperiment ASDEX Upgrade*. PhD thesis, Technische Universität München, 2001. Also: IPP Report 10/19.
224. S. Geisser. *Predictive Inference: An Introduction*. Number 55 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1993.
225. G.J. Gelpke and J.D.F. Habbema. *User's Manual for the INDEP–SELECT Discriminant Analysis Program*. Leiden University, Department of Medical Statistics, Leiden, 1981.
226. L. Giannone, A.C.C. Sips, O. Kardaun, F. Spreitler, W. Suttrop, and the ASDEX Upgrade Team. Regime identification in ASDEX Upgrade. *Plasma Physics and Controlled Fusion*, 46:835–856, 2004.
227. J.W. Gibbs. *Elementary Principles in Statistical Mechanics*. Edward Arnold, London, 1902. Reprinted: Dover, New York, 1960 and Ox Bow Press, Woodbridge, Conn., 1981.
228. P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization*. Academic Press, London, 1981.
229. N.W. Gillham. *A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics*. Oxford University Press, 2001.
230. R. Gilmore. *Catastrophe Theory for Scientists and Engineers*. Wiley, New York, 1981. Reprinted: Dover, New York, 1993.
231. I. Glazmann and Y. Liubitch. Конечномерный линейный анализ б задачах. Nauka, Moscow, 1972. Translation into French by H. Samadian: *Analyse Linéaire dans les Espaces de Dimensions Finies*, 1972, Mir, Moscow.
232. B.W. Gnedenko. Курс теории вероятностей. Nauka, Moscow, sixth edition, 1988. First edition: Государственное издателство технико-теоретической литературы, 1950, Москва; translated into English by B.R. Seekler: The Theory of Probability, Chelsea, 1962, New York; translated into German by H.-J. Roßberg and G. Laue: Einführung in die Wahrscheinlichkeitstheorie (mit einem Anhang des Hrsg. über positiv definite Verteilungsdichten), Akademie Verlag GmbH, 1991, Berlin.
233. R.J. Goldston. Energy confinement scaling in tokamaks: Some implications of recent experiments with ohmic and strong auxiliary heating. *Plasma Physics and Controlled Fusion*, 26:87–103, 1984.
234. R.J. Goldston and P.H. Rutherford. *Introduction to Plasma Physics*. Institute of Physics Publishing, 1995. German translation: 1998, Vieweg, Wiesbaden.
235. I.J. Good. *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press, Minneapolis, Minn., 1983.

236. I.J. Good. Some statistical applications of Poisson's work. *Statistical Science*, 1:157–170, 1986.
237. C. Gordon, H.-W. Bartels, T. Honda, M. Iseli, K. Moshonas, H. Okada, J. Raeder, N. Taylor, and L. Topilski. An overview of results in the ITER Generic Site Safety Report (GSSR). In *Fusion Energy 2002* (*Proc. 19th Int. Conf. Lyon, 2002*), Vienna, 2003. IAEA. IAEA-CN-94/CT/P-17, available on internet, URL=http://www.iaea.org/programmes/ripc/physics/fec2002/-html/node179.htm.
238. C. Gordon, H.-W. Bartels, T. Honda, M. Iseli, J. Raeder, L. Topilski, K. Moshonas, and N. Taylor. Lessons learned from the ITER safety approach for future fusion facilities. *Fusion Engineering and Design*, 54:397–403, 2001.
239. M. Gosh, N. Mukhopadhyay, and P.K. Sen. *Sequential Estimation*. Wiley, 1997.
240. S. Gottwald. *The Stanford Encyclopedia of Philosophy* (*Winter 2001 Edition*), *E.N. Zalta* (*editor*), chapter Many-Valued Logic. The Metaphysics Research Laboratory at the Center for the Study of Language and Information, Stanford, CA, 2001. Available on internet, URL = http://plato.stanford.edu/-archives/win2001/entries/logic-manyvalued.
241. H. Götze and H. Sarkowski. *Springer–Verlag: History of a Scientic Publishing House Part 1: 1842–1945. Foundation – Maturation – Adversity; Part 2: 1945–1992. Rebuilding – Opening Frontiers – Securing the Future*. Springer–Verlag, 1996. Translated into English by G. Graham and M. Schäfer.
242. Z. Govindarajulu. *The Sequential Statistical Analysis of Hypothesis Testing, Point and Interval Estimation and Decision Theory*. American Sciences Press Inc., Columbus, Ohio, 1981.
243. J.C. Gower. Multivariate analysis and multidimensional geometry. *Statistician*, 17:13–28, 1967.
244. I.S. Gradshteyn, I.M. Ryzhik, and A. Jeffrey. *Table of Integrals, Series, and Products*. Academic Press, sixth edition, 2000. First edition: Государственное Издателство, 1963.
245. M. Greenwald. Density limits in toroidal plasmas. *Plasma Physics and Controlled Fusion*, 44:R27–R80, 2002.
246. M. Greenwald, S. Ejima, M.G. Bell, G.H. Neilson, et al. A new look at density limits in tokamaks. *Nuclear Fusion*, 28:2199–2207, 1988.
247. J.A. Greenwood and D. Durand. Aids for fitting the gamma distribution by maximum likelihood. *Technometrics*, 2:55–65, 1960.
248. W. Greiner, L. Neise, and H. Stöcker. *Thermodynamics and Statistical Mechanics*. Springer–Verlag, Heidelberg, 1995.
249. G. Grieger, C. Beidler, H. Maaßberg, et al. Physics and engineering studies for Wendelstein 7–X. *Proceedings of the 13th Conference on Plasma Physics and Controlled Nuclear Fusion Research, Washington 1990*, 3:525–532, 1991. IAEA–CN–53/G–1–6.
250. J.E. Grizzle and D.M. Allen. Analysis of growth and dose response curves. *Biometrics*, 25:357–381, 1969.
251. R. Groebner, editor. *Special issue: Papers from the 9th IAEA Technical Committee Meeting on H–Mode Physics and Transport Barriers, California, USA, 24–26 September 2003*, volume 46 of *Plasma Physics and Controlled Fusion*. Institute of Physics Publishing, May 2004. Supplement 5A.

252. Academia Groningana. *MDCXIV–MCMXIV: gedenkboek ter gelegenheid van het derde eeuwfeest der universiteit te Groningen, uitgegeven in opdracht van den academischen senaat*. Noordhoff, Groningen, 1914 and 1916.
253. O. Gruber. Confinement regimes in ohmically and auxiliary heated Plasmas. In M.Q. Tran and R.J. Verbeek, editors, *Proceedings of the International Conference on Plasma Physics* (*ICPP*) – *Invited Papers*, volume I, pages 67–96, Lausanne, 1984. Commission of the European Communities / Centre de Recherches en Physique des Plasmas, Ecole Polytechnique Fédérale de Lausanne.
254. O. Gruber for the ASDEX Upgrade Team. Overview of ASDEX Upgrade results. *Nuclear Fusion*, 41:1369–1389, 2001.
255. J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer–Verlag, New York, second edition, 1997. First edition: Springer–Verlag, 1983.
256. W. Gulden, J. Raeder, and I. Cook. SEAFP and SEAL: Safety and environmental aspects. *Fusion Engineering and Design*, 51–52:419–434, 2000.
257. R.F. Gunst and J.W. White. Latent root regression: Large sample analysis. *Technometrics*, 21:481–488, 1979.
258. S. Günter, A. Gude, M. Maraschek, Q. Yu, and the ASDEX Upgrade Team. Influence of neoclassical tearing modes on energy confinement. *Plasma Physics and Controlled Fusion*, 41:767–774, 1999.
259. S. Günter, S. Schade, M. Maraschek, S.D. Pinches, E. Strumberger, R. Wolf, Q. Yu, and the ASDEX Upgrade Team. MHD phenomena in reversed shear discharges on ASDEX Upgrade. *Nuclear Fusion*, 40:1541–1548, 2000.
260. J.D.F. Habbema and G.J. Gelpke. A computer program for selection of variables in diagnostic and prognostic problems. *Computer Programs in Biomedicine*, 13:251–270, 1981.
261. T.S. Hahm. Physics behind transport barrier theory and simulations. *Plasma Physics and Controlled Fusion*, 44:A87–A101, 2002.
262. J. Hajek, Z. Šidak, and P.K. Sen. *Theory of Rank Tests*. Academic Press, 1999. First edition: 1967.
263. P.R. Halmos. *Measure Theory*. Van Nostrand Reinhold, Toronto, 1950. Reprinted: Springer–Verlag, 1974.
264. F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics, The Approach based on Influence Functions*. Wiley, New York, 1986.
265. G. Harding, A.N. Dellis, A. Gibson, B. Jones, D.J. Lees, R.W.P. McWhirter, S.A. Ramsden, and S. Ward. Diagnostic techniques used in controlled thermonuclear research at Harwell. In C. Longmire, J.L. Tuck, and W.B. Thompson, editors, *Plasma Physics and Thermonuclear Research*, volume 1, New York, 1959. Pergamon Press. Geneva Conference Paper (1958) P/1520.
266. W. Härdle. *Smoothing Techniques with Implementation in S*. Springer Series in Statistics. Springer–Verlag, New York, 1990.
267. T. Hastie and W. Stützle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.
268. T. Hastie and R. Tibshirani. *Generalized Additive Models*. Number 43 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1990.

269. T. Hatae, M. Sugihara, A.E. Hubbard, Yu. Igitkhanov, Y. Kamada, et al. Understanding of H–mode pedestal characteristics using the multimachine pedestal database. *Nuclear Fusion*, 41:285–293, 2001.
270. H. Haupt and W. Oberhofer. Fully resticted linear regression: A pedagogical note. *Economics Bulletin*, 3:1–7, 2002.
271. F. Hausdorff. Dimension und äußeres Maß. *Mathematische Annalen*, 79:157–179, 1919.
272. M.J.R. Healy. *GLIM: An Introduction*. Clarendon Press, 1988.
273. R. Hermann. *Fusion: The Search for Endless Energy*. Cambridge University Press, Cambridge, MA, 1991. Book reviewed by C.M. Braams in Nuclear Fusion **31** (1991) 2397-2398.
274. A. Herrmann. Overview on stationary and transient divertor heat loads. *Plasma Physics and Controlled Fusion*, 44:883–903, 2002.
275. A. Herrmann, W. Junker, K. Günther, S. Bosch, M. Kaufmann, J. Neuhauser, G. Pautasso, Th. Richter, and R. Schneider. Energy flux to the ASDEX Upgrade divertor plates determined by thermography and calorimetry. *Plasma Physics and Controlled Fusion*, 37:17–36, 1995.
276. F. Herrnegger, F. Rau, and H. Wobig (editors). Contributions to the Wendelstein 7-X and the Helias Reactor 1991–1998. Technical Report IPP 2 / 343, Max–Planck–Institut für Plasmaphysik, Garching, 1999. A collection of publications dedicated to Prof. Günter Grieger on the occasion of his 68th birthday.
277. L.T.M.E. Hillegers. *The Estimation of Parameters in Functional Relationship Models*. PhD thesis, Eindhoven Technical University, 1986.
278. K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments*. Wiley, New York, 1994.
279. F.L. Hinton and R.D. Hazeltine. Theory of plasma transport in toroidal plasmas. *Review of Modern Physics*, 48:239–308, 1976.
280. O. Hölder. Über einen Mittelwerthsatz. *Nachrichten von der königlichen Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, 1889:38–47, 1889.
281. P.W. Holland. Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81:945–970, 1986.
282. M. Hollander and D.A. Wolfe. *Nonparametric Statistical Methods*. Wiley, second edition, 1999. First edition: 1973.
283. A.F. Holleman, E.H. Büchner, E. Wiberg, and N. Wiberg. *Lehrbuch der Anorganischen Chemie*. Walter de Gruyter, 34th edition, 1995. First Dutch edition: 1898, first German edition: 1900, English translation: Academic Press, 2001.
284. J. Honerkamp. *Statistical Physics*. Springer–Verlag, second edition, 2002.
285. P.M. Hooper. Flexible regression modeling with adaptive logistic basis functions. *The Canadian Journal of Statistics*, 29:343–378, 2001.
286. W. Hoppe, W. Lohmann, H. Markl, and H. Ziegler, editors. *Biophysik, Ein Lehrbuch mit Beiträgen zahlreicher Fachwissenschaftler*. Springer–Verlag, Heidelberg, second edition, 1982. First edition: 1977.
287. L.D. Horton, T. Hatae, A. Hubbard, G. Janeschitz, Y. Kamada, B. Kurzan, L. Lao, P.J. McCarthy, D. Mossessian, T.H. Osborne, S.D. Pinches, S. Saarelma, M. Sugihara, W. Suttrop, K. Thomsen, and H. Urano. Dependence of H–mode pedestal parameters on plasma magnetic geometry. *Nuclear Fusion*, 44:A273–A278, 2002.

288. H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.
289. H. Hotelling. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society, Series B*, 15:193–232, 1953.
290. P. Hougaard. *Analysis of Multivariate Survival Analysis*. Springer–Verlag, Heidelberg, 2000.
291. K. Huang. *Statistical Mechanics*. Wiley, second edition, 1987. First edition: Wiley, 1963.
292. P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.
293. S. Huet, A. Bouvier, and E. Jolivet. *Statistical Tools for Nonlinear Regression*. Springer–Verlag, Heidelberg, second edition, 2003. First edition: 1996.
294. G.E. Hughes and M.J. Cresswell. *A New Introduction to Modal Logic*. Routledge, 1996.
295. J. Hugill. Transport in tokamaks – a review of experiment. *Nuclear Fusion*, 23:331–373, 1983.
296. K.M.S. Humak. *Statistical Inference in Linear Models*. Wiley, New York, 1986.
297. A. Hummel. *Bifurcations of Periodic Points*. PhD thesis, Groningen University, Groningen, 1979.
298. G.T.A. Huysmans, H.J. de Blank, W. Kerner, J.P. Goedbloed, and M.F.F. Nave. MHD stability models of edge localized modes in JET discharges. In *Controlled Fusion and Plasma Heating* (*Proc. 19th Eur. Conf. Innsbruck 1992*), volume 16B, Part I, pages 247–250, Geneva, 1993. European Physical Society.
299. K. Imre, K.S. Riedel, and B. Schunke. A hierarchy of empirical models of plasma profiles and transport. *Physics of Plasmas*, 2:1614–1622, 1995.
300. S. Inoue Itoh, K. Itoh, and Y. Terashima. A scaling law for high density tokamaks and its application to J.I.P.P. T–II device. Technical Report IPPJ–318, Institute of Plasma Physics (Nagoya University), Nagoya, 1977.
301. K. Ioki, V. Barabasch, A. Cardella, et al. Design and material selection for ITER first wall/blanket, divertor and vacuum vessel. *Journal of Nuclear Materials*, 258–263:74–84, 1998.
302. R.B. Israel. *Convexity in the Theory of Lattice Gases*. Princeton University Press, Princeton, NJ, 1979.
303. K. Itoh. Summary: Theory of magnetic confinement. *Nuclear Fusion*, 43:1710–1719, 2003. Dedicated to the memory of Prof. M. Wakatani and Prof. M.N. Rosenbluth.
304. K. Itoh, A. Fukuyama, Itoh S.-I., and M. Yagi. Self-sustained magnetic braiding in toroidal plasmas. *Plasma Physics and Controlled Fusion*, 37:707–713, 1995.
305. K. Itoh and S.-I. Itoh. The role of the electric field in confinement (review article). *Plasma Physics and Controlled Fusion*, 38:1–49, 1996.
306. K. Itoh, S.-I. Itoh, and A. Fukuyama. *Transport and Structural Formation in Plasmas*. Institute of Physics Publishing, 1999.
307. K. Itoh, S.-I. Itoh, and K. Fukuyama. The impact of improved confinement on fusion research. *Fusion Engineering and Design*, 15:297–308, 1992.
308. S.-I. Itoh and K. Itoh. Statistical theory and transition in multiple–scale–length turbulence in plasmas. *Plasma Physics and Controlled Fusion*, 43:1055–1102, 2001.

309. S.-I. Itoh, K. Itoh, and A. Fukuyama. Model of a giant ELM. *Plasma Physics and Controlled Fusion*, 38:1367–1371, 1996.
310. S.-I. Itoh, H. Maeda, and Y. Miura. Improved operating mode and the evaluation of confinement improvement. *Fusion Engineering and Design*, 15:343–352, 1992.
311. S.-I. Itoh, S. Toda, M. Yagi, K. Itoh, and A. Fukuyama. Physics of collapses: Probabilistic occurrence of ELMs and crashes. *Plasma Physics and Controlled Fusion*, 40:737–740, 1998.
312. J.D. Jackson. *Classical Electrodynamics*. Wiley, New York, third edition, 1998. First edition: 1962.
313. R. Jaenicke for the W7-AS Team. A new quasi-stationary, very high density plasma regime on the W7–AS stellarator. *Plasma Physics and Controlled Fusion*, 44:B193–B205, 2002.
314. E. Jahnke, F. Emde, and F. Lösch. *Tafeln höherer Funktionen*. Teubner–Verlag, Stuttgart, seventh edition, 1966. First edition: 1938.
315. M. Jammer. *Concepts of Mass in Classical and Modern Physics*. Harvard University Press, first edition, 1961. Reprinted: Dover, New York, 1997.
316. G. Janeschitz, A. Antipenkov, G. Federici, C. Ibbott, A. Kukushkin, P. Ladd, E. Martin, and R. Tivey. Divertor design and its integration to ITER. *Nuclear Fusion*, 42:14–20, 2002.
317. G. Janeschitz, P. Barabaschi, G. Federici, K. Ioki, P. Ladd, V. Mukhovatov, M. Sugihara, and R. Tivey. The requirements of a next-step large steady-state tokamak. *Nuclear Fusion*, 40:1197–1221, 2000.
318. K. Jänich. *Analysis für Physiker und Ingenieure*. Springer–Verlag, Heidelberg, fourth edition, 2001.
319. K. Jänich. *Lineare Algebra*. Springer–Verlag, Heidelberg, ninth edition, 2002. First edition: 1981.
320. A. Jaun, J. Vaclavik, and L. Villard. Stability of global drift-kinetic Alfvén eigenmodes in DIII-D. *Physics of Plasmas*, 4:1110–1116, 1997.
321. E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, 2003. Edited by G.L. Bretthorst.
322. T.J. Jech. *The Axiom of Choice*. North–Holland, Amsterdam, 1973.
323. H. Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, third edition, 1998. Re-issue from 1961; first edition: 1939.
324. R.I. Jennrich and M.D. Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42:805–820, 1986.
325. S. Johansen. Introduction to the theory of regular exponential families. Technical report, The Institute of Mathematical Statistics, Copenhagen, 1979.
326. S. Johansen. *Functional Relations, Random Coefficients and Non-Linear Regression – with Applications to Kinetic Data*, volume 22 of *Lecture Notes in Statistics*. Springer–Verlag, Heidelberg, 1983.
327. N.L. Johnson and S. Kotz, editors. *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*. Wiley, New York, 1997.
328. N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley, New York, second edition, 1994.
329. N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 2. Wiley, second edition, 1995.
330. N.L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. Wiley, New York, 1997.

331. N.L. Johnson, S. Kotz, and W. Kemp. *Discrete Univariate Distributions*. Wiley, New York, second edition, 1992.
332. I.T. Jolliffe. *Principal Component Analysis*. Springer–Verlag, Heidelberg, second edition, 2002.
333. B. Jörgensen. *Statistical Properties of the Generalized Inverse Gaussian Distribution*. Springer–Verlag, Heidelberg, 1982.
334. J. Jurečková and P.K. Sen. *Robust Statistical Inference: Asymptotics and Interrelations*. Wiley, New York, 1996.
335. B.B. Kadomtsev. *Tokamak Plasmas: A Complex Physical System*. Institute of Physics Publishing, 1992.
336. B.B. Kadomtsev and O.P. Pogutse. *Reviews of Plasma Physics*, volume 5, chapter Turbulence in Toroidal Systems, pages 249–400. Consultants Bureau, New York, 1970. M.A. Leontovich, editor.
337. J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, New York, second edition, 2002. First edition: 1980.
338. O. Kallenberg. *Random Measures*. Academic Press, New York, 1976.
339. O. Kallenberg. *Foundations of Modern Probability Theory*. Springer–Verlag, Heidelberg, second edition, 2002.
340. L.N. Kanal and J.F. Lemmer, editors. *Uncertainty in Artificial Intelligence*. North–Holland, Amsterdam, 1986.
341. J.W.P.F. Kardaun. *Contributions to a Rational Diagnosis and Treatment of Lumbar Disk Herniation*. PhD thesis, Erasmus University, Rotterdam, 1990. Also: ISBN 90–5166–139–8, Eburon, Delft, 1990.
342. J.W.P.F. Kardaun and O.J.W.F. Kardaun. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Methods of Information in Medicine*, 29:12–22, 1991.
343. O. Kardaun, P.J. McCarthy, K. Lackner, K.S. Riedel, and O. Gruber. A statistical approach to profile invariance. In *A. Bondeson, E. Sindoni, F. Troyon (editors): Theory of Fusion Plasmas*, pages 435–444, Bologna, 1988. Commission of the European Communities, Società Italiana di Fisica. EUR 11336 EN.
344. O. Kardaun, P.J. McCarthy, K. Riedel, and F. Wagner. IPP annual report 1987, section 1.5.4 (statistical analysis of plasma profiles). Technical report, Max–Planck–Institut für Plasmaphysik, Garching, 1988.
345. O. Kardaun, K. Thomsen, J. Christiansen, J. Cordey, N. Gottardi, M. Keilhacker, K. Lackner, P. Smeulders, and the JET Team. On global H–mode scaling laws for JET. In *Controlled Fusion and Plasma Heating, (Proc. 16th Eur. Conf., Venice, 1989)*, volume 13 B, Part I, pages 253–256, Geneva, 1990. European Physical Society.
346. O. Kardaun, K. Thomsen, J. Cordey, F. Wagner, the JET Team, and the ASDEX Team. Global H–mode scaling based on JET and ASDEX data. In *Controlled Fusion and Plasma Heating, (Proc. 17th Eur. Conf., Amsterdam, 1990)*, volume 14 B, Part I, pages 110–113, Geneva, 1991. European Physical Society.
347. O.J.W.F. Kardaun. *On Statistical Survival Analysis and its Applications in Medicine*. PhD thesis, Groningen University, Groningen, 1986. Chaps. 2 and 3 appeared as Chaps. 14 and 15 in: Handbook of Statistics, Vol. 8, Rao C.R. and Chakraborty, R. (editors), Elsevier, Amsterdam, 1991.

348. O.J.W.F. Kardaun. Scaling investigations and statistical profile analysis. IPP–Internet–Report IPP–IR 91/5 1.2, Max–Planck–Institut für Plasmaphysik, Garching bei München, 1991. Available on internet, URL = http://www.-ipp.mpg.de/netreports.
349. O.J.W.F. Kardaun. Interval estimation of global H–mode energy confinement in ITER. *Plasma Physics and Controlled Fusion*, 41:429–469, 1999.
350. O.J.W.F. Kardaun. On estimating the epistemic probability of realizing $Q = P_{fus}/P_{aux}$ larger than a specified lower bound in ITER. *Nuclear Fusion*, 42(7):841–852, 2002.
351. O.J.W.F. Kardaun, S.-I. Itoh, K. Itoh, and J.W.P.F. Kardaun. Discriminant analysis to predict the occurrence of ELMs in H-mode discharges. NIFS-Report 252, National Institute for Fusion Science, Toki-shi, Gifu-ken, Japan, 1993.
352. O.J.W.F. Kardaun, J.W.P.F. Kardaun, S.-I. Itoh, and K. Itoh. Discriminant analysis of plasma fusion data. In Y. Dodge and Whittaker J., editors, *Computational Statistics X* (*Proc. 10th Symposium on Computational Statistics, Neuchâtel 1992*), pages 163–170, Heidelberg, 1992. Physica–Verlag. Also: NIFS report 156.
353. O.J.W.F. Kardaun, J.W.P.F. Kardaun, S.-I. Itoh, and K. Itoh. Catastrophe–type models to fit non-linear plasma response functions. In *Controlled Fusion and Plasma Physics,* (*Proc. 25th Eur. Conf., Prague, 1998*), volume 22 C, pages 1975–1978, Geneva, 1998. European Physical Society. Available on internet, URL=http://epsppd.epfl.ch/Praha/WEB/AUTHOR_K.HTM.
354. O.J.W.F Kardaun and A. Kus. Basic probability theory and statistics for experimental plasma physics. Technical Report IPP 5/68, Max–Planck–Institut für Plasmaphysik, Garching, 1996.
355. O.J.W.F. Kardaun, A. Kus, and the H– and L–mode Database Working Group. Generalising regression and discriminant analysis: Catastrophe models for plasma confinement and threshold data. In A. Prat, editor, *Computational Statistics XII* (*Proc. 12th Symposium on Computational Statistics, Barcelona 1996*), pages 313–318, Heidelberg, 1996. Physica–Verlag.
356. O.J.W.F. Kardaun, Y. Miura, T. Matsuda, and H. Tamai. Introduction to SAS on VAX. Technical Report 91–098, JAERI–M, Ibaraki–ken, Naka–gun, Tokai–mura, Japan, 1991.
357. O.J.W.F. Kardaun, K.S. Riedel, P.J. McCarthy, and K. Lackner. A statistical approach to plasma profile analysis. Technical Report IPP 5/35, Max–Planck–Institut für Plasmaphysik, Garching, 1990.
358. O.J.W.F. Kardaun, D. Salomé, W. Schaafsma, A.G.M. Steerneman, J.C. Willems, and D.R. Cox. Reflections on fourteen cryptic issues concerning the nature of statistical inference (with discussion). *International Statistical Review*, 71(2):277–318, 2003.
359. O.J.W.F. Kardaun and W. Schaafsma. *Distributional Inference*. Preprint version, available on request, Groningen, 2003.
360. O.J.W.F. Kardaun for the H-mode Database Working Group. ITER: Analysis of the H–mode confinement and threshold databases. *Proceedings of the 14th Conference on Plasma Physics and Controlled Nuclear Fusion Research, Würzburg 1992*, 3:251–270, 1993. IAEA–CN–56/F–1–3.
361. O.J.W.F. Kardaun for the H-mode Database Working Group. Offset–linear scalings based on the ITER H–mode confinement database. In *Controlled Fu-*

sion and Plasma Physics, (*Proc. 21th Eur. Conf., Montpellier, 1994*), volume 18B, Part I, pages 90–94, Geneva, 1995. European Physical Society.
362. O.J.W.F. Kardaun for the International Confinement Database Working Group. Next step tokamak physics: Confinement–oriented global database analysis. In *Fusion Energy 2000* (*Proc. 18th Int. Conf. Sorrento, 2000*), Vienna, 2001. IAEA-CN-77-ITERP/04. CD-ROM (ISSN 1562–4153), also available on internet, URL=http://www.iaea.org/programmes/ripc/physics/fec2000/-html/node238.htm.
363. R.E. Kass and P.W. Vos. *Geometrical Foundations of Asymptotic Inference.* Wiley, New York, 1997.
364. V.J. Katz. *A History of Mathematics: An Introduction.* Addison Wesley, New York, second edition, 1998.
365. M. Kaufmann. *Plasmaphysik und Fusionsforschung: Eine Einführung.* Teubner–Verlag, Wiesbaden, 2003.
366. M. Kaufmann for the ASDEX Upgrade Team. Overview of ASDEX Upgrade results. *Proceedings of the 16th Conference on Fusion Energy, Montreal 1996*, 1:79–94, 1997. IAEA–CN–64/O1–5.
367. H. Kawamura, E. Ishitsuka, T. Tsuchiya, et al. Development of advanced blanket materials for a solid breeder blanket of a fusion reactor. *Nuclear Fusion*, 43:675–680, 2003.
368. S.M. Kaye, C.W. Barnes, and M.G. Bell. Status of global energy confinement studies. *Physics of Fluids B*, 2:2926–2940, 1990.
369. S.M. Kaye and R.J. Goldston. Global energy confinement scaling for neutral–beam–heated tokamaks. *Nuclear Fusion*, 25:65–69, 1985.
370. S.M. Kaye for the ITER Confinement Database Working Group. ITER L–mode confinement database (special topic). *Nuclear Fusion*, 37:1303–1330, 1997.
371. E.S. Keeping. *Introduction to Statistical Inference.* Van Nostrand Reinhold, New York, first edition, 1962. Reprinted: Dover, New York, 1995.
372. M. Keilhacker, A. Gibson, C. Gormezano, and P.-H. Rebut. The scientific success of JET. *Nuclear Fusion*, 41:1925–1966, 2001.
373. M. Keilhacker for the ASDEX Team. Confinement studies in L and H–type ASDEX discharges. *Plasma Physics and Controlled Fusion*, 26:49–63, 1984.
374. R. Kempter, W. Gerstner, and J.L. van Hemmen. Hebbian learning and spiking neurons. *Physical Review E*, 59:4498–4514, 1999.
375. M. Kendall. Conditions for uniqueness in the problem of moments. *Annals of Mathematical Statistics*, 11:402–409, 1940.
376. M. Kendall. *Multivariate Analysis.* Griffin, London, second edition, 1980.
377. M. Kendall and K. Ord. *Time Series.* Edward Arnold, 1990.
378. M. Kendall, A. Stuart, and K. Ord. *The Advanced Theory of Statistics*, volume I, II and III. Griffin, London, fourth edition, 1987.
379. W.S. Kendall. Computer algebra in probability and statistics. *Statistica Neerlandica*, 47:9–25, 1993.
380. W. Kerner and H. Tasso. Tearing mode stability for arbitrary current distribution. *Plasma Physics*, 24:97–107, 1982.
381. C.G. Khatri. A note on the MANOVA model applied to problems in growth curve. *Annals of the Institute of Statistical Mathematics*, 18:75–86, 1966.
382. A.I. Khinchin. *Mathematical Foundations of Statistical Mechanics.* Dover, New York, 1949.

383. A. Khutoretsky, H.-U. Fahrbach, O.J.W.F. Kardaun, J. Stober, W. Herrmann, and Yu.N. Dnestrovskij. Recovery of ion temperature profiles from the analysis of energy-resolved neutral flux measurements. Technical Report 5/99, Max–Planck–Institut für Plasmaphysik, 2002. Also: Volkswagen Grant Report 1/70032.
384. V.I. Khvesyuk and A.Yu. Chirkov. Low–radioactivity D–^{3}He fusion fuel cycles with ^{3}He production. *Plasma Physics and Controlled Fusion*, 44:253–260, 2002.
385. M. Kline. *Mathematical Thought from Ancient to Modern Times*. Oxford University Press, New York, 1972.
386. G.J. Klir and T.A. Folger. *Fuzzy Sets, Uncertainty, and Information*. Prentice–Hall, Englewood Cliffs, NJ, 1988.
387. S. Klose, W. Bohmeyer, M. Laux, H. Meyer, G. Fußmann, and the PSI-team. Investigation of ion drift waves in the PSI–2 using Langmuir–probes. *Contributions to Plasma Physics*, 41:467–472, 2001.
388. Ya.L. Koleshnichenko, V.V. Parail, and G.V. Pereverzev. *Reviews of Plasma Physics*, volume 17, chapter Generation of Noninductive Current in a Tokamak, pages 1–191. Consultants Bureau, New York, 1992. M.A. Leontovich, editor.
389. A. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer–Verlag, Heidelberg, first edition, 1933. Reprinted: Springer-Verlag, 1973. English translation: Foundations of Probability Theory, 1950, 1960, Chelsea, New York.
390. A. Kolmogorov and S. Fomine. Элементы теории функций и функционалъного анализа. Nauka, Moscow, 1957. Translated into French by M. Dragnev: *Eléments de la Théorie des Fonctions et de l'Analyse Fonctionelle*, 1974, Mir, Moscow. English edition: 1975, Dover.
391. B.P. Kooi. The Monty Hall dilemma. Master's thesis, Department of Philosophy, Groningen University, 1999.
392. B.O. Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39:399–409, 1936.
393. E.W. Kopf. Florence Nightingale as statistician. *Publications of the American Statistical Association*, 15:388–404, 1916.
394. W. Köppendörfer for the ASDEX Upgrade Team. Results of the first operational phase of ASDEX Upgrade. *Proceedings of the 14th Conference on Plasma Physics and Controlled Nuclear Fusion Research, Würzburg 1992*, 1:127–140, 1993. IAEA–CN–56/A–2–3.
395. S. Kotz, N. Balakrishnan, and N.L. Johnson. *Continuous Multivariate Distributions*. Wiley, second edition, 2000. First edition (by Johnson and Kotz): Wiley, 1972.
396. S. Kotz and N.L. Johnson, editors. *Breakthroughs in Statistics*, volume I, II and III. Springer–Verlag, Heidelberg, 1993–1997.
397. C.J. Kowalski. On the effects of non-normality on the distribution of the sample product–moment correlation coefficient. *Applied Statistics*, 21:1–12, 1972.
398. A. Krause and M. Olsen. *The Basics of S and S–PLUS*. Springer–Verlag, second edition, 2000.
399. P.R. Krishnaiah and L.N. Kanal, editors. *The Handbook of Statistics*, volume 1. North–Holland, Amsterdam, 1982.

400. A.H. Kroese. *Distributional Inference: A Loss Function Approach*. PhD thesis, Groningen University, Groningen, 1994.
401. A.H. Kroese, E.A. Van der Meulen, K. Poortema, and Schaafsma W. Distributional inference. *Statistica Neerlandica*, 49:63–82, 1995.
402. W.J. Krzanowksi and F.H. Marriott. *Multivariate Analysis, Part I: Ordination and Inference*. Kendall's Advanced Theory of Statistics. Wiley, New York, 1994.
403. W.J. Krzanowksi and F.H. Marriott. *Multivariate Analysis, Part II: Classification, Covariance Structures and Repeated Meassurements*. Kendall's Advanced Theory of Statistics. Wiley, New York, 1995.
404. A.M. Kshirsagar. *Multivariate Analysis*. Marcel Dekker, New York, 1972.
405. A.M. Kshirsagar and W.B. Smith. *Growth Curves*. Marcel Dekker, New York, 1995.
406. A.S. Kukushkin, H.D. Pacher, G. Janeschitz, A. Loarte, D.P. Coster, G. Matthews, D. Reiter, and R. Schneider. Basic divertor operation in ITER–FEAT. *Nuclear Fusion*, 42:187–191, 2002.
407. H.E. Kyburg Jr. *The Logical Foundations of Statistical Inference*. Reidel, Dordrecht, 1974.
408. P.A. Lachenbruch. *Discriminant Analysis*. Hafner, New York, 1975.
409. K. Lackner. Computation of ideal MHD equilibria. *Computer Physics Communications*, 12:33–44, 1976.
410. K. Lackner, R. Andreani, D. Campbell, M. Gasparotto, D. Maisonnier, and M.A. Pick. Long-term fusion strategy in Europe. *Journal of Nuclear Materials*, 307–311:10–20, 2002.
411. K. Lackner, O. Gruber, F. Wagner, et al. Confinement regime transitions in ASDEX. *Plasma Physics and Controlled Fusion*, 31:1629–1648, 1989.
412. K. Lackner and the ASDEX Upgrade Team. Recent results from divertor operation in ASDEX Upgrade. *Plasma Physics and Controlled Fusion*, 36:B79–B92, 1994.
413. K. Lackner and W. Wobig. A comparison of energy confinement in tokamaks and stellarators. *Plasma Physics and Controlled Fusion*, 29:1187–1204, 1987.
414. J.W. Lamperti. *Stochastic Processes: A Survey of the Mathematical Theory*. Springer–Verlag, Heidelberg, 1977.
415. D. Landers and L. Rogge. *Nichtstandard Analysis*. Springer–Verlag, Heidelberg, 1994.
416. S. Lang. *Algebra*, volume 211 of *Graduate Texts in Mathematics*. Springer–Verlag, New York, revised third edition, 2002.
417. R. Lange, T.A. Oliva, and S.R. McDade. An algorithm for estimating multivariate catastrophe models: GEMCAT II. *Studies in Nonlinear Dynamics and Econometrics*, 4(3), 2000.
418. P. Langevin. Magnétisme et théorie des électrons. *Annales de Chémie et de Physique*, 5:70, 1905.
419. P.S. Laplace. Mémoire sur la probabilité des causes par évènements. *Mémoires de l'Académie Royale des Sciences présentés par divers Savans et lus dans ses Assemblées*, 6:621–656, 1774. Reprinted in: *Oeuvres Complètes*, 8: 27–65.
420. P.S. Laplace. *Théorie Analytique des Probabilités*. Courcier, Paris, third edition, 1820. First edition: Courcier, Paris, 1812, Reprinted: Editions Jacques Gabay, 1995, Paris.
421. K.D. Lawrence and J.L. Arthur. *Robust Regression*. Marcel Dekker, New York, 1990.

422. L. LeCam. The central limit theorem around 1935. *Statistical Science*, 1:78–91, 1986.
423. C. Lechanoine-Leluc, M. Martin, and H. Wind. Method for the determination of the momentum of a particle from spark chamber data used in connection with an analysing magnet. Technical report, CERN, Geneva, 1968.
424. E.L. Lehmann. *Testing Statistical Hypotheses*. Springer–Verlag, Heidelberg, third edition, 1997. First edition: Wiley, 1959.
425. E.L. Lehmann. *Nonparametrics: Statistical Methods based on Ranks*. Prentice Hall, New York, revised, 1/e edition, 1998.
426. E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer–Verlag, Heidelberg, second edition, 1998. First edition: Wiley, 1983.
427. E.L. Lehmann and H. Scheffé. Completeness, similar regions and unbiased estimation. *Sankhyā*, 10:305–340 and 15:219–236, 1950 and 1955. Correction: Sankhyā 17:250, 1956.
428. G.G. Letac. Introduction to Morris (1982) natural exponential families with quadratic variance functions. In S. Kotz and N.L. Johnson, editors, *Breakthroughs in Statistics, Vol. III*, Springer Series in Statistics, pages 369–373. Springer–Verlag, 1997. C.N. Morris' original paper: *Annals of Statistics*, 10: 65-80, 1982.
429. R.A. Lew. An approximation to the cumulative normal distribution with simple coefficients. *Applied Statistics*, 30:299–301, 1981.
430. H.N. Linssen and L.T.M.E. Hillegers. Approximate inference in multivariate nonlinear functional relationships. *Statistica Neerlandica*, 43:141–156, 1989.
431. G.G. Lister. FAFNER: a fully 3–D neutral beam injection code using Monte Carlo methods. Technical Report IPP 4/222, Max–Planck–Institut für Plasmaphysik, Garching, 1985. Modified for ASDEX Upgrade by A. Teubel.
432. M. Loève. *Probability Theory I, II*. Springer–Verlag, New York, fourth edition, 1977. First edition: Van Nostrand, Princeton, 1955.
433. H. Maaßberg, R. Brakel, R. Burhenn, U. Gasparino, P. Grigull, M. Kick, G. Kühner, H. Ringler, F. Sardei, U. Stroth, and A. Weller. Transport in stellarators. *Plasma Physics and Controlled Fusion*, 35:B319–B332, 1993.
434. H.L. MacGillivray. Skewness and asymmetry: Measures and orderings. *Annals of Statistics*, 14:994–1011, 1986.
435. J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, New York, second edition, 1999. First edition: 1988.
436. W. Magnus, F. Oberhettinger, and R.P. Soni. *Formulas and Theorems for the Special Functions of Mathematical Physics*. Springer–Verlag, Heidelberg, 1966.
437. A.A. Mamun. Nonlinear propagation of ion-acoustic waves in a hot magnetized plasma with vortexlike electron distribution. *Physics of Plasmas*, 322–324, 1998.
438. M. Maraschek, S. Günter, T. Kass, B. Scott, H. Zohm, and the ASDEX Upgrade Team. Observation of toroidicity–induced Alfvén eigenmodes in Ohmically heated plasmas by drift wave excitation. *Physical Review Letters*, 79:4186–4189, 1997.
439. F.B. Marcus et al. A power step-down approach to extended high performance of ELM-free H-modes in JET. *Nuclear Fusion*, 37:1067–1080, 1997.
440. K.V. Mardia. Assessment of multinormality and the robustness of Hotelling's T^2 test. *Applied Statistics*, 24:163–171, 1975.

441. K.V. Mardia and P.E. Jupp. *Directional Statistics*. Wiley, 1999.
442. K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.
443. T. Marshall, M.T. Porfiri, L. Topilski, and B. Merill. Fusion safety codes: International modeling with MELCOR and ATHENA–INTRA. *Fusion Engineering and Design*, 63–64:243–249, 2002.
444. W.F. Massey. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60:234–256, 1965.
445. R. Matera and G. Federici. Design requirements for plasma facing materials in ITER. *Journal of Nuclear Materials*, 233–237:17–25, 1996.
446. R. Mathis and J. Sapper. Design and engineering aspects of the main components for the Wendelstein VII-AS stellarator experiment. *Fusion Engineering and Design*, 11:399–422, 1990.
447. G. Matthews, editor. *Special Issue on ELMs*, volume 45 of *Plasma Physics and Controlled Fusion*. Institute of Physics Publishing, September 2003.
448. P. Mattila. *Geometry of Sets and Measures in Euclidean Spaces*. Cambridge University Press, Cambridge, 1995.
449. P.J. McCarthy. *An Integrated Data Interpretation System for Tokamak Discharges*. PhD thesis, University College Cork, Cork, 1992.
450. P.J. McCarthy, P. Martin, and Schneider W. The CLISTE interpretative equilibrium code. Technical Report IPP 5/85, Max–Planck–Institut für Plasmaphysik, 1999.
451. P.J. McCarthy, K.S. Riedel, O.J.W.F. Kardaun, H. Murmann, and K. Lackner. Scalings and plasma profile parameterisation of ASDEX high density ohmic discharges. *Nuclear Fusion*, 31:1595–1634, 1991.
452. P.J. McCarthy and M.C. Sexton. Plasma profile recovery by function parameterisation. Technical Report 5/12, Max–Planck–Institut für Plasmaphysik, 1986.
453. K. McCormick, P. Grigull, et al. New advanced operational regime on the W7–AS stellarator. *Physical Review Letters*, 89:015001–1–015001–4, 2002.
454. P. McCullagh. *Tensor Methods in Statistics*. Number 29 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1987.
455. P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Number 37 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, second edition, 1989.
456. D.C. McDonald, J.G. Cordey, E. Righi, F. Ryter, G. Saibene, R. Sartori, et al. ELMy H–modes in JET helium-4 plasmas. *Plasma Physics and Controlled Fusion*, 46:519–534, 2004.
457. G.J. McLachan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.
458. R.W.P. McWhirter. *Plasma Diagnostic Techniques*, chapter Spectral Intensities, pages 201–264. Academic Press, New York, 1965. R.H. Huddlestone and S.L. Leonard (editors).
459. D.M. Meade, C.E. Kessel, G.W. Hammett, S.C. Jardin, M.A. Ulrickson, P. Titus, et al. Exploration of burning plasmas in FIRE. In *Fusion Energy 2002* (*Proc. 19th Int. Conf. Lyon, 2002*), Vienna, 2003. IAEA. IAEA-CN-94/FT-2-6, available on internet, URL=http://www.iaea.org/programmes/ripc/physics/fec2002/html/node340.htm.
460. D. Merkl. *'Current Holes' and other Structures in Motional Stark Effect Measurements*. PhD thesis, Technische Universität München, 2004.

461. V. Mertens, K. Borrass, J. Gafert, M. Laux, J. Schweinzer, and ASDEX Upgrade Team. Operational limits of ASDEX Upgrade H–mode discharges in the new closed divertor II configuration. *Nuclear Fusion*, 40:1839–1843, 2000.
462. V. Mertens et al. High density operation close to Greenwald limit and H–mode limit in ASDEX Upgrade. *Nuclear Fusion*, 37:1607–1614, 1997.
463. A.J. Miller. *Subset Selection in Regression*. Number 95 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, second edition, 2002. First edition: 1990.
464. R.L. Miller, Y.R. Lin-Liu, T.H. Osborne, and T.S. Taylor. Ballooning mode stability for self-consistent pressure and current profiles at the H–mode edge. *Plasma Physics and Controlled Fusion*, 40:753–756, 1998.
465. R.G. Miller Jr. *Simultaneous Statistical Inference*. Springer–Verlag, Heidelberg, second edition, 1981. First edition: 1966.
466. R.G. Miller Jr. *Beyond ANOVA*. Wiley, New York, 1986. Reprinted: Chapman and Hall, 1997.
467. H. Minkowski. *Geometrie der Zahlen*. Teubner–Verlag, Leipzig, first edition, 1896, 1910. Reprinted: Johnson, New York, 1968.
468. S.V. Mirnov and I.B. Semenov. Investigation of the plasma string in the tokamak–3 system by means of a correlation method. Атомная Энергия, 30:20–27, 1971. English translation: Soviet Atomic Energy, 30, 22–29.
469. Y. Miura, H. Aikawa, K. Hoshino, et al. Studies on improved confinement on JFT–2M. *Proceedings of the 13th Conference on Plasma Physics and Controlled Nuclear Fusion Research, Washington 1990*, 1:325–333, 1991. IAEA–CN–53/A–IV–6.
470. Y. Miura, T. Takizuka, H. Tamai, T. Matsuda, N. Suzuki, M. Mori, H. Maeda, K. Itoh, S.-I. Itoh, and O.J.W.F. Kardaun. Geometric dependence of the energy confinement time scaling for H–mode discharges. *Nuclear Fusion*, 32:1473–1479, 1992.
471. K. Miyamoto. *Plasma Physics for Nuclear Fusion*. MIT Press, Cambridge MA, second edition, 1989. First edition (in Japanese): Iwanami Shoten, Tokyo, 1976.
472. R.W. Moir. Liquid first wall for magnetic fusion energy configurations. *Nuclear Fusion*, 37:557–566, 1997.
473. D.C. Montgomery, E.A. Peck, and G.G. Vining. *Introduction to Linear Regression Analysis*. Wiley, New York, third edition, 2001.
474. D.S. Moore. *Studies in Statistics*, volume 19 of *Studies in Mathematics*, chapter Chi–Square Tests, pages 66–106. The Mathematical Association of America, 1978. R.V. Hogg (editor).
475. G.H. Moors. *Zermelo's Axiom of Choice: Its Origin, Development and Influence*. Springer–Verlag, Heidelberg, 1982.
476. J.J.A. Moors. A quantile alternative for kurtosis. *Statistician*, 37:25–32, 1988.
477. J.J.A. Moors, T.Th.A. Wagemakers, V.M.J. Coenen, R.M.J. Heuts, and M.J.B.T. Janssens. Characterising systems of distributions by quantile measurements. *Statistica Neerlandica*, 50:417–430, 1996.
478. I. Moret, G. Capodaglio, G. Scarponi, and M. Romanazzi. Statistical evaluation of the group structures of five Venetian wines from chemical measurements. *Analytica Chemica Acta*, 191:331–350, 1986.
479. B.J.T. Morgan. *Analysis of Quantal Response Data*. Number 46 in Monographs on Applied Statistics and Applied Probability. Chapman and Hall, 1992.

480. D.F. Morrison. *Multivariate Statistical Methods*. McGraw–Hill, New York, third edition, 1990.
481. F. Mosteller and T.C. Chalmers. Meta–analysis: Methods for combining independent studies. Some progress and problems in meta–analysis of clinical trials. *Statistical Science*, 7:227–236, 1992.
482. V.S. Mukhovatov. Итоги науки и техники, серия физика Плазмы, chapter Токамаки. Академия наук СССР, Москва, 1980. Book 1, Part 1.
483. V.S. Mukhovatov, D. Boucher, N. Fujisawa, G. Janeschitz, V. Leonov, H. Matsumoto, A. Polevoi, M. Shimada, and G. Vayakis. RTO/RC ITER plasma performance: inductive and steady-state operation. *Plasma Physics and Controlled Fusion*, 42:A223–A230, 2000.
484. V.S. Mukhovatov, M. Shimada, A.N. Chudnovskij, A.E. Costley, Y. Gribov, G. Federici, O. Kardaun, A.S. Kukushkin, A. Polevoi, V.D. Pustinov, Y. Shimomura, T. Sugie, M. Sugihara, and G. Vayakis. Overview of physics basis for ITER. *Plasma Physics and Controlled Fusion* (*special issue*), 45:A235–A252, 2003.
485. V.S. Mukhovatov, Y. Shimomura, A. Polevoi, M. Shimada, M. Sugihara, G. Bateman, J. Cordey, O. Kardaun, G. Pereverzev, I. Voitsekhovich, J. Weiland, O. Zolotukhin, A. Chudnovskij, A. Kritz, A. Kukushkin, T. Onjun, A. Pankin, and F.W. Perkins. Comparison of ITER performance predicted by semi-empirical and theory-based transport models. *Nuclear Fusion*, 43:942–948, 2003.
486. M. Murakami, J.D. Callen, and L.A. Berry. Some observations on maximum densities in tokamak experiments. *Nuclear Fusion*, 16:347–348, 1976.
487. R.H. Myers and D.C. Montgomery. *Response Surface Methodology: Process and Product Optimization using Designed Experiments*. Wiley, second edition, 2002. First edition: 1995.
488. J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statististical Society A*, 135:370–384, 1972.
489. A. Netushil, editor. *Theory of Automatic Control*. Mir, Moscow, 1973. English translation by A. Parnakh. Original Russian edition: Теория автоматического управления, 1972, Mir, Moscow.
490. R. Neu. Tungsten as a plasma facing material in fusion devices. Technical Report IPP 10/25, Max–Planck–Institut für Plasmaphysik, Garching, 2003. Habilitation Thesis, Tübingen University.
491. R. Neu, K. Asmussen, G. Fussmann, P. Geltenbort, G. Janeschitz, K. Schönmann, U. Schramm, G. and Schumacher, and the ASDEX Upgrade Team. Monitor for the carbon and oxygen impurities in the ASDEX Upgrade tokamak. *Review of Scientific Instruments*, 67(5):1829–1833, 1996.
492. J. Neuhauser, H.-S. Bosch, D. Coster, A. Herrmann, and A. Kallenbach. Edge and divertor physics in ASDEX Upgrade. *Fusion Science and Technology*, 44:659–681, 2003. (Special issue on ASDEX Upgrade, edited by A. Herrmann).
493. J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A*, 236:333–380, 1937.
494. J. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, 231:289–337, 1933. Reprinted in: Johnson, N.L. and Kotz,

S. (editors), *Breakthroughs in Statistics*, Vol I, Springer–Verlag (1992) 73–108, with an introduction by E.L. Lehmann.
495. Data Analysis Division of MathSoft Inc. *S–PLUS 6.0 Guide to Statistics for UNIX/Linux*, volume 1,2. MathSoft Inc., Seattle, WA, 2000.
496. A. O'Hagan. *Kendall's Advanced Theory of Statistics*, volume 2B: Bayesian Inference. Edward Arnold, London, 1994.
497. M.O. Ojo and A.K. Olapade. On the generalized logistic and log-logistic distributions. *Kragujevac Journal of Mathematics*, 25:65–73, 2003.
498. T. A. Oliva, W.S. Desarbo, D.L. Day, and K. Jedidi. GEMCAT: A general multivariate methodology for estimating catastrophe models. *Behavioral Science*, pages 121–137, 1987.
499. D.V. Orlinskij and G. Magyar. Plasma diagnostics on large tokamaks (review paper). *Nuclear Fusion*, 28:611–697, 1988.
500. S. Ortolani. Reversed field pinch confinement physics. *Plasma Physics and Controlled Fusion*, 31:1665–1683, 1989.
501. W. Ott, E. Speth, and A. Stäbler. Slowing-down of fast ions in a plasma: Energy transfer, charge exchange losses and wall sputtering. Technical report, Max–Planck–Institut für Plasmaphysik, 1977. Report 4/161.
502. E. Page. Approximations to the cumulative normal function and its inverse for use on a pocket calculator. *Applied Statistics*, 26:75–76, 1977.
503. V.V. Parail. Energy and particle transport in plasmas with transport barriers. *Plasma Physics and Controlled Fusion*, 44:A63–A85, 2002.
504. K.R. Parthasarathy. *Probability Measures on Metric Spaces.* Academic Press, New York, 1967.
505. G. Pautasso, K. Büchl, J.C. Fuchs, O. Gruber, A. Herrmann, K. Lackner, P.T. Lang, K.F. Mast, M. Ulrich, and H. Zohm. Use of impurity pellets to control energy dissipation during disruption. *Nuclear Fusion*, 36:1291–1297, 1996.
506. G. Pautasso, A. Herrmann, and K. Lackner. Energy balance during disruption associated with vertical displacement events. *Nuclear Fusion*, 34:455–458, 1994.
507. G. Pautasso, C. Tichmann, S. Egorov, T. Zehetbauer, et al. On-line prediction and mitigation of disruptions in ASDEX Upgrade. *Nuclear Fusion*, 42:100–108, 2002.
508. K. Pearson. Contributions to the mathematical theory of evolution, part II: Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London*, 186:343–414, 1895.
509. K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.
510. K. Pearson. On the distribution of severity of attack in cases of smallpox. *Biometrika*, 4:505–510, 1906.
511. K. Pearson. The fundamental problem of practical statistics. *Biometrika*, 13:1–16, 1920.
512. F.W. Perkins, A. Bondeson, R.J. Buttery, J.D. Callan, J.W. Connor, et al. Neoclassical islands, β–limits, error fields and ELMs in reactor scale tokamaks. *Nuclear Fusion*, 39:2051–2054, 1999.
513. S.E. Pestchanyi, H. Würz, and I.S. Landman. Impurity production and edge plasma pollution during ITER–FEAT ELMs. *Plasma Physics and Controlled Fusion*, 44:845–853, 2002.
514. V.V. Petrov. *Limit Theorems of Probability Theory.* Clarendon Press, Oxford, 1995.

515. J. Pfanzagl. Characterizations of conditional expectations. *Annals of Mathematical Statistics*, 38:415–421, 1967.
516. E.J.G. Pitman. Sufficient statistics and intrinsic accuracy. *Proceedings of the Cambridge Philosophical Society*, 32:567–579, 1936.
517. S.D. Poisson. *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile, précédés des Règles Générales du Calcul des Probabilités*. Bachelier, Paris, 1837.
518. J.W. Polderman and J. Willems. *Introduction to Mathematical Systems Theory: A Behavioral Approach*. Springer–Verlag, Heidelberg, 1998.
519. K. Poortema. *On the Statistical Analysis of Growth*. PhD thesis, Groningen University, Groningen, 1989.
520. K.R. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1972. First edition: Logik der Forschung, Julius Springer, Wien, 1935, tenth edition: Mohr, Tübingen, 1994.
521. T. Poston and I. Stuart. *Catastrophe Theory and its Applications*. Pitman, London, 1978. Reprinted: Dover, New York, 1996.
522. R. Potthoff and S.N. Roy. A generalized multivariate analysis of variance models useful especially for growth curve problems. *Biometrika*, pages 313–326, 1964.
523. R. Pratap. *Getting Started with MATLAB: A Quick Introduction for Scientists and Engineers*. Oxford University Press, Oxford, 2002.
524. R.L. Prentice. A generalisation of the probit and logit methods for dose response curves. *Biometrics*, 32:761–768, 1976.
525. R. Preuss, V. Dose, and W. von der Linden. Dimensionally exact form-free energy confinement scaling in W7–AS. *Nuclear Fusion*, 39:849–862, 1999.
526. B.N. Pšeničnyi and Ju.M. Danilin. *Numerische Methoden für Extremalaufgaben*. VEB, 1982.
527. F. Pukelsheim. *Optimal Design of Experiments*. Wiley, New York, 1993.
528. S.G. Rabinovich. *Measurement Errors and Uncertainties*. Springer–Verlag, New York, second edition, 1999.
529. J. Raeder, K. Borrass, R. Bünde, W. Dänner, R. Klingelhöfer, L. Lengyel, F. Leuterer, and M. Soll. *Kontrollierte Kernfusion*. Teubner–Verlag, Stuttgart, 1981.
530. J. Raeder for the SEAFP Team. Report on the European safety and environmental assessment of fusion power (SEAFP). *Fusion Engineering and Design*, 29:121–140, 1995.
531. J. Raeder for the SEAFP Team. Safety and environmental assessment of fusion power (SEAFP), report of the SEAFP project. Commission of the European Communities, Directorate General XII – Fusion Programme, Brussels, June 1995. EURFUBRU XII-217/95.
532. A.R. Raffray, G. Federici, V. Barabasch, et al. Beryllium application in ITER plasma facing components. *Fusion Engineering and Design*, 1:261–286, 1997.
533. J. Ranke. *Beiträge zur Anthropologie und Urgeschichte Baierns*, volume VIII, chapter Beiträge zur physischen Anthropologie der Baiern. Riedel, München, 1883.
534. C.R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945. Reprinted in: Johnson, N.L. and Kotz, S. (editors), *Breakthroughs in Statistics*, Vol I, Springer–Verlag (1992) 236–247, with an introduction by P.K. Pathak.

535. C.R. Rao. Some statistical methods for comparison of growth curves. *Biometrics*, 14:1–17, 1958.
536. C.R. Rao. Least squares theory using an estimated dispersion matrix and its application to measurement of signals. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:355–372, 1967.
537. C.R. Rao. Simultaneous estimation of parameters in different linear models and applications in biometric problems. *Biometrics*, 31:545–554, 1975.
538. C.R. Rao. *Linear Statistical Inference and Its Application*. Wiley, New York, second edition, 2001. Reprint from 1973; first edition: 1959.
539. C.R. Rao and M.B. Rao. *Matrix Algebra and Its Application to Statistics and Econometrics*. World Scientific, Singapore, 1998.
540. D. Rasch. *Mathematische Statistik*. Johann Ambrosius Barth Verlag, Heidelberg, 1995.
541. G. Rasch. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 19:49–56, 1966.
542. W. Rautenberg. *Klassische und nichtklassische Ausssagenlogik*. Vieweg, Braunschweig, 1979.
543. J.O. Rawlings. *Applied Regression Analysis*. Wadsworth, Belmont, CA, 1988.
544. P.-H. Rebut, P.P. Lallia, and M.L. Watkins. The critical temperature gradient model of plasma transport: Applications to JET and future tokamaks. *Proceedings of the 12th Conference on Plasma Physics and Controlled Nuclear Fusion Research, Nice, 1988*, 2:191–200, 1989. IAEA–CN–50/D–IV–1.
545. H. Reichenbach. *Wahrscheinlichkeitslehre: eine Untersuchung über die logischen und mathematischen Grundlagen der Wahrscheinlichkeitsrechnung*. Vieweg, 1994. First edition: 1935; Translated into English by M. Reichenbach and E. Hutten: The Theory of Probability: An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability, Berkeley, University of California Press, 1949, 1971.
546. N. Reid. The roles of conditioning in inference. *Statistical Science*, 10:138–157, 1995.
547. N. Reid. Asymptotics and the theory of inference (The 2000 Wald memorial lectures). *The Annals of Statistics*, 31:1695–1731, 2003.
548. D. Reiter, C. May, D. Coster, and Schneider R. Time dependent neutral gas transport in tokamak edge plasmas. *Journal of Nuclear Materials*, pages 987–992, 1995.
549. H. Renner et al. Divertor concept for the W7–X stellarator and mode of operation. *Plasma Physics and Controlled Fusion*, 44:1005–1019, 2002.
550. N. Rescher. *Many–valued Logic*. McGraw Hill, New York, 1969.
551. V. Riccardo, P. Andrew, L.C. Ingesson, and G. Maddaluno. Disruption heat loads on the JET MkIIGB divertor. *Plasma Physics and Controlled Fusion*, 44:905–929, 2002.
552. H. Richter. *Wahrscheinlichkeitstheorie*. Die Grundlehren der Mathematischen Wissenschaften in Einzeldarstellungen, Band 86. Springer–Verlag, Heidelberg, 1956.
553. F.G. Rimini et al. Combined heating experiments in ELM–free H-modes in JET. *Nuclear Fusion*, 39:1591–1603, 1999.
554. H.E. Robbins. The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics*, 35:1–20, 1964.
555. C.P. Robert. *The Bayesian Choice*. Springer–Verlag, second edition, 2001.

556. T. Robertson, F.T. Wright, and R.L. Dykstra. *Order Restricted Statistical Inference*. Wiley, New York, 1988.
557. A. Robinson. *Non–Standard Analysis*. Princeton University Press, first (revised) edition, 1996. Previous editions: 1966, 1974, 1980, North–Holland, Amsterdam.
558. J.L Rodgers and Nicewander W.A. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42:59–66, 1988.
559. C.A. Rogers. *Hausdorff Measures*. Cambridge University Press, second edition, 1999. First edition: 1970.
560. L.J. Rogers. An extension of a certain theorem in inequalities. *Messenger of Mathematics*, 17:145–150, 1888.
561. H. Röhr, K.-H. Steuer, H. Murmann, and D. Meisel. Periodic multichannel Thomson scattering in ASDEX. Technical Report IPP III/121 B, Max–Planck–Institut für Plasmaphysik, Garching, 1987.
562. H. Röhr, K.-H. Steuer, G. Schramm, K. Hirsch, and H. Salzmann. First high–repetition–rate Thomson scattering for fusion plasmas. *Nuclear Fusion*, 22:1099–1102, 1982.
563. A.E. Ronner. *p–Norm Estimators in a Linear Regression Model*. PhD thesis, Groningen University, Groningen, 1977.
564. C. Rose and M.D. Smith. *Mathematical Statistics with Mathematica*. Springer–Verlag, New York, 2002.
565. D.W. Ross, R.V. Bravenec, W. Dorland, M.A. Beer, G.W. Hammett, G.R. McKee, R.J. Fonck, M. Murakami, K.H. Burrell, G.L. Jackson, and G.M. Staebler. Comparing simulation of plasma turbulence with experiment. *Physics of Plasmas*, 9:177–184, 2002.
566. P. Rousseeuw and K. van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.
567. P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.
568. W. Rudin. *Real and Complex Analysis*. McGraw–Hill, third edition, 1987.
569. F. Ryter for the Threshold Database Working Group. H–mode power threshold database for ITER (special topic). *Nuclear Fusion*, 36:1217–1264, 1996.
570. D.D. Ryutov. Environmental aspects of fusion energy. *Plasma Physics and Controlled Fusion*, 34:1805–1815, 1992.
571. A.D. Sacharov. Theory of the magnetic thermonuclear reactor, part 2. In *Plasma Physics and the Problem of Controlled Thermonuclear Reactions*, pages 20–30. Pergamon Press, Oxford, 1961. Work done in 1951.
572. R.M. Sakia. The Box–Cox transformation technique: A review. *The Statistician*, 41:169–178, 1992.
573. D. Salomé. *Statistical Inference via Fiducial Methods*. PhD thesis, Groningen University, Groningen, 1998.
574. M.D. Sammel, L.M. Ryan, and J.M. Legler. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society*, 59:667–678, 1997.
575. L. Sanathanan and S. Blumenthal. The logistic model and estimation of latent structure. *Journal of the American Statistical Association*, 73:794–799, 1978.
576. J. Sapper and H. Renner. Stellarator Wendelstein VII-AS: Physics and engineering design. *Fusion Technology*, 17:62–75, 1990.
577. W.S. Sarle. Neural networks and statistical models. In *Proceedings of the 19th Annual SAS Users Group International Conference*, Cary, NC, 1994.

578. SAS Institute Inc., Cary, NC. *SAS/STAT User's Guide*, fourth edition, 1989. Version 6.
579. SAS Institute Inc. *Proceedings of the Fifteenth Annual SAS Users Group International Conference*, Cary, NC, 1990.
580. SAS Institute Inc., Cary, NC. *SAS/IML User's Guide*, 1999. Version 8.
581. SAS Institute Inc., Cary, NC. *SAS/STAT User's Guide*, first edition, 2000. Version 8.
582. L.J. Savage. *The Foundations of Statistics*. Dover, New York, second edition, 1972. First edition: Wiley, 1954.
583. W. Schaafsma. *Hypothesis Testing Problems with the Alternative restricted by a Number of Inequalities*. PhD thesis, Groningen University, Groningen, 1966.
584. W. Schaafsma. Minimax risk and unbiasedness for multiple decision problems of type I. *Annals of Mathematical Statistics*, 40:1684–1720, 1969.
585. W. Schaafsma. Me and the anthropologist. In *Proceedings of the Seventh Conference on Probability Theory, held at Brasov, Romania*, pages 333–343, Bucharest, 1984. The Centre of Mathematical Statistics of the National Institute of Metrology, Editura Academiei Republicii Socialiste Romania.
586. W. Schaafsma and G.N. Van Vark. Classification and discrimination problems with applications, part I. *Statistica Neerlandica*, 31:25–45, 1977.
587. W. Schaafsma and G.N. Van Vark. Classification and discrimination problems with applications, part II. *Statistica Neerlandica*, 33:91–126, 1979.
588. J. Scheffel and D.D. Schnack. Confinement scaling laws for the conventional reversed-field pinch. *Physical Review Letters*, 85:322–325, 2000.
589. T.J. Schep and M. Venema. Collisionless drift modes in a toroidal configuration. *Plasma Physics and Controlled Fusion*, 27:653–671, 1985.
590. M.J. Schervish. P values: What they are and what they are not. *The American Statistician*, 50:203–206, 2001.
591. D. Schissel for the H–mode Database Working Group. Analysis of the ITER H–mode confinement database. In *Controlled Fusion and Plasma Physics, (Proc. 20th Eur. Conf., Lisbon, 1993)*, volume 17 C, pages 103–106, Geneva, 1994. European Physical Society.
592. A. Schlüter. Fusion at Venice, remembered 32 years later. *Plasma Physics and Controlled Fusion*, 31(10):1725–1726, 1989.
593. P.I.M. Schmitz. *Logistic Regression in Medical Decision Making and Epidemiology*. PhD thesis, Erasmus University, Rotterdam, 1986.
594. R. Schneider. Plasma edge physics for tokamaks. Technical Report IPP 12/1, Max–Planck–Institut für Plasmaphysik, 2001.
595. J.R. Schott. Some tests for common principal components in several groups. *Biometrika*, 78:771–777, 1991.
596. E. Schrödinger. Zur Theorie der Fall– und Steigversuche an Teilchen mit Brownscher Bewegung. *Physikalische Zeitschrift*, 16, 1915.
597. F.C. Schüller. Disruptions in tokamaks. *Plasma Physics and Controlled Fusion*, 37:A135–A163, 1995.
598. U. Schumacher. *Fusionsforschung: Eine Einführung*. Wissenschaftliche Buchgesellschaft, Darmstadt, 1993.
599. R.A. Schumacker and A. Akers. *Understanding Statistical Concepts using S–PLUS*. Lawrence Erlbaum Associates, Mahwah, NJ, 2001.
600. B. Schunke, K. Imre, and K.S. Riedel. Profile shape parameterization of JET electron temperature and density profiles. *Nuclear Fusion*, 37:101–117, 1997.

601. L. Schwartz. *Théorie des Distributions*. Hermann, Paris, troisième edition, 1966. First edition: 1950, 1951.
602. L. Schwartz. *Analyse III*. Hermann, Paris, deuxième edition, 1998. First edition: 1993.
603. B. Scott. Three dimensional computation of collisional drift wave turbulence and transport in tokamak geometry. *Plasma Physics and Controlled Fusion*, 39:471–504, 1997.
604. S.R. Searle. *Linear Models*. Wiley, 1971.
605. A.F. Seber and C.J. Wild. *Nonlinear Regression*. Wiley, New York, 1989. Reprinted: Wiley–Interscience paperback Series, 2003.
606. U. Seifert and L.C. Jain, editors. *Self-Organizing Neural Networks*. Number 78 in Studies in Fuzziness and Soft Computing. Springer–Verlag, Heidelberg, 2002.
607. R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, first edition, 1980. Reprinted 2001.
608. V. Seshadri. *The Inverse Gaussian Distribution –A Case Study in Exponential Families–*. Clarendon Press, 1993.
609. G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
610. J. Shao. *Mathematical Statistics*. Springer Texts in Statistics. Springer–Verlag, Heidelberg, second edition, 2003. First edition: 1999.
611. A. Shapiro and M.W. Browne. Analysis of covariance structures under elliptical distributions. *Journal of the American Statistical Association*, 82:1092–1097, 1987.
612. S.S. Shapiro and R.S. Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67:215–216, 1972.
613. E.N. Shaw. *Europe's Experiment in Fusion: the JET Joint Undertaking*. Elsevier, Amsterdam, 1990.
614. Y. Shimomura, R. Aymar, V.A. Chuyanov, M. Huguet, H. Matsumoto, T. Mizoguchi, Y. Murakami, A.R. Polevoi, M. Shimada, ITER Joint Central Team, and ITER Home Teams. ITER–FEAT operation. *Nuclear Fusion*, 41:309–316, 2001.
615. Y. Shimomura, Y. Murakami, A. Polevoi, P. Barabaschi, V. Mukhovatov, and M. Shimada. ITER: Opportunity of burning plasma studies. *Plasma Physics and Controlled Fusion*, 43:A385–A394, 2001.
616. Y. Shimomura and K. Odajima. Empirical scaling of incremental energy confinement time of L–mode plasma and comments on improved confinement in tokamaks. *Comments on Plasma Physics and Controlled Fusion*, 10:207–215, 1987.
617. H. Shirai, T. Takizuka, M. Kikuchi, M. Mori, T. Nishitani, S. Ishida, et al. Non-dimensional transport scaling and its correlation with local transport properties in JT-60U plasmas. *Proceedings of the 15th Conference on Fusion Energy, Seville 1994*, 1:355–364, 1995. IAEA–CN–60/A2–17.
618. H. Shore. Simple approximations for the inverse cumulative function, the density function and the loss integral of the normal distribution. *Applied Statistics*, 31(2):108–114, 1982.
619. S. Siegel and N.J. Castellan. *Nonparametric Statistics for the Behavioural Sciences*. McGraw Hill, 1988.
620. A.E. Siegman. *Lasers*. University Science Books, Mill Valley, CA, 1986.

621. D. Siegmund. *Sequential Analysis*. Springer Series in Statistics. Springer–Verlag, New York, 1985.
622. D.J. Sigmar and H.S. Hsu. Comments on Shimomura–Odajima scaling. *Comments on Plasma Physics and Controled Fusion*, 12:15–34, 1998.
623. M.J. Silvapulle and P.K. Sen. *Constrained Statistical Inference: Order, Inequality and Shape Constraints*. Wiley, New York, 2004.
624. E.E. Simmett and the ASDEX Team. Statistical analysis of the global energy confinement time in Ohmic discharges in the ASDEX tokamak. *Plasma Physics and Controlled Fusion*, 38:689–704, 1996.
625. T.A.B. Snijders. *Asymptotic Optimality Theory for Testing Problems with Restricted Alternatives*. PhD thesis, Groningen University, Groningen, 1979. Also: MC tract 113, Mathematical Centre, Amsterdam.
626. J.A. Snipes for the International H–mode Threshold Database Working Group. Latest results on the H–mode threshold using the international H-mode threshold database. *Plasma Physics and Controlled Fusion*, 42:A299–A308, 2000.
627. M.I. Solonin. Materials science problems of blankets in Russian concept of fusion reactor. *Journal of Nuclear Materials*, 258–263:30–46, 1998.
628. R.M. Solovay. A model of set–theory in which every set of reals is Lebesgue measurable. *Annals of Mathematics*, 92:1–56, 1970.
629. K.H. Spatschek. *Theoretische Plasmaphysik*. Teubner–Verlag, Stuttgart, 1990.
630. L. Spitzer. The stellarator concept. *Physics of Fluids*, 1:253–264, 1958.
631. A. Stäbler, K. McCormick, V. Mertens, E.R. Müller, J. Neuhauser, H. Niedermeyer, K.-H. Steuer, H. Zohm, et al. Density limit investigations on ASDEX. *Nuclear Fusion*, 32:1557–1583, 1992.
632. W.M. Stacey. *Fusion Plasma Analysis*. Wiley, first edition, 1981.
633. A.J. Stam. Statistical problems in ancient numismatics. *Statistica Neerlandica*, 41(3):151–173, 1987.
634. R.D. Stambaugh, S.M. Wolfe, and R.K. Hawryluk. Enhanced confinement in tokamaks. *Physics of Fluids B*, 2:2941–2960, 1990.
635. R.D. Stambaugh for the DIII-D Team. DIII-D research program progress. *Proceedings of the 13th Conference on Plasma Physics and Controlled Nuclear Fusion Research, Washington 1990*, 3:69–90, 1991. IAEA–CN–53/A–I–4.
636. P.C. Stangeby. *The Plasma Boundary of Magnetic Fusion Devices*. Institute of Physics Publishing, Bristol, 2000.
637. A.W. Steerneman. *On the Choice of Variables in Discriminant Analysis and Regression Analysis*. PhD thesis, Groningen University, Groningen, 1987.
638. W. Stegmüller. *Hauptströmungen der Gegenwartsphilosophie*. Kröner, Stuttgart, 1987.
639. K.-H. Steuer, H. Röhr, and B. Kurzan. Bremsstrahlung measurements in the near infrared on ASDEX. *Review of Scientific Instruments*, 61:3084–3086, 1990.
640. S.M. Stigler. Napoleonic statistics: The work of Laplace. *Biometrika*, 62:503–517, 1975.
641. S.M. Stigler. Poisson on the Poisson distribution. *Statistics and Probability Letters*, 1:33–35, 1982.
642. S.M. Stigler. Thomas Bayes's Bayesian inference. *Journal of the Royal Statistical Society*, 145:250–258, 1982.
643. S.M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Havard University Press, Cambridge, Mass., 1986.

644. S.M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1:359–363, 1986.
645. S.M. Stigler. *Statistics on the Table: The History of Statistical Concepts and Methods*. Havard University Press, Cambridge, Mass., 1999.
646. T.H. Stix. *Waves in Plasmas*. Springer–Verlag, New York, 1992. First edition: 1962.
647. J. Stober, M. Maraschek, G.D. Conway, et al. Type II ELMy H-modes on ASDEX Upgrade with good confinement at high density. *Nuclear Fusion*, 41:1123–1134, 2002.
648. J. Stober and the ASDEX Upgrade Team. Improvement of ion–temperature–profile determination from charge exchange measurements by inclusion of total–neutral–flux data. *Plasma Physics and Controlled Fusion*, 39:1145–1152, 1997.
649. J. Stoer and R. Burlisch. *Einführung in die Numerische Mathematik II*. Springer–Verlag, Heidelberg, second edition, 1978.
650. C.J. Stone, M.H. Hansen, C. Kooperberg, and Y.K. Truong. Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics*, 25:1371–1470, 1997.
651. J.F. Strenio, H.I. Weisberg, and A.S. Bryk. Empirical Bayes estimation of individual growth–curve parameters and their relationship to covariates. *Biometrics*, 39:71–86, 1983.
652. U. Stroth. A comparative study of transport in stellarators and tokamaks. *Plasma Physics and Controlled Fusion*, 40:9–74, 1998.
653. U. Stroth, M. Murakami, R.A. Dory, et al. Energy confinement scaling from the international stellarator database. *Nuclear Fusion*, 36:1063–1077, 1996.
654. J.W. Strutt (Lord Rayleigh). On an anomaly encountered in determination of the density of nitrogen gas. *Proceedings of the Royal Society of London*, 55:340–344, 1894.
655. A. Stuart and J.K. Ord. *Kendall's Advanced Theory of Statistics*, volume I: Distribution Theory. Edward Arnold, London, sixth edition, 1994.
656. T.A. Stukel. Generalised logistic models. *Journal of the American Statistical Association*, 83:426–431, 1988.
657. W. Suttrop. The physics of large and small edge localized modes. *Plasma Physics and Controlled Fusion*, 42:A1–A14, 1999.
658. A.A. Sveshnikov. *Problems in Probability Theory, Mathematical Statistics and Theory of Random Functions*. Dover Publications, New York, 1978.
659. D.F. Swayne, D. Cook, and A. Buja. Xgobi: Interactive dynamic graphics in the X–window system. *Journal of Computational and Graphical Statistics*, pages 110–130, 1998.
660. T.H. Szatrowski and J.J. Miller. Explicit maximum likelihood estimates from balanced data in the mixed model of the analysis of variance. *Annals of Statistics*, 8:811–819, 1980.
661. A. Takayama. *Mathematical Economics*. Cambridge University Press, Cambridge, second edition, 1985.
662. T. Takizuka. An offset nonlinear scaling for ELMy H–mode confinement. *Plasma Physics and Controlled Fusion*, 40:851–855, 1998.
663. T. Takizuka for the ITER Confinement Database Working Group. ITER: Analysis of the H–mode confinement and threshold databases. *Proceedings of the 16th Conference on Fusion Energy, Montreal 1996*, 2:795–806, 1997. IAEA–CN–64/F–5.

664. T. Takizuka for the ITPA H-mode Power Threshold Database Working Group. Roles of aspect ratio, absolute B and effective Z of the H–mode power threshold in tokamaks of the ITPA database. *Plasma Physics and Controlled Fusion*, 46:A227–A233, 2004.
665. I.E. Tamm. Theory of the magnetic thermonuclear reactor, parts 1 and 3. In *Plasma Physics and the Problem of Controlled Thermonuclear Reactions*, pages 3–19 and 31–41. Pergamon Press, Oxford, 1961. Work done in 1951.
666. A.S. Tanenbaum. *Structured Computer Organization*. Prentice-Hall, Upper Saddle River, NJ, fourth edition, 1999. First edition: 1976.
667. W.M. Tang. Microinstability theory in tokamaks: A review. *Nuclear Fusion*, 18:1089–1160, 1978.
668. ASDEX Team. The H–mode of ASDEX. *Nuclear Fusion*, 29:1959–2040, 1989.
669. NET Team. NET (Next European Torus) status report. Commission of the European Communities, Directorate General XII – Fusion Programme, Brussels, December 1985. (NET report 51).
670. N.M. Temme. *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*. Wiley, New York, 1996.
671. ITER Expert Groups and ITER Physics Basis Editors. ITER Physics Basis. *Nuclear Fusion*, 39:2137–2664, 1999.
672. ITER Joint Central Team and ITER Home Teams. *ITER Technical Basis*. Number 24 in ITER EDA Documentation Series. IAEA, Vienna, 2002.
673. L. Thierney. *LISP–STAT, A Statistical Environment Based on the XLISP Language*. Wiley, New York, 1990.
674. H.C. Thode Jr. *Testing for Normality*. Marcel–Dekker, Heidelberg, 2002.
675. H.C.S. Thom. *Direct and Inverse Tables of the Gamma Distribution*. Environmental Data Service, Silver Spring, MD, 1968.
676. R. Thom. *Mathematiker über die Mathematik*, chapter Die Katastrophen–Theorie: Gegenwärtiger Stand und Aussichten, pages 125–137. Springer–Verlag, Heidelberg, 1974. M. Otte (editor).
677. R. Thom. *Modèles Mathématiques de la Morphogénèse*. Bourgois, Paris, 1981. Translated into English, several publishers 1983/1984.
678. K. Thomsen for the H-mode Database Working Group. ITER confinement database update (special topic). *Nuclear Fusion*, 34:131–167, 1994.
679. K. Thomsen for the International Confinement Database Working Group. The international global H–mode confinement database: public release of global confinement data from 17 tokamaks. Technical report, EFDA CSU Garching, 2001. Available on internet, URL=http://efdasql.ipp.mpg.de/HmodePublic.
680. A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-posed Problems*. Wiley, New York, 1977.
681. N.H. Timm. *Applied Multivariate Analysis*. Springer–Verlag, Heidelberg, 2002.
682. T. Tjur. *Probability Based on Radon Measures*. Wiley, New York, 1980.
683. M. Toda, R. Kubo, and N. Saitô. *Statistical Physics*, volume I (Equilibrium Statistical Mechanics) of *Springer Series in Solid–State Sciences, Vol. 30*. Springer–Verlag, Heidelberg, second edition, 1992. First edition: 1983.
684. K. Tomabechi, J.R. Gilleland, Yu.A. Sokolov, and R. Toschi. ITER conceptual design (special item). *Nuclear Fusion*, 31:1135–1224, 1991.
685. L.N. Topilski, X. Masson, M.T. Porfiri, T. Pinna, L.-L. Sponton, et al. Validation and benchmarking in support of ITER–FEAT safety analysis. *Fusion Engineering and Design*, 54:672–633, 2001.

686. H. Toutenburg. *Lineare Modelle*. Physica–Verlag, second edition, 2002. First edition: 1992.
687. M.Q. Tran. Propagation of solitary waves in a two ion species plasma with finite ion temperature. *Plasma Physics*, 16:1167–1175, 1974.
688. C. Tricot. Two definitions of fractal dimensions. *Mathematical Proceedings of the Cambridge Philosophical Society*, 91:57–74, 1982.
689. C. Tricot. *Curves and Fractal Dimensions*. Springer–Verlag, Heidelberg, 1995.
690. M.C.K. Tweedie. Inverse statistical variates. *Nature*, 155:453, 1945.
691. UCLA, Los Angeles. *Journal of Statistical Software*, from 1996 onwards. Available on internet, URL = http://www.jstatsoft.org/.
692. H. Urano, Y. Kamada, H. Shirai, T. Takizuka, S. Ide, F. Fujita, and T. Fukuda. Thermal energy confinement properties of ELMy H-mode plasmas in JT–60U. *Nuclear Fusion*, 42:76–85, 2002.
693. E.A. Van der Meulen. *Assessing Weights of Evidence for Discussing Classical Statistical Hypotheses*. PhD thesis, Groningen University, Groningen, 1992.
694. D.M. Van der Sluis, W. Schaafsma, and A.W. Ambergen. *POSCON user manual, A Decision Support System in Diagnosis and Prognosis*. Groningen University, Gronigen, 1989.
695. A.W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.
696. B.L. Van der Waerden. *Mathematische Statistik*. Die Grundlehren der Mathematischen Wissenschaften in Einzeldarstellungen, Band 87. Springer–Verlag, Heidelberg, first edition, 1957. English translation: Springer–Verlag, New York, 1969.
697. B.L. Van der Waerden. *Algebra*, volume I and II. Springer–Verlag, Heidelberg, ninth and sixth edition, 1993. First edition: Springer–Verlag, 1930; English translation: Springer–Verlag, New York, 1991.
698. J.L. Van Hemmen. The map in your head: How does the brain represent the outside world? *European Journal of Chemical Physics and Physical Chemistry*, 3:291–298, 2002.
699. G.N. Van Vark and W.W. Howells, editors. *Multivariate Statistical Methods in Physical Antropology, A Review of Recent Advances and Current Developments*. Reidel, 1984.
700. W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S–PLUS*. Springer–Verlag, Heidelberg, fourth edition, 2004. First edition: 1999.
701. M. Venema. Collisional electrostatic drift modes in toroidal geometry. *Plasma Physics and Controlled Fusion*, 28:1083–1092, 1986.
702. H. Verbeek, O. Heinrich, R. Schneider, H.-U. Fahrbach, W. Herrmann, J. Neuhauser, U. Stroth, and the ASDEX Team. Ion temperature profiles from the plasma center to the edge of ASDEX combining high and low energy CX–diagnostics. *Journal of Nuclear Materials*, 196–198:1027–1031, 1992.
703. R. Verbrugge. *The Stanford Encyclopedia of Philosophy* (*Summer 2003 Edition*), *E.N. Zalta* (*editor*), chapter Provability Logic. The Metaphysics Research Laboratory at the Center for the Study of Language and Information, Stanford, CA, 2003. Available on internet, URL = http://plato.stanford.edu/archives/sum2003/entries/logic-provability.
704. F. Verhulst. *Nonlinear Differential Equations and Dynamical Systems*. Springer–Verlag, Heidelberg, second edition, 1997. First edition: Springer–Verlag, 1990.

705. R. Vierhaus and B. Brocke, editors. *Forschung im Spannungsfeld von Politik und Gesellschaft: Geschichte und Struktur der Kaiser–Wilhelm–/Max–Planck–Gesellschaft aus Anlaß ihres 75jährigen Bestehens*. Deutsche Verlags–Anstalt GmbH, Stuttgart, 1990.
706. O. Vollmer, F. Ryter, A. Stäbler, and P.J. McCarthy. Scaling of thermal energy confinement in ASDEX Upgrade. In *Controlled Fusion and Plasma Physics* (*Proc. 24th Eur. Conf., Berchtesgaden, 1997*), volume 21A, Part IV, pages 1485–1489, Geneva, 1997. European Physical Society.
707. L. Von Bortkiewicz. *Das Gesetz der kleinen Zahlen*. Teubner–Verlag, Leipzig, 1898.
708. R. Von Mises. Über die 'Ganzzahligkeit' der Atomgewichte und verwandte Fragen. *Physikalische Zeitschrift*, 7:153–159, 1918.
709. R. Von Mises. *Mathematical Theory of Probability and Statistics*. Academic Press, 1964.
710. R. Von Mises. *Probability, Statistics and Truth*. Dover, London, second (reprint) edition, 1981. Second English edition: Allan and Unwin, London 1957, translated from the third German edition, Julius Springer, Vienna 1951; first German edition: Julius Springer, Vienna 1928.
711. B. Von Querenburg. *Mengentheoretische Topologie*. Springer–Verlag, Heidelberg, third edition, 2001. First edition: 1973.
712. A. Wachsmuth, L. Wilkinson, and G. Dallal. Galton's bend: A previously undiscovered nonlinearity in Galton's family stature regression data. *The American Statistician*, 57:190–192, 2003.
713. F. Wagner, F. Ryter, A.R. Field, et al. Recent results of H–mode studies at ASDEX. *Proceedings of the 13th Conference on Plasma Physics and Controlled Nuclear Fusion Research, Washington 1990*, 1:277–290, 1991.
714. F. Wagner for the ASDEX Team. Regime of improved confinement and high beta in neutral-beam-heated divertor discharges of the ASDEX tokamak. *Physical Review Letters*, 49:1408–1412, 1982.
715. F. Wagner for the W7–AS Team. Major results from the Wendelstein 7–AS stellarator. In *Fusion Energy 2002* (*Proc. 19th Int. Conf. Lyon, 2002*), Vienna, 2002. IAEA–CN–94–OV/2–4. Available on internet, URL=http://www.iaea.-org/programmes/ripc/physics/fec2002/html/node32.htm.
716. S. Wagon. *The Banach–Tarski Paradox*. Cambridge University Press, 1985.
717. M. Wakatani. *Stellarator and Heliotron Devices*. Oxford University Press, Oxford, 1998.
718. A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 14:426–482, 1943.
719. A. Wald. *Sequential Analysis*. Wiley, New York, 1947. Reprinted: Dover, New York, 1973.
720. A. Wald. Statistical decision functions. *Annals of Mathematical Statistics*, 29:165–205, 1949. Reprinted in: Johnson, N.L. and Kotz, S. (editors), *Breaktroughs in Statistics*, Vol. I, Springer–Verlag (1992) 342–357, with an introduction by L. Weiss.
721. P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Number 42 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1991.
722. P. Walley. Inferences from multinomial data: Learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society*, 58:3–57, 1996.

723. Z. Wang and G.J. Klir. *Fuzzy Measure Theory*. Plenum Press, New York, 1992.
724. J.T. Webster, R.F. Gunst, and R.L. Mason. Latent root analysis. *Technometrics*, 16:513–522, 1974.
725. R.W.M. Wedderburn. Quasi–likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, 61:439–447, 1974.
726. J. Weiland. *Collective Modes in Inhomogeneous Plasma: Kinetic and Advanced Fluid Theory*. Institute of Physics Publishing, Bristol, 2000.
727. S. Weisberg. *Applied Linear Regression*. Wiley, New York, 1985.
728. K.A. Werley, J.N. DiMarco, R.A. Krakowski, and C.G. Bathke. Energy confinement and future reversed field pinches. *Nuclear Fusion*, 36:629–642, 1996.
729. J. Wesson. *Tokamaks*. Clarendon Press, Oxford, second edition, 1997.
730. J. Whitehead. *The Design and Analysis of Sequential Clinical Trials*. Wiley, Chichester, second edition, 1997.
731. J.E. Whitesitt. *Boolean Algebra and its Applications*. Addison-Wesley, Reading, Mass., first edition, 1961. Reprinted: Dover, New York, 1995.
732. E.T. Whittaker and G.N. Watson. *A Course of Modern Analysis*. Cambridge University Press, fourth edition, 1997. First edition: 1902.
733. A.S. Whittemore. Transformations to linearity in binary regression. *SIAM Journal of Applied Mathematics*, 43:703–710, 1983.
734. R. Wilhelm. The strengths, needs and possible drawbacks of different heating and current drive systems in relation to ITER. *Plasma Physics and Controlled Fusion*, 40:A1–A12, 1998.
735. J.H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, 1988. First edition: Oxford University Press, 1965.
736. S.S. Wilks. *Mathematical Statistics*. Wiley, 1963. First edition: Princeton University Press, 1943.
737. H. Wind. Function parameterization. In *Proceedings of the 1972 CERN Computing and Data Processing School*, number 72–21 in Yellow Reports, pages 53–106, Geneva, 1972. CERN.
738. H. Wind. Principal component analysis, pattern recognition for track finding, interpolation and functional representation (revised edition). Technical Report EP-Int-81-12-rev., CERN, Geneva, 1982.
739. B.B. Winter. An alternate development of conditioning. *Statistica Neerlandica*, 33:197–212, 1979.
740. H. Wobig. On radiative density limits and anomalous transport in stellarators. *Plasma Physics and Controlled Fusion*, 42:931–948, 2000.
741. H. Wobig. Concept of a Helias ignition experiment. *Nuclear Fusion*, 43:889–898, 2003.
742. H. Wobig and F. Rau (editors). Stellarator experiments at IPP Garching. Technical Report IPP 2 / 311, Max–Planck–Institut für Plasmaphysik, Garching, 1991. A collection of publications on stellarator experiments in Garching, dedicated to Dr. Günter Grieger on the occasion of his 60th birthday.
743. S. Wolfram. *The Mathematica Book*. Wolfram Media Inc., fifth edition, 2003.
744. E. Wong and B. Hajek. *Stochastic Processes in Engineering Systems*. Springer–Verlag, Heidelberg, 1985.
745. K.L. Wong, R.J. Fonck, S.F. Paul, D.R. Roberts, et al. Excitation of toroidal Alfvén eigenmodes in TFTR. *Physical Review Letters*, 66:1874–1877, 1991.

746. J.W. Wu, W.L. Hung, and H.M. Lee. Some moments and limit behaviors of the generalized logistic distribution with aplications. *Proceedings of the National Science Council, Republic of China* (*Series A*), 24:7–14, 2000.
747. P.N. Yushmanov, T. Takizuka, K. Riedel, O.J.W.F. Kardaun, J.G. Cordey, S.M. Kaye, and D.E. Post. Scalings for tokamak energy confinement. *Nuclear Fusion*, 30:1999–2008, 1990.
748. E.N. Zalta, editor. *The Stanford Encyclopedia of Philosophy* (*Spring 2004 Edition*). The Metaphysics Research Lab at the Center for the Study of Language and Information, Stanford, CA, 2004. Available on internet, URL = http://plato.stanford.edu/archives.
749. E. Zermelo. Neuer Beweis für die Möglichkeit einer Wohlordnung. *Mathematische Annalen*, 65:107–128, 1908.
750. H. Zohm. Edge localised modes (ELMs). *Plasma Physics and Controlled Fusion*, 38:105–128, 1996.
751. H. Zohm, K. Lackner, and C. Ludescher. Statistical analysis of disruptions in ASDEX. *Nuclear Fusion*, 33:655–662, 1993.
752. H. Zohm, T.H. Osborne, K.H. Burrell, M.S. Chu, et al. ELM studies on DIII–D and a comparison to ASDEX results. *Nuclear Fusion*, 35:543–550, 1995.
753. H. Zohm, F. Wagner, M. Endler, et al. Studies of edge localized modes on ASDEX. *Nuclear Fusion*, 32:489–494, 1992.
754. K.H. Zou, K. Tuncali, and S.G. Silverman. Correlation and simple linear regression. *Radiology*, 227(3):617–622, 2003.

Index