Lecture Notes in Statistics 217
Proceedings

Anestis Antoniadis
Jean-Michel Poggi
Xavier Brossat *Editors*

Modeling and Stochastic Learning for Forecasting in High Dimensions

Springer

Lecture Notes in Statistics 217

Edited by P. Bickel, P. Diggle, S.E. Fienberg, U. Gather,
I. Olkin, S. Zeger

More information about this series at
http://www.springer.com/series/694

Anestis Antoniadis • Jean-Michel Poggi •
Xavier Brossat
Editors

Modeling and Stochastic Learning for Forecasting in High Dimensions

Springer

Editors
Anestis Antoniadis
Department of Statistics
University Joseph Fourier
Grenoble, France

Jean-Michel Poggi
Laboratoire de Mathématiques
Université Paris-Sud
Orsay Cedex, France

Xavier Brossat
Electricité de France R & D, OSIRIS
Clamart Cedex, France

ISSN 0930-0325 ISSN 2197-7186 (electronic)
Lecture Notes in Statistics
ISBN 978-3-319-18731-0 ISBN 978-3-319-18732-7 (eBook)
DOI 10.1007/978-3-319-18732-7

Library of Congress Control Number: 2015941890

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Forecasting and time series prediction have seen a great deal of development and attention over the last few decades, fostered by an impressive improvement in observational capabilities and measurement procedures. Time series prediction is a challenge in many fields. In finance, one forecasts stock exchange or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the electric load, and hydrologists forecast river floods. Many methods designed for time series prediction and forecasting perform well (depending on the complexity of the problem) on a rather short-term horizon but are rather poor on a longer-term one. This is due to the fact that these methods are usually designed to optimize the performance at short-term prediction using relatively low dimensional models. For long-term prediction, linear and nonlinear methods and tools for the analysis and predictive modeling of high-dimensional phenomena are necessary and more useful.

From an industrial point of view, forecasting the electricity/natural gas consumption of consumers is yet another challenge. In order to be able to provide accurate forecasts on different horizons (from short term, the next hour, to middle term, several years) and at different levels of aggregation (from complete portfolio to local demand), these forecasts have to be done in an evolving environment. Indeed, national demand is increasing, uses are changing, the number of available data is significantly growing (thanks to smart meters), the individual renewable electricity generation is becoming important, and smart grids are developing. On the other side of electricity markets, renewables provide a bigger and bigger part of electricity generation. Forecasting wind or solar power is also difficult because of their inner variability, the need for accurate and local meteorological forecasts at any horizon – from a few hours up to several days ahead – as well as the rapid evolution of operating installations.

On June 5–7, 2013, an international workshop on Industry Practices for Forecasting was held in Paris, France, organized and supported by the OSIRIS Department of EDF R&D, located in Clamart, France. OSIRIS stands for Optimization Simulation Risks and Statistics for energy markets. The meeting was the second in the series

of the WIPFOR conferences and was attended by several researchers (academics, industrial and professionals, and other interested parties) from several countries. Following tradition, both theoretical statistical results and practical contributions of this active field of statistical research and on forecasting issues in a fast evolving industrial environment were presented. The program and abstracts are available on the conference website (http://conferences-osiris.org/wipfor/13-Main-page).

The editors of this volume hope that these lecture notes reflect the broad spectrum of the conference, as it includes 16 articles contributed by specialists in various areas in this field. The material compiled is fairly wide in scope and ranges from the development of results on forecasting in industry and in time series, on nonparametric and functional methods, on online machine learning for forecasting, and on tools for high-dimensional and complex data analysis.

The articles are arranged and numbered in alphabetical order by author rather than subject matter. Papers 1, 3, 5, and 9 are dedicated to nonparametric techniques for short-term load forecasting in the industry and include classical curve linear regression, sparse functional regression based on dictionaries, as well as a new estimation procedure based on iterative bias reduction. Papers 11 and 14 focus on electrical system changes: the first is dedicated to large-scale electrical load simulation for smart grids using GAM modeling; the second focuses on space-time trajectories of wind power generation including parameterized precision matrices based on a Gaussian copula approach. Paper 7 provides flexible and dynamic modeling of dependencies via copulas. Papers 6, 8, and 13 explore different aspects of online learning ranging from the most operational for online residential baseline estimation to the more theoretical, which focuses on oracle bounds for prediction errors related to aggregation strategies. The third one is dedicated to the aggregation of experts proposing some resampling ideas to enlarge the basic family of the so-called experts. Papers 2, 4, 7, 12, and 16 study some general approaches to high-dimensional and complex data (inference in high-dimensional models, graphical models and model selection, adaptive spot volatility estimation for high-frequency data, functional classification and prediction). Finally, papers 10 and 15 deal with some special topics in time series, namely, modeling and prediction of time series arising on a graph and optimal reconciliation of contemporaneous hierarchical time series forecasts.

We would like to address our gratitude to the keynote speakers and all the contributors for accepting to participate in these Lecture Notes, and we greatly appreciate the time that they have taken to prepare their papers.

To conclude, we would like to acknowledge the following distinguished list of reviewers who helped improve the papers by providing general and focused feedback to the authors: U. Amato, R. Becker, R. Cao, J. Cugliari, G. Fort, P. Fryzlewicz, F. Gamboa, B. Ghattas, I. Gijbels, S. Girard, E. Gobet, Y. Goude, R. Hyndman, I. Kojadinovic, S. Lambert-Lacroix, J.-M. Loubes, E. Matzner-Løber, G. Oppenheim, P. Pinson, G. Stoltz, A. Verhasselt, A. Wigington, and Q. Yao. We thank them again for their work, dedication, and professional expertise.

Finally, to the editorial office of Springer-Verlag and particularly to Eva Hiripi, we are grateful for their efficient assistance.

Grenoble, France — Anestis Antoniadis
Clamart, France — Xavier Brossat
Paris, France — Jean-Michel Poggi
Paris, June 2014.

Contents

Short Term Load Forecasting in the Industry for Establishing Consumption Baselines: A French Case 1
José Blancarte, Mireille Batton-Hubert, Xavier Bay, Marie-Agnès Girard, and Anne Grau

Confidence Intervals and Tests for High-Dimensional Models: A Compact Review 21
Peter Bühlmann

Modelling and Forecasting Daily Electricity Load via Curve Linear Regression 35
Haeran Cho, Yannig Goude, Xavier Brossat, and Qiwei Yao

Constructing Graphical Models via the Focused Information Criterion 55
Gerda Claeskens, Eugen Pircalabelu, and Lourens Waldorp

Fully Nonparametric Short Term Forecasting Electricity Consumption 79
Pierre-André Cornillon, Nick Hengartner, Vincent Lefieux, and Eric Matzner-Løber

Forecasting Electricity Consumption by Aggregating Experts; How to Design a Good Set of Experts 95
Pierre Gaillard and Yannig Goude

Flexible and Dynamic Modeling of Dependencies via Copulas 117
Irène Gijbels, Klaus Herrmann, and Dominik Sznajder

Online Residential Demand Reduction Estimation Through Control Group Selection 147
Leslie Hatton, Philippe Charpentier, and Eric Matzner-Løber

Forecasting Intra Day Load Curves Using Sparse Functional Regression 161
Mathilde Mougeot, Dominique Picard, Vincent Lefieux, and Laurence Maillard-Teyssier

Modelling and Prediction of Time Series Arising on a Graph 183
Matthew A. Nunes, Marina I. Knight, and Guy P. Nason

Massive-Scale Simulation of Electrical Load in Smart Grids Using Generalized Additive Models 193
Pascal Pompey, Alexis Bondu, Yannig Goude, and Mathieu Sinn

Spot Volatility Estimation for High-Frequency Data: Adaptive Estimation in Practice 213
Till Sabel, Johannes Schmidt-Hieber, and Axel Munk

Time Series Prediction via Aggregation: An Oracle Bound Including Numerical Cost 243
Andres Sanchez-Perez

Space-Time Trajectories of Wind Power Generation: Parametrized Precision Matrices Under a Gaussian Copula Approach ... 267
Julija Tastu, Pierre Pinson, and Henrik Madsen

Game-Theoretically Optimal Reconciliation of Contemporaneous Hierarchical Time Series Forecasts 297
Tim van Erven and Jairo Cugliari

The BAGIDIS Distance: About a Fractal Topology, with Applications to Functional Classification and Prediction 319
Rainer von Sachs and Catherine Timmermans

Short Term Load Forecasting in the Industry for Establishing Consumption Baselines: A French Case

José Blancarte, Mireille Batton-Hubert, Xavier Bay, Marie-Agnès Girard, and Anne Grau

Abstract The estimation of baseline electricity consumptions for energy efficiency and load management measures is an essential issue. When implementing real-time energy management platforms for Automatic Monitoring and Targeting (AMT) of energy consumption, baselines shall be calculated previously and must be adaptive to sudden changes. Short Term Load Forecasting (STLF) techniques can be a solution to determine a pertinent frame of reference. In this study, two different forecasting methods are implemented and assessed: a first method based on load curve clustering and a second one based on signal decomposition using Principal Component Analysis (PCA) and Multiple Linear Regression (MLR). Both methods were applied to three different sets of data corresponding to three different industrial sites from different sectors across France. For the evaluation of the methods, a specific criterion adapted to the context of energy management is proposed. The obtained results are satisfying for both of the proposed approaches but the clustering based method shows a better performance. Perspectives for exploring different forecasting methods for these applications are considered for future works, as well as their application to different load curves from diverse industrial sectors and equipments.

J. Blancarte (✉)
EDF R&D, Département Eco-efficacité et Procédés Industriels, Avenue des Renardières, 77818 Moret-Sur-Loing, France

Ecole Nationale Supérieure des Mines, UMR 6158, EMSE-Fayol, Saint-Etienne F-42023, France
e-mail: jose.blancarte@edf.fr; blancarte@emse.fr

M. Batton-Hubert • X. Bay • M.-A. Girard
Ecole Nationale Supérieure des Mines, UMR 6158, EMSE-Fayol, Saint-Etienne F-42023, France
e-mail: batton@emse.fr; bay@emse.fr; girard@emse.fr

A. Grau
EDF R&D, Département Eco-efficacité et Procédés Industriels, Avenue des Renardières, 77818 Moret-Sur-Loing, France
e-mail: anne.grau@edf.fr

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_1

1 Introduction

Establishing a baseline is the starting point to evaluate the potential as well as the results of different climate change mitigation related programs [31]. A baseline is the point of comparison to evaluate the behavior of different systems or procedures and allows to determine over or under performances. Thus, determining an energy consumption baseline is a key issue when implementing energy efficiency measures, deploying energy management programs, analyzing energy performance, and evaluating demand side management programs [8, 24, 31, 32].

When trying to determine if an industrial site or equipment is working under normal conditions, it is important to be able to compare their energy consumption with a "business as usual" forecasted one. This business as usual energy consumption is considered as the baseline scenario for comparison. This concept is particularly important in energy performance and efficiency contracts. The baseline allows the detection of abnormal consumption behaviors and/or overconsumption of equipments. Real-time monitoring of energy consumption helps an industrial site to optimize its energy consumption, reduce its costs, and adapt to changing electricity prices.

Energy efficiency has become a key parameter to be monitored by plant operators and managers aiming at optimizing their costs and reducing their energy losses [11]. Nowadays, most of the existing energy management platforms in the industry have a rather static nature, not adapting to real-time variability and having fixed thresholds, alarms and follow-up procedures. Energy management platforms should allow industrials to monitor their energy consumption and thus optimize their costs and detect abnormal behaviors on a real-time basis [16, 30].

Industrial sites are eager to implement energy efficiency recommendations. However, industry consumption may vary enormously from site to site and from sector to sector, and companies may deal with energy efficiency measures differently [1]. Added to this, there is a lack of relevant scientific literature for integrating energy performance in production management [4]. Baselines need to be consistently defined [31] and data analysis shall be as close as possible to standardized procedures in order to deploy energy management protocols faster and thus, reach as much industries as possible to increase the economical impacts due to energy efficiency [24].

Real-time energy consumption follow-up belongs to Automatic Monitoring and Targeting (AMT) techniques. AMT can be improved by the enhancement of the capabilities of the intrinsic data analysis methods used within an energy management platform. Adaptive methods for real-time energy consumption monitoring and analysis will lead to new methods of forecasting for establishing consumption baselines and thus, better energy consumption follow-up.

The main objective of this research study is to propose two different Short Term Load Forecasting (STLF) approaches for establishing a specific electricity consumption baseline on industrial data. The proposed techniques are applied for forecasting the power consumption of three different industrial sites from France,

from different sectors, at different moments of the day, and for short term forecasting horizons (2 h).

2 Materials and Data

For the purpose of this study, electricity consumption data was collected from three industrial sites from different regions in France (hereafter identified as sites A, B, and C). The three industrial sites belong to three different sectors and present different consumption patterns that are described below.

A big issue when implementing energy management programs is data availability. Generally speaking, energy consumption data at an industrial site level is monitored for billing and accounting purposes. This is not always the case with disaggregated data at workshop or equipment level, where metering instruments may be scarce. Other influential parameters are also not always monitored and thus not available on a first approach.

The only available monitored variable for the three sites is electricity consumption issued from historical billing data. The collected data is a 10 min interval load curve (each value being the average power consumption over a fixed 10 min interval). Each one of these intervals corresponds to each 10 min of the day from 12:00 am to 11:50 pm, which means 144 power consumption values for every available day. For the implementation of the different methods, the `R` software is used (N.B.: Due to confidentiality issues, orders of magnitude of the load curves have been omitted).

Site A

This particular site operates on an 8-h shift mainly from Monday to Friday, and in some occasions, on Saturdays. Not all weekdays present an operating electrical consumption activity, due to operational constraints of the site. Data is available for almost 2 years of electricity consumption. Figure 1 shows a 4-week interval of the load curve. Three main electricity consumption equipments are present at this site, which are turned on once the site is operating. Different equipment arrangements are operated as reflected in the load curve. For site A, 702 days of electricity consumption are available for analysis.

Site B

The second industrial site operates in a continuous 24 h cover, comprising three 8-h shifts. As it can be observed in Fig. 1, the consumption level might have big variations, since different workshops and equipments are engaged at different times

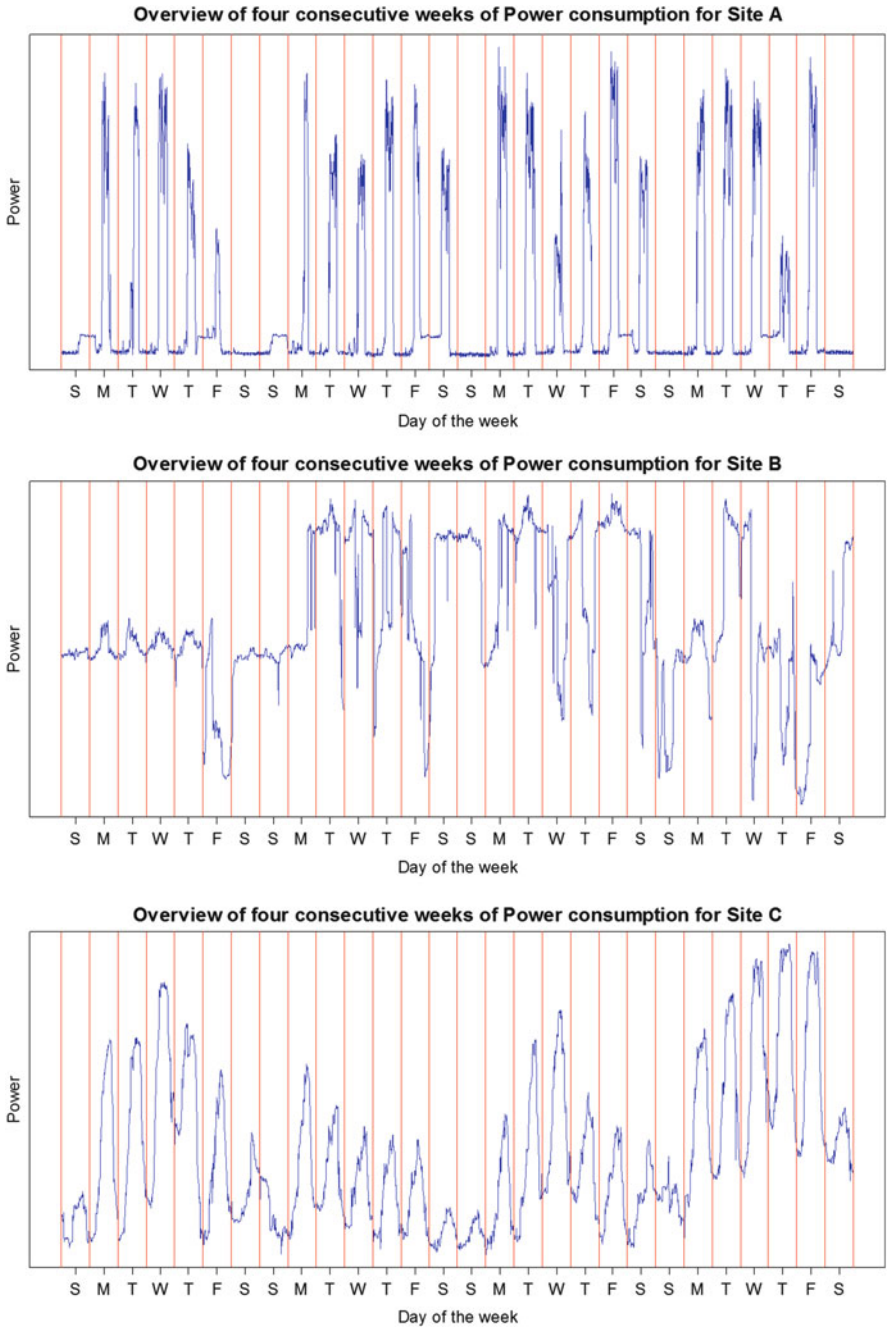


Fig. 1 Four weeks of electricity consumption for sites A, B and C

of the day. The load curve shape is significantly different than that of site A. For site B, 665 days of electricity consumption are available for analysis.

Site C

Site C is also a continuous 24 h cover industrial site. The consumption pattern is dependent on a daily activity as it is reflected in the load curve, shown in Fig. 1. During weekends, energy consumption is different than during weekdays. However, the electricity consumption base represents the biggest part of the consumed energy. Equipments keep consuming electricity during the night and during weekends in order to ensure certain operations at the site. For site C, 770 days of electricity consumption are available for analysis.

3 Forecasting Characteristics and Methods

Current electricity consumption forecasts are generally performed at a regional or national level, since their main interest is to ensure the efficient management of existing electrical power systems. National electricity loads have been at the core of electricity load forecasting for many years, and many techniques and methods have been proposed and assessed, as reviewed by many different authors [13, 15, 26, 27]. The different existing load forecasting methods can be classified into three main families: time-series analysis, multivariate analysis, and data-mining techniques [19]. However, when forecasting electricity consumption of industrial sites, the consumption may differ enormously in form, variability and influencing parameters for every single different site.

There is a lack of scientific literature focused on applying forecasting methods for establishing consumption baselines at lower consumption levels. Typical forecasting methods tend to be not well adapted when applied at an industrial site level. Seasonality, calendar events, and weather dependency are parameters usually taken into account when modeling a national electricity consumption curve [3]. However, due to the radically different nature of industrial sites, these parameters are inconsistent from site to site and may not be reflected in the consumption curve. Innovative approaches shall be followed to standardize the methods and have a larger reach and impact, as it was previously discussed.

When deploying energy management platforms in the industry, one of the main assumptions shall be that predictive models shall work with as little parameters as possible. As previously discussed, for industrial sites A, B, and C, the only available variable is historical electricity consumption. This section presents two different proposed forecasting techniques as well as the methods chosen to evaluate their relevance.

Table 1 lists all the symbols and parameters used in the text.

Table 1 List of symbols used in the text

Symbol	Definition
	Generic symbols
ξ	Gross energy deviations of the forecasted energy baseline with respect to the real energy consumption during the evaluation period (2 h)
i	Time period identifier
n	Time period at which a forecast is launched
t_n	Starting time of the forecast
P_i	Real power consumption at time period i
$\hat{P}_i$	Forecasted power consumption at time period i
θ	Forecast evaluation period (2 h)
N	Number of power consumption values during the evaluation periods (12)
TV_n	Truncated test vector up to the nth interval.
p	Dimension size of the individuals (144 power consumption points)
v	Number of intervals used to construct the adjustment factor
FAJ_v	Adjustment factor using v intervals
	Method 1
M_1,M_2	Dimensions of the SOM grid
m	Number of neurons
k	Identifier of the neuron
Cl_k	Coordinate vectors of the different neurons
$Cl_{k,h}$	Coordinate of the hth element of the kth neuron
$Cl_{k,n}$	Coordinates vector of the truncated cluster vectors up to the nth interval.
WN	Reference vector corresponding to the winning neuron (Also known as the BMU)
X_{tr}	Vector corresponding to the chosen element from the training data
λ	Number of iterations of the SOM algorithm
s	Current iteration step of the SOM algorithm
α	Learning rate of the SOM algorithm
σ	Radius of the neighborhood of the SOM algorithm
D_k	Distance of the updated node to the winning unit
Cl_W	Winning reference vector
	Method 2
Λ	Eigenvectors matrix
U	Eigenvalues matrix of the principal components
A	Covariance matrix
j	Number of principal components explaining 90 % of the data variability
q	Principal component identifier
U_q	Coordinates of the qth principal component
$\hat{C}_q$	MLR coefficients for the qth principal component
$\hat{\varepsilon}$	MLR disturbance coefficient
Un_q	Coordinates vector of the truncated principal component up to the nth interval

3.1 Forecasting Characteristics

For constructing the different models, every day is considered as an independent element (a vector) composed of 144 values of power consumption. Days can be considered to be independent for practical purposes: forecasts are performed intra-day and consumption cycles present daily patterns in most of the cases, corresponding to different consumption modes. Simple data splitting [22] is used for model validation. Eighty-five percent of the data (85 % of the available days for analysis) is used as the training dataset. The remaining 15 % (test dataset) is used to test the models and compare the performance of both methods. Data sampling of the days is performed randomly on a stratified manner at two levels, in order to obtain a distribution of different seasonal variabilities related to time parameters: day of the week and month of the year.

In order to test the different methods for power consumption forecasting at the site level and at different moments of the day, different parameters and characteristics for the forecasts have to be defined. For each test day, the baseline load is estimated at each hour from 9:00 am to 5:00 pm for site A and from 9:00 am to 9:00 pm for sites B and C. In order to evaluate the performance of the forecasting methods, the forecasting periods are fixed to be the following 2 h (called forecast evaluation period, identified by θ, composed of 12 power consumption intervals), considered as short term load forecasting (STLF). In short the different methods will forecast the power and energy consumption from a specific time-step (called t_n, which will be varied from 9:00 am to 5:00 pm or 9:00 pm, depending on the site) for a specific number of time intervals (called N, which has been defined as 12) that corresponds to 2 h.

In total, for site A, 882 simulations will be performed (98 test days, 9 simulations per day from 9:00 am to 5:00 pm), 1,170 for site B (90 test days, 13 simulations per day from 9:00 am to 9:00 pm), and 1,339 for site C (103 test days, 13 simulations per day, similar as for site B). The simulation results will be compared according to the performance indicator defined further on.

3.2 Proposed Forecasting Methods

If the objective is to analyze as many sites as possible, methods shall be easy to deploy and should not require much human input or expertise. Also, they shall demand low calculation times in order to easily standardize the procedures. To overcome these problems, two different approaches for electricity consumption forecasting are proposed, based on the nature of the examined data:

- A first method using load curves clustering in order to detect consumption patterns that will be used as electricity consumption forecasts.

- A second method based on signal decomposition in order to detect the variability of the daily behavior of the curves, using the eigenvectors issued from a Principal Component Analysis (PCA) of the training dataset.

3.2.1 Method 1: Electricity Consumption Forecasting by Pattern Recognition Using Load Curve Clustering

Electrical load curve clustering has attracted much attention in recent years for its application in client profiling and electricity pricing [6, 10, 20]. The capacity of clustering techniques for handling large amounts of time-series data has been assessed in the past [23]. Diverse clustering techniques have been used in the past, as reviewed by Chicco in [5]. From the different assessed clustering techniques, K-means and Self-organizing Maps (SOM) are the best performing ones. SOM is not a direct clustering method, as explained in [6], however, it produces a visually understandable projection of the original data into a map. In this study, SOM has been chosen as the clustering technique due to its prior application for forecasting purposes, as proposed by different authors [7, 18, 25]. Nevertheless, these previous work were focused in forecasting national electricity demand.

SOM [17, 23, 28] is an unsupervised neural network that projects a p-dimensional input dataset onto a reduced dimension space (one or two-dimensional in most cases). SOM is composed of a predefined grid of $M1 \times M2$ elements called neurons (m number of neurons). Each neuron (identified by k) is also p-dimensional. Neurons have to be initialized, this means, the p-dimensions of the k neurons have to be previously defined by a reference vector Cl_k, as in expression (1), where $1 \leq k \leq m$, and $P_{k,i}$ is the value of power consumption for element i of neuron k, where $1 \leq i \leq p$. Initialization of the SOM algorithm can be done in different manners (randomly or data analysis based initialization) as described in [2].

$$Cl_k = [P_{k,1}, P_{k,2}, \ldots P_{k,p}] \tag{1}$$

All neurons are associated to neighboring neurons of the map by a neighborhood relation, which determines the "area" of influence within the defined space. Neurons are calculated through a competitive algorithm that recalculates the weights of the winning neuron and the weights of its neighboring neurons proportionally inverse to their distance. The neighborhood size will be reduced at each iteration during the map training process, starting with almost the full topography and ending in single neurons.

Once all Cl_k reference vectors have been initialized, SOM training starts. The algorithm will be run a predefined number of iterations, represented by λ. At each iteration (represented by $s \leq \lambda$), an input vector X_{tr} (as described in formula (2)) issued from the training data set is chosen randomly, where tr goes from 1 to the number of individuals in the training dataset, and $P_{tr,i}$ is the value of power consumption for element i of the tr individual and where $1 \leq i \leq p$. Euclidean distances between the chosen X_{tr} and all the Cl_k vectors are calculated. The closest

reference vector is known as the winning neuron (*WN*) or best matching unit (BMU), as in expression (3).

$$X_{tr} = [P_{tr,1}, P_{tr,2}, \ldots, P_{tr,p}] \tag{2}$$

$$WN = argmin \left\{ \sqrt{\sum_{i=1}^{i=p} (X_{tr,i} - Cl_{k,i})^2)} \right\}; 1 \leq k \leq m \tag{3}$$

The coordinates of *WN* and its neighboring neurons are adjusted then towards the coordinates of the input vector X_{tr}, as in expression (4), where $\alpha(s)$ is the learning rate which decays with each iteration, and $\theta(s)$ is the neighborhood function, represented in expression (5). The radius $\sigma(s)$ is also updated at every iteration, shrinking over time. D_k is the distance of the updated node to *WN* (the winning neuron).

$$Cl_k(s+1) = Cl_k(s) + \alpha(s)\theta(s) \cdot (X_{tr}(s) - Cl_k(s)) \tag{4}$$

$$\theta(s) = exp(-\frac{D_k^2}{2\sigma^2(s)}) \tag{5}$$

The proposed methodology based on pattern recognition using SOMs is described below and divided into three steps:

1. Once the data splitting has been performed as described previously, the training dataset will be used to construct the reference vectors.

The SOM algorithm is launched considering the daily load curves as individuals for analysis ($p = 144$). As defined previously, the SOM algorithm needs a number of clusters (m) before its initialization. Tsekouras et al. [29] have determined that for large electricity customers 8–12 clusters are necessary for a satisfactory description of the daily load curves. Different numbers of clusters will be tested in order to determine a good description of the different possible load curves. The algorithm is performed on non-reduced data, as suggested in [10]. For the purpose of this study, in order to converge to the same solution, a linear initialization is used. A rectangular configuration of the neighborhood is chosen due to its visualization effectiveness. The chosen number of iterations is $\lambda = 100$. The chosen learning rate is a linear function from 0.05 to 0.01 over the 100 iterations for which it was found that the algorithm converges rapidly. The neighborhood radius is varied from a value of two thirds of all unit to unit distances to its negative value, linearly through the different iterations. Once the neighborhood gets smaller than one individual, only the *WN* reference vector is changed.

The resulting Cl_k reference vectors of each neuron are then kept and assigned to the neuron according to the described SOM algorithm. Every cluster is then represented by a vector composed of 144 variables identified as Cl_k, the identifier of the cluster. $Cl_{k,h}$ contains the value of the hth variable of the kth neuron. Every

variable represents a specific power consumption point of the day, as defined previously.

2. Once the SOM algorithm has been performed, the Cl_k centroid vectors will be used. At the time (t_n) a forecast is simulated for a chosen individual of the test dataset, the test element is truncated to a vector (identified as TV_n) composed of n elements as shown in expression (6) (where $n \leq p$)

$$TV_n = [P_{n,1}, P_{n,2}, \ldots, P_{n,n}] \tag{6}$$

The different Cl_k vectors will be then truncated up to the nth element and called $Cl_{k,n}$. Euclidean distances will then be calculated for the TV_n vector to the different $Cl_{k,n}$ vectors. The vector corresponding to the minimum distance is then considered the winning vector, identified by Cl_W as in expression (7).

$$Cl_W = argmin \left\{ \sqrt{\sum_{i=1}^{i=n} (Cl_{k,i} - TV_{n,i})^2)} \right\}; 1 \leq k \leq m \tag{7}$$

3. The coordinates of the cluster Cl_W corresponding to that vector will be proposed as the forecast for the following consumption points. The forecasted power consumption points $\hat{P}_i$ will correspond to those the elements with the same index i of the closest Cl_W vector, represented by $Cl_{W,i}$ as expressed in formula (8).

$$\hat{P}_i = Cl_{W,i} \tag{8}$$

3.2.2 Method 2: Electricity Consumption Forecasting by Signal Decomposition Using Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate analysis technique used in many different areas for analyzing large sets of data [9, 14]. The pertinence of PCA coupled with other techniques for forecasting energy consumption has been assessed by some authors [21, 27]. However, for different applications, the PCA is used as a tool among others to produce a specific predictor, adapted to the nature of the data.

It can be assumed that the electricity consumption at the site level is a function composed of different signals. The changes and variability in electricity consumption can be explained by different variables. The PCA allows to obtain the eigenvalues (matrix Λ) that explain most of the variability of the data and the eigenvectors (matrix U) of the principal components which are obtained by the decomposition of covariance matrix A in ${}^tU\Lambda U$ that are uncorrelated to each other.

For the purpose of this study, the PCA is performed on the training data set (on non-reduced data). The coordinates of the j first eigenvectors explaining 90 % of the variability are preserved. These coordinates have in fact a meaning according to a specific power consumption interval of the day, since the reduced variables are

in fact power consumption time intervals. U_q are the coordinates of the different j eigenvectors, where $1 \leq q \leq j$, as expressed in (9).

$$U_q = [P_{q,1}, P_{q,2}, \ldots, P_{q,144}] \tag{9}$$

The preserved principal components are used to build a predictor based on their linear combination in order to predict the variability of the data. This is done by using a Multiple Linear Regression (MLR). At the time t_n a prediction is launched for a chosen element of the test dataset, the different U_q vectors are then truncated up to the point n of the forecast and represented by Un_q, as expressed in formula (10).

$$Un_q = [P_{q,1}, P_{q,2}, \ldots, P_{q,n}] \tag{10}$$

A multiple linear regression is used in order to find the coefficients for the principal components, for which the combination of these components fit the data of the chosen element of the test dataset. For this purpose, an ordinary least squares model is used. The vector TV_n is defined as a function ($P(i)$) determining the power consumption value at timestep i. Along with the different j truncated Un_q eigenvectors, $P(i)$ is used to fit a linear model as in expression (11). The coefficients $\hat{C}_q$ and the intercept term $\hat{\varepsilon}$ are obtained through the MLR.

$$P(i) = \sum_{q=1}^{j} \hat{C}_q Un_q(i) + \hat{\varepsilon} \tag{11}$$

These coefficients and the eigenvectors coordinates U_q are used for predicting the power consumption of the site for the rest of the day for every consumption point $\hat{P}_i$, as in expression (12).

$$\hat{P}_i = \sum_{q=1}^{j} \hat{C}_q U_{q,i} + \hat{\varepsilon} \tag{12}$$

3.3 Adjustment Factor

An adjustment factor can be used to improve the forecasts of different techniques. Method 1 is based on pattern recognition, and the proposed forecast is a typical energy consumption mode. However, even if the a pattern has been correctly recognized, the forecast may under or over estimate the actual consumption level corresponding to a specific day. An adjustment factor may deal with this problem by adjusting the forecast to the correct level of energy consumption. The adjustment factor deals as well with the issue of trends, in case they exist. Method 2 forecasts

are issued from a Multiple Linear Regression and thus, the different coefficients fit a model to the actual consumption level and no adjustment factor is needed.

Different forms of adjustment factors exist, but the most important ones can be classified in two different categories for univariate methods: scalar and additive, as described by different authors [8, 12]. Since only electricity consumption information is available for the concerned industrial sites, weather or other related adjustment factors will not be considered in this study.

To deal with the mentioned issues, a scalar adjustment factor is used to improve the forecasts of Method 1. The proposed adjustment factor is calculated as in expression (13), and corresponds to the average of the ratios of the v previous real power consumption values to predicted ones. P_i represents the real power consumption at time interval i, $\hat{P}_i$ is the forecasted power consumption at time interval i, and v the number of intervals used to construct the adjustment factor. The chosen number of intervals (v) for the adjustment factor is one, since better results are obtained in terms of the chosen performance indicator and since calculation times are reduced.

$$FAJ_v = [\frac{P_{i-1}}{\hat{P}_{i-1}} + \cdots + \frac{P_{i-v}}{\hat{P}_{i-v}}] \times \frac{1}{v} \tag{13}$$

3.4 Performance Indicator

The main indicators used in literature to evaluate the performance of forecasting methods are the Mean Absolute Percentage Error (MAPE) and the Mean Squared Error (MSE). These parameters are adequate when evaluating the resemblance of a forecast compared to a real curve. These indicators are adapted to situations where the goal is to optimize the use of production means to meet an electricity demand, or residual demand curves calculation in competitive electricity markets. This is not the case when evaluating the forecasting performance in an industrial site for energy efficiency purposes. Most energy efficiency programs have an economic constraint and are rewarded or penalized economically if objectives are met or not [31]. For this reason, a specific performance criterion is proposed and used which is easily transformed into an economic indicator.

This criterion is directly linked to the site's global energy consumption. It is based on the difference, in energy (kWh), between forecasted energy consumptions issued from the models and real energy consumptions issued from the data. The indicator is based on gross energy differences (hereafter referred as Gross Energy Deviation, GED) through the time period of the forecast (2 h), and represented by symbol ξ. GED and its distribution will be used to evaluate the relevance of each method. Expression (14) formalizes the way of calculating these deviations, where P_i is the actual power consumption at time-step i, and $\hat{P}_i$ is the forecasted power consumption at that same time-step. θ and N were previously defined.

$$\xi = \left[\sum_{i=t_n}^{t_n+N} P_i - \sum_{i=t_n}^{t_n+N} \hat{P}_i \right] \times \frac{\theta}{N} \tag{14}$$

GED allows to evaluate the distribution of the forecasts in terms of how much is the forecast above or below an energy threshold which is the real energy consumption during that time period. This approach is useful to set operational parameters and thresholds in the industry, and to easily translate them into an economic indicator.

4 Results and Discussion

The results for each of the implemented methods are described below. A focus is made on the evaluation of the performance of both forecasting methods presented above, according to the defined criteria.

4.1 Method 1: Electricity Consumption Forecasting Using Self-Organizing Maps

For the implementation of Method 1, different tests were carried out varying the neurons number from 8 to 12 (as in [29]), selecting the lowest number of neurons for which the GED distribution is does not vary greatly if another neuron is proposed. Twelve neurons were selected for site A, 9 for site B, and 12 for site C. The graphical representation of the different reference vectors for the sites can be seen in Fig. 2. The different identified patterns correspond to the different typical consumption modes of the sites. These typical load curves are used for forecasting as explained previously.

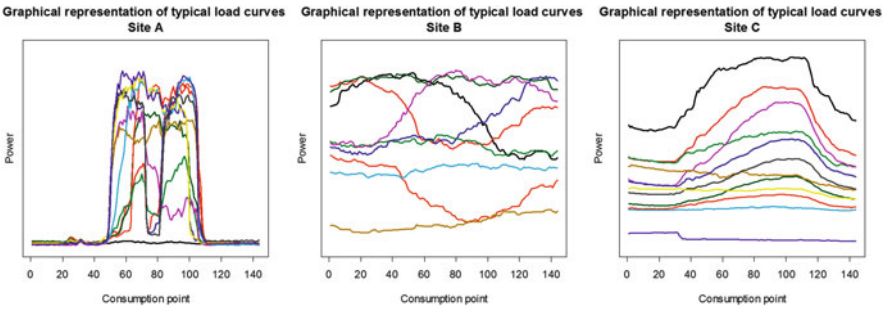


Fig. 2 Resulting curves after applying the SOM algorithm to the test dataset for the three sites

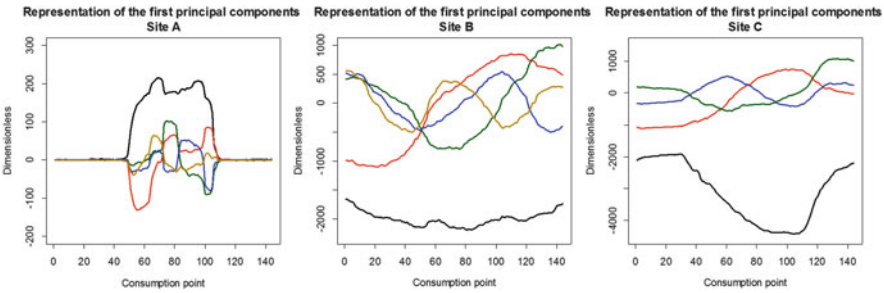


Fig. 3 Graphical representation of the main principal components coordinates for the three sites

4.2 Method 2: Electricity Consumption Forecasting Using Principal Component Analysis

For the implementation of Method 2, in order to explain 90 % of the variability of the data, the first 5 principal components are selected for sites A and B, and 4 principal components are selected for site C. Adding more principal components or increasing the 90 % threshold would increase calculation times, which is to be taken into account when monitoring energy consumption in real time. The footprint of the different components for the different sites can be seen in Fig. 3. These principal components are the ones used to run the MLR that will determine the coefficients for the forecasting models.

4.3 Results by Site

Results obtained using both methods for each of the studied sites are presented below.

Site A

The distribution of the different obtained GED for site A with Method 1 is presented in Fig. 4. The "y" axis represents the gross energy deviation and the "x" axis is the energy that was actually consumed during that period, in order to relativize the error of the forecast in terms of energy. Points outside the dashed lines are above a 50 % GED threshold, and points outside the solid lines are above a 10 % threshold. In order to evaluate the performance of the methods at differents times of the day, different hours were grouped into four different time-spans: from 9:00 am to 11:00 am (morning), identified by the solid green squares; from 12:00 pm to 2:00 pm (noon), identified by the solid pink circles; from 3:00 pm to 5:00 pm (early

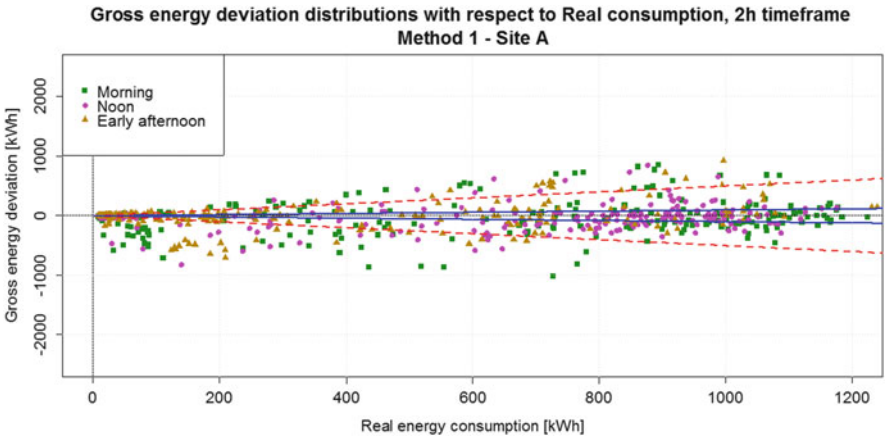


Fig. 4 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 1 for site A

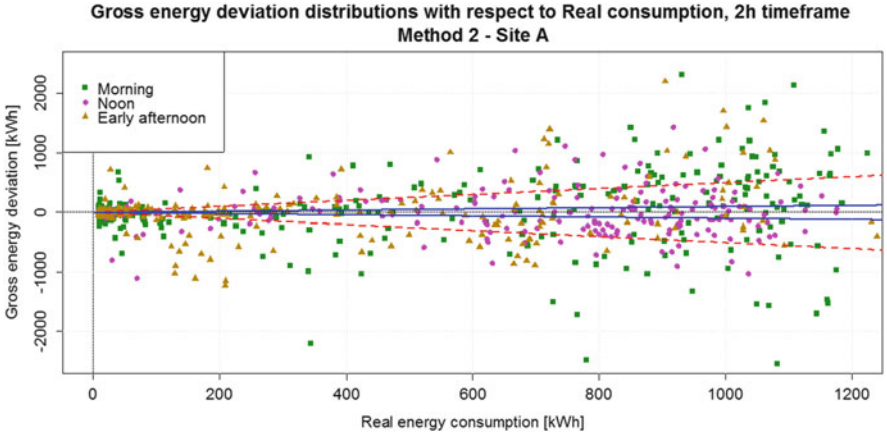


Fig. 5 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 2 for site A

afternoon), identified by the solid yellow triangles, and from 6:00 pm to 9:00 pm (late afternoon), identified by black crosses.

For Method 1, 226 simulation points are inside the solid lines and 642 are between the dashed lines. The total simulation points for this site are 882. Figure 5 represents the GED distributions for site A using Method 2. Solid lines and dashed lines represent the same thresholds as in Fig. 4. For Method 2, only 104 simulation points are inside the solid lines, while 410 are between the dashed lines.

Looking at the dispersion of the points and the number of them outside of the defined thresholds, of Figs. 4 and 5, Method 1 clearly outperforms Method 2 for this particular industrial site. Regarding the distribution of the different timespans,

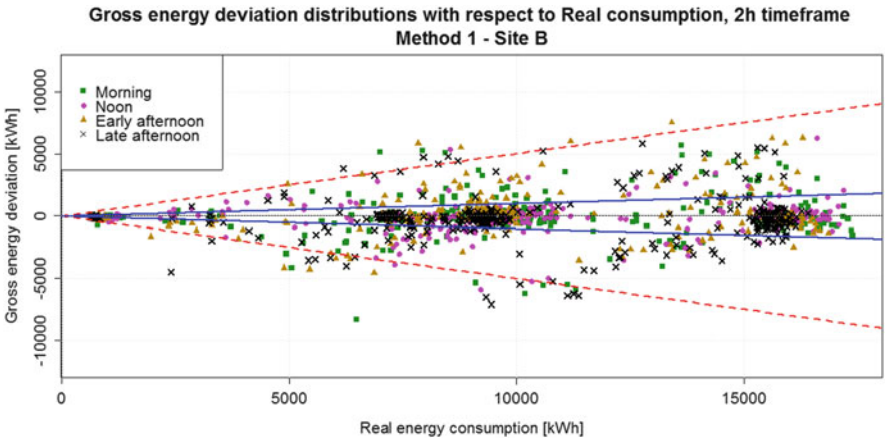


Fig. 6 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 1 for site B

besides a slightly wider distribution for the morning period, no significant difference can be observed for the different hours of the day.

Site B

Figure 6 shows the GED distribution for site B using Method 1 as in Fig. 4. For this site, 767 simulation points are inside the solid lines and 1,123 are between the dashed lines. The total simulation points for this site are 1,170.

Figure 7 represents the GED distributions for site B using Method 2. Solid lines and dashed lines represent the same thresholds as in previous figures. For Method 2, 467 simulation points are inside the solid lines, and 1,001 are between the dashed lines.

For this industrial site, Method 1 also outperforms Method 2. As for the distribution regarding the different timespans, no significant difference can be observed to conclude a strong influence of the hours of the day for both methods.

Site C

Figure 8 shows the GED distribution for site C using Method 1 as in Fig. 6. For this site, 1,146 simulation points are inside the solid lines, which represent less than 10 % in error, and 1,334 are between the dashed lines that represent less than 50 % in energy error. The total simulation points for this site are 1,339. It is important to notice that only five points are outside the dashed line boundaries in this particular case.

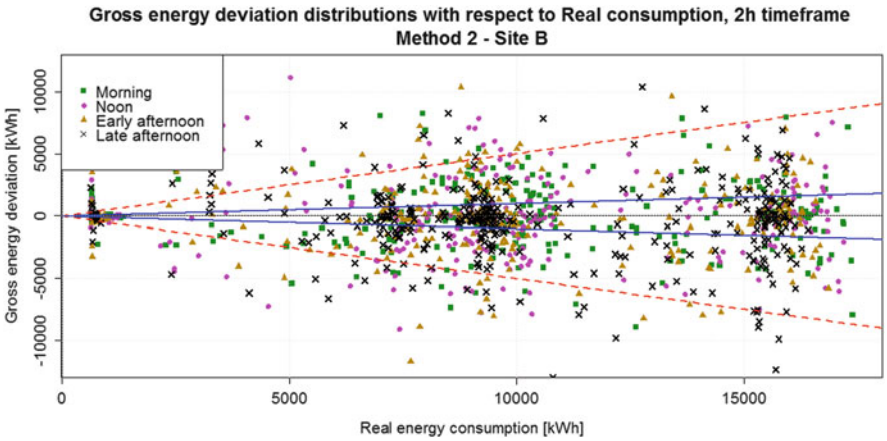


Fig. 7 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 2 for site B

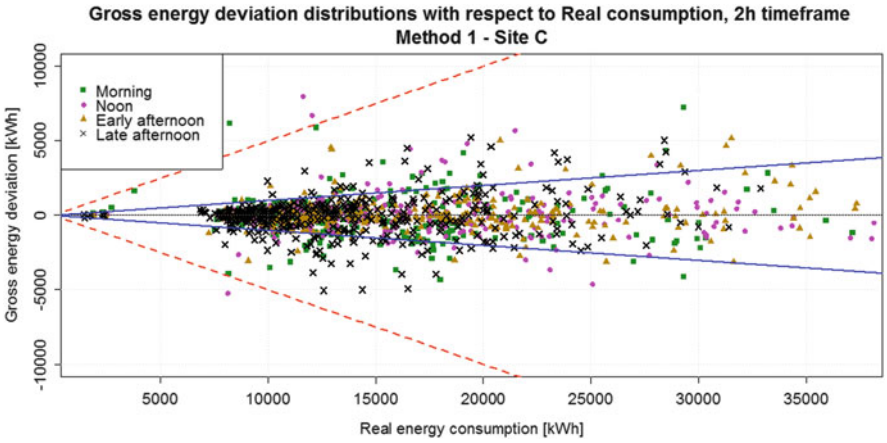


Fig. 8 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 1 for site C

Figure 9 represents the GED distributions for site C using Method 2. Solid lines and dashed lines represent the same thresholds as in previous figures. For Method 2, 884 simulation points are inside the solid lines, and 1,319 are below the 50 % threshold represented by the dashed lines.

Even though results can be considered satisfactory for Method 2 applied to industrial site C, Method 1 still shows better performances. As well as for sites A and B, the hour of the day does not seem to influence greatly the performance of the methods, since the GED distributions are evenly distributed for all of the timespans.

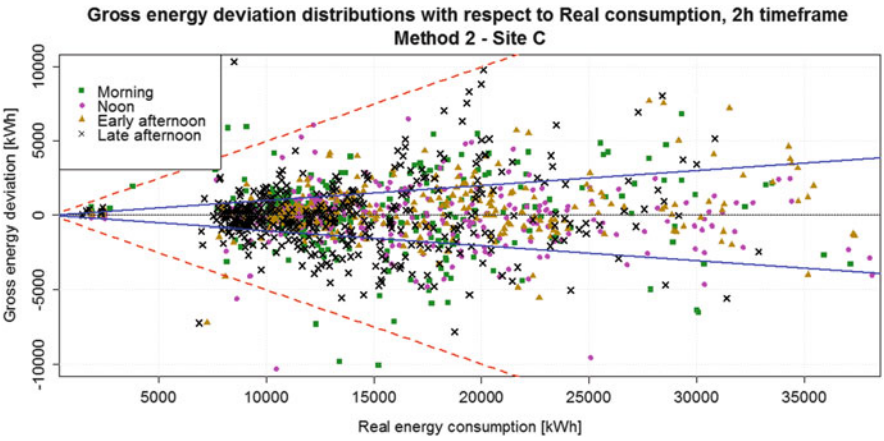


Fig. 9 GED distribution with respect to real energy consumption over a timeframe of 2 h using Method 2 for site C

5 Conclusions and Perspectives

Two different methods for establishing short-term electricity consumption baselines were proposed and assessed. From the obtained results, Method 1 outperforms Method 2 when forecasting the short term electricity consumption for the three presented industrial sites, according to the chosen performance indicator. Added to this, the hour of the day does not significantly influence the performance of the methods.

Subsequent works will focus on specific industrial equipments that are installed at the industrial sites and that contribute to most of their power consumption. The aggregation of industrial equipments allows a more flexible and adaptable energy consumption follow-up, since information can be lost at the industrial site level. In order to ensure the validity and repeatability of the obtained results for their generalisation, future research works will focus on the construction of a bootstrapping procedure.

Perspectives to improve the forecasting potential for Method 2, could be the integration of weighing factors for the coefficients and studying the errors obtained for the different forecasts at different times of the day.

Model combination could be a clue to improve the performance of the forecasts, since it could integrate different approaches (such as form recognition and Bayesian inference) in order to overcome the deficiencies of the different methods.

It is important to point out that due to the variability of the data, the differences from site to site and from sector to sector, standardizing the methods to build energy consumption baselines can be a hard task. The use of additional variables shall be considered when possible, which will make the methods more adaptable. Univariate

methods could rapidly reach a limit of performance. The main problem which may persist will be data availability.

Energy management can be improved by the utilization of different methods to calculate energy consumption baselines for the diverse energy management applications. Performing bottom-up approaches provides more precise information and makes energy consumption flexibility fast and reactive.

References

1. Alhourani, F., & Saxena, U. (2009). Factors affecting the implementation rates of energy and productivity recommendations in small and medium sized companies. *Journal of Manufacturing Systems, 28*(1), 41–45. doi:10.1016/j.jmsy.2009.04.001.
2. Attik, M., Bougrain, L., & Alexandre, F. (2005). Self-organizing map initialization. In *Artificial neural networks: biological inspirations – ICANN 2005*, Warsaw (pp. 357–362).
3. Bunn, D. W., & Farmer, E. D. (1985). *Comparative models for electrical load forecasting*. Chichester/New York: Wiley.
4. Bunse, K., Vodicka, M., Schönsleben, P., Brülhart, M., & Ernst, F. O. (2011). Integrating energy efficiency performance in production management – Gap analysis between industrial needs and scientific literature. *Journal of Cleaner Production, 19*(6–7), 667–679. doi:10.1016/j.jclepro.2010.11.011.
5. Chicco, G. (2012). Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy, 42*(1), 68–80. doi:10.1016/j.energy.2011.12.031.
6. Chicco, G., Napoli, R., & Piglione, F. (2006). Comparisons among clustering techniques for electricity customer classification. *IEEE Transactions on Power Systems, 21*(2), 933–940. doi:10.1109/TPWRS.2006.873122.
7. Cottrell, M. (2003). Some other applications of the SOM algorithm: How to use the Kohonen algorithm for forecasting. In *Invited lecture at the international work-conference on artificial neural networks, IWANN 2003:* Maó, Menorca, Spain.
8. Coughlin, K., Piette, M. A., Goldman, C., & Kiliccote, S. (2009). Statistical analysis of baseline load models for non-residential buildings. *Energy and Buildings, 41*(4), 374–381.
9. Daultrey, S. (1976). *Principal components analysis*. Norwich: Geo Abstracts.
10. Fidalgo, J. N., Matos, M. A., & Ribeiro, L. (2012). A new clustering algorithm for load profiling based on billing data. *Electric Power Systems Research, 82*(1), 27–33. doi:10.1016/j.epsr.2011.08.016.
11. Giacone, E., & Manc, S. (2012). Energy efficiency measurement in industrial processes. *Energy, 38*(1), 331–345. doi:10.1016/j.energy.2011.11.054.
12. Goldberg, M. L., & Kennedy Agnew, G. (2003). *Protocol development for demand response calculation: Findings and recommendations* (Technical report). KEMA-Xenergy.
13. Hahn, H., Meyer-Nieberg, S., & Pickl, S. (2009). Electric load forecasting methods: Tools for decision making. *European Journal of Operational Research, 199*(3), 902–907
14. Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary: Sas Institute.
15. Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems, 16*(1), 44–55.
16. Hu, S., Liu, F., He, Y., & Hu, T. (2012). An on-line approach for energy efficiency monitoring of machine tools. *Journal of Cleaner Production, 27*, 133–140. doi:10.1016/j.jclepro.2012.01.013.
17. Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*(9), 1464–1480. doi:10.1109/5.58325.

18. Lendasse, A., Lee, J., Wertz, V., & Verleysen, M. (2002). Forecasting electricity consumption using nonlinear projection and self-organizing maps. *Neurocomputing, 48*(1), 299–311.
19. Li, D. C., Chang, C. J., Chen, C. C., & Chen, W. C. (2012). Forecasting short-term electricity consumption using the adaptive grey-based approach – An Asian case. *Special Issue on Forecasting in Management Science, 40*(6), 767–773. doi:10.1016/j.omega.2011.07.007.
20. Mahmoudi-Kohan, N., Moghaddam, M. P., & Sheikh-El-Eslami, M. (2010). An annual framework for clustering-based pricing for an electricity retailer. *Electric Power Systems Research, 80*(9), 1042–1048. doi:10.1016/j.epsr.2010.01.010.
21. Manera, M., & Marzullo, A. (2005). Modelling the load curve of aggregate electricity consumption using principal components. *Environmental Modelling & Software, 20*(11), 1389–1400. doi:10.1016/j.envsoft.2004.09.019.
22. McLachlan, G. J. (2004). Discriminant analysis and statistical pattern recognition. Hoboken: Wiley-Interscience.
23. Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., & Kolehmainen, M. (2010). Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy, 87*(11), 3538–3545. doi:10.1016/j.apenergy.2010.05.015.
24. Reichl, J., & Kollmann, A. (2010). Strategic homogenisation of energy efficiency measures: An approach to improve the efficiency and reduce the costs of the quantification of energy savings. *Energy Efficiency, 3*(3), 189–201.
25. Rousset, P. (1999). Applications des algorithmes d'auto-organisation à la classification et àă la prévision. PhD thesis, Université Paris I, Paris.
26. Soliman, S. Ah., & Al-Kandari, A. M. (2010). *Electrical load forecasting: Modeling and model construction.* New York: Elsevier
27. Taylor, J. W., De Menezes, L. M., & McSharry, P. E. (2006) A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting, 22*(1), 1–16.
28. Thang, K., Aggarwal, R., McGrail, A., & Esp, D. (2003). Analysis of power transformer dissolved gas data using the self-organizing map. *IEEE Transactions on Power Delivery, 18*(4), 1241–1248. doi:10.1109/TPWRD.2003.817733.
29. Tsekouras, G., Kotoulas, P., Tsirekis, C., Dialynas, E., & Hatziargyriou, N. (2008). A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. *Electric Power Systems Research, 78*(9), 1494–1510. doi:10.1016/j.epsr.2008.01.010.
30. Vijayaraghavan, A., & Dornfeld, D. (2010). Automated energy monitoring of machine tools. *CIRP Annals Manufacturing Technology, 59*(1), 21–24. doi:10.1016/j.cirp.2010.03.042.
31. Vine, E. (2008). Breaking down the silos: The integration of energy efficiency, renewable energy, demand response and climate change. *Energy Efficiency, 1*(1), 49–63.
32. Vine, E. L., & Sathaye, J. A. (2000). The monitoring, evaluation, reporting, verification, and certification of energy-efficiency projects. *Mitigation and Adaptation Strategies for Global Change, 5*(2), 189–216.

Confidence Intervals and Tests for High-Dimensional Models: A Compact Review

Peter Bühlmann

Abstract We present a compact review of methods for constructing tests and confidence intervals in high-dimensional models. Links to theory, finite sample performance results and software allows to obtain a "quick" but sufficiently deep overview for applying the procedures.

1 Introduction

We review some methods for assigning significance of (co-)variables or for confidence intervals of a parameter in a high-dimensional regression-type model. Our major focus is for a high-dimensional linear model

$$Y = \mathbf{X}\beta^0 + \varepsilon \tag{1}$$

with $n \times 1$ response vector Y, $n \times p$ design matrix $\mathbf{X}$, $p \times 1$ regression vector β^0 and $n \times 1$ error vector ε having i.i.d. components with $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2$ and ε_i uncorrelated from $\mathbf{X}_i$. We also discuss some extensions, including generalized linear models. While there is much literature on convergence rates for parameter estimation and prediction (cf. [6]), only recent work addresses the problem of constructing confidence intervals or tests. Some recent reviews on this topic include Bühlmann et al. [5] with a focus on applications in biology, and Dezeure et al. [8] who present a much more detailed and broader treatment. The current work aims to provide a very compact and "fast to read" access to the topic, yet it still contains the main ideas and hints to software.

P. Bühlmann (✉)
Seminar for Statistics, ETH Zürich, Zürich, Switzerland
e-mail: buhlmann@stat.math.ethz.ch

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_2

2 High-Dimensional Linear Model and Some Methods for Inference

Consider the high-dimensional linear model in (1). The goal is to test null-hypotheses $H_{0,j} : \beta_j^0 = 0$ versus $H_{A,j} : \beta_j^0 \neq 0$ (or a one-sided alternative) for individual variables with index $j \in \{1, \ldots, p\}$, or to construct a confidence interval for β_j^0. In the high-dimensional setting, these tasks are non-trivial since standard least squares methodology cannot be used.

2.1 De-sparsified Lasso

Zhang and Zhang [26] propose a method based on low-dimensional regularized projection using the Lasso. A motivation can be derived from standard least squares: in the low-dimensional setting with $p < n$ and $\mathbf{X}$ having full rank, it is well-known that the ordinary least squares estimator satisfies:

$$\hat{\beta}_{\mathrm{OLS},j} \text{ is the projection of } Y \text{ onto the residuals of } Z_{\mathrm{OLS,j}},$$

where the $n \times 1$ residual vector $Z_{\mathrm{OLS,j}}$ arises from OLS regression of X_j versus all other co-variables $\mathbf{X}_{-j}$ (which is the design matrix without the jth column). In the high-dimensional setting, the projection is ill-defined since the residual vector $Z_{\mathrm{OLS,j}} \equiv 0$. The idea is to replace the residuals by a regularized version: we fit X_j versus $\mathbf{X}_{-j}$ with the Lasso and denote the corresponding residuals by Z_j (when doing this for all j's, this is the nodewise Lasso from Meinshausen and Bühlmann [18]). We then look at the projection

$$Z_j^T Y / Z_j^T X_j = \beta_j^0 + \sum_{k \neq j} \beta_k^0 Z_j^T X_k / Z_j^T X_j + Z_j^T \varepsilon / Z_j^T X_j.$$

The first term on the right-hand side is what we aim for, the second one is a bias, and the third one is the noise component with mean zero. To get rid of the bias, we employ a bias correction using (again) the Lasso: this leads to a new estimator

$$\hat{b}_j = Z_j^T Y / Z_j^T X_j - \sum_{k \neq j} \hat{\beta}_k Z_j^T X_k / Z_j^T X_j \quad (j = 1, \ldots, p), \tag{2}$$

where $\hat{\beta}$ denotes the Lasso estimator for the regression of Y versus $\mathbf{X}$. A typical choice for the regularization parameter involved in Z_j and for $\hat{\beta}$ is based on cross-validation of the corresponding Lasso estimations. The estimator $\hat{b}$ is not sparse and hence the name "de-sparsified Lasso". One can show that the error in bias estimation

is asymptotically negligible [10, 24, 26] on the $1/\sqrt{n}$-scale, and one then obtains

$$\sqrt{n}(\hat{b}_j - \beta_j^0) \Rightarrow \mathcal{N}(0, \sigma_\varepsilon^2 \Omega_{jj}) \ (n \to \infty), \quad \Omega_{jj} = \frac{\|Z_j\|_2^2/n}{(Z_j^T X_j/n)^2}. \tag{3}$$

The convergence as $n \to \infty$ encompasses that the dimension $p = p(n) \gg n$ tends to infinity as well, at a potentially much faster rate than sample size. We thus have an asymptotic pivot and we can then construct p-values for $H_{0,j}$ or confidence intervals by plugging in an estimate for σ_ε^2, see Sect. 2.3. In fact, the asymptotic variance is the smallest possible (among regular estimators) and it reaches the Cramér-Rao lower bound [24]: thus, statistical tests and confidence intervals derived from (3) are asymptotically optimal. Furthermore, the convergence in (3) to a Gaussian limit is uniform for a large part of the parameter space and thus, we obtain honest confidence intervals [11].

It is important to outline the assumptions which are used to establish the result in (3). Assume that the design $\mathbf{X}$ consists of (possibly fixed realizations of) i.i.d. rows whose distribution has a $p \times p$ covariance matrix Σ. The main conditions are as follows:

(A1) The rows of $\mathbf{X}$ have a (sub-)Gaussian distribution and the smallest eigenvalue of Σ is bounded away from zero.
(A2) The matrix Σ^{-1} is row-sparse: the maximal number of non-zero entries in each row is bounded by $o(\sqrt{n/\log(p)})$.
(A3) The linear model is sparse: the number of non-zero entries of β^0 is $o(\sqrt{n}/\log(p))$.
(A4) The error ε has a (sub-) Gaussian distribution.

We note that these assumptions imply the ones in van de Geer et al. [24]. The most restrictive conditions are (A2) regarding the design and (A3) saying that the linear model needs to be rather sparse.

2.2 Ridge Projection

The estimator in (2) is has a linear part and a non-linear bias correction. A similar construction can be made based on the Ridge estimator:

$$\hat{\beta}_{\text{Ridge}} = (n^{-1}\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} n^{-1}\mathbf{X}^T Y. \tag{4}$$

A main message is that the Ridge estimator has substantial bias when $p \gg n$: in fact, it estimates a projected parameter

$$\theta^0 = P\beta^0, \quad P = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^-\mathbf{X},$$

where $(\mathbf{X}\mathbf{X}^T)^-$ denotes a generalized inverse of $\mathbf{X}\mathbf{X}^T$ [22].

The bias for θ^0 can be made arbitrarily small by choosing λ sufficiently small, and a quantitative bound is given in Bühlmann [3]. A potentially substantial bias occurs, however, due to the difference between θ^0 and the target β^0. Since

$$\frac{\theta^0}{P_{jj}} = \beta_j^0 + \sum_{k \neq j} \frac{P_{jk}}{P_{jj}} \beta_k^0,$$

this bias can be estimated and corrected with

$$\sum_{k \neq j} \frac{P_{jk}}{P_{jj}} \hat{\beta}_k,$$

where $\hat{\beta}$ is the Lasso estimator. Thus, we construct a bias corrected Ridge estimator

$$\hat{b}_{R;j} = \frac{\hat{\beta}_{\text{Ridge};j}}{P_{jj}} - \sum_{k \neq j} \frac{P_{jk}}{P_{jj}} \hat{\beta}_k, \; j = 1, \ldots, p. \tag{5}$$

A typical choice of the regularization parameter in (4) for $\hat{\beta}_{\text{Ridge}}$ is $\lambda = \lambda_n = n^{-1}$ and we can use cross-validation for the regularization parameter in the Lasso $\hat{\beta}$. This estimator has the following property [3]:

$$\begin{aligned}
&\sigma_\varepsilon^{-1} \Omega_{R;jj}^{-1/2} (\hat{b}_{R;j} - \beta_j^0) \approx Z + \Delta_j, \;\; Z \sim \mathcal{N}(0,1), \\
&\Omega_R = (\hat{\Sigma} + \lambda)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda)^{-1}, \;\; \hat{\Sigma} = n^{-1} \mathbf{X}^T \mathbf{X}, \\
&|\Delta_j| \leq \sigma_\varepsilon^{-1} \max_{k \neq j} \Omega_{R;jj}^{-1/2} \left| \frac{P_{jk}}{P_{jj}} \right| \|\hat{\beta} - \beta^0\|_1.
\end{aligned} \tag{6}$$

Here, the "$\approx$" symbol represents an approximation which becomes exact as $\lambda \searrow 0^+$. The problem here is that the behavior of $|P_{jk}/P_{jj}|$ and of the diagonal elements $\Omega_{R;jj}$ are not easily under control, but they are observed for fixed design $\mathbf{X}$ so that it is possible to construct an upper bound as discussed next.

2.2.1 Inference Based on an Upper Bound

Assuming the so-called compatibility condition on the design $\mathbf{X}$ [6, Ch.6.2], we obtain that

$$|\Delta_j| \leq \Omega_{R;jj}^{-1/2} \max_{k \neq j} \left| \frac{P_{jk}}{P_{jj}} \right| O_P(s_0 \sqrt{\log(p)/n}),$$

and in practice, we use an upper bound of the form

$$\Delta_j^{\text{bound}} := \Omega_{R;jj}^{-1/2} \max_{k \neq j} \left| \frac{P_{jk}}{P_{jj}} \right| (\log(p)/n)^{1/2 - \xi}, \tag{7}$$

for some small $0 < \xi < 1/2$, typically $\xi = 0.05$; this bound is motivated via an implicit assumption that $s_0 \leq (n/\log(p))^{\xi}$.

Inference can then be based on (6) with the upper bound in (7). For example, for testing $H_{0,j} : \beta_j^0 = 0$ against the two-sided alternative $H_{A,j} : \beta_j^0 \neq 0$ we use the upper bound for the p-value

$$2(1 - \Phi((\sigma_\varepsilon^{-1}\Omega_{R;jj}^{-1/2}|\hat{b}_{R;j}| - \Delta_j^{\text{bound}})_+)),$$

and an analogous construction can be used for a two-sided $1-\alpha$ confidence interval for β_j^0:

$$[\hat{b}_{R;j} - a, \hat{b}_{R;j} + a],$$
$$a = (\Phi^{-1}(1 - \alpha/2) + \Delta_j^{\text{bound}})\sigma_\varepsilon \Omega_{R;jj}^{1/2}.$$

The main conditions used for proving consistency of the Ridge-based inference method are as follows:

(B1) As assumption (A1).
(B2) The linear model is sparse: for $0 < \xi < 1/2$ which is used in (7), the number of non-zero entries of β^0 is $O((n/\log(p))^{\xi})$.
(B3) The error ε has a Gaussian distribution.

It is expected that assumption (B3) could be relaxed to sub-Gaussian distributions as in (A4). No condition is required in terms of sparsity of Σ^{-1} as in (A2), but typically the method does not lead to optimality as with the de-sparsified Lasso estimator from Sect. 2.1.

2.3 Estimation of the Error Variance

The de-sparsified Lasso and the Ridge projection method in Sects. 2.1 and 2.2 require an estimate of σ_ε for construction of tests or confidence intervals.

The scaled Lasso [23] leads to a consistent estimate of the error variance: it is a fully automatic method which does not need a user-specific choice of a tuning parameter. Reid et al. [21] present an empirical comparison of various estimators which suggests that the alternative scheme of residual sum of squares of a cross-validated Lasso solution exhibits has good finite-sample performance.

2.4 Multi Sample Splitting

Sample splitting is a generic method for construction of p-values. The sample is randomly split in two halves with corresponding indices from disjoint sets $I_1, I_2 \subset$

$\{1, \ldots, n\}$, $I_1 \cup I_2 = \{1, \ldots, n\}$ with $|I_1| = \lfloor n/2 \rfloor$ and $|I_2| = n - \lfloor n/2 \rfloor$. A variable selection technique $\hat{S} \subseteq \{1, \ldots, p\}$ is used on the first half I_1, denoted by $\hat{S}(I_1)$: a prime example is the Lasso where $\hat{S} = \{j;\ \hat{\beta}_j \neq 0\}$, and other selectors $\hat{S}$ can be derived from a sparse estimator in the same way. With the fewer variables from $\hat{S}$, we can obtain p-values based on the second half I_2 and using classical t-tests from ordinary least squares: that is, we only use the subsample $(Y_{I_2}, \mathbf{X}_{I_2,\hat{S}})$ of the data, with obvious notational meaning of the sub-indices. Such a procedure is implicitly contained in Wasserman and Roeder [25]. Sample splitting avoids that we would use the data twice for selection and inference which would lead to over-optimistic p-values.

It is rather straightforward to see that such a principle works if

$$\begin{aligned} &\hat{S}(I_1) \supseteq S_0 = \{j;\ \beta_j^0 \neq 0\}, \\ &|\hat{S}(I_1)| < n/2, \end{aligned} \tag{8}$$

where $\hat{S}(I_1)$ denotes the selector based on the subsample with indices I_1. Furthermore, multiple testing adjustment over all components $j = 1, \ldots, p$ (see Sect. 3.2) can be done in a powerful way, e.g., Bonferroni correction only needs an adjustment with a factor $|\hat{S}(I_1)|$ which is often much smaller than p. A drawback of the method is its severe sensitivity of how the sample is split: Meinshausen et al. [20] propose repeated splitting of the sample (multi sample splitting) and show how to combine the corresponding dependent p-values. The latter is of independent interest and the procedure is described below in Sect. 2.4.1.

Such a multi sample splitting method leads to p-values which are already adjusted for multiple testing, either for the familywise error rate or the false discovery rate. The main conditions which are required for the method are (8): when using the Lasso as a screening method (typically with either a cross-validated choice of λ or taking a fixed fraction of the variables entering the Lasso path first), they are implied by the following:

(C1) As assumption (A1).
(C2) beta-min assumption:

$$\min_{j \in S_0} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n},$$

and $s_0 = o(n/\log(p))$ where $s_0 = |S_0|$ denotes the number of non-zero entries of β^0.
(C3) As assumption (A4).

The beta-min assumption in (C2) is rather unpleasant since, for example, we would like to find out with significance testing whether a regression coefficient is large or smallish (or zero): thus, an a-priori assumption excluding smallish coefficients is unpleasant. The condition can be somewhat relaxed to "zonal assumptions" which

still require that there is a gap between large and smallish coefficients and restrict the number of smallish coefficients [4].

2.4.1 Aggregation of p-Values

With the multi sample splitting approach described above we obtain the following: for testing the null-hypothesis $H_{0,j} : \beta_j^0 \neq 0$, when repeating the sample splitting B times, we get p-values

$$P_j^{(1)}, \ldots, P_j^{(B)}.$$

The problem, in general, is how to aggregate many p-values which can be arbitrarily dependent to a single p-value P_j. The following Lemma is very general and might be of interest in other problems.

Lemma 1 ([20]) *Assume that we have B p-values $P^{(1)}, \ldots, P^{(B)}$ for testing a null-hypothesis H_0, i.e., for every $b \in \{1, \ldots, B\}$ and any $0 < \alpha < 1$, $\mathbb{P}_{H_0}[P^{(b)} \leq \alpha] \leq \alpha$. Consider for any $0 < \gamma < 1$ the empirical γ-quantile of the values $\{P^{(b)}/\gamma;\ b = 1, \ldots, B\}$:*

$$Q(\gamma) = \min\left(\textit{empirical } \gamma\textit{-quantile } \{P^{(1)}/\gamma, \ldots, P^{(B)}/\gamma\}, 1\right).$$

Furthermore, consider a suitably corrected minimum value of $Q(\gamma)$ over a range which is lower bounded by a positive constant $\gamma_{\min}$:

$$P = \min\left((1 - \log(\gamma_{\min})) \min_{\gamma \in (\gamma_{\min}, 1)} Q(\gamma), 1\right). \tag{9}$$

Then, both $Q(\gamma)$ (for any $\gamma \in (0, 1)$) and P are conservative p-values satisfying for any $0 < \alpha < 1$: $\mathbb{P}_{H_0}[Q(\gamma) \leq \alpha] \leq \alpha$ or $\mathbb{P}_{H_0}[P \leq \alpha] \leq \alpha$, respectively.

A simple generic aggregation rule is with $\gamma = 1/2$: multiply the raw p-values by the factor 2 and take the sample median. Potential power improvement is possible with an adaptive version searching for the best γ as in (9) but paying a price in terms of the factor $(1 - \log(\gamma_{\min}))$ (which e.g. is $\approx$3.996 for $\gamma_{\min} = 0.05$).

2.5 Stability Selection

Stability Selection [19] is an even (much) more generic method than the multi sample splitting from Sect. 2.4. It can be applied to any structure estimation problem such as edges in a graph: variable selection in a regression problem is a special case thereof which we discuss now a bit further.

As with multi sample splitting, we randomly split the sample in two halves with indices I_1 and I_2, respectively, and we consider a variable selection method $\hat{S} \subseteq \{1, \ldots, p\}$. The idea is to analyze the stability of $\hat{S}(I_1)$, based on the half-sample I_1, when subsampling the data, and in fact, we do not make any use of the other half of the sample I_2. Thus, denote by I^* a random subsample of size $\lfloor n/2 \rfloor$. We consider the event that a single variable j is selected by $\hat{S}(I^*)$ based on the subsample I^*, $j \in \hat{S}(I^*)$, and we compute its probability

$$\pi(j) = \mathbb{P}^*[j \in \hat{S}(I^*)].$$

In practice, this probability is computed based on $B \approx 100$ random subsamples and calculating empirical relative frequencies.

The main problem is to determine a threshold $1/2 < \tau_{\text{thr}} \leq 1$ such that $\pi(j) \geq \tau$ implies that variable j is selected in a "stable way". This can be formalized as follows: denote by $V = |\cup_{j \in S_0^c} \{\pi(j) \geq \tau\}|$, that is, the number of false positive selections. Then, assuming some conditions as outlined below, the following formula holds [19]:

$$\mathbb{E}[V] \leq \frac{1}{2\tau_{\text{thr}} - 1} \frac{q^2}{p}, \tag{10}$$

where $q \geq |\hat{S}(I^*)|$ (almost surely). For example, q can be specified as the top q variables of a ranking (or selection) scheme, e.g., the q variables having largest regression coefficients in absolute value (if there are fewer than q coefficients with non-zero values, just take all of them). For the Lasso based on the first half-sample, since it selects at most $\lfloor n/2 \rfloor$ variables, a good value of q might be in the range of $n/10$ to $n/3$.

The formula (10) can then be inverted to determine a threshold τ_{thr} for a given bound of $\mathbb{E}[V]$ and a given q (which specifies the selection method $\hat{S}$). For example, by tolerating $\mathbb{E}[V] \leq 5$, a specified $q = 30$ and $p = 1{,}000$ we choose

$$\tau_{\text{thr}} = (1 + \frac{q^2}{p} \frac{1}{5})/2 = (1 + \frac{30^2}{1{,}000} \frac{1}{5})/2 = 0.59$$

and such a choice then satisfies $\mathbb{E}[V] \leq 5$. When using the tolerance bound $\mathbb{E}[V] \leq \alpha$, the corresponding threshold τ_{thr} leads to a procedure where

$$\mathbb{P}[V > 0] \leq \mathbb{E}[V] \leq \alpha,$$

and hence, with control of the familywise error rate.

The main assumptions for validity of (10) are here sketched only:

(D1) The selector $\hat{S}$ is performing better than random guessing.
(D2) An exchangeability condition holds implying that it is equally likely that a noise variable is selected by $\hat{S}$.

The formal assumptions are given in Meinshausen and Bühlmann [19]. In fact, assumption (D1) is a mild condition while (D2) is rather restrictive: however, it was shown empirically that formula (10) approximately holds even for scenarios where (D2) does not hold. Interestingly, a beta-min assumption as in (C2) is not required for Stability Selection.

2.6 A Summary of an Empirical Study

We briefly summarize the results from a fairly large empirical study in Dezeure et al. [8]. An overall conclusion is that the multi sample splitting and the Ridge projection method are often somewhat more reliable for familywise error control (type I error control) than the de-sparsified Lasso procedure; on the other hand, the de-sparsified Lasso has often (a bit) more power in comparison to multi sample splitting and Ridge projection. However, these findings depend on the particular case and they are not consistent among all considered settings. Figure 1 illustrates the familywise error control and power of various methods for 96 different scenarios, varying over different covariate designs, sparsity degrees and structure of active sets, and signal to noise ratios.

From a practical point of view, if one is primarily concerned about false positive statements, the multi sample splitting method might be preferable: especially for logistic linear models (see Sect. 3.1), the adapted version of multi sample splitting was found to be most "robust" for reliable error control.

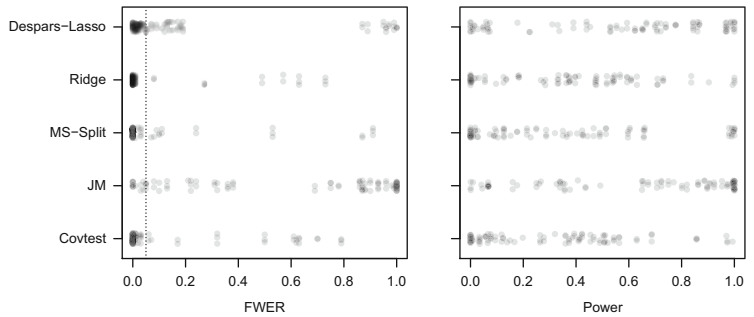


Fig. 1 Ninety-six different simulation scenarios, all with $p = 500$ and $n = 100$, with varying covariate design, sparsity and structure of the active set, and signal to noise ratio. Each *dot* represents a scenario, shown with jittered plotting. Five methods: De-sparsified Lasso (Despars-Lasso, as in Sect. 2.1), Ridge projection (Ridge, as in Sect. 2.2), Multi sample splitting (MS-Split, as in Sect. 2.4), a method from Javanmard and Montanari [10] (JM), covariance test from Lockhart et al. [13] (Covtest). *Left panel*: familywise error rate (FWER) with nominal level at 0.05 indicated by the *dotted line*; *right panel*: power (Power) representing the fraction of correctly identified active variables with non-zero regression coefficients. The figure is similar to some graphical representations in Dezeure et al. [8]

2.6.1 Supporting Theoretical Evidence and Discussion of Various Assumptions

Supporting evidence from theory, for the performance results in the empirical study, can be based by discussing the main assumptions underlying the different methods. The de-sparsified Lasso method is expected to work well and is most powerful if the design matrix is sparse in terms of its corresponding row-sparsity of Σ^{-1} (assumption (A2)) and if the linear model is rather sparse as well (assumption (A3)). The Ridge projection method does allow for designs with non-sparse rows of Σ^{-1}; however, the less restrictive assumption come with a price in that there is no optimality results in terms of power. The multi sample splitting method, which performs empirically quite reliably, has a theoretical drawback as it requires a zonal or the stronger beta-min assumption for the underlying regression coefficients (assumption (C2)); in terms of sparsity for the linear model, the multi sample splitting method is justified for a broader regime, allowing for $s_0 = o(n/\log(p))$ (assumption (C2)), than the required $s_0 = o(\sqrt{n}/\log(p))$ in assumption (A2) for the de-sparsified Lasso.

Stability Selection is controlling the number of false positives $\mathbb{E}[V]$ and not e.g. the familywise error rate (except when controlling $\mathbb{E}[V]$ at a very low level α which implies familywise error control at level α). The restrictive theoretical assumption is the exchangeability condition (D2): however, it seems that this condition is far from necessary. Stability Selection does not require a beta-min assumption as in (C2).

2.7 Other Methods

Very much related to the de-sparsified Lasso in Sect. 2.1 is a proposal by Javanmard and Montanari [10]. Their method is proved to be asymptotically optimal without requiring sparsity of the design as in condition (A2). Empirical evidence suggests though that the error control is not very reliable, see Fig. 1.

Bootstrap methods have been suggested to construct confidence intervals and p-values [7, 12]. They seem to work well for the components where the true parameter value equals zero, but they are often poor for the other components with non-zero parameters. Furthermore, multiple testing adjustment often requires a huge number of bootstrap replicates for reasonable computational approximation of tail events.

The covariance test [13] has been recently proposed as an "adaptive" method for assigning significance for the Lasso. Asymptotic validity of the test was shown under rather restrictive assumptions, in particular a restrictive beta-min assumption in the spirit of condition (C2). Empirical results of the covariance test are illustrated in Fig. 1, indicating that its power is comparably poor and error control is less reliable than for example for the Ridge projection or multi sample split method.

Another interesting proposal is due to Meinshausen [17]: we outline more details in Sect. 3.4.

3 Extensions and Further Topics

We briefly discuss here important extensions and additional issues.

3.1 Generalized Linear Models

Generalized linear models can be immediately treated with the multi sample splitting method or Stability Selection. Instead of e.g. the Lasso, we use ℓ_1-norm regularized maximum likelihood estimation for the selector $\hat{S}$, and low-dimensional inference (for the multi sample splitting method) is then based on maximum likelihood methodology.

The de-sparsified Lasso or the Ridge projection method are most easily adapted via additional weights as in iteratively reweighted least squares estimation [15]. The weights $w_i = w_i(\beta^0)$ $(i = 1, \ldots n)$ can be estimated by plugging in the ℓ_1-norm regularized maximum likelihood estimate; we can then proceed with new weighted data

$$\tilde{Y} = WY, \quad \tilde{\mathbf{X}} = W\mathbf{X}, \quad W = \mathrm{diag}(w_1, \ldots, w_n),$$

and apply the procedures from Sects. 2.1 and 2.2.

3.2 Multiple Testing Correction

Adjustment to multiple testing can be based using standard procedures which require valid p-values for individual tests as input: even under arbitrary dependence among the p-values, we can use e.g. the Bonferroni-Holm method for controlling the familywise error rate or the procedure from Benjamini and Yekutieli [1] to control the false discovery rate.

For the de-sparsified Lasso or Ridge projection method, one can use a simulation-based method which is less conservative than Bonferroni-Holm in presence of dependence: the details are given in Bühlmann [3].

We note that the multi sample splitting method from Sect. 2.4 as in the software package `hdi` (see Sect. 3.3) yields p-values which are adjusted for controlling the familywise error or false discovery rate.

3.3 R-Package `hdi`

The R-package `hdi` [16] contains implementations of various methods, namely the de-sparsified Lasso, the Ridge projection, the multi sample splitting method and of Stability Selection. We refer to to Dezeure et al. [8] how to use the procedures and what the various R-functions can do.

3.4 Testing Groups of Parameters

There might be considerable interest in testing the null-hypothesis $H_{0,G} : \beta_j^0 = 0$ for all $j \in G$, where $G \subseteq \{1, \ldots, p\}$ corresponds to a group of variables. The alternative is $H_{A,G}$: there exists $j \in G$ with $\beta_j^0 \neq 0$.

Based on the de-sparsified Lasso or Ridge projection method, one can use a simulation-based procedure to obtain an approximate distribution of $\max_{j \in G} |\hat{b}_j|$ under the null-hypothesis $H_{0,G}$. We refer to Bühlmann [3] for the details. The multi sample splitting method can be modified for testing $H_{0,G}$, as described in Mandozzi and Bühlmann [14].

An interesting and very different proposal is given by Meinshausen [17] which can be used for testing individual but also groups of variables (and the latter is the main motivation in that work): the procedure does not even require an identifiability condition in terms of the design matrix $\mathbf{X}$ as it automatically determines whether a parameter or a group of parameters is identifiable.

3.5 Selective Inference

Especially with confidence intervals, one would typically report only for a few selected variables. An interesting approach to account for the selection effect, in terms of the false coverage rate, is presented in Benjamini and Yekutieli [2]. Their procedure can be applied for confidence intervals from e.g. the de-sparsified Lasso or the Ridge projection method from Sects. 2.1 or 2.2.

3.6 Some Thoughts on Bayesian Methods

For expository simplicity, consider a Gaussian linear model with Gaussian prior for the regression coefficients $\beta = (\beta_1, \ldots, \beta_p)$:

$$\begin{aligned} &\beta_1, \ldots, \beta_p \text{ i.i.d. } \sim \mathcal{N}(0, \tau^2), \\ &Y|\beta \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2). \end{aligned} \tag{11}$$

The maximum a-posteriori estimator is then the Ridge estimator

$$\hat{\beta}_{\mathrm{MAP}} = \mathrm{argmin}_\beta \|Y - \mathbf{X}\beta\|_2^2/n + \frac{\sigma^2}{\tau^2 n}\|\beta\|_2^2.$$

For τ^2 large, this is the Ridge estimator with small regularization parameter λ as in Sect. 2.2.

Denote by β^* a realization from the prior distribution, and we are interested in constructing an interval which contains β^* with high probability. Alternatively, when adopting the frequentist Bayesian viewpoint (cf. [9]), we assume that the data is generated from a true parameter β^0, and we are interested to construct an interval which covers β^0 with high probability, based on a Bayesian model in (11). As discussed in Sect. 2.2, we know that for τ^2 large or σ^2 very small, $\hat{\beta}_{\mathrm{MAP}}$ is essentially unbiased for $\theta^* = P\beta^*$ (or $\theta^0 = P\beta^0$), where P is as in Sect. 2.2, but it can be severely *biased* for β^* (or β^0) in the high-dimensional scenario with $p \gg n$. Thus, the standard (Gaussian prior) Bayesian credible region centered around $\hat{\beta}_{\mathrm{MAP}}$ seems rather flawed for covering β^* or β^0 in the frequentist Bayesian paradigm.

Of course, in the classical Bayesian inference paradigm, such a bias does not occur, even when $p \gg n$, since the distribution of $\beta|Y$ is Gaussian with mean $\mathbb{E}[\beta|Y] = \hat{\beta}_{\mathrm{MAP}}$.

4 Conclusions

We provide a compact review of some methods for constructing tests and confidence intervals in high-dimensional models. The main assumptions underlying each method as well as a summary of empirical results are presented: this helps to understand, also from a comparative perspective, the strengths and weaknesses of the different approaches. Furthermore, a link to the R-package hdi is made. Thus, the user and practitioner obtains a "quick" but sufficiently deep overview for applying the procedures.

References

1. Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*, 1165–1188.
2. Benjamini, Y., & Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association, 100*, 71–81.
3. Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli, 19*, 1212–1242.
4. Bühlmann, P., & Mandozzi, J. (2013). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Computational Statistics*. Published online doi:10.1007/s00180-013-0436-3.

5. Bühlmann, P., Meier, L., & Kalisch, M. (2014). High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and Its Applications, 1*, 255–278.
6. Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Heidelberg/New York: Springer.
7. Chatterjee, A., & Lahiri, S. (2013). Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *Annals of Statistics, 41*, 1232–1259.
8. Dezeure, R., Bühlmann, P., Meier, L., & Meinshausen, N. (2014). High-dimensional inference: Confidence intervals, p-values and software `hdi`. Preprint arXiv:1408.4026.
9. Diaconis, P., & Freedman, D. (1986). On the consistency of Bayes estimates (with discussion). *Annals of Statistics, 14*, 1–63.
10. Javanmard, A., & Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. arXiv:1306.3171.
11. Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Annals of Statistics, 17*, 1001–1008.
12. Liu, H., & Yu, B. (2013). Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression. arXiv:1306.5505.
13. Lockhart, R., Taylor, J., Tibshirani, R., & Tibshirani, R. (2014). A significance test for the Lasso. *Annals of Statistics, 42*(2), 413–468.
14. Mandozzi, J., & Bühlmann, P. (2013). Hierarchical testing in the high-dimensional setting with correlated variables. arXiv:1312.5556.
15. McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.
16. Meier, L. (2013). *hdi: High-dimensional inference*. R package version 0.0-1/r2.
17. Meinshausen, N. (2013). Assumption-free confidence intervals for groups of variables in sparse high-dimensional regression. arXiv:1309.3489.
18. Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics, 34*, 1436–1462.
19. Meinshausen, N., & Bühlmann, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society Series B, 72*, 417–473.
20. Meinshausen, N., Meier, L., & Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association, 104*, 1671–1681.
21. Reid, S., Tibshirani, R., & Friedman, J. (2013). A study of error variance estimation in Lasso regression. arXiv:1311.5274.
22. Shao, J., & Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *Annals of Statistics, 40*, 812–831.
23. Sun, T., & Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika, 99*, 879–898.
24. van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics, 42*(3), 1166–1202.
25. Wasserman, L., & Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics, 37*, 2178–2201.
26. Zhang, C.-H., & Zhang, S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society Series B, 76*, 217–242.

Modelling and Forecasting Daily Electricity Load via Curve Linear Regression

Haeran Cho, Yannig Goude, Xavier Brossat, and Qiwei Yao

Abstract In this paper, we discuss the problem of short-term electricity load forecasting by regarding electricity load on each day as a curve. The dependence between successive daily loads and other relevant factors such as temperature, is modelled via curve linear regression where both the response and the regressor are functional (curves). The key ingredient of the proposed method is the dimension reduction based on the singular value decomposition in a Hilbert space, which reduces the curve linear regression problem to several ordinary (i.e. scalar) linear regression problems. This method has previously been adopted in the hybrid approach proposed by Cho et al. (J Am Stat Assoc 108:7–21, 2013) for the same purpose, where the curve linear regression modelling was applied to the data after the trend and the seasonality were removed by a generalised additive model fitted at the weekly level. We show that classifying the successive daily loads prior to curve linear regression removes the necessity of such a two-stage approach as well as resulting in reducing the forecasting error by a great margin. The proposed methodology is illustrated using the electricity load dataset collected between 2007 and mid-2012, on which it is compared to the hybrid approach and other available competitors. Finally, various ways for improving the forecasting performance of the curve linear regression technique are discussed.

H. Cho (✉)
School of Mathematics, University of Bristol, Bristol, UK
e-mail: haeran.cho@bristol.ac.uk

Y. Goude • X. Brossat
Électricité de France, Paris, France

Q. Yao
Department of Statistics, London School of Economics, London, UK

Guanghua School of Management, Peking University, Beijing, China

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_3

1 Introduction

While there are means for storing and discharging electricity, they cause extra costs as well as being limited to a small capacity compared to the overall electric power consumption. Therefore, it is of great importance for electricity providers to model and forecast electricity loads accurately over short-term (from 1 h to 1 month ahead) and middle-term (from 1 month to 5 years ahead) horizons. The electricity load forecast is an essential entry to the optimisation tools adopted by many energy companies for power system scheduling, and a small improvement in load forecasting can bring in substantial benefits from reducing production costs. Besides, there are further advantages to be gained in the electricity trading market, especially during the peak periods.

The French energy company Électricité de France (EDF) manages a large panel of production units across Europe, which includes water dams, nuclear plants, wind turbines, coal and gas plants. Figure 1 shows the electricity consumption of their customers measured over half an hour intervals between 2007 and mid-2012. Note that for confidentiality, we only report the ratio between the load over each half-hour interval, and the maximum load during the period throughout the paper. Based on the vast knowledge on French electricity consumption patterns accumulated over 20 years, EDF has developed a forecasting model which consists of complex regression models based on past loads, temperature, date and calendar events, coupled with classical time series models such as the seasonal ARIMA (SARIMA) [4]. This operational model performs very well, attaining about 1.4 % mean absolute percentage error (see (8)) in forecasting the consumption of EDF customers over one day horizon. Due to its complexity, however, the model may not be well-adapted to constant changes in electricity consumption habits resulted from the opening of new electricity markets, technological innovations and social and economic changes, to name a few.

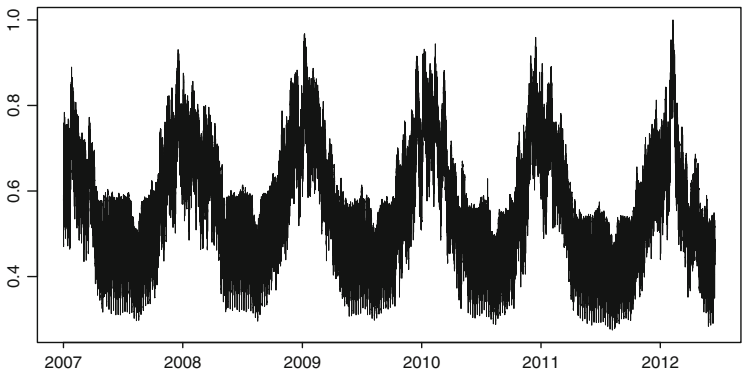


Fig. 1 Electricity consumption of the French customers of EDF measured every half an hour between 2007 and mid-2012

Cho et al. [6] recognised the strategic importance of a forecasting model which was more adaptive to ever-changing electricity consumption environment. Electricity loads exhibit several interesting features at more than one level, as can be seen in Fig. 1, and addressing such multi-level nature of the data, they proposed a hybrid approach which consisted of the following two building blocks:

- Modelling the overall trend and seasonality in the data by fitting a generalised additive model (GAM) to the *weekly* averages of the load, with meteorological factors (e.g., temperature and nebulosity) as explanatory variables;
- Modelling the dependence across successive, de-trended *daily* loads via curve linear regression, where both the response and the regressor are functional (curves), with the load curve on the next day as the response and that on the current day, jointly with the temperature forecast, as the regressor.

By regarding each daily load and temperature as a curve, the proposed curve linear regression modelling takes advantage of the continuity of the curve data in statistical modelling. Moreover, it embeds some nonstationary features, such as daily patterns of electricity loads (see Fig. 2), into a stationary framework in a functional space. Its key ingredient is the dimension reduction based on the singular value decomposition in a Hilbert space, which effectively reduces the curve linear regression problem to several ordinary linear regression problems. Compared to the EDF operational model, the hybrid method does not incorporate much of the data-specific knowledge, while maintaining competitive prediction accuracy when applied to the French electricity consumption data.

While the hybrid approach represents a determined effort in developing an adaptive and widely-applicable forecasting model, it is conceivable that the two-stage procedure may carry over the estimation and the forecasting errors from the

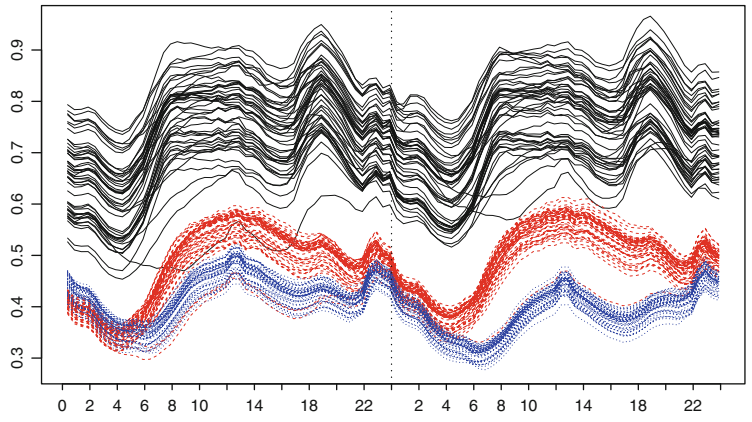


Fig. 2 Electricity loads on Mondays–Tuesdays in January and December (*solid*), Mondays–Tuesdays in June and July (*dashed*) and Saturdays–Sundays in June and July between 2007 and 2012 (*dotted*)

first stage to the next stage, and thus lead to greater forecasting errors. Besides, even after the trend and the seasonality are removed at the weekly level, the daily loads exhibit dependency on calendar variables, such as the corresponding days of a week and the months of a year, both in their profiles and the covariance structure between successive loads. As a solution, [6] proposed to classify the pairs of daily loads into (approximately) homogeneous sub-groups prior to fitting a curve linear regression model, which, as we show, renders the weekly level modelling unnecessary.

Therefore, we focus on the curve linear regression method and its application to the one-day ahead forecasting problem in conjunction with the daily load classification, and investigate whether this simplified approach improves the accuracy and the adaptivity of the forecasting model when compared to the hybrid approach. Besides, the ways of further enhancing its forecasting performance are discussed, such as aggregating several forecasting models resulting from varying choices for the curve regressor.

The rest of the paper is organised as follows. In Sect. 2, we describe the dimension-reduction based curve linear regression technique in a generic setting. Section 3 discusses the application of the proposed approach to electricity load modelling, including the problem of classifying the successive daily load curves. We conduct a comparative study in Sect. 4, where our method and other competitors are applied to predict the daily electricity consumption of EDF customers in France. Finally, we conclude the paper with some remarks on the future research.

2 Curve Linear Regression via Dimension Reduction

Every day at noon, EDF forecasts the half-hourly consumption of electricity for the next 24 h. Viewing that the 48 half-hourly loads are sampled from a curve, we may regard the loads for the next 24 h from the noon of day i as a curve response ($\equiv Y_i(\cdot)$), and let the curve regressor ($\equiv X_i(\cdot)$) contain information such as the loads observed up to the noon of the same day, as well as observed and predicted daily temperature. Then the following curve linear regression model can be adopted to model the dependence between such $Y_i(\cdot)$ and $X_i(\cdot)$:

$$Y_i(u) = \mu_Y(u) + \int_{\mathscr{I}_2} \{X_i(v) - \mu_X(v)\}\beta(u, v)dv + \varepsilon_i(u) \quad \text{for} \quad u \in \mathscr{I}_1, \quad (1)$$

where $\mu_Y(u) = \mathbb{E}\{Y_i(u)\}$, $\mu_X(v) = \mathbb{E}\{X_i(v)\}$ and $\mathscr{I}_1$ and $\mathscr{I}_2$ denote the supports of $Y_i(\cdot)$ and $X_i(\cdot)$, respectively. The linear operator β is a regression coefficient function defined on $\mathscr{I}_1 \times \mathscr{I}_2$, and $\varepsilon_i(\cdot)$ is noise satisfying $\mathbb{E}\{\varepsilon_i(u)\} = 0$ for all $u \in \mathscr{I}_1$.

The conventional approach to the linear regression problem in (1) is based on decomposing $Y_i(\cdot)$ and $X_i(\cdot)$ using the Karhunen-Loève expansion, which has been featured predominantly in the functional data analysis literature for dimension reduction. Then the terms from such expansions are modelled using simple linear regression, which is equivalent to the dimension reduction based on principal

component analysis in multivariate analysis. For further references on functional linear models, see e.g. [20, 25] and [12].

Since the principal components do not necessarily represent the directions in which $X_i(\cdot)$ and $Y_i(\cdot)$ are most correlated, [6] presented a novel approach where the singular value decomposition (SVD) in a Hilbert space was adopted to single out the directions upon which the projections of $Y_i(\cdot)$ were most correlated with $X_i(\cdot)$. While closely related to the functional canonical regression method proposed in [15], this approach focuses on regressing $Y_i(\cdot)$ on $X_i(\cdot)$ and thus the two curves are not treated on an equal footing which is different from, and much simpler than, the latter method. In what follows, we lay out the details of the SVD-based curve linear regression method in a generic setting.

Let $\{Y_i(\cdot), X_i(\cdot)\}$, $i = 1, \ldots, n$, be a random sample where $Y_i(\cdot) \in \mathscr{L}_2(\mathscr{I}_1)$, $X_i(\cdot) \in \mathscr{L}_2(\mathscr{I}_2)$, and let $\mathscr{I}_1$ and $\mathscr{I}_2$ be two compact subsets of $\mathbb{R}$. We denote by $\mathscr{L}_2(\mathscr{I})$ the Hilbert space consisting of all the square integrable curves defined on the set $\mathscr{I}$, which is equipped with the inner product $\langle f, g\rangle = \int_{\mathscr{I}} f(u)g(u)du$ for any $f, g \in \mathscr{L}_2(\mathscr{I})$. For now, it is assumed that $\mathbb{E}\{Y_i(u)\} = 0$ for all $u \in \mathscr{I}_1$ and $\mathbb{E}\{X_i(v)\} = 0$ for all $v \in \mathscr{I}_2$. The covariance function between $Y_i(\cdot)$ and $X_i(\cdot)$ is denoted by $\Sigma(u, v) = \mathrm{cov}\{Y_i(u), X_i(v)\}$. Under the assumption

$$\int_{\mathscr{I}_1} \mathbb{E}\{Y_i(u)^2\}du + \int_{\mathscr{I}_2} \mathbb{E}\{X_i(v)^2\}dv < \infty, \tag{2}$$

Σ defines the following two bounded operators between $\mathscr{L}_2(\mathscr{I}_1)$ and $\mathscr{L}_2(\mathscr{I}_2)$,

$$f_1(u) \to \int_{\mathscr{I}_1} \Sigma(u, v)f_1(u)du \in \mathscr{L}_2(\mathscr{I}_2) \text{ and } f_2(v) \to \int_{\mathscr{I}_2} \Sigma(u, v)f_2(v)dv \in \mathscr{L}_2(\mathscr{I}_1)$$

for any $f_l(\cdot) \in \mathscr{L}_2(\mathscr{I}_l)$, $l = 1, 2$.

Performing the SVD on Σ, we obtain a triple sequence $\{\lambda_j, \varphi_j(\cdot), \psi_j(\cdot)\}$, $j = 1, 2, \ldots$ which satisfies

$$\Sigma(u, v) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \varphi_j(u)\, \psi_j(v), \tag{3}$$

where $\{\varphi_j(\cdot)\}$ is an orthonormal basis of $\mathscr{L}_2(\mathscr{I}_1)$, $\{\psi_j(\cdot)\}$ is that of $\mathscr{L}_2(\mathscr{I}_2)$, and the squared singular values $\{\lambda_j\}$ are ordered in a decreasing manner as

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0.$$

Further, it holds that for $u \in \mathscr{I}_1$, $v \in \mathscr{I}_2$ and $j = 1, 2, \ldots,$

$$\int_{\mathscr{I}_1} M_1(u, u')\, \varphi_j(u')\, du' = \lambda_j\, \varphi_j(u), \qquad \int_{\mathscr{I}_2} M_2(v, v')\, \psi_j(v')\, dv' = \lambda_j\, \psi_j(v),$$

where $M_l,\ l = 1, 2$ are non-negative operators defined on $\mathscr{L}_2(\mathscr{I}_l)$ as

$$M_1(u, u') = \int_{\mathscr{I}_2} \Sigma(u, w)\, \Sigma(u', w)\, dw, \qquad M_2(v, v') = \int_{\mathscr{I}_1} \Sigma(w, v)\, \Sigma(w, v')\, dw.$$

It is clear from the above that λ_j is the j-th largest eigenvalue of M_1 and M_2, with $\varphi_j(\cdot)$ and $\psi_j(\cdot)$ as the respective eigenfunctions. See [23] for further discussion on the SVD in a Hilbert space.

Since $\{\varphi_j(\cdot)\}$ and $\{\psi_j(\cdot)\}$ are the orthonormal bases of $\mathscr{L}_2(\mathscr{I}_1)$ and $\mathscr{L}_2(\mathscr{I}_2)$, we may write

$$Y_i(u) = \sum_{j=1}^{\infty} \xi_{ij}\varphi_j(u), \qquad X_i(v) = \sum_{k=1}^{\infty} \eta_{ik}\psi_k(v), \tag{4}$$

where ξ_{ij} and η_{ik} are random variables defined as $\xi_{ij} = \langle Y_i, \varphi_j\rangle$ and $\eta_{ik} = \langle X_i, \psi_k\rangle$. From (3), it is straightforward to derive that

$$\mathrm{cov}(\xi_{ij},\ \eta_{ik}) = \mathbb{E}(\xi_{ij}\eta_{ik}) = \begin{cases} \sqrt{\lambda_j} & \text{when } j = k, \\ 0 & \text{when } j \neq k. \end{cases} \tag{5}$$

The dimensionality of the functional data has been defined in various contexts, e.g. see [13] and [2]. A correlation dimension between the two curves $Y_i(\cdot)$ and $X_i(\cdot)$ was defined in [6] with the squared singular values λ_j.

Definition 1 If $\lambda_r > 0$ and $\lambda_{r+1} = 0$, the (linear) correlation between $Y_i(\cdot)$ and $X_i(\cdot)$ is r-dimensional.

When the correlation between $Y_i(\cdot)$ and $X_i(\cdot)$ is r-dimensional, it follows from (5) that $\mathrm{cov}\{\xi_{ij},\ X_i(v)\} = 0$ for all $j > r$ and $v \in \mathscr{I}_2$, from which we can conclude that the curve linear regression model (1) has an equivalent representation by r (scalar) linear regression models, as summarised in the following theorem.

Theorem 1 (Theorem 1 of [6]) *Let the linear correlation between $Y_i(\cdot)$ and $X_i(\cdot)$ be r-dimensional. Assume that*

- *The regression coefficient operator β is in the Hilbert space $\mathscr{L}_2(\mathscr{I}_1 \times \mathscr{I}_2)$, and*
- *$\varepsilon_i(\cdot)$ are i.i.d. with $\mathbb{E}\{\varepsilon_i(u)\} = 0$ and $\mathbb{E}\{X_i(v)\varepsilon_j(u)\} = 0$ for any $u \in \mathscr{I}_1,\ v \in \mathscr{I}_2$ and $i, j \geq 1$.*

Then the curve regression model (1) may be represented equivalently by

$$\begin{aligned} \xi_{ij} &= \textstyle\sum_{k=1}^{\infty} \beta_{jk}\eta_{ik} + \varepsilon_{ij} && \text{for } j = 1, \ldots, r, \\ \xi_{ij} &= \varepsilon_{ij} && \text{for } j = r+1, r+2, \ldots, \end{aligned} \tag{6}$$

where $\varepsilon_{ij} = \int_{\mathscr{I}_1} \varphi_j(u)\varepsilon_i(u)du$, and $\beta_{jk} = \int_{\mathscr{I}_1\times\mathscr{I}_2} \varphi_j(u)\psi_k(v)\beta(u, v)dudv$.

The above theorem implies that the SVD-based approach provides a framework to define and exploit the correlation dimension between a pair of curves, and to model the functional linear regression relationship between the pair using a finite number of ordinary (scalar) linear regression models. In this framework, as described in Sect. 3.2 below, the prediction is achieved directly from the estimated ordinary linear regression models.

Taking into account the fact that $\text{var}(\eta_{ik}) \to 0$ as $k \to \infty$ (see (2) and (4)), we may include only the first Q terms $\eta_{ik},\ k = 1, \ldots, Q$ in the r multiple linear regression models, and obtain the ordinary least squares (OLS) estimator of the finite number of linear coefficients. Note that, while the OLS estimator of β_{jk} is unbiased, its variance tends to increase with Q in finite sample performance. That is, if Q is selected too large, we may end up with a model which fits the data too closely but performs poorly in prediction.

As noted in [6], Theorem 1 holds for any valid expansion $X_i(v) = \sum_k \eta_{ik}\psi_k(v)$, provided $\{\xi_{ij}\}$ are obtained from the SVD. Let $X_i(\cdot)$ be of finite dimension in the sense that its Karhunen-Loève decomposition has q terms only, i.e. $X_i(v) = \sum_{k=1}^{q} \zeta_{ik}\gamma_k(v)$ where $q(\geq r)$ is a finite integer, $\{\gamma_k(\cdot)\}_{k=1}^{q}$ are q orthonormal functions in $\mathscr{L}_2(\mathscr{I}_2)$ and $\zeta_{i1}, \ldots, \zeta_{iq}$ are uncorrelated random variables with $\text{var}(\zeta_{ik}) > 0$. Then, decomposing $X_i(\cdot)$ with respect to $\{\psi_k(\cdot)\}_{k=1}^{q}$ from the SVD of Σ, the corresponding $\{\eta_{ik}\}$ satisfy $\text{cov}(\eta_{ik}, \eta_{il}) = 0$ for any $k \neq l$. This, together with (5) and (6), implies that $\beta_{jk} = 0$ for all $j \neq k$ and thus (6) is reduced to r simple linear regression problems

$$\begin{aligned} \xi_{ij} &= \beta_{jj}\eta_{ij} + \varepsilon_{ij} \qquad && \text{for } j = 1, \ldots, r, \\ \xi_{ij} &= \varepsilon_{ij} && \text{for } j = r+1, r+2, \ldots. \end{aligned}$$

2.1 Estimation

Given the observed pairs of curves $\{Y_i(\cdot),\ X_i(\cdot)\},\ i = 1, \cdots, n$, let

$$\hat{\Sigma}(u, v) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(u) - \bar{Y}(u)\}\{X_i(v) - \bar{X}(v)\},$$

where $\bar{Y}(u) = n^{-1}\sum_i Y_i(u)$ and $\bar{X}(v) = n^{-1}\sum_i X_i(v)$. Performing the SVD on $\hat{\Sigma}(u, v)$, we obtain the estimators $\{\hat{\lambda}_j, \hat{\varphi}_j(\cdot), \hat{\psi}_j(\cdot)\}$ for $\{\lambda_j, \varphi_j(\cdot), \psi_j(\cdot)\},\ j = 1, 2, \ldots$ in (3). Note that the SVD can be achieved by performing eigenanalysis on the non-negative operators

$$\hat{M}_1(u, u') = \int_{\mathscr{I}_2} \hat{\Sigma}(u, w)\hat{\Sigma}(u', w)dw \ \text{ and } \ \hat{M}_2(v, v') = \int_{\mathscr{I}_1} \hat{\Sigma}(w, v)\hat{\Sigma}(w, v')dw,$$

which may be transformed into the eigenanalysis of non-negative definite matrices, see Section 2.2.2 of [2].

Adapting Theorem 1 of [2] to the current setting, we can show the consistency of $\hat{\lambda}_j$. We first assume that

- $\{Y_i(\cdot), X_i(\cdot)\}$ is strictly stationary and ψ-mixing with the mixing coefficients $\psi(k)$ satisfying the condition

$$\sum_{k\geq 1} k\psi(k)^{1/2} < \infty.$$

- $\mathbb{E}\{\int_{\mathscr{I}_1} Y_i(u)^2 du + \int_{\mathscr{I}_2} X_i(v)^2 dv\}^2 < \infty$.
- $\lambda_1 > \cdots > \lambda_r > 0 = \lambda_{r+1} = \lambda_{r+2} = \cdots$.

Then we have $|\hat{\lambda}_k - \lambda_k| = O_p(n^{-1/2})$ for $1 \leq k \leq r$, and $|\hat{\lambda}_k| = O_p(n^{-1})$ for $k > r$, as $n \to \infty$.

This result implies that the ratios $\hat{\lambda}_{j+1}/\hat{\lambda}_j$ for $j < r$ are asymptotically bounded away from 0, while $\hat{\lambda}_{r+1}/\hat{\lambda}_r \to 0$ in probability. Therefore, one way of determining the correlation dimensionality is to employ the following ratio-based estimator

$$\hat{r} = \arg\max_{1\leq j\leq d} \hat{\lambda}_j/\hat{\lambda}_{j+1},$$

where d is a pre-specified upper bound on r. However, this estimator should be used with caution as different components of the SVD can have different degrees of "strength" in the sense that, there may exist some $k < r$ for which non-zero $\lambda_j \neq 0$, $j > k$ are considerably smaller than $\lambda_{j'}$, $j' \leq k$. Further discussion on this point in the framework of factor analysis can be found in [17]. Heuristically, we may estimate r as

$$\hat{r} = \max\{1 \leq j \leq d : \hat{\lambda}_j/\hat{\lambda}_{j+1} > M\}, \tag{7}$$

for sufficiently chosen M to avoid neglecting such smaller non-zero eigenvalues.

Alternatively, [6] proposed the following information criterion based on the estimated eigenvalues, which extended the information criterion introduced in [14] for high-dimensional time series analysis:

$$IC(q) = \log\left(c_* + \frac{1}{d^2}\sum_{k=q+1}^{d} \hat{\lambda}_k\right) + \tau q \cdot g(n),$$

where $c_*, \tau > 0$ are constants and $g(\cdot)$ is a function of n satisfying $n \cdot g(n) \to \infty$ and $g(n) \to 0$ as $n \to \infty$. While $IC(\cdot)$ was shown to be consistent in identifying r asymptotically, the choices of τ and $g(\cdot)$ played a significant role in finite sample performance. Therefore, it was proposed to fix $g(n)$ as $g(n) = n^{-1/2}$, obtain $q^* = \arg\min_q IC(q;\tau)$ over a grid of values for τ, and choose r as the most frequently

returned among q^*. For the full description of this majority voting scheme, see Section 3.2 of [6].

3 Application to Electricity Load Modelling

In this section, we discuss applying the proposed curve linear regression method to electricity load modelling and forecasting. The load data example (plotted in Fig. 1) contains electricity loads consumed by the French customers of EDF between 2007 and mid-2012. We first highlight some time-varying features exhibited by the daily electricity load curves, which makes it difficult to assume that the entire data can be modelled as being stationary. Then, we introduce a simple classification rule which divides the pairs of load curves into homogeneous sub-groups, such that the curve linear regression modelling is applicable to each sub-group separately. Finally, the combined procedure of classification and curve linear regression is illustrated using a real electricity load forecasting example.

3.1 Classification of Daily Electricity Load Curves

In electricity load data, there exist systematic discrepancies in the profiles and the variability of daily load curves observed on different days of a week or in different months. Figure 2 shows that, while successive daily loads on Mondays–Tuesdays in June and July behave similarly, they are distinctively different from those observed on Saturdays–Sundays in June and July, and also from those observed on Mondays–Tuesdays in January and December. Those profile discrepancies are reflected predominantly in the locations and magnitudes of daily peaks. Typically in France, daily peaks occur at noon in summer and in the evening in winter, due to economic cycle as well as the usage of electrical heating or cooling and lighting. Hence, the profiles and (presumably) the dynamic structure of successive daily curves vary over different days within a week, and also over different months within a year. It has been noted that these systematic discrepancies persist even after the weekly level de-trending step of the hybrid approach (see Section 4.1 of [6]), which implies that the classification of daily loads is an essential step prior to curve linear regression modelling with or without the weekly level modelling.

According to the experts at EDF, in the case of French electricity consumption data, load curves on the same day of a week tend to have similar profiles. Therefore it is reasonable to assign a day type (DT) to each daily load as summarised in Table 1.

Table 1 Daily classification of daily load curves

Day type	0	1	2	3	4	5	6
Day of a week	Mon	Tue	Wed	Thu	Fri	Sat	Sun

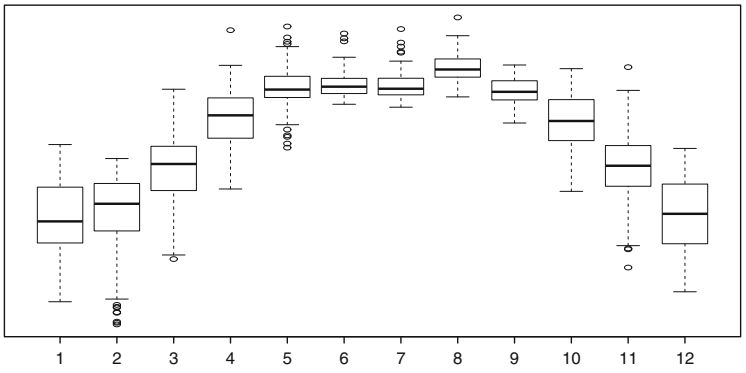


Fig. 3 Boxplots of $\hat{\zeta}_{i1}$ from different months

Table 2 Seasonal classification of daily load curves

Seasonal class	1	2	3	4	5	6	7	8
Month	Jan–Feb, Nov–Dec	Mar	Apr	May	Jun–Jul	Aug	Sep	Oct

To gain an insight into the possible seasonal variation present in the covariance between successive daily loads, as well as in their profiles, we decompose the daily load curves (denoted by $Z_i(\cdot)$ for the loads on the i-th day) as follows. Performing the SVD on the sample covariance function between successive daily curves $Z_{i+1}(\cdot)$ and $Z_i(\cdot)$, we obtain the first left singular function $\hat{\gamma}_1(\cdot)$ and decompose each $Z_{i+1}(\cdot)$ as $\hat{\zeta}_{i1} = \langle Z_{i+1}, \hat{\gamma}_1 \rangle$; see Fig. 3. We note that each $Z_i(\cdot)$ has been de-meaned with the mean curve obtained by averaging out all the observations of the same DT. If the dependence structure between the pairs of curves undergoes seasonal changes, we expect such seasonality to be reflected in the behaviour of $\hat{\zeta}_{i1}$ over the span of 1 year. Indeed, this is the case as observable in the boxplots of $\hat{\zeta}_{i1}$ from different months and based on this, we choose to create 8 seasonal classes (SC) as in Table 2.

Combining the two classification rules, we classify each pair of successive daily loads into sub-groups of (approximately) homogeneous dependence structure, according to the corresponding DTs and SCs. While it lacks rigorous statistical ground, the forecasting models estimated within such sub-groups perform well as demonstrated in Sect. 4. Besides, the problem of classifying electricity load curves and functional data in general can stand alone as an independent research problem, and it has attracted considerable attention, see e.g. [5, 21, 22] and [16] for functional data clustering, and [1] for that in the context of electricity loads classification.

3.2 An Illustration

We illustrate the application of curve linear regression with an example where our aim is to predict the electricity load curve for the next 24 h (48 half-hours), denoted by $Y(\cdot)$, at the noon of Tuesday 12 June 2012. Note that in (1), $Y_i(\cdot)$ and $X_i(\cdot)$ are allowed to have different supports as $\mathscr{I}_1$ and $\mathscr{I}_2$, such that we have flexibility in the choice of the curve regressor. Therefore we consider the following three choices:

- $X^{(1)}(\cdot)$: load curve for the 24 h up to the midday of 12 June 2012.
- $X^{(2)}(\cdot)$: $X^{(1)}(\cdot)$ joined with the temperature forecast ($\equiv T^F(\cdot)$) for the next 24 h.
- $X^{(3)}(\cdot)$: $X^{(2)}(\cdot)$ joined with the temperature curve ($\equiv T(\cdot)$) observed over the same 24 h interval as $X^{(1)}(\cdot)$.

We have used the temperature forecasts from meteoFrance in our study. As discussed in Sect. 3.1, $\{Y_i(\cdot), X_i^{(m)}(\cdot)\}$, $m = 1, 2, 3$ are collected as all the observed pairs of curves corresponding to {(DT 1, SC 5), (DT 0, SC 5)} between 1 January 2007 and the midday 12 June 2012. In total, there are $n = 38$ observations, which are plotted in Fig. 4 along with their respective mean curves. It may be noted that, due to the classification step, the regressor curves $\{X_i^{(1)}(\cdot),\ i = 1, \ldots, n\}$ and the response curves $\{Y_i(\cdot),\ i = 1, \ldots, n\}$ do not satisfy the relationship

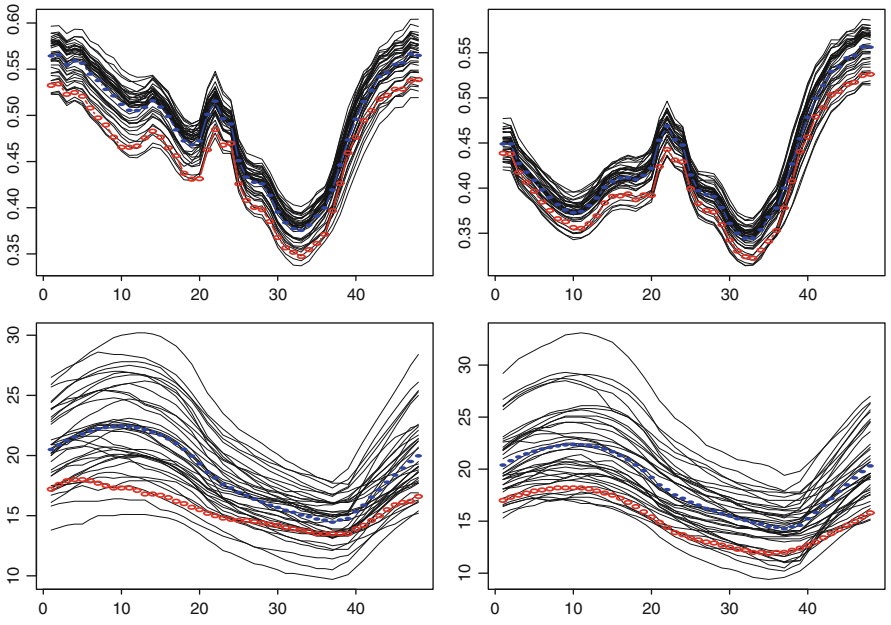


Fig. 4 The n curve observations $Y_i(\cdot)$ (*top left*), $X_i^{(1)}(\cdot)$ (*top right*), $T_i(\cdot)$ (*bottom left*) and $T_i^F(\cdot)$ (*bottom right*), together with their respective mean curves (*filled circle*) as well as $Y(\cdot)$, $X^{(1)}(\cdot)$, $T(\cdot)$ and $T^F(\cdot)$ (*empty circle*)

$X_{i+1}^{(1)}(\cdot) = Y_i(\cdot)$. Hence, even with $X_i(\cdot) \equiv X_i^{(1)}(\cdot)$, the curve linear regression model (1) is distinguished from the autoregressive Hilbertian process of order 1 (ARH(1)) proposed in [3].

Note that for $X_i^{(2)}(\cdot)$ and $X_i^{(3)}(\cdot)$ which join the observed loads with the temperature, different components have different scales since $X_i^{(1)}(\cdot)$ range in tens of thousands (MW), while $T_i(\cdot)$ and $T_i^F(\cdot)$ range in a far smaller scale between 6 and 33 (^{o}C). Since the SVD-based method is not scale-invariant, we apply a simple standardisation step to have different components of the regressor curves in a similar scale.

From the observed curves, we estimate the sample covariance function

$$\hat{\Sigma}^{(m)}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(u) - \bar{Y}(u)\}\{X_i^{(m)}(v) - \bar{X}^{(m)}(v)\}, \qquad m = 1,2,3,$$

and perform the SVD on $\hat{\Sigma}^{(m)}(u,v)$ to obtain $\{\hat{\lambda}_j^{(m)}, \hat{\varphi}_j^{(m)}(\cdot), \hat{\psi}_j^{(m)}(\cdot)\}$, $j = 1,2,\ldots$. Applying (7) to the estimated eigenvalues with $M = 5$, the correlation dimensions are estimated as $\hat{r}^{(m)} = 4$ for all $m = 1,2,3$. Defining $\hat{\xi}_{ij}^{(m)} = \langle Y_i - \bar{Y}, \hat{\varphi}_j^{(m)}\rangle$ and $\hat{\eta}_{ik}^{(m)} = \langle X_i^{(m)} - \bar{X}^{(m)}, \hat{\psi}_k^{(m)}\rangle$ analogously as $\hat{\xi}_{ij}$ and $\hat{\eta}_{ik}$, the next step is to estimate the linear coefficients $\beta_{jk}^{(m)}$ in the following scalar linear regression models

$$\hat{\xi}_{ij}^{(m)} = \sum_{k=1}^{Q}\beta_{jk}^{(m)}\hat{\eta}_{ik}^{(m)} + \varepsilon_{ij}^{(m)}$$

for $m = 1,2,3$. We set $Q = 15$ to preserve the prediction accuracy by having sufficient number of terms, while attaining the numerical stability of the OLS estimator of $\beta_{jk}^{(m)}$. Then the predictor of $Y(u)$ takes the following form

$$\hat{Y}^{(m)}(u) = \bar{Y}(u) + \sum_{j=1}^{\hat{r}^{(m)}}\hat{\xi}_j^{(m)}\hat{\varphi}_j^{(m)}(u),$$

where $\hat{\xi}_j^{(m)}$ are predicted as

$$\hat{\xi}_j^{(m)} = \sum_{k=1}^{Q}\hat{\beta}_{jk}^{(m)}\hat{\eta}_k^{(m)}, \qquad j = 1,\ldots,\hat{r}^{(m)},$$

with $\hat{\eta}_k^{(m)} = \langle X^{(m)} - \bar{X}^{(m)}, \hat{\psi}_k^{(m)}\rangle$.

For each m, we obtain two other predictors besides $\hat{Y}^{(m)}(\cdot)$, the oracle and the base predictors. The oracle predictor is of the form

$$\tilde{Y}^{(m)}(u) = \bar{Y}(u) + \sum_{j=1}^{\hat{r}^{(m)}} \tilde{\xi}_j^{(m)} \hat{\varphi}_j^{(m)}(u),$$

which is similar to $\hat{Y}^{(m)}(u)$ except that $\hat{\xi}_j^{(m)}$ are replaced by $\tilde{\xi}_j^{(m)} = \langle Y - \bar{Y}, \hat{\varphi}_j^{(m)} \rangle$. We use the term "oracle" to emphasise the fact that $\tilde{\xi}_j^{(m)}$ require the prior knowledge of $Y(\cdot)$ and thus are unavailable in practice. The base predictor is set simply as $\bar{Y}^{(m)}(\cdot) = \bar{Y}(\cdot)$, ignoring the dynamic dependence between the response and the regressor curves.

To evaluate the performance of different predictors, we employ the following two error measures

$$\text{RMSE} = \left\{ \frac{1}{N} \sum_{t=1}^{N} (\hat{f}_t - f_t)^2 \right\}^{1/2} \quad \text{and} \quad \text{MAPE} = \frac{1}{N} \sum_{t=1}^{N} \left| \frac{\hat{f}_t - f_t}{f_t} \right|, \tag{8}$$

where $\hat{f}_t$ and f_t denote the predicted and the true loads in the t-th half-hour interval and N denotes the forecasting horizon ($N = 48$ in this case). The MAPE and RMSE for the above predictors are reported in Table 3.

As expected, the oracle predictors return smaller prediction errors than the SVD-based predictors, and the base predictor returns the largest error. Based on this, we can conclude that (a) there is much to be accounted for by the dependence between the regressor and the response curves, as observable from the poor performance of $\bar{Y}(\cdot)$, and (b) the reduced dimension captures such dependence structure well, as demonstrated by the superior performance of $\tilde{Y}^{(m)}(\cdot)$.

$\hat{Y}^{(m)}(\cdot)$ perform as competitively as $\tilde{Y}^{(m)}(\cdot)$, attaining RMSE as small as 292 MW without any prior knowledge on the true load $Y(\cdot)$. This fact is also confirmed in Fig. 5, where all $\hat{Y}^{(m)}(\cdot)$ and $\tilde{Y}^{(m)}(\cdot)$ are quite close to $Y(\cdot)$ throughout the forecasting horizon. Among the three $\hat{Y}^{(m)}(\cdot)$, $m = 1, 2, 3$, the choice of $X^{(2)}(\cdot)$ returns the best forecast.

Finally, when the aim is to produce multi-step ahead forecasts, we simply replace the curve regressor $X^{(1)}(\cdot)$ by one of the forecasts $\hat{Y}^{(m)}(\cdot)$ and repeatedly apply the above procedure until the desired multi-step ahead prediction is achieved. Note that the corresponding multi-step ahead temperature forecast may not be available and in such a case, $X^{(1)}(\cdot)$ is the only possible choice as a regressor curve. Thus-produced

Table 3 RMSE and MAPE of the different predictors

Predictor	$\hat{Y}^{(1)}$	$\hat{Y}^{(2)}$	$\hat{Y}^{(3)}$	$\tilde{Y}^{(1)}$	$\tilde{Y}^{(2)}$	$\tilde{Y}^{(3)}$	$\bar{Y}$
RMSE (MW)	361	292	327	189	218	220	2,440
MAPE (%)	0.77	0.64	0.73	0.41	0.46	0.46	6.65

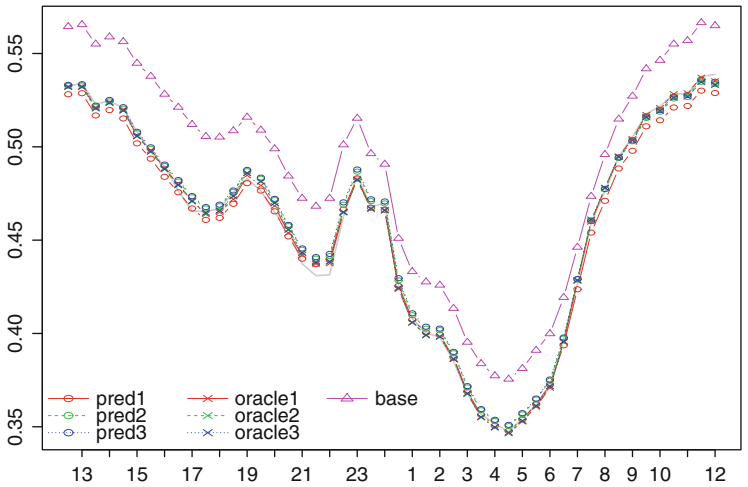


Fig. 5 Different predictors against the true load curve (*grey, solid*) for the next 24 h at noon 12 June 2012

two-day ahead predictor at the noon of 12 June 2012 attains 464 MW RMSE and 1.20 % MAPE, with $X^{(1)}(\cdot)$ replaced by $\hat{Y}^{(2)}(\cdot)$.

4 Forecasting Daily Electricity Consumption of EDF Customers

In this section, we perform one-day ahead forecasting for daily electricity loads consumed by the French customers of EDF from 1 September 2011 to 15 June 2012. As with the example in Sect. 3.2, the forecast is produced every day at noon. Hence, when forecasting the load curve for the next 24 h on day t, we assume the accessibility of the load and the temperature observations from 1 January 2007 up to the noon of day t, as well as the temperature forecast for the next 24 h. During this period, there are certain days (e.g., bank holidays) on which the load observations have not been validated and excluding such days, load forecasts are produced for 234 days in total. Also, when the temperature forecast ($T_i^F(\cdot)$) is not available, we assume that the true one-day ahead temperature ($T_{i+1}(\cdot)$) is known for convenience.

Recalling the notations from Sect. 3.2, we denote the forecasting models with the three regressors $X^{(m)}(\cdot)$, $m = 1, 2, 3$ by P1–P3, respectively, and the corresponding oracle forecasting models by O1–O3. We also consider the predictors from the hybrid approach ([6], H1–H3), where we employ the same regressors at the curve linear regression stage. At the weekly GAM stage, the explanatory variables are lagged weekly average load, weekly average temperature, weekly average cloud cover and two calendar variables representing the yearly trend and the seasonality.

Table 4 RMSE and MAPE of the daily electricity load forecasts between 1 September 2011 and 15 June 2012

	P1	P2	P3	P4	O1	O2	O3	Base	H1	H2	H3	GAM
RMSE (MW)	1,250	853	872	804	336	312	312	6,164	1,917	1,812	1,813	832
MAPE (%)	1.97	1.47	1.50	1.37	0.53	0.50	0.51	10.75	2.91	2.72	2.75	1.40

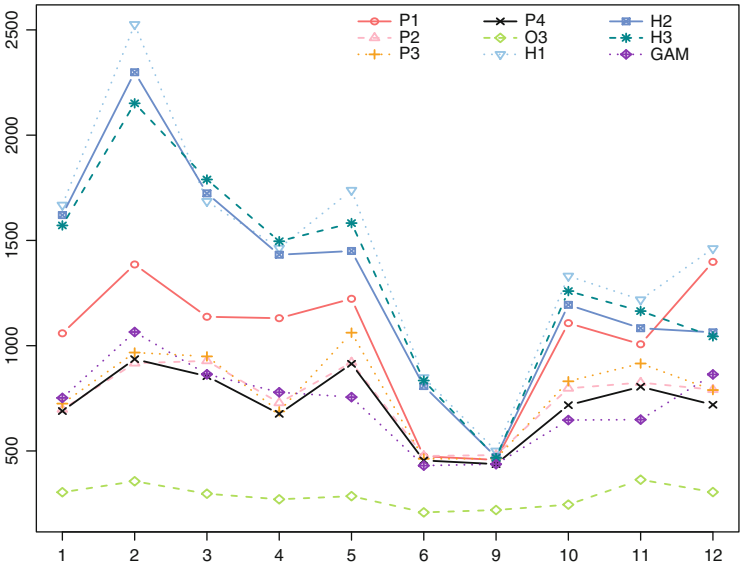


Fig. 6 RMSE from P1–P4, O3, H1–H3 and GAM with respect to different months

Finally, the results from the GAM model provided by the EDF R&D department are presented ("GAM") for comprehensive comparative study.

Additive models for short-term electricity load forecasting have been studied e.g., in [19] and [11], where the proposed models were shown to be well-adapted to non-linear behaviour of the electricity load. The GAM included in our study models the relationship between each half-hour interval load and several explanatory variables such as the lagged load, calendar events, temperature and cloud cover forecasts. For further information, see [24] and [18]. The EDF operational model is not included in our study. In practice, the true consumption of the EDF customers is not known in real time unlike our assumption above, and therefore the operational model cannot be compared with other models on an equal footing.

The RMSE and the MAPE from different models are reported in Table 4, and Fig. 6 shows the plot of RMSE averaged within each month. For brevity and better representation, only O3 is included among the oracle predictors and the base predictor is omitted.

Overall, the forecasting performance of any model considered, including the oracle predictors, is better in summer than in winter as can be seen in Fig. 6. The

relative difficulty of forecasting French electricity loads in winter has been noted in [9, 10] and [7]. This may be accounted for by higher variability among the daily loads in winter, which is markedly greater than that in summer as demonstrated in Fig. 2.

Also, it is observable from Fig. 6 that among P1–P3, different models outperform the others in different months. For instance, in June and September, P1 performs as well as P2 and P3 or even slightly better, but its performance is considerably worse during colder seasons. In general, the efficacy of having temperature included in the regressor is likely to depend on the homogeneity of the observed temperature curves within each class and the quality of the temperature forecasts. Therefore, we may achieve improved forecasting performance by combining these predictors in an adaptive way, either by selecting one predictor out of the three, or by assigning some data-driven weights to the three predictors on each day. Indeed, by selecting the best forecast out of the three a posteriori (i.e. assuming that the true future load is known), we can reduce the overall RMSE to 660 MW.

Without attempting to be theoretically rigorous, we produced a new predictor (P4) by averaging two out of the three each day, where the two predictors were chosen as those two closest to each other. This additional step can be achieved without any prior knowledge of the future load, yet succeeds in reducing prediction errors by a considerable margin as reported in Table 4. Also, P4 universally outperforms P1–P3 in terms of RMSE in any month of a year. We note that there is a growing interest in the problem of aggregating multiple expert advices in the context of short-term electricity load forecasting. For example, [8] investigate this problem by sequentially updating the convex weights applied to various forecasting models based on the past performance.

The performance of hybrid approaches (H1–H3) is substantially worse than that of their simplified counterparts (P1–P3). It can be explained by the fact that the errors from fitting and predicting the weekly average loads at the weekly level modelling (see Fig. 7), are carried over to the daily level curve linear regression modelling. We note that the electricity load dataset studied here covers the consumption of the customers of EDF only, rather than that of the entire French population as in [6]. Therefore its weekly average loads are more prone to digress from the overall trend or the seasonality estimated from the past observations due to e.g., the departure and the arrival of customers. This leads to greater variance in modelling the linear relationship between $\hat{\xi}_{ij}$ and $\hat{\eta}_{ik}$, $k = 1, \ldots, Q$ (see Fig. 8), even when the same classification rule has been applied to the daily loads, and thus to worse prediction models.

The superior performance of P1–P3 to H1–H3 indicates that the classification of successive daily loads effectively handles the dependency of the trend and the seasonality of electricity load data on the calendar variables. While we have used a simple classification rule combining the DT and the SC in this study, existing functional data clustering methods such as [5] may be applied to divide the successive daily loads into sub-groups of homogeneous profiles and covariance

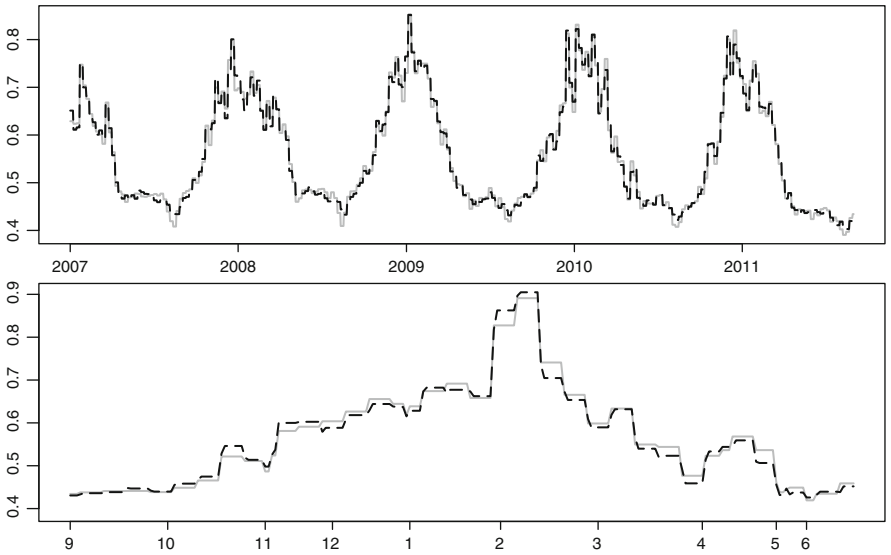


Fig. 7 Weekly average loads (*grey, bold*) between 1 January 2007 and 31 August 2011 and their fitted values (*black, dashed*) (*top*); weekly average loads between 1 September 2011 and 15 June 2012 and their forecasts (*bottom*)

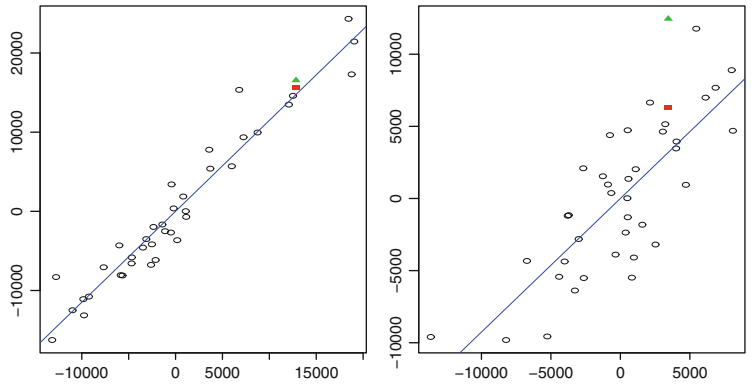


Fig. 8 Relationship between $\hat{\eta}_{i1}$ and $\hat{\xi}_{i1}$ obtained from P2 (*left*) and H2 (*right*) at noon 12 June 2012, along with the respective $(\hat{\eta}_1, \hat{\xi}_1)$ (predicted, *filled square*) and $(\hat{\eta}_1, \tilde{\xi}_1)$ (oracle, *filled triangle*)

structure in a more data-driven way, rather than relying on any prior knowledge on electricity consumption patterns which may change over time.

Also, there are certain factors which are known to have substantial influence on daily electricity loads yet have not been incorporated into our forecasting framework. An example of such factors is the special tariff options offered by EDF to large businesses on certain days in January–March and November–December, with the purpose of reducing heavy electricity consumption during winter. This

scheme is known to affect not only the daily consumption on the special tariff days but also that on the days preceding and following. A data-driven classification tool may be able to identify such influence of the scheme without being furnished with the exact dates or any other information on the load patterns on the relevant days, and thus further improve the quality of forecasts.

According to the overall prediction errors, GAM performs better than P2 and worse than P4 by a small margin, and the breakdown of RMSE with respect to different months in Fig. 6 does not reveal any patterns so as to the relative performance of our method and GAM in different months. The oracle predictors attain the minimum errors throughout the year, which further validates the previous statement that the SVD-based dimension reduction method is successful in capturing the dependence between the regressor and the response curves. It also supports our observations that there is a scope for improvement, e.g. via adaptive aggregation of different forecasting models and data-dependent classification of successive daily loads.

5 Conclusions

In this article, we addressed the problem of daily electricity load forecasting via curve linear regression, with emphasis on the adaptivity of the proposed method to ever-changing electricity consumption environment. The curve linear regression technique was introduced in a generic setting, where the singular value decomposition in a Hilbert space reduced the curve linear regression model to a finite number of scalar linear regression models.

Although it had previously been proposed by [6] as the second stage of the hybrid method for daily load forecasting, we showed that the curve linear regression technique could be applied directly to the data without any preliminary trend and seasonality modelling, based on the following rationale.

- The trend and the seasonality depend on the calendar variables which can be used as classification criteria, and when equipped with such a classification step, the weekly level modelling is redundant.
- In the hybrid approach, the prediction error from the first stage is carried over to the second stage, which leads to the increased variance in curve linear regression modelling and thus to significantly deteriorated prediction performance.

Also, the reduced approach requires less human intervention and is more adaptive to the time-varying nature of the data, and its superior prediction performance has been demonstrated with a real data example. Besides, within the reduced framework, it is more straightforward to carry out further statistical analysis such as obtaining a prediction interval around the forecast. By focusing exclusively on curve linear regression, some interesting topics for further improving the methodology have been made clearer throughout the real data analysis.

Firstly, as seen in Sect. 3.1, clustering the daily loads into homogeneous subgroups, in terms of both their profiles and dependence structure, plays a key role in

electricity load data analysis. Data-driven classification of the successive daily loads can greatly improve the forecasting results, as well as providing interesting insights on the data itself. There is an active interest on developing functional data clustering techniques, and adapting these methods to electricity load data is a problem which requires our immediate attention.

Further, since the curve linear regression framework allows flexible choice of regressor, we can have a number of forecasting models with different regressors. Therefore, it is of interest to see whether we can achieve improved forecasting performance by adaptively aggregating multiple forecasts. As briefly explored in Sect. 4, a simple adjustment in this direction can enhance the prediction performance substantially. Also on a more general note, an automatic selection of the regressor in curve linear regression may be widely adopted as a functional data analysis tool beyond the context of electricity load forecasting.

References

1. Antoniadis, A., Brossat, X., Cugliari, J., & Poggi, J. M. (2013). Functional clustering using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing, 11*, 1350003(30 pages).
2. Bathia, N., Yao, Q., & Ziegelmann, F. (2010). Identifying the finite dimensionality of curve time series. *Annals of Statistics, 38*, 3352–3386.
3. Bosq, D. (2000). *Linear processes in function spaces: Theory and applications* (Lecture Notes in Statistics, Vol 149). New York: Springer.
4. Bruhns, A., Deurveilher, G., & Roy, J. S. (2005). A non linear regression model for mid-term load forecasting and improvements in seasonality. In *Proceedings of the 15th power systems computation conference*, Liège (pp. 22–26).
5. Chiou, J. M., & Li, P. L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B, 69*, 679–699.
6. Cho, H., Goude, Y., Brossat, X., & Yao, Q. (2013). Modelling and forecasting daily electricity load curves: A hybrid approach. *Journal of the American Statistical Association, 108*, 7–21.
7. Cugliari, J. (2011). Prévision non paramétrique de processus à valeurs fonctionnelles: Application à la consommation d'électricité. PhD thesis. University Paris XI, Paris, France.
8. Devaine, M., Gaillard, P., Goude, Y., & Stoltz, G. (2013). Forecasting electricity consumption by aggregating specialized experts. *Machine Learning, 90*, 231–260.
9. Dordonnat, V., Koopman, S. J., Ooms, M., Dessertaine, A., Collet, J. (2008). An hourly periodic state space model for modelling french national electricity load. *International Journal of Forecasting, 24*, 566–587.
10. Dordonnat, V., Koopman, S. J., Ooms, M., Dessertaine, A., & Collet, J. (2012). Dynamic factors in periodic time-varying regressions with an application to hourly electricity load modelling. *Computational Statistics and Data Analysis, 56*, 3134–3152.
11. Fan, S., & Hyndman, R. J. (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems, 27*, 134–141.
12. Hall, P., & Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics, 35*, 70–91.
13. Hall, P., & Vial, C. (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society: Series B, 68*, 689–705.
14. Hallin, M., & Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association, 102*, 603–617.

15. He, G., Müller, H. G., Wang, J. L., & Yang, W. (2010). Functional linear regression via canonical analysis. *Bernoulli, 16*, 705–729.
16. James, G. M., & Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association, 98*, 397–408.
17. Lam, C., & Yao, Q. (2012). Factor modelling for high-dimensional time series: inference for the number of factors. *Annals of Statistics, 40*, 694–726.
18. Nedellec, R., Cugliari, J., & Goude, Y. (2014). Gefcom2012: Electricity load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting, 30*(2), 375–381.
19. Pierrot, A., & Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. In *Proceedings of the 16th international conference on intelligent system application to power systems*, Hersonissos (pp. 22–26)
20. Ramsay, J. O., & Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B, 53*, 539–572.
21. Ray, S., & Mallick, B. (2006). Functional clustering by bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B, 68*, 305–332.
22. Serban, N., & Wasserman, L. (2005). CATS: Clustering after transformation and smoothing. *Journal of the American Statistical Association, 100*, 990–999.
23. Smithies, F. (1937). The eigenvalues and singular values of integral equations. *Proceedings of the London mathematical society, 43*(2), 255–279.
24. Wood, S., Goude, Y., & Shaw, S. (2015). Generalized additive models for large datasets. *Journal of Royal Statistical Society: Series C, 64*(1), 139–155.
25. Yao, F., Müller, H. G., & Wang, J. L. (2005). Functional linear regression analysis for longitudinal data. *Annals of Statistics, 33*, 2873–2903.

Constructing Graphical Models via the Focused Information Criterion

Gerda Claeskens, Eugen Pircalabelu, and Lourens Waldorp

Abstract A focused information criterion is developed to estimate undirected graphical models where for each node in the graph a generalized linear model is put forward conditioned upon the other nodes in the graph. The proposed method selects a graph with a small estimated mean squared error for a user-specified focus, which is a function of the parameters in the generalized linear models, by selecting an appropriate model at each node. For situations where the number of nodes is large in comparison with the number of cases, the procedure performs penalized estimation with quadratic approximations to several popular penalties. To show the procedure's applicability and usefulness we have applied it to two datasets involving voting behavior of U.S. senators and to a clinical dataset on psychopathology.

1 Introduction

We propose a focused search method of estimating an undirected graph when the distribution of the random variables associated with each node is a member of an exponential family of distributions, including the Gaussian, Poisson and binomial distributions as special cases. The graph is constructed nodewise, hence instead of solving one multivariate optimization problem which in this case is difficult in general, we proceed by optimizing many univariate problems (one for each node) and then 'glue together' all the pieces of information.

By the *focus* of the research we mean a predefined function of the model parameters, such as the mean of a response variable in a regression model. This focus we wish to estimate well in the sense of having a low mean squared error

G. Claeskens (✉) • E. Pircalabelu
ORSTAT and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
e-mail: gerda.claeskens@kuleuven.be; eugen.pircalabelu@kuleuven.be

L. Waldorp
Department of Psychological Methods, University of Amsterdam, Weesperplein 4, 1018 Amsterdam, The Netherlands
e-mail: waldorp@uva.nl

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_4

(MSE). In the nodewise approach we fit at each node a generalized linear model (GLM), implying that selecting the neighboring nodes, or equivalent, selecting the edges in the graph, is nothing but a variable selection problem in a generalized linear model. Obviously, different models give rise to different bias and variance quantities for the focus estimator, and thus searching a model which produces a small MSE for an estimator is a sensible thing to do. Moreover, a researcher can have different focuses which reflect his/her scientific interests and thus one can estimate using a given dataset several (possibly different) graphs which serve the corresponding research purposes. With this in mind, we point out that the focused approach may take more carefully the domain knowledge into account that is available to the researcher when defining the focus, and outputs a model fine-tuned for that specific focus. Thus this approach moves from a 'one model for all purposes' scheme, to a 'one model per purpose' approach.

Graphical models visualize relations that exist between components of a multivariate random vector, say $\boldsymbol{X} = (X_1, \ldots, X_p)$. In a graph, each component of this vector is identified with a node, and a relation between two components is visualized by drawing a connection, an edge, between the corresponding nodes. For an example, see Fig. 3. Different types of relations between the random components may be represented by different types of edges (with or without arrowheads).

The most common types of graphical models to be encountered in the literature are Bayesian networks and Markov networks. In terms of graphical representations the Bayesian networks are based on directed graphs (lines with arrowheads) while Markov networks are based on undirected graphs. Both types of models try to graphically encode the conditional independencies that hold between variables that are represented here by nodes in the graph. In the case of directed graphs drawing a directed edge as $i \rightarrow j$ is to be understood that node i influences node j or that node i 'causes' in some sense node j. For example one might represent a relation between Age and Income in a graph as $Age \rightarrow Income$ with a clear message that one's income depends on ones age as on average older people earn more than younger, but refute the relation $Age \leftarrow Income$ as this makes probably no 'causal' sense. On the other hand if one faces a situation where a causal effect cannot be assumed in any of the two directions then one can find it useful to place undirected edges between the two nodes, in order to signalize that there exists an association between these two nodes though without a precise directionality effect. In a genomics study one might assume that gene i is correlated with gene j, and represent it by an undirected edge as it might be implausible that any of the genes has a direct effect on the other. For a comprehensive explanation about graphical models, we refer to Lauritzen [12] or Cox and Wermuth [6].

We here concentrate on undirected graphical models. For Gaussian random vectors, having an edge in an undirected graph yields the following interpretation. Random variable X_i is dependent on X_j conditioned on all remaining variables in the multivariate vector, if and only if there is an edge in the graph between the nodes representing these variables X_i and X_j. Equivalently, there is a zero entry at the crossing of row i and column j in the inverse covariance matrix, also called the concentration matrix, if and only if no edge is drawn between the corresponding

nodes i and j in the graph. Thus, estimating a graphical model in the sense of drawing edges between nodes, is equivalent to determining the positions of the zero and non-zero entries in the concentration matrix, see Dempster [7] and Lauritzen [12]. Thus, in case one discovers an entry in the concentration matrix that is zero, or equivalently, one finds conditional independencies, there is a simpler way of writing the joint distribution of the multivariate vector X, that adequately describes the relations between the components of X.

Let us consider a sample of n multivariate random vectors $X_k = (X_{k1}, \ldots, X_{kp})$, $k = 1, \ldots, n$, each consisting of p components. One way to estimate the non-zero entries in the concentration matrix is through nodewise regression models [16]. In turn, each random variable associated with a single node (say node i) is taken as the response variable and the variables corresponding to the other nodes act as covariates (predictors). A non-zero entry is considered to exist at row i and column j $(i, j = 1, \ldots, p)$ when, for Gaussian data, the coefficient $\beta_{ij} \neq 0$ in the regression model with the variable corresponding to node i as the response

$$X_{ki} = \beta_{i0} + \sum_{l=1,\ldots,p, l \neq i} \beta_{il} X_{kl} + \varepsilon_{ki}, \tag{1}$$

and at the same time, $\beta_{ji} \neq 0$ in the regression model with the variable corresponding to node j as the response variable

$$X_{kj} = \beta_{j0} + \sum_{l=1,\ldots,p, l \neq j} \beta_{jl} X_{kl} + \varepsilon_{kj}, \tag{2}$$

with ε_{ki} and ε_{kj} independent normal random variables with zero mean and $k = 1, \ldots, n$ observations. This is referred to as an 'AND' rule. One could also use an 'OR' rule that includes an edge between nodes i and j when either β_{ij} '*or*' β_{ji} is nonzero (we refer to Meinshausen and Bühlmann [16], for an application based on the two rules). Throughout the paper the 'OR' rule is applied for constructing the graphs, due to the high-dimensionality of the problem and the greedy manner in which nodewise models are constructed. The 'OR' rule might overfit by including spurious edges, but one would rather have a model that overfits (i.e. not missing some important edges) than a model that underfits (missing important edges).

In Pircalabelu et al. [18] we propose to use the focused information criterion (FIC, [3]), which is driven by the mean squared error, to select the variables in the above nodewise regression models. Once the neighbors for all variables in the nodewise regression models are selected by FIC, we can draw the selected graph. This is referred to as the *FIC selected graph*. See Sect. 4.2. This idea is extended to larger graphs in Pircalabelu et al. [17], using penalized estimation methods, see also Sect. 4.3.

A main reason for using the focused information criterion and not any other variable selection method, is that this criterion makes it possible to obtain tailor-made graphs. For instance, graphs representing interconnectivity in mental symptoms (see

e.g., [2]) can provide predictions of the development of a disorder in patients. Such predictions are optimal whenever the graph used to represent the disorder is tuned to certain types of predictions (see the example in Sect. 2 on psychopathology for more details). In a statistical sense, a good estimator is one with a low mean squared error (MSE), which is defined as the sum of the squared bias and the variance of the estimator.

For each node as a 'response variable' we have for each remaining variable the choice to include it or not to include it as a covariate, resulting in a list of possible models. In each such model we can estimate the focus. Underlying the FIC are estimators of the mean squared error of the focus estimators in each of the different regression models under consideration. Minimizing the FIC is equivalent to minimizing the estimated MSE of the focus over the different models. Thus, we select for each node a regression model and use this in a next step to construct a graph, that is aimed to give a low MSE for the estimated focus.

The tailor-made aspect of FIC is easily understood. Specifying a different focus, will result in different focus estimators and thus in different MSE values, and consequently different FIC values. Hence, different focuses may lead to different graphs. Each time, we select that graph that scores best in estimated MSE (that is, FIC) for that focus. More details are given in Sect. 4.

In this chapter we extend the methodology of the FIC for graphs based on Gaussian random variables, to graphs for multivariate random vectors where the nodes may be fit through generalized linear models [15].

2 Data Examples

2.1 Data Example on 'Dynamics of Psychopathology'

The data used in this subsection come from a study of van Borkulo et al. [21] and consist of a series of measurements for two subjects: a rapid cycling bipolar patient and a healthy control case. A bipolar patient has episodes of mania (energetic, highly productive, etc.) and/or depression; on average 0.5 episodes per year. A rapid cycling bipolar patient has at least four such episodes per year. Both subjects were asked to rate their feelings during 93 days on the Positive and Negative Affect scale (PANAS, [25]). The scale consists of 22 feelings or emotions and during each day the two subjects were asked to rate on a 5 point Likert scale (ranging from 'not at all' to 'extremely') to what extent the feeling pertains to them. All variables have been discretized to 0/1 binary values where 0 indicated 'not at all', while 1 indicated all other categories. Afterwards, positive affect items were reversed scored, such that for a positive affect item a '0' value represents the presence of the positive feeling while on the negative affect item the same value represents the absence of a negative value. The purpose of the recoding was to concentrate on subjects that have had positive feelings compared to subjects that lacked to have these feelings. For

example, it means that if for a subject the value 0 is recorded for 'feeling interested' and 0 is recorded for 'feeling distressed and unhappy' then the subject is likely to have felt interested but *not* distressed and unhappy.

All 22 feelings were considered as nodes in a network that influence each other.

The main goal of those authors was to numerically quantify differences between the patient and control in the contact process framework (see [9]). In the contact process an infected node (determined by a value of 1) at time t can infect its immediate neighbors, which in turn can infect their other immediate neighbors. As time passes some of the previously infected nodes can also recover (switching from 1 to 0). Two independent Poisson processes are assumed: spontaneous recovery of infected nodes (with rate μ) and infection of healthy nodes (with rate proportional to λ). The estimated ratio $\rho = \lambda/\mu$, called the 'basic reproduction number' (BRP) is then used to quantify the differences between the two subjects. The analysis in van Borkulo et al. [21] suggests that for the bipolar patient the BRP is much higher than that of the control, meaning that for the patient the network will continue to be infected indefinitely.

One of the main assumptions of the model is that the researcher has a network at his/her disposal on which the infections and recoveries can be observed, and as such we wish to put forward possible networks after which, based on the estimated networks, we will estimate and compare the BRP for both subjects.

2.2 Data Example on U.S. Voting Behavior

The data set used here encodes the U.S. senate voting records data from the 109th congress between 2004 and 2006 (see [1]). It contains only binary 0/1 variables where a '0' represents a 'No' vote for the proposed bill and a '1' marks a 'Yes' vote. There are 100 variables, corresponding to 100 senators (64 of them being Democrats and 36 being Republican) and 542 cases, corresponding to 542 bills and amendments put to vote. As in the original paper, all missing votes per bill have been recoded as 'No' votes. The aim of the analysis is to estimate an undirected graph structure where each node represents a senator and each edge between two nodes represents a form of interaction between senators such that the voting behavior of one senator could be used as a predictor for the behavior of another senator. The entire dataset corresponding to 100 senators and 542 bills has been used in the analysis.

We are interested in the describing how the voting behavior of the senators depends on the voting behavior of all other senators. Therefore, we use as a focus the expected value of a node conditioned on the values of all other nodes. To this end we will use the voting pattern of all senators for the 'Flag Desecration' amendment sometimes referred to in the media as the 'flag-burning' amendment. The initiative proposed a constitutional amendment that would allow the U.S. Congress to outlaw the physical desecration of the flag of the United States. A vivid debate was started between supporters of the freedom of speech and supporters of national symbols,

and the attempt to adopt such an amendment failed by only one vote. All senators have given their vote on the bill and there was no missing information for this focus. We wish to estimate the undirected graphical structure that provides the smallest MSE of the focus estimator at each node, using the procedure described in Sect. 4. Since there are 100 nodes in this example, we will use the penalized approaches of Sect. 4.3.

2.3 Data Example on Hunting Spider Species

The data come from a study of van der Aart and Smeenk-Enserink [22] and consists of abundances (numbers trapped over a 60 week period) of hunting spiders in a Dutch dune area. There were 28 sites where data on 12 spider species were collected. In addition, the dataset contains measurements on 6 extra environmental variables for each studied area. The interest here lies in knowing whether and how selected graphs differ for two locations from the dataset. It is of interest to know whether environmental characteristics influence the structure of the estimated networks, as it is expected that some species might prefer to inhabit one type of environment while others might be less influenced by area characteristics. For this purpose we use the observed counts for each species at two locations for which the amount of fallen leaves, moss or the herb layer and the reflection of the soil surface are quite different (see Fig. 4). The hypothesis is that if the abundance of spiders was not related to area characteristics, the spider counts would be similar at the two locations and the estimated graphical models for these two focuses would be quite similar. Differences between the two estimated graphs can thus be linked to the effects area characteristics have on the presence of spiders.

3 Generalized Linear Models and Graphs

A p-variate random variable $\boldsymbol{X} = (X_1, \ldots, X_p)$ may be represented by a *graph* $\mathcal{G}$. A graph is mathematically defined by a pair of sets $(\mathcal{E}, \mathcal{V})$ where $\mathcal{V}$ is the set of nodes $\{1, \ldots, p\}$, each node j is identified with a univariate variable X_j, $j = 1, \ldots, p$, and where the set of edges $\mathcal{E}$ is a subset of $\mathcal{V} \times \mathcal{V}$ consisting of pairs of distinct nodes.

While in a Gaussian graph $\boldsymbol{X}$ follows a multivariate normal distribution, other distributions may be assumed. We here consider the situation that each component of $\boldsymbol{X}$ has a distribution belonging to an exponential family, such that we may fit nodewise generalized linear models, extending upon the linear models as in (1).

In a generalized linear model, the response Y has a distribution of the type

$$f(y; \vartheta, \phi) = \exp\Big\{\frac{y\vartheta - b(\vartheta)}{a(\phi)} + c(y, \phi)\Big\}, \tag{3}$$

where ϑ and ϕ are unknown parameters and where the functions a, b and c are known. The parameter ϕ is a scale parameter, and ϑ is the main parameter of interest, since it holds that $\mathrm{E}(Y) = \partial b(\vartheta)/\partial\vartheta = b'(\vartheta)$. Another interesting aspect of such a distribution is that $\mathrm{Var}(Y) = a(\phi)\partial^2 b(\vartheta)/\partial\vartheta_i^2$ (see e.g., [15]).

Common examples of this exponential family include the normal, Poisson, binomial and gamma distributions. For regression models where each observation may comprise of a vector of covariate values, the parameter ϑ may be taken differently for each observation.

While in a linear model the mean of the response $\mathrm{E}(Y|\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{\beta}$ is a linear function, in a generalized linear model there is a monotone and smooth link function denoted by g such that $\mathrm{g}\{\mathrm{E}(Y|\boldsymbol{x})\} = \boldsymbol{x}^t\boldsymbol{\beta}$. The special choice of $\mathrm{g}(\cdot) = (b')^{-1}(\cdot)$ is referred to as the canonical link. For Bernoulli distributions the logistic link is canonical, the identity function is canonical for normal distributions and for Poisson data it is the log-function.

While it would lead too far to construct a complete list of the existing work on Gaussian and 0/1 binary data for graph construction, we refer to some recent work of Yang et al. [26], Lee and Hastie [13], Jalali et al. [11] and Loh and Wainwright [14] who construct procedures oriented towards either situations where $\boldsymbol{X}$ is a discrete random variable, or situations where the distribution of $\boldsymbol{X}$ is a member of the more general exponential family of distributions. The above mentioned works are relevant to our case for several reasons. First, they work nodewise, where models are first selected at the level of the nodes and then everything is 'glued' together, and more importantly they also suggest that such nodewise constructions have merit because under certain conditions they are able to recover aspects of the true underlying graph.

Starting from a general form of a univariate exponential distribution, Yang et al. [26] formulate the problem as follows. The joint density (or probability mass function) of a p-dimensional random vector X is characterized by parameters ϑ that depend on the edges $(s,t) \in \mathscr{E}$, for all $s,t \in \mathscr{V}$, similar to a representation of the Ising model where it is assumed that the interactions between random variables X_i are of first and second order [23]. The density of a particular node x_s conditioned upon all remaining nodes, can then be determined, based on their modelling approach as

$$f(x_s|x_{\mathscr{V}\setminus s}) = \exp\{\vartheta_s x_s + \sum_{t\in\mathscr{N}_s} \vartheta_{st} x_s x_t - b(\vartheta, x_{\mathscr{V}\setminus s}) + c(x_s)\},$$

where $b(\vartheta, x_{\mathscr{V}\setminus s})$ is a log-normalizing constant, $c(x_s)$ is a 'base measure' and $\mathscr{N}_s$ is the neighborhood of node s, namely the set of nodes that are directly connected to node s.

Given independent and identically distributed samples and the above conditional densities, Yang et al. [26] then proceed by minimizing an ℓ_1-regularized conditional log likelihood (see also Sect. 4.3) at each of the nodes, estimating sets of neighbors for each node. The merit of such an approach is that under general 'ℓ_1' regularity conditions the estimated neighbors correspond with high probability to the ones in

the underlying, unknown graph, thus making the effort worthwhile and at the same time justifying why a simple nodewise approach is a sensible thing to do.

A second important interest lies in knowing if for non-Gaussian graphs, the missing edges in $\mathcal{G}$ can be translated into a '0' entry in the inverse of a covariance matrix, mimicking the behavior encountered for Gaussian graphical models. This is the topic of Loh and Wainwright [14]. Unfortunately this property of having 0's on position (s,t) and (t,s) in a general inverse of a covariance matrix if an edge is missing between nodes s and t does not hold for general graphs. Corollary 2 in their paper asserts that the inverse covariance matrix is graph structured only for graphs with singleton separator sets. Outside this condition, one can still observe 0's in a inverse covariance matrix constructed not on the original nodes but on an 'augmented' set of nodes where one includes also higher order interactions between the nodes in the set

$$S(s;d) := \{U \subseteq \mathcal{V} \setminus s, |U| = d\}$$

where d denotes an upper bound on the degree of node s (i.e. the number of edges connecting node s to any other node in the graph). As such, Corollary 3 in Loh and Wainwright [14] asserts that the inverse of the augmented covariance matrix contains 0's on positions (s,t) for all nodes $t \notin \mathcal{N}_s$. It is in a sense a weaker result than desired, but nonetheless quite useful in understanding which conditional independencies can be read from the graph and how these are translated in 0 entries in a more familiar and easier to use generalized inverse covariance matrix. The main conclusion of the above line of work is that nodewise models can still enjoy good theoretical properties and that, as in the Gaussian case, a missing edge in $\mathcal{G}$ can still correspond to a 0 element in an augmented covariance matrix.

While in Yang et al. [26] sparsity constraints were included and large graphs were considered, we start in Sect. 4 with unconstrained estimation for small to moderately sized graphs.

Extending upon the nodewise linear regression models in (1) and (2), we will include an edge in the graph between nodes i and j when using the focused information criterion, see Sect. 4, results in including variable X_j in the generalized linear model using X_i as a response variable with a non-zero coefficient β_{ij},

$$\mathrm{g}\{\mathrm{E}(X_i|\{X_j : j \in \mathcal{V} \setminus i\})\} = \beta_{i0} + \sum_{l \in \mathcal{V} \setminus i} \beta_{il} X_l,$$

and vice versa, when β_{ji} is nonzero in the generalized linear model when X_j is the response variable. In the case that X_i is a binary random variable, logistic regression models may be used to model the log-odds

$$\log\left\{\frac{P(X_i = 1|\{X_j : j \in \mathcal{V} \setminus i\})}{1 - P(X_i = 1|\{X_j : j \in \mathcal{V} \setminus i\})}\right\} \equiv \log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_{i0} + \sum_{l \in \mathcal{V} \setminus i} \beta_{ik} X_l.$$

The 'response' node is referred to as the 'child' and the 'covariate' nodes with non-zero coefficient are commonly called the 'parents' of that node.

4 The Focused Information Criterion for Graphs

As a definition of a focus, the current theoretical derivation allows for any function μ of the nodewise model parameters $\boldsymbol{\beta}$ that is differentiable with respect to these parameters, at least in a neighborhood of the true but unknown parameter values $\boldsymbol{\beta}_0$. We will here define the focus nodewise such that we can readily apply existing search algorithms for nodewise variable selection. The mean squared errors of nodewise focus estimators are summed to yield the graph-wise mean squared error [18].

4.1 Model Notation and Local Misspecification

Consider a sample of n observations of the p-variate vector $\boldsymbol{X}_k = (X_{k1}, \ldots, X_{kp})$, with $k = 1, \ldots, n$. For each node $j = 1, \ldots, p$ and for each observation $k = 1, \ldots, n$, we have that

$$X_{kj} | \{X_{ki} : i \in \mathscr{V} \setminus j\}$$

follows a generalized linear model as in (3).

We define the vector $\boldsymbol{\theta}_j$ to contain the parameters that should be estimated in all models for this node and that are always included. One example is the scale parameter ϕ when not already specified by the particular exponential family distribution. The vector θ_j might also include the coefficient corresponding to parent nodes that are forced to be in the graph, often based on domain knowledge or on theoretical grounds. Note that $\boldsymbol{\theta}_j$ may be empty (absent).

We further define for each node $j \in \mathscr{V}$ the vector $\boldsymbol{\gamma}_j$ of length $p - 1$ with ith element, $i \in \mathscr{V}$, equal to

$$\gamma_{ji} = \begin{cases} \beta_{ji} & \text{if } X_i \text{ is a parent of } X_j \\ 0 & \text{otherwise.} \end{cases}$$

Note that the vector $\boldsymbol{\gamma}_j$ does not have any overlap in parameters with $\boldsymbol{\theta}_j$, that is, model parameters are either included in $\boldsymbol{\gamma}_j$ or in $\boldsymbol{\theta}_j$. Thus for each node $j \in \mathscr{V}$ the vector of unknown parameters $\boldsymbol{\beta}_j = (\boldsymbol{\theta}_j, \boldsymbol{\gamma}_j)$.

This notation assumes that for every node a full model is fit, with all other nodes as parents. This results in a full graph, where all nodes are connected to all other nodes. It is the task of a model selection method such as the FIC that we will use, to properly select the parents of each node, and as such, to reduce the fully connected graph to a simpler graph.

For this purpose we introduce notation for submodels. For each node $j \in \mathscr{V}$, when using a submodel $S \subset \mathscr{V} \setminus j$, we denote by γ_S the subvector of γ formed by the components $\{\gamma_{ji} : i \in S\}$. In the submodel defined by S, other components γ_{jk}

with $k \notin S$ are taken to be zero. Such a selection of components may algebraically be defined through multiplication with projection matrices selecting the wanted components, see Claeskens and Hjort [5, sec. 6.1].

The nodewise focus μ_j that we wish to estimate with low mean squared error can be written as $\mu_j(\theta_j, \gamma_j; x)$. One example is a nodewise expectation $\mu_j(\theta_j, \gamma_j; x) = \boldsymbol{x}^t\gamma_j$ for a user-specified vector x. We will estimate μ_j in a submodel S by $\hat{\mu}_{j,S} = \mu_j(\hat{\theta}_{j,S}, \hat{\gamma}_{j,S}; x)$ using maximum likelihood estimators in the submodel. Note that no selection of components takes place for θ_j, though its estimated value might in general depend on which components of γ_j are included in S.

In order to estimate the mean squared error of $\hat{\mu}_S$, we employ a *local misspecification framework* where the true parameter vector has the form $(\boldsymbol{\theta}_{j,0}, \boldsymbol{\gamma}_{j,0} + \boldsymbol{\delta}_j/\sqrt{n})$, for some unknown vector δ. This construction will result in squared biases of estimators that are of the same order as variances, thus resulting in mean squared error values that are not driven by bias or variance only, as the sample size grows. Working under a fixed true model (not depending on the sample size) would lead to suggest to always use the full model since in that case the bias would dominate, see Claeskens and Hjort [5, sec. 5.2].

4.2 FIC for Small to Moderate Graphs

The strategy for estimation of the mean squared error of $\hat{\mu}_{j,S}$ in each considered model S is as follows. By taking the mean and variance from the asymptotic distribution of the estimator $\hat{\mu}_{j,S}$, the mean squared error is easily formed. This expression is estimated in a next step to form the focused information criterion. For the asymptotic distribution of the estimators $\hat{\mu}_{j,S}$ under local misspecification in the specific case of generalized linear models, see Claeskens and Hjort [4].

Let us consider the general situation where there is an unknown scale parameter ϕ in the exponential family distribution and where some of the parents are protected from variable selection. For node $j \in \mathcal{V}$, denote the 'protected' parents, those that are forced to be present in the graph, by $\boldsymbol{U}_j$ and those that are subject to model selection by $\boldsymbol{Z}_j$; remark that nodes are either protected or unprotected, not both, hence $\boldsymbol{U}_j$ and $\boldsymbol{Z}_j$ do not contain common components. Likewise, we write $x = (u, z)$.

Define $J_{n,\phi} = -n^{-1}\sum_{k=1}^{n} E[\partial^2 \log f(X_{kj}; \vartheta_k, \phi)/\partial\phi^2]$ using the exponential family density function as in (3). Then, the information matrix corresponding to the full model for the jth node is given by

$$J_n = \begin{pmatrix} J_{n,\phi} & 0 & 0 \\ 0 & n^{-1}a(\phi)^{-1}\boldsymbol{U}^t\boldsymbol{V}\boldsymbol{U} & n^{-1}a(\phi)^{-1}\boldsymbol{U}^t\boldsymbol{V}\boldsymbol{Z} \\ 0 & n^{-1}a(\phi)^{-1}\boldsymbol{Z}^t\boldsymbol{V}\boldsymbol{U} & n^{-1}a(\phi)^{-1}\boldsymbol{Z}^t\boldsymbol{V}\boldsymbol{Z} \end{pmatrix},$$

for which we assume that a limit J exists for n tending to infinity; this condition could also have been phrased in terms of conditions on the design matrices $\boldsymbol{U}$ and $\boldsymbol{Z}$.

We assume that J_n and J are invertible. The matrix V that is used in J_n is a diagonal matrix $\mathrm{diag}\{v_1, \ldots, v_n\}$ with $v_k = [b''(\vartheta_k)\{g'(\xi_k)\}^2]^{-1}$ and $\xi_k = E[X_{jk}|U_{jk}, Z_{jk}] = b'(\vartheta_k)$. The vector $\boldsymbol{\theta}_j$ consists of ϕ and of the coefficients ν_j belonging to the protected variables $\boldsymbol{U}_j$. Thus $\theta_j = (\phi, \nu_j)$. Denote the length of $\boldsymbol{\theta}_j$ by p_θ and the number of elements in S by $|S|$.

Standard maximum likelihood methods yield that as n tends to infinity,

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_{j,S} - \theta_{j,0}) \\ \sqrt{n}\hat{\gamma}_{j,S} \end{pmatrix} \xrightarrow{d} N_{p_\theta+|S|}\left(\begin{pmatrix} \mathbf{0} \\ \boldsymbol{\delta} \end{pmatrix}, J_S^{-1}\right).$$

We denoted by J_S^{-1} the inverse of the $(p_\theta + |S|) \times (p_\theta + |S|)$ matrix J_S that is formed by selecting from J those rows and columns indexed by S.

Since we are interested in the asymptotic distribution for the focus estimator at the jth node $\mu_j(\hat{\theta}_{j,S}, \hat{\gamma}_{j,S}; x)$, we use the differentiability of μ_j with respect to the parameters (θ, γ) to first define

$$\omega = \boldsymbol{Z}^t\boldsymbol{V}\boldsymbol{U}(\boldsymbol{U}^t\boldsymbol{V}\boldsymbol{U})^{-1}\frac{\partial\mu_j}{\partial\boldsymbol{\nu}_j} - \frac{\partial\mu_j}{\partial\boldsymbol{\gamma}_j},$$

$$\tau_0^2 = \frac{1}{J_\phi}\left(\frac{\partial\mu_j}{\partial\phi}\right)^2 + na(\phi)\left(\frac{\partial\mu_j}{\partial\boldsymbol{\nu}_j}\right)^t(\boldsymbol{U}^t\boldsymbol{V}\boldsymbol{U})^{-1}\left(\frac{\partial\mu_j}{\partial\boldsymbol{\nu}_j}\right),$$

where all partial derivatives are evaluated at $(\theta_0, 0)$. Then, see Claeskens and Hjort [4, 5, Chapter 6], as n tends to infinity,

$$\sqrt{n}(\hat{\mu}_{jS} - \hat{\mu}_{j\mathrm{true}}) \xrightarrow{d} \Lambda_S,$$

where $\mathrm{E}(\Lambda_S) = \omega^t(I_{p-1} - G_S)\delta$ and $\mathrm{Var}(\Lambda_S) = \tau_0^2 + \omega^t Q_S^0\omega$ with Q the limit of $Q_n = a(\phi)n\{\boldsymbol{S}^t\boldsymbol{V}(I_n - \boldsymbol{U}(\boldsymbol{U}^t\boldsymbol{V}\boldsymbol{U})^{-1}\boldsymbol{U}^t\boldsymbol{V})\boldsymbol{Z}\}^{-1}$, I_n a square identity matrix with n rows and G_S the limit of $G_{n,S} = Q_{n,S}^0 Q_n^{-1}$. The matrix $Q_{n,S}^0$ is defined as follows. Take from Q_n^{-1} the submatrix consisting of those rows and columns indexed by S. We invert the obtained matrix and place its matrix elements in a $(p-1) \times (p-1)$ matrix in the rows and columns indexed by S, and set the other matrix elements equal to zero. In words, Q^{-1} is premultiplied with part of its inverse such that $G_{n,S}$ is a zero matrix when S is the empty set and $G_{n,S}$ is the identity matrix for the full model when $S = \mathscr{V} \setminus j$. Since $G_{n,S}$, $Q_{n,S}$ and $Q_{n,S}^0$ are all defined via submatrices of J_n, the existence of a limit matrix for $n \to \infty$ is guaranteed via the existence of the limit matrix J and of its inverse J^{-1}.

We now obtain the mean squared error for $\mu_j(\hat{\theta}_{j,S}, \hat{\gamma}_{j,S})$ by adding its variance and its squared bias as

$$\mathrm{MSE}(\hat{\mu}_{jS}) = \tau_0^2 + \omega^t Q_S^0\omega + \omega^t(I_{p-1} - G_S)\delta\delta^t(I_{p-1} - G_S)^t\omega, \tag{4}$$

where Q_S^0 is the limit of $Q_{n,S}^0$ and I_{p-1} represents a square identity matrix with $p-1$ rows. The best choice of parents to use in the nodewise regression model is that set S for which $\text{MSE}(\hat{\mu}_{jS})$ is as small as possible. Since this expression contains several unknown quantities, we insert estimators for unknowns, indicated by a 'hat' notation, where for example $\hat{Q}$, $\hat{Q}_S^0$ and $\hat{G}_S$ represent the empirical estimates of the corresponding matrices, resulting in an expression for the focused information criterion, FIC.

In particular, we estimate $\delta\delta^t$ unbiasedly by $\hat{\gamma}_{j,w}\hat{\gamma}_{j,w}^t - \hat{Q}$ where $\hat{\gamma}_{j,w}$ is the estimator of γ_j in the wide, or full, model using $S = \mathscr{V} \setminus j$, and an empirical information is used with parameters estimated at the full model. This results in defining the focused information criterion for node $j \in \mathscr{V}$ using subset S as parents:

$$\text{FIC}(S, \mu_j) = \hat{\tau}_0^2 + 2\hat{\omega}^t \hat{Q}_S^0 \hat{\omega} + n\hat{\omega}^t (I_{p-1} - \hat{G}_S)\hat{\gamma}_{j,w}\hat{\gamma}_{j,w}^t (I_{p-1} - \hat{G}_S)^t \hat{\omega} - \hat{\omega}^t \hat{Q} \hat{\omega}. \quad (5)$$

Note that since the first and the last term do not depend on the particular submodel S, these terms may be omitted when nodewise ranking the values of $\text{FIC}(S, \mu_j)$ for different sets S. Further, note that in these nodewise regression models, also the matrix Q_n, and as a consequence also ω, $Q_{n,S}^0$ and $G_{n,S}$ are nodewise defined.

The value of the FIC for the complete graph is defined by Pircalabelu et al. [18] as the nodewise summation of the FIC values for each node given in (5),

$$\text{FIC}(\mathscr{S}; \mathscr{G}) = \sum_{j=1}^{p} \text{FIC}(S_j, \mu_j), \quad (6)$$

where $\mathscr{S} = \{(S_1, \ldots, S_p) : S_1 \subseteq \{\mathscr{V} \setminus 1\}; \ldots; S_p \subseteq \{\mathscr{V} \setminus p\}\}$, and each S_j corresponds to the selected nodes that minimize the FIC score of the estimated focus at node j.

The best graph in estimated MSE sense according to the focused information criterion is given by that selection of nodewise parents S_j ($j = 1, \ldots, p$) for which the combined FIC value $\text{FIC}(S; \mathscr{G})$ is the smallest over all considered sets. In the case it happens that two choices of $\mathscr{S}$ would give identical FIC values, other aspects of modeling, e.g. parsimony considerations, might help decide the final selection, in the same way as is done for model selection via other information criteria. Model averaging might also be an option when prediction is the objective.

4.3 FIC for Large Graphs

While the FIC in (6) relies on maximum likelihood estimation, this no longer is feasible when many nodes are involved. For situations with many unknown parameters (including the situations where there are more unknown parameters than observed cases), penalized estimation methods are appropriate.

In such case the estimators are maximizers of the penalized objective function

$$Q(\theta,\gamma) = \frac{1}{n}\sum_{k=1}^{n} \log f(y_k|w_k,z_k,\theta,\gamma) - \frac{1}{n}\sum_{j=1}^{p-1} \psi_\lambda(|\gamma_j - \gamma_{j0}|), \tag{7}$$

with respect to θ and γ for a given penalty function ψ that is twice differentiable in 0 and that depends on the penalty constant λ. This $\lambda \geq 0$ is a user-determined value, which may be obtained in a data-driven fashion, and γ_{j0} is the value of the coefficient γ_j in the narrow model. The effect of the penalty is that the estimators are shrunk towards zero. Typical choices are ℓ_2 (sum of squares), ℓ_1 (sum of absolute values) or ℓ_0 (hard thresholding) penalties.

By adding a penalty to the estimation Meinshausen and Bühlmann [16] propose using a series of nodewise Lasso regression models, using an ℓ_1 penalty, to estimate large graphical models. See also Wainwright et al. [24] and Schmidt et al. [19] among many others. Neighborhoods of different nodes can be connected in an undirected graphical structure by means of an 'AND' rule, or an 'OR' rule, in the same way as for unpenalized nodewise regression models,

$$\hat{\mathcal{E}}_\lambda^{\text{AND}} = \{(i,j) : i \in \hat{\mathcal{N}_j}(\lambda) \text{ AND } j \in \hat{\mathcal{N}_i}(\lambda)\},$$
$$\hat{\mathcal{E}}_\lambda^{\text{OR}} = \{(i,j) : i \in \hat{\mathcal{N}_j}(\lambda) \text{ OR } j \in \hat{\mathcal{N}_i}(\lambda)\}.$$

For non-differentiable penalty functions, such as the ℓ_1 or ℓ_0 penalties, which are not differentiable at zero, Fan and Li [8] suggest a local quadratic approximation. This had lead Pircalabelu et al. [17] to use the following approximations to $\psi_\lambda(|\gamma_j - \gamma_{j0}|)$, $\psi'_\lambda(|\gamma_j - \gamma_{j0}|)$ and $\psi''_\lambda(|\gamma_j - \gamma_{j0}|)$, where γ_{japx} is a value close to $|\gamma_j - \gamma_{j0}|$,

$$\begin{aligned}
\psi_\lambda(|\gamma_j - \gamma_{j0}|) &\approx \psi_\lambda(\gamma_{j\,\text{apx}}) + \tfrac{1}{2}\frac{\psi'_\lambda(|\gamma_{j\text{apx}}|)}{|\gamma_{j\text{apx}}|}\left[(\gamma_j - \gamma_{j0})^2 - \gamma^2_{j\text{apx}}\right];\\
\psi'_\lambda(|\gamma_j - \gamma_{j0}|) &\approx \frac{\psi'_\lambda(|\gamma_{j\text{apx}}|)}{|\gamma_{j\text{apx}}|}(\gamma_j - \gamma_{j0});\\
\psi''_\lambda(|\gamma_j - \gamma_{j0}|) &\approx \frac{\psi'_\lambda(|\gamma_{j\text{apx}}|)}{|\gamma_{j\text{apx}}|}
\end{aligned}$$

The above quadratic approximations have been used on the one hand to 'ease' the optimization problem by making use of a relatively fast iterative procedure in order to obtain estimated coefficients. On the other hand, more importantly, they have been introduced to satisfy the existence of a second derivative at zero, needed in (8), which is not generally satisfied by most penalty functions. Working with non-differentiable expressions might lead to an alternative approach to obtain the MSE that avoids such approximations, however, this is not addressed here.

In the practical computations, the value $\gamma_{j\text{apx}}$ is arbitrarily at the start and is updated in an iterative Newton-Raphson scheme.

Often used examples of penalty function that can be used in (7) with these approximations include

- Lasso [20] with $\psi_\lambda^l(|\gamma_j - \gamma_{j0}|) = \lambda|\gamma_j - \gamma_{j0}|$;
- Bridge [10] with $\psi_\lambda^b(|\gamma_j - \gamma_{j0}|) = \lambda|\gamma_j - \gamma_{j0}|^\alpha$; $\alpha > 0$;
- Hard thresholding: $\psi_\lambda^h(|\gamma_j - \gamma_{j0}|) = \lambda^2 - (|\gamma_j - \gamma_{j0}| - \lambda)^2 I(|\gamma_j - \gamma_{j0}| < \lambda)$;
- Adaptive lasso [27] with $\psi_\lambda^{al}(|\gamma_j - \gamma_{j0}|) = \lambda w_j|\gamma_j - \gamma_{j0}|$ with w_j being a set of weights corresponding to each node in the graph;
- Smoothly clipped absolute deviation (SCAD, [8]) for which the first derivative is defined as
  $\psi_\lambda^{'s}(|\gamma_j - \gamma_{j0}|) = I(|\gamma_j - \gamma_{j0}| \leq \lambda) + \frac{(a\lambda - |\gamma_j - \gamma_{j0}|)_+}{(a-1)\lambda} I(|\gamma_j - \gamma_{j0}| > \lambda); a > 2.$

The nodewise MSE for the estimator for μ_j in model S can for penalized estimation be written as [17]

$$\begin{aligned} \text{MSE}(\hat{\mu}_{jS}) = \tau_0^2 + \omega^t Q_S^0 \omega + \omega^t \{(I_{p-1} - G_S)\delta\delta^t(I_{p-1} - G_S^t)\}\omega + \\ + \omega^t \{Q_S^0 cc^t (Q_S^0)^t - 2(I - G_S)\delta c^t (Q_S^0)^t\}\omega, \end{aligned} \tag{8}$$

where $c = n^{-1/2}\psi_\lambda^{''}(0)1_{p-1}$. Note that (8) reduces to (4) when there is no penalty, thus $\lambda = 0$. The FIC for penalized estimation results by inserting in (8) estimators obtained in the full model for unknowns, leading to

$$\begin{aligned} \text{FIC}(S, \mu_j; \lambda) = \hat{\tau}_0^2 + 2\hat{\omega}^t \hat{Q}_S^0 \hat{\omega} + n\hat{\omega}^t (I_{p-1} - \hat{G}_S)\hat{\gamma}_{j,w}\hat{\gamma}_{j,w}^t (I_{p-1} - \hat{G}_S)^t \hat{\omega} - \hat{\omega}^t \hat{Q}\hat{\omega} \\ + \hat{\omega}^t \{\hat{Q}_S^0 cc^t (\hat{Q}_S^0)^t - 2(I - \hat{G}_S)n^{1/2}\hat{\gamma}_{j,w} c^t (\hat{Q}_S^0)^t\}\hat{\omega}. \end{aligned} \tag{9}$$

Note that the value of the FIC depends on the choice of λ (which is contained in c). Since the above procedure is applied to each node, the modeling strategy allows thus different amounts of penalization at each node. In practice, for each node a value from a grid of λ values is chosen based on a three-fold cross-validation procedure on the deviance of the GLM.

5 Computational Aspects

While for small graphs only containing a small number of nodes it might be possible to investigate an all subsets search for each node, this is not feasible for moderate to large sized graphs.

For large graphs, at each node a penalized GLM based on all the other nodes is fitted from which one obtains immediately the penalized maximum likelihood estimator $(\hat{\theta}, \hat{\gamma})$ in the full model as well as the empirical Fisher information matrix J_n and the weights for the 'working-variables' once the Newton-Raphson algorithm converges. By allowing for a quadratic approximation of the penalty, the problem which was originally a convex problem, now enjoys first and second order

differentiability properties as well, making the optimization based on Newton-type methods easy to implement.

Once all the necessary quantities have been estimated from the full model, we start building collections of models S in an incremental fashion. We start first from an intercept-only model (for which the cardinality is 1) and compute its FIC score. In a second step, all models that include one extra neighbor (having thus cardinality 2) are compared to the benchmark model, namely the intercept model. The model with the lowest FIC score then becomes the new benchmark. If none of the models provides lower FIC values than the intercept model, the procedure stops. Otherwise, all models of cardinality 3 that include the benchmark model, are compared to the benchmark model. If any of these models, improves the FIC score we retain it and then search for models with higher cardinality, otherwise the procedure stops and outputs the model with the best attained score so far. Since this is a greedy local search algorithm and since the relation between the FIC scores and cardinality is non-linear, we do not restrict to making every time the hard decision of stopping the search if the score is not improved at each step, but test also some locally non-optimal models which at the next stage due to the inclusion of other neighbors, might offer a better FIC value than if we would have stopped at the best model from the previous stage.

6 Data Analysis

We here return to data examples stated in Sect. 2.

6.1 Dynamics of Psychopathology

Due to the binary recoding of the data, 7 of the 22 items in that PANAS resulted in having constant values (all zero, or all one) across the 93 days, these items have been excluded from the analysis. After this elimination we have treated each of the remaining items as a node in an undirected network, where the goal was to discover the edges that provide the lowest MSE for the logit(π_i) where π_i is the probability that item (or node) i indicates a tendency towards negative feelings.

The datasets for the two subjects were treated separately, in two distinct applications of the same FIC procedures. The observed sequence of emotions provides information on how a patient (or control) is doing. Therefore, we specified the focus point as being the observed sequence of emotions at each day. Afterwards, for each of the specified focuses we estimate a network and based on that network we estimated the basic reproduction number ρ, see Sect. 2.1. This resulted (due to missing observations) in having specified 90 different focuses and so 90 different networks (for each network a value of ρ is estimated) for the patient and 88 different focuses and networks for the control subject.

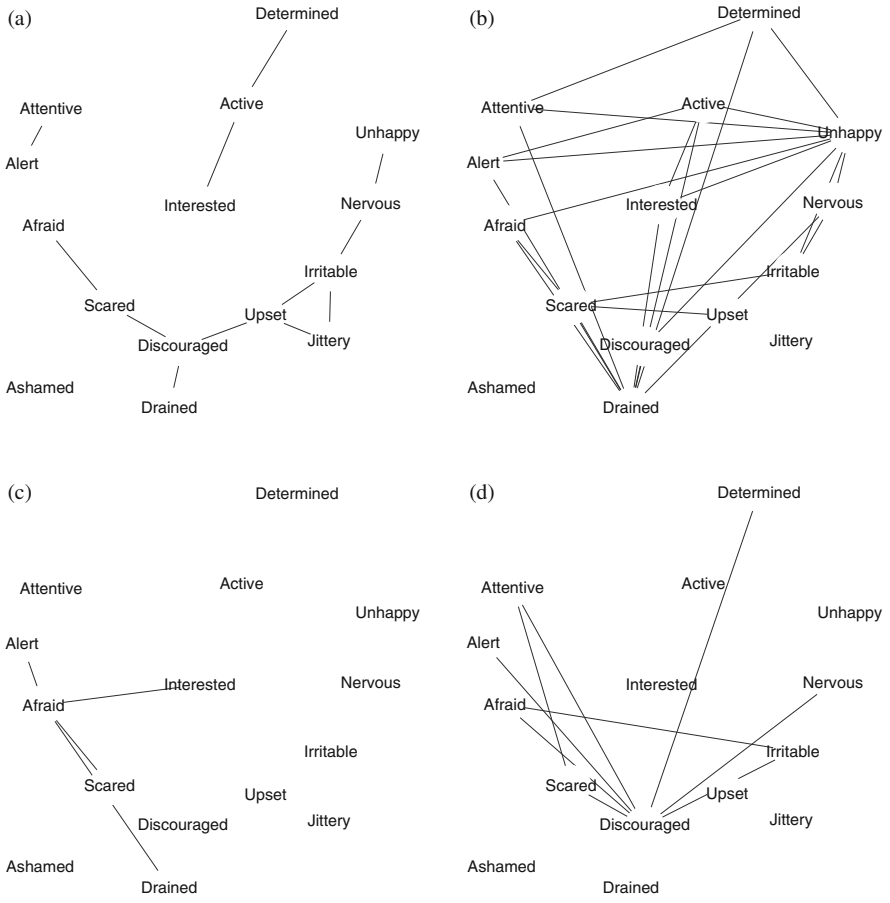


Fig. 1 PANAS data. Visual representation of the graphical structure estimated using a local quadratic approximation to an ℓ_1 penalty when the focus point is the sequence of emotions observed at time points 1 (panels **a**, **c**) and 70 (panels **b**, **d**) for the patient (panels **a**, **b**) and control (panels **c**, **d**). The corresponding estimated BRP rates are equal to 5.81 and 1.31 for the patient and 0.99 and 0.78 for the control. The *black* (resp. *gray*) colors reflect the positive (resp. negative) affect aspect of the node

The questions for which we want to find an answer can be formulated as follows: having the entire dataset of observations for the patient (likewise for the control), and assuming that tomorrow we observe a sequence of emotions that corresponds to what we have observed at time $t \in \{1, \ldots, 90\}$, what is the topology of the network which would generate a low MSE of the focus at each of the nodes? For example, Fig. 1 presents four estimated networks corresponding to the sequence of emotions that were observed for both the patient and the control, at time points 1 and 70. It is apparent that for the first time point in the estimated graphs for the patient there is a higher tendency to separate the negative affects from the positive ones, whereas

in the graphs estimated at the second time point there is a tendency to have a higher density of edges and to also positive and negative affects get linked with each other much more often blurring in a sense the separation between the two categories of feelings. As expected the network topology plays a crucial role, as for instance the four estimated ρ's based on these networks are quite different, especially for the patient which for these two focuses exhibits higher BRPs.

Since one can estimate thus a multitude of networks, each pertaining to the sequence of emotions observed on a particular day, one might also be interested in how 'stable' the patient tends to have a high BRP. Is this phenomenon stable across sequences (one for each day) of affects observed at each time point or was the above conclusion largely due to the effect of the particular focuses? To answer this question we have plotted the estimated BRPs for both subjects as a function of time. The results are presented in the upper row of Fig. 2 and it is apparent that the levels of the observed trends are almost always larger for the patient than for the control across many such observed sequences. Quite interestingly, this analysis seems to support the conclusions of the original authors concerning the fact that the patient exhibits higher BRP rates than the control, though coming from a conceptually different stand point with respect to estimating an unknown hidden undirected network.

A further investigation concentrates on 'confusing' the FIC procedure in the following sense. Up to now both the data used for the estimation as well as the focus would come from the same subject in a sense making the problem somewhat easier. As such we were interested in the discriminatory power of the procedure when the data came from the patient, but the focus point came from the control. This would correspond to the situation where based on the behavior seen so far, if for a brief moment the patient would exhibit normal behavior (situation summarized by the focus point), can he still be categorized as being patient based on the estimated ρ? Or vice versa, if based on what was observed so far, if a healthy subject exhibits for a moment a sequence of emotions similar to what the patient exhibited, do we still estimate networks for which ρ is relatively large? To answer this question we have proceeded as in the above application, but with the major difference that now the focuses are coming from what was observed for the other subject. The bottom row of Fig. 2 illustrates the findings and supports the conclusion that even though the FIC procedure estimated graphs that exhibited generally higher BRP ratios for the patient than for the control, it is still able to discriminate between the two subjects based on the proposed ratio, even when the focuses are probably not in line with the data used for estimation.

6.2 U.S. Voting Behavior

Since the vote is coded with a binary value, we fit at each node (i.e. Senator) a penalized logistic regression model with the vectors of votes for all remaining senators used as predictors. Based on this full model we construct the estimate $\hat{\omega}_w$ using the estimated vector of regression coefficients $\hat{\beta}$, corresponding to the

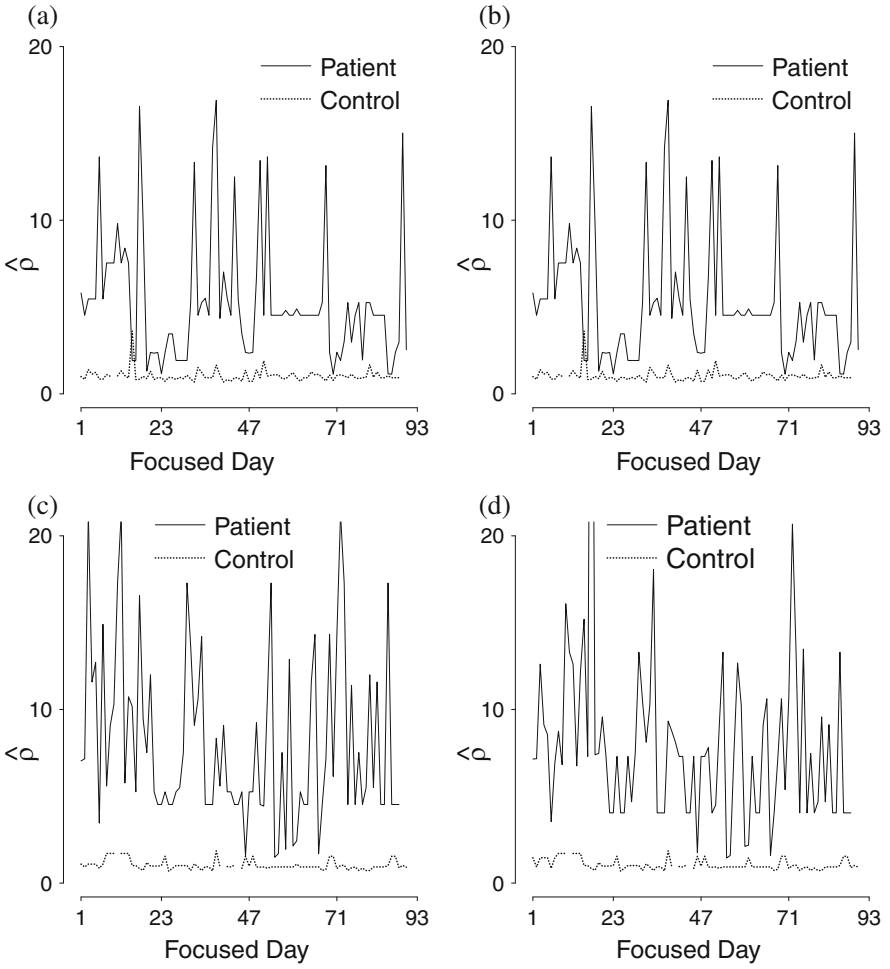


Fig. 2 PANAS data. Plotted is the $\hat{\rho}$ for each of the two subjects, when different observed sequence of emotions (corresponding to a day on the x-axis) are chosen as focus points. In the *upper row*, the focus points come from the same subject as the training data, whereas in the *second row* the focus points come from the emotion pattern displayed by the other subject. The local quadratic approximation to an ℓ_1 penalty (*left column*) and to a SCAD penalty (*right column*) have been used for estimating the corresponding networks for the two subjects

influence of each 'covariate' or 'parent' node on the probability of a 'Yes' vote for the dependent node. The intercept of the model, β_{i0}, acts as the protected parameter θ.

In order to fix notation, let $X_{ki} \sim \text{Bernoulli}(\pi_i)$ with

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_{i0} + \sum_{l \in \mathcal{V}\setminus i} \beta_{il} X_{kl} - \frac{1}{n} \sum_{l \in \mathcal{V}\setminus i} \psi_\lambda(|\beta_{il}|)$$

where X_{ki} denotes the result of a vote of Senator i on bill k and X_{kl}, for $l \in \mathcal{V} \setminus i$, represent the voting results for the remaining senators on the same bill. A constraining penalty function is placed on the vector of unknown β parameters. In the narrow model all coefficients corresponding to the unprotected nodes are set equal to zero, since in the narrow model none of them is included.

In a subsequent step we proceed as in Sect. 4.3. At the dependent node we select from the set of potential neighbors, the ones which minimize the MSE($\hat{\mu}_S$). Once the set of neighbors is selected for a particular node, we move to another node and proceed in the same fashion by estimating its set of neighbors. We perform the same procedure at each of the $p = 100$ nodes in the graph, obtain p sets of neighbors and afterwards combine all the information by drawing an edge between two nodes i and j if i belongs to the set of neighbors of j or vice versa. Notationally, this amounts to $(i, j) \in \mathcal{E}$ if $i \in \hat{\mathcal{N}}_j$ OR $j \in \hat{\mathcal{N}}_i$.

Figure 3 illustrates a few interesting patterns. First of all, it seems that the 'party vote' had a major role to play as most of the edges in the graph link two senators that belong to the same party. Second, within the Democratic party, the graph suggests that senators opposing the amendment are more likely to get linked to other democrats opposing the amendment than to the democrats in favor of the amendment. Lastly, the graph suggests also that there are some between-party edges, although they appear less frequently than the within-party edges.

Since at each node i, neighbors are selected on the basis that the model provides the lowest estimated MSE for logit(π_i) where π_i is the probability that bill i receives a favorable vote, one might be interested in the performance of such a classifier for the focus for which it was constructed. In this case this corresponds to predicting for each senator his vote on the bill. Based on the graph presented in Fig. 3 we estimate the correct vote for 78 % (or 84 % for SCAD) of the senators, whereas predicting based on average vote for all other bills (not incorporating any knowledge about the relations between senators) resulted in a correct prediction in only 46 % of the cases. These predictions are slightly optimistic since they are within sample, as the information which we are predicting is also used for constructing the graph.

6.3 Hunting Spider Species

Since the number of captured spiders is observed per location, we estimate an interactions network where connected nodes indicate that the two species are co-occurring. At each node a Poisson model is fitted where $X_{ki} \sim \text{Poisson}(\xi_i)$ with

$$\log(\xi_i) = \beta_{i0} + \sum_{l \in \mathcal{V} \setminus i} \beta_{il} X_{kl} - \frac{1}{n} \sum_{l \in \mathcal{V} \setminus i} \psi_\lambda(|\beta_{il}|)$$

where X_{ki} denotes the number of spiders at location i coming from species k and X_{kl}, for $l \in \mathcal{V} \setminus i$, represents the number of spiders at the remaining locations coming

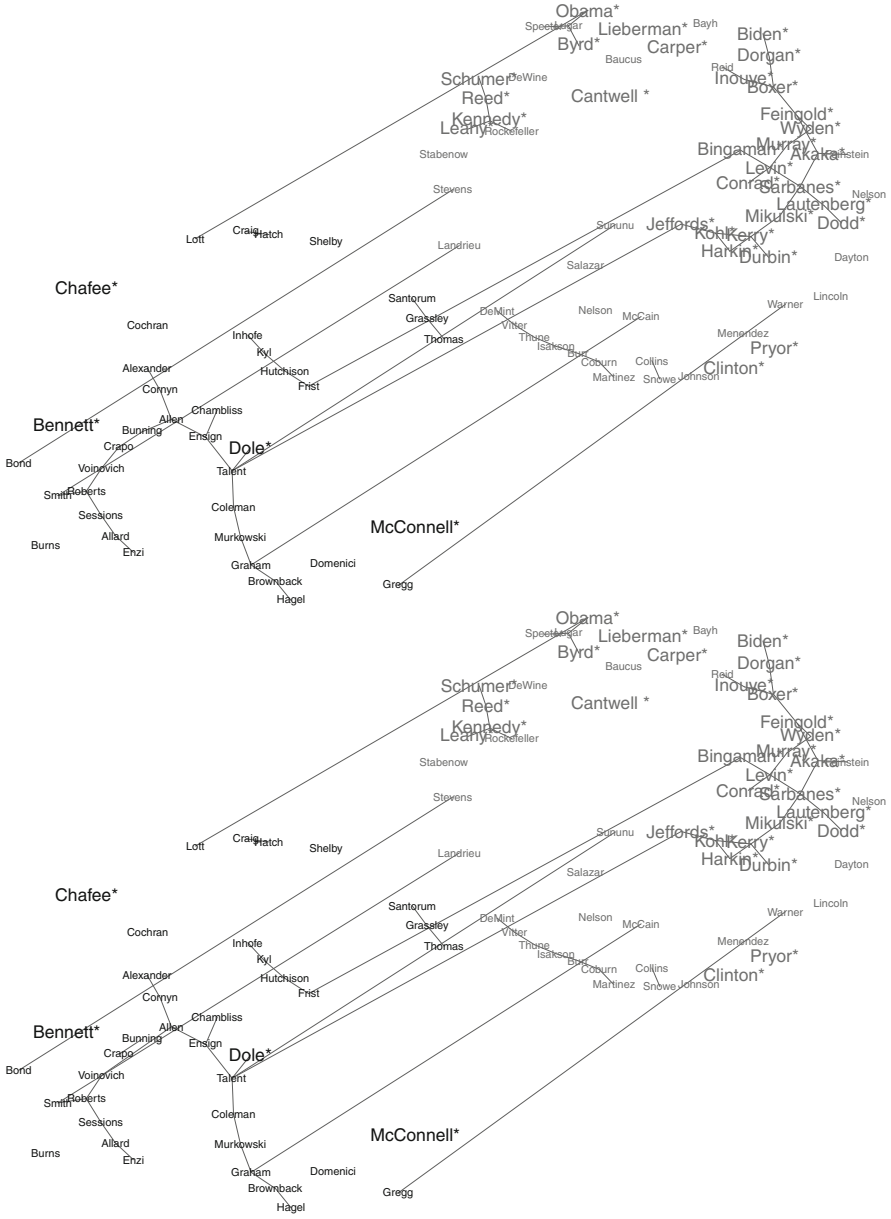


Fig. 3 Senate vote data. Visual representation of the graphical structure estimated using a local quadratic approximation to an ℓ_1 penalty (*top*) and to a SCAD penalty (*bottom*). In each figure, *black nodes* denote the Republican senators (*lower left quadrant*) and the *gray nodes* denote the Democrat senators (*upper right quadrant*). Within each party, the nodes accompanied by a $\star$ symbol denote senators that have opposed the amendment

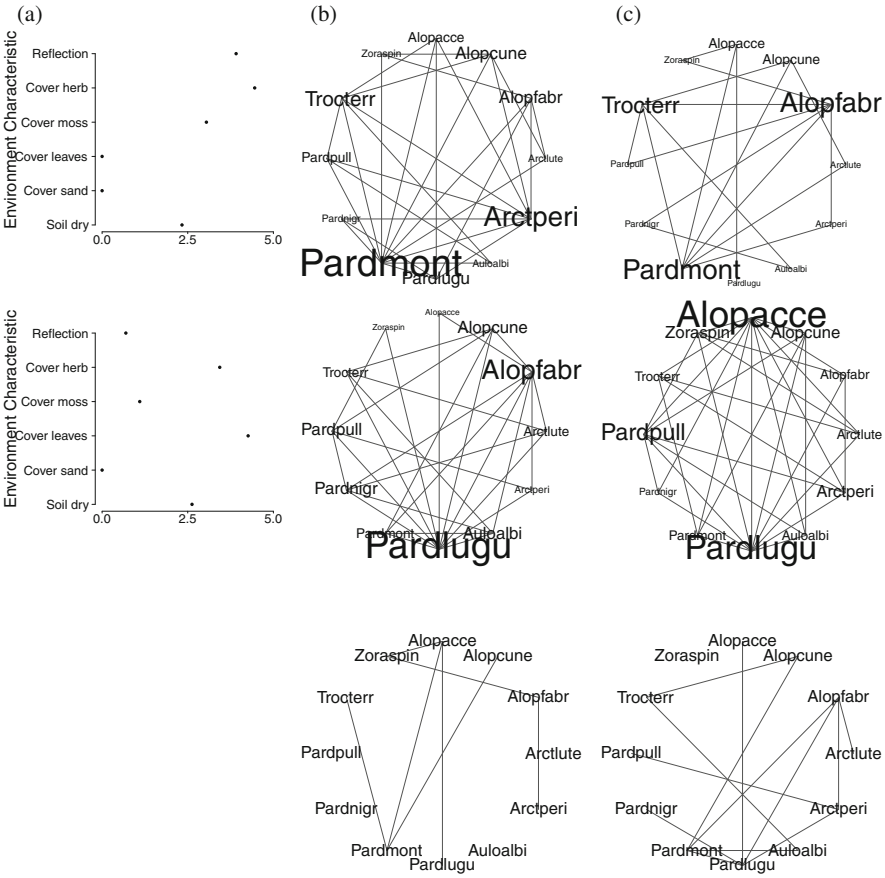


Fig. 4 Spider data. FIC estimated graphs (*first two lines*) based on two focuses (the corresponding environmental characteristics are shown in the *left column*). The ℓ_2, LQA Bridge denote the ψ function used in the estimation. Larger labels correspond to highly connected nodes and the *bottom line* presents the common edges across estimated graphs per focus. (**a**) Focus. (**b**) FIC LQA Bridge. (**c**) FIC ℓ_2

from the same species. A constraining penalty function is placed again on the vector of unknown β parameters, and in the narrow model all coefficients corresponding to the unprotected nodes are set equal to zero.

Figure 4 shows the estimated undirected graphs for two focuses when (i) a quadratic approximation of the Bridge penalty and (ii) an ℓ_2 penalty is used. The immediate conclusion is that the characteristics of the environment have an influence on the structure of the estimated networks as does the type of penalty that is being used. For the first focus both graphs suggest that the Pardmont species tends to co-occur most often with the other species, while for the second focus the Pardlugu species is the highly connected species. The ℓ_2 graph identifies for the second focus

also the Alopacce species as being highly connected to other species as well. Quite interesting is the fact that regardless of the environment conditions for the two cases studied the pairs Pardlugu-Alopacce, Pardmont-Allopcune and Alopfabr-Arctperi are present in all graphs and their abundance seems to be related.

7 Discussion

In this chapter we presented an extension of our method to construct graphs from the focused information criterion to generalized linear models. The three main advantages of using the FIC to construct graphs are: (i) a focus of interest can be defined incorporating prior knowledge of the system under investigation, (ii) the mean squared error of the focus is minimized which balances squared bias and variance of the estimator and increases generalizability, and (iii) the framework of local misspecification is used relaxing the assumption of having the correct model.

We showed that the combination of the GLM and FIC leads to an easily interpretable Fisher information matrix, separating the two types of parameters, ones that are always included and ones that are to be determined. This in turn was seen to lead to an estimate of the mean squared error that is used to determine the FIC.

The three examples shown in this chapter indicate the richness of the method. In the first example data from a bipolar patient and a control were contrasted suggesting different patterns of predictions for whether symptoms of bipolar disorder would remain or not. Especially interesting was the fact that using a sequence of emotional items (knowledge of the system), the basic reproduction number ρ, resulting from the estimated graph, was seen to vary strongly in the patient but not the control. And even using an emotional item sequence of the control in the bipolar patient resulted in a largely varying pattern of values of ρ. These results show that a network of emotional states which influence each other can be obtained, from which behavior of symptoms can be predicted.

The second example using data from the voting behavior of U.S. senators showed that for the 'Flag Desecration' amendment predicting voting behavior using estimated relations between senators may result in higher accuracy than using previous voting behavior of senators, while in the same time discovering that intra-party cooperation is dominant (the voting pattern of a senator can best be described by patterns of colleagues from the same party), the cluster of opposing democrats stands out in this respect, but also that cross-party cooperations is not negligible.

The third examples uses Poisson distributions to model counts of different species of spiders at different locations.

Because of the extensions to the more general exponential family of distributions enlarge the range of applicability of this procedure to the more realistic situations where one has at disposal binary or count data, the estimation of connections is not limited just to Gaussian data.

In conclusion, there are many possibilities of using the focused information criterion to obtain meaningful graphs of many kinds of systems, as shown by the

examples presented here which showed that the presented FIC procedure can be useful for estimating graph structures by taking the researcher's objectives more closely into account and outputting a model that comes closer to his goals. Since we can easily incorporate knowledge of a system through the focus, the method is flexible and useful.

Acknowledgements The authors wish to thank Prof. J.-H. Kamphuis for the PANAS data. The authors acknowledge the support of the Fund for Scientific Research Flanders, KU Leuven grant GOA/12/14 and of the IAP Research Network P7/06 of the Belgian Science Policy.

References

1. Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research, 9*, 485–516.
2. Borsboom, D., Cramer, A. O. J., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The small world of psychopathology. *PLoS ONE, 6*(11), e27407.
3. Claeskens, G., & Hjort, N. (2003). The focused information criterion. *Journal of the American Statistical Association, 98*, 900–916. With discussion and a rejoinder by the authors.
4. Claeskens, G., & Hjort, N. (2008). Minimising average risk in regression models. *Econometric Theory, 24*, 493–527.
5. Claeskens, G., & Hjort, N. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.
6. Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation*. London: Chapman & Hall.
7. Dempster, A. (1972). Covariance selection. *Biometrics, 28*(1), 157–175.
8. Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.
9. Fiocco, M., & van Zwet, W. (2004). Maximum likelihood estimation for the contact process. *Institute of Mathematical Statistics Lecture Notes - Monograph Series, 45*, 309–318.
10. Frank, I. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics, 35*(2), 109–135.
11. Jalali, A., Ravikumar, P., Vasuki, V., & Sanghavi, S. (2010). On learning discrete graphical models using group-sparse regularization. In *Proceedings of the 13th international conference on artificial intelligence and statistics*, Chia Laguna Resort, Sardinia, Italy.
12. Lauritzen, S. (1996). *Graphical models*. New York: Oxford University Press.
13. Lee, J., & Hastie, T. (2012). Learning mixed graphical models. Preprint arXiv:1205.5012v3.
14. Loh, P. L., & Wainwright, M. J. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics, 41(6)*, 3022–3049.
15. McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. London: Chapman & Hall.
16. Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics, 34*(3), 1436–1462.
17. Pircalabelu, E., Claeskens, G., Jahfari, S., & Waldorp, L. (2013). *Focused information criterion for graphical models. Large p, small n considerations* (Technical report). KBI, Faculty of Economics and Business, KU Leuven.
18. Pircalabelu, E., Claeskens, G., & Waldorp, L. (2012). *Structure learning using a focused information criterion in graphical models* (Technical report). KBI, Faculty of Economics and Business, KU Leuven.

19. Schmidt, M., Niculescu-Mizil, A., & Murphy, K. (2007). Learning graphical model structure using ℓ_1-regularization paths. In *Proceedings of the 22nd national conference on artificial intelligence*, Vancouver, Canada (Vol. 2, pp. 1278–1283).
20. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (SERIES B), 58*, 267–288.
21. van Borkulo, C. D., Kamphuis, J. H., & Waldorp, L. J. (2013). *Predicting behaviour of networks of mental disorders: The contact process as a model for dynamics of psychopathology* (Technical report). University of Amsterdam.
22. van der Aart, P. J. M., & Smeenk-Enserink, N. (1975). Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology, 25*, 1–45.
23. Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning, 1*(1–2), 1–305.
24. Wainwright, M. J., Ravikumar, P., & Lafferty, J. D. (2007). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems*, 20 (NIPS 2007), (Vol. 19, pp. 1465–1472). Cambridge: MIT Press.
25. Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070.
26. Yang, E., Ravikumar, P. K., Allen, G. I., & Liu, Z. (2012). Graphical models via generalized linear models. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems*, Curran Associates, Inc. (Vol. 25, pp. 1367–1375).
27. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Fully Nonparametric Short Term Forecasting Electricity Consumption

Pierre-André Cornillon, Nick Hengartner, Vincent Lefieux, and Eric Matzner-Løber

Abstract Electricity Transmission System Operators (TSO) are responsible for operating, maintaining and developing the high and extra high voltage network. They guarantee the reliability and proper operation of the power network. Anticipating electricity demand helps to guarantee the balance between generation and consumption at all times, and directly influences the reliability of the power system. In this paper, we focus on predicting short term electricity consumption in France. Several competitors such as iterative bias reduction, functional nonparametric model or non-linear additive autoregressive approach are compared to the actual SARIMA method. Our results show that iterative bias reduction approach outperforms all competitors both on Mean Absolute Percentage Error and on the percentage of forecast errors higher than 2,000 MW.

1 Introduction

Electricity Transmission System Operators (TSO) shall ensure the balance of electricity flows on the network at all times, as well as the reliability, safety and efficiency of the network, taking into account the technical constraints to which it is subject. The daily coordination is facilitated by having short term demand-supply balance predictions on a day ahead horizon, whereas longer term predictions are

P.-A. Cornillon (✉)
University Rennes 2, Rennes, France
e-mail: pierre-andre.cornillon@uhb.fr

N. Hengartner
Los Alamos National Laboratory, Los Alamos, NM, USA
e-mail: nickh@lanl.gov

V. Lefieux
RTE-EPT & UPMC-ISUP, Paris, France
e-mail: vincent.lefieux@rte-france.com

E. Matzner-Løber
University Rennes 2 & Agrocampus Ouest, Rennes, France
e-mail: eml@uhb.fr

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_5

useful for network safety studies, network maintenance and planning. In this paper, we focus on predicting short term electricity consumption in France, where Réseau de Transport d'Electricité (RTE) is the unique TSO.

Each day, RTE is given by the producers the electricity production plans for the following day. While informative, these predictions are attached with uncertainties. To better manage its network, RTE drafts its own load curve forecasting that takes into account historical consumption patterns, weather forecast and daily pricing information. The French national dispatchers make the final load balancing decisions, taking into account the most recent information, including unexpected modifications of consumption patterns, as for example: strikes, national sporting events and even weather conditions (like heavy snow fall) which can impact how and where electricity is consumed. RTE draws up its global load forecast for France and calculates the values of the reserves required to cover the two types of contingencies: load contingencies and generation contingencies. With regards to the latter, generation facilities may be affected, as regards to their operation, by a number of fortuitous events and/or limitations giving rise, in real time, to the unscheduled unavailability of a certain volume of generation. The primary, secondary and tertiary controls in real time allow to manage the supply-demand balance, by using reserves set aside for this purpose. RTE evaluates the reserves that are effectively available and, if they are insufficient, performs adjustments on the generation facilities. This is the main reason why RTE must avoid large forecasting errors. The evaluation of a model is based at the same time on its average quality and its ability not to generate large errors. In this paper we consider a threshold of 2,000 MW based on feedback from RTE.

Modifying forecasting tools for a TSO is very sensitive and costly, and acting such modifications should rely on strong evidences that the new proposed model is better in terms of forecasts (accuracy, robustness) but also in term of maintaining and computational time. Strong relationships between utility industry and academic researchers are encouraged in France and forecasting models for the electricity consumption in France have been continuously developed over the last 30 years. Many models and approaches have been considered. Poggi [19] used nonparametric kernel estimators for prediction, [16] used Kalman filters and more recently [14] developed a semi-parametric model, [10] aggregated different predictors and [4] proposed a robust SARIMA model. Additive models are also widely used in this context as explained in [18] or [9]. Considering that the discretization of the daily consumption curve is dense enough, [8, 23] or [2] used linear functional forecasting methods.

The paper provides the feedback from RTE with a new modeling paradigm. In Sect. 2, we discuss the French electricity consumption data and the previous operational forecasting model. Section 3 presents iterative bias reduction (IBR) procedure together with the practical. The results are presented and compared to the current state-of-the-art modeling at RTE and to various competitors in Sect. 4. Finally, in Section 5, we provide a discussion and conclusions about our approach.

2 RTE Current Forecasting Model

A cursory look at the electricity demand curve in Fig. 1 shows both weekly and seasonal variations. Such data are available on RTE webpage (www.rte-france.com). Annual consumption patterns are usually explained by seasonal change in climate (temperature, cloud cover) and daylight duration. We refer to [22] for a comprehensive analysis of seasonal patterns of electricity loads over 10 European countries. The French global consumption increases in winter due to the relatively low price of electricity (compared to other energy sources), which has promoted the use of electrical heating since the seventies. As a measure of penetration of the electrical heating in France, it is estimated that at 7 pm in winter, the decrease in temperature of 1 °C increases the demand of 2,400 MW. These last years, the French global consumption also increased in summer because of the air-conditioning.

Actually, in order to take into account the different tariffs and the climate (temperature and cloud cover measured at 32 weather stations), RTE uses a complex nonlinear parametric regression model to "correct" the half hourly load curve. This model has around 1,000 coefficients which are estimated every March and September using the true climatic data. Subtracting the climate and tariff dependent part of the load curve estimated by the regression, the load curve (Fig. 1) becomes the "corrected load curve" presented in Fig. 2.

The corrected consumption (denoted from now on by Z_t) shows traditional shape of weekly load diagram that can not be discerned on Fig. 2 because the number of data shown but which can be readily seen in the excerpt (during year 2006) shown in Fig. 3. A typical week presents a similar form for each day, with a first peak in the morning and a second one in the evening; the levels of Saturday and Sunday are lower than other days.

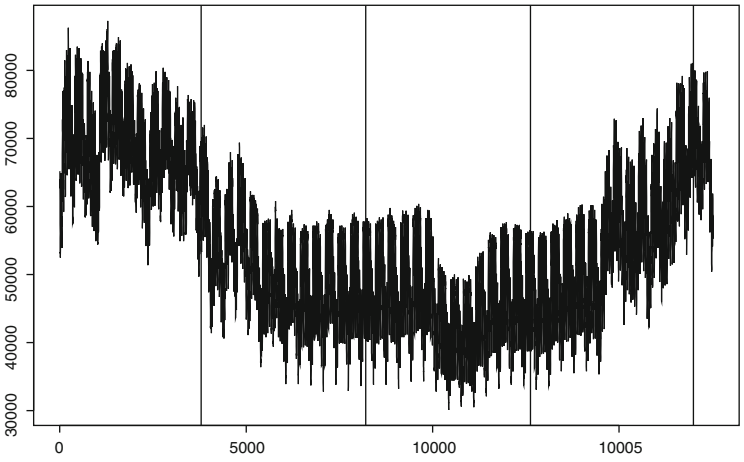


Fig. 1 One year of French consumption measured every half hour (from January to December)

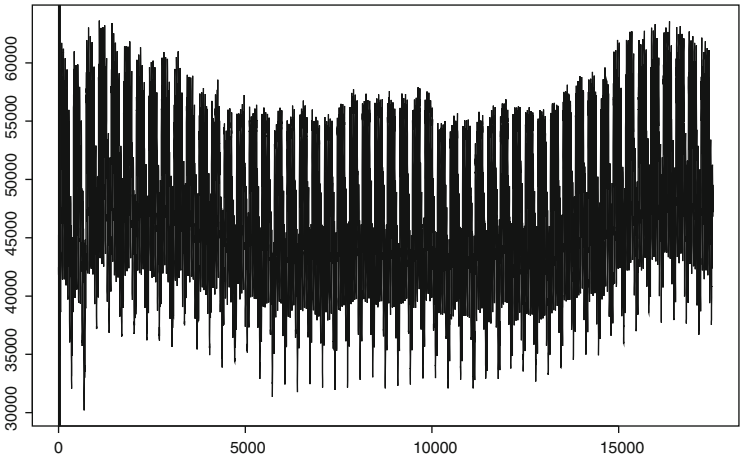


Fig. 2 Corrected load curve (corrected French consumption)

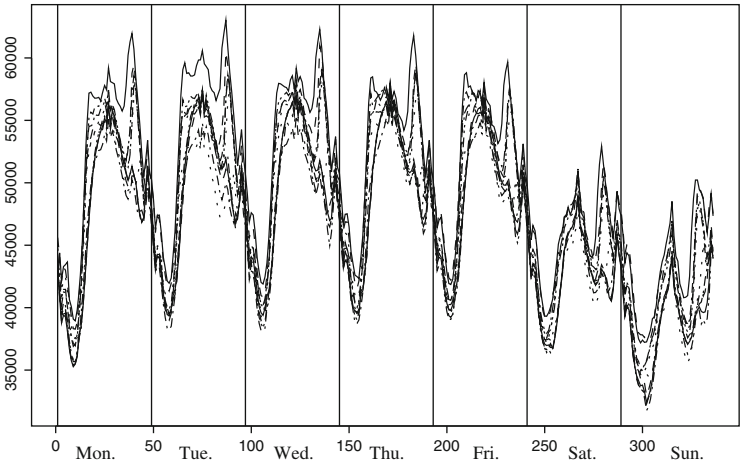


Fig. 3 Typical weeks of corrected consumption (excerpt from 2006). Each line type indicates a different week

In a second step, RTE uses a SARIMA model on the "corrected" time series of the form:

$$(1 - \varphi B)\, \nabla_{48} \nabla_{336} Z_t = \left(1 - \theta_{48} B^{48}\right) \left(1 - \theta_{336} B^{336}\right) \eta_t,$$

where B and ∇ are the classic lag operator $B^i Z_t = Z_{t-i}$ and $\nabla_i = 1 - B^i$. The series values are observed every 30 min. The operator ∇_{48} corresponds to a daily differentiation and ∇_{336} to a weekly differentiation operator. The parameters φ, θ_{48} and θ_{336}, are estimated using the procedures defined by Box and Jenkins.

In summary, RTE uses a three steps procedure to do the short-term forecast:

Step 1 RTE "corrects" the half hourly load curve by modeling the impact of climate and prices, in order to work on a time series that doesn't depend on exogenous variables. This first step is done by using a regression model with dependent variables based on climate and tariff. We denote the corrected series by Z_t.

Step 2 RTE uses a SARIMA model to forecast Z_t at the horizon H: $\hat{Z}_{t+H}$.

Step 3 RTE adds the forecasts given in Step 2 with the estimation given by the regression model using prices and forecasts for the temperature and cloud cover.

Even if operationally the results were correct, the complexity of the model does not allow incorporation (or test) of a new variable within a production framework. This kind of operation must be conducted by a senior statistician with an computer engineer in order to make the good assumptions and to program them accordingly. In order to improve the current methodology, RTE has chosen two approaches. The first one consists of improving the SARIMA modeling: with a daily or a robust modeling [4]. The second one consists in replacing the SARIMA modeling by the nonparametric approach that will be presented in the following section.

3 A New Nonparametric Forecasting Tool

The classical approach for forecasting a univariate time series $\{Z_t\}_{1\leq t\leq T}$ is to postulate a parametric model, estimate its coefficients, and compute the forecasts. The most popular model is the autoregressive (AR) model of order p

$$Z_t = \sum_{i=1}^{p} \alpha_i Z_{t-i} + \varepsilon_t, \tag{1}$$

where $\{\alpha_i\}$ are the unknown parameters of the model and ε_t is white noise. Model (1) represents the current state of Z_t through its immediate past p values. In practice, the sample mean from the data is subtracted before fitting, which allows to ensure $\mathbb{E}(Z_t) = 0$.

The limitation of that type of linear models is well known and many forms of nonlinear models have been explored since the 1980s. The developments in nonparametric regression provide flexible techniques for modeling time series through the following expression:

$$Z_t = f(Z_{t-1}, \ldots, Z_{t-p}) + \varepsilon_t.$$

However, when p is large, the nonparametric approach suffers from the "curse of dimensionality" and a natural simplification is the nonlinear additive autoregressive models which assume that the unknown function f from $\mathbb{R}^p$ to $\mathbb{R}$ could be written as

a sum of univariate functions:

$$Z_t = f_1(Z_{t-1}) + \ldots + f_p(Z_{t-p}) + \varepsilon_t.$$

Additive models are very useful for approximating the high-dimensional autoregressive function $f(.)$ given above. Their extensions have become one of the widely used nonparametric techniques and references on their applications to load curves can be found in [9, 17, 18] for example. However despite their interpretability and their small estimation error, additive models could have a large approximation error (and by the way a large prediction mean square error) if the underlying function is not additive. In recent papers, [5, 6] propose a practical adaptive nonparametric regression method with good results. A R package is also available on CRAN. For sake of completeness we summarize this method below.

3.1 Iterative Bias Reduction

Suppose the data $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ are related via the following regression model

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where the errors $\{\varepsilon_i\}$ are independent of all the covariates $(X_1, \ldots, X_n)$. It is helpful to rewrite Eq. (2) in vector form. Let us denote $Y = (Y_1, \ldots, Y_n)^t$, the column vector of observations of the dependent variable, $m = (m(X_1), \ldots, m(X_n))^t$ the column vector of unknown function m at data points and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$ the column vector of errors (where ^{t} denotes matrix transposition) to get

$$Y = m + \varepsilon.$$

Linear smoothers typically depend on a tuning parameter, denoted by λ, that governs the trade-off between the smoothness of the estimate and the goodness-of-fit of the smoother to the data. It controls the effective size of the local neighborhood of the explanatory variables over which the responses are averaged. For example, the tuning parameter λ is the bandwidth for kernel smoother. Linear smoothers can be written as

$$\hat{m} = S_\lambda(X)Y,$$

where $S_\lambda(X)$ is an $n \times n$ smoothing matrix and $\hat{m} = \hat{Y} = \left(\hat{Y}_1, \ldots, \hat{Y}_n\right)^t$ denotes the vector of fitted values. We parametrize the smoothing matrix so that large values of λ will produce very smooth curves while small λ will produce a more wiggly curve that tends to interpolate the data. Much has been written about how to select "optimally" the tuning parameter, see for instance [12, 13, 15, 20]. The

IBR procedure [5] does not try to select an optimal λ, instead a large λ is chosen, resulting in a (pilot) smoother with a substantial bias. In the initial step, this pilot smoother is applied to the data (oversmoothing the data) leading to

$$\hat{m}_1 = S_\lambda(X)Y.$$

The second step consists in estimating the bias $\mathbb{E}(\hat{m}_1|X) - m = (S_\lambda(X) - I)m$ by replacing m by a smooth linear estimate (possibly using the same pilot smoother):

$$(S_\lambda(X) - I)S_\lambda(X)Y = S_\lambda(X)(S_\lambda(X) - I)Y = S_\lambda(X)(\hat{m}_1 - Y)$$

and correcting the initial estimator by removing the estimated bias:

$$\hat{m}_2 = \hat{m}_1 + S_\lambda(X)(Y - \hat{m}_1) = [S_\lambda(X) + S_\lambda(X)(I - S_\lambda(X))]Y.$$

The bias estimation and bias correction steps can be iterated to generate a sequence of bias corrected smoothers, which gives at step k

$$\begin{aligned}\hat{m}_k &= S_\lambda(X)Y + S_\lambda(X)(I - S_\lambda(X))Y + \ldots + S_\lambda(X)(I - S_\lambda(X))\ldots(I - S_\lambda(X))Y \\ &= [I - (I - S_\lambda(X))^k]Y. \end{aligned} \tag{3}$$

This closed form is interesting and shows that the qualitative behavior of the iterated estimator is governed by the spectrum of $I - S$. For splines smoothers and kernel smoothers with a positive define kernel, the spectrum lies in the unit interval $[0, 1]$ [5]. It follows that as the number of iterations k goes to infinity, the sequence of iterated smoothers $\hat{m}_k$ tends to reproduce the raw data Y. Thus, iterating the algorithm until convergence is not desirable. However, since each iteration reduces the bias and increases the variance, a few iterations of the algorithm will often produce a better smoother than the pilot smoother. This brings up the important question of how to decide when to stop the iterative bias correction process. Viewing the latter question as a model selection problem suggests stopping rules based on Akaike Information Criterion, AIC, [1], Bayesian Information Criterion, BIC, [21], and Generalized Cross Validation, GCV, [7]. These selectors, all implemented in the **ibr** package. Theoretical results [5] and simulations advocated for GCV defined here by

$$GCV(k) = \log \widehat{\sigma_k}^2 - 2\log\left(1 - \frac{\mathrm{tr}([I - (I - S_\lambda(X))^k])}{n}\right),$$

where $\widehat{\sigma_k}^2$ corresponds to the estimated variance of the current residuals at step k and $\mathrm{tr}(A)$ denotes the trace of the matrix A.

Strongly related to the number of iterations is the smoothness of the pilot smoother. One has to be sure that the pilot smoother oversmooths. We choose one bandwidth for each explanatory variable X_i. This choice is done such as the effective

degree of freedom for the one-dimensional smoothing matrix related to X_i has a trace equal to a given number chosen by the user (by default that value is just bigger than one – for instance 1.1 or 1.01 – so the initial pilot smoother has less equivalent degree of freedom than a linear model). For such values, the pilot smoothers always oversmooths.

The linear smoother defined by (3) predicts the conditional expectation of responses only at the design points. It is useful to extend regression smoothers to enable predictions at arbitrary locations $x \in \mathbb{R}^d$ of the covariates. Rewrite as follows

$$\begin{aligned}\hat{m}_k &= S_\lambda(X)[Y + (I - S_\lambda(X))Y + \ldots + (I - S_\lambda(X)) \ldots (I - S_\lambda(X))Y] \\ &:= S_\lambda(X)\hat{\beta}_k.\end{aligned}$$

This formulation allows an evaluation of the estimator at any point x via

$$\hat{m}_k(x) = S_\lambda(x)\hat{\beta}_k.$$

Such an extension allows us to use IBR in forecasting settings.

3.2 Practical Implementation

We consider the following RTE case-study: data are known half hourly until 12 am, and we want to predict the 48 values of the consumption of the following day. Having $(Z_1, \cdots, Z_T)$ (Z_T is the data obtained at 12 am) we want to predict $Z_{T+25}, \cdots, Z_{T+72}$. In order to consider the relation between Z_{T+25} and the available data, let us consider the model:

$$Z_{(T-48i)+25} = f(Z_{T-48i}, \ldots, Z_{T-48i-p}) + \varepsilon \quad i > 0$$

where the lagged variable p will be discuss later. Once the unknown function f is estimated, the predicted value is obtained by

$$\hat{Z}_{T+25} = \hat{f}(Z_T, \ldots, Z_{T-p}).$$

As classically done in nonparametric forecasting methods, to forecast at different horizons, we might consider different models. For any $H \in [25, \cdots, 72]$, we have

$$Z_{(T-48i)+H} = f_H(Z_{T-48i}, \ldots, Z_{T-48i-p}) + \varepsilon_H \quad i > 0. \tag{4}$$

where again T corresponds to the point 12 am and p is the memory. The size of p will be discussed later on, but if p is bigger than 24, part of the data used correspond to the previous day. In order to keep the modeling as simple as possible, we decide

to use the same p for all the different models and estimate the unknown function f_H. Using the same notation as in the previous section, we consider

$$X = \begin{pmatrix} Z_{T-48} & \cdots & Z_{T-48-p} \\ Z_{T-96} & \cdots & Z_{T-96-p} \\ \vdots & \vdots & \vdots \end{pmatrix} \qquad Y_H = \begin{pmatrix} Z_{T-48+H} \\ Z_{T-96+H} \\ \vdots \end{pmatrix}$$

The forecast model in vector form can be written (for a given horizon H) as

$$Y_H = f_H(X) + \varepsilon_H.$$

We can apply the IBR procedure to estimate this model. This is done using the R package **ibr** (available on CRAN) with the default value: a Gaussian kernel smoother with the smoothing parameter (bandwidth) $\lambda = (\lambda_1, \ldots, \lambda_{p+1})$ chosen such as each univariate Gaussian (product) kernel smoother have a degree of freedom (the trace of the smoother) equal to 1.01. This last value is chosen to ensure that the pilot smoother $S_\lambda(X)$ is smooth enough to get bias at the beginning of the procedure. This can be checked easily by looking at the number of iterations selected by GCV which have to be at least a few hundred. This number is chosen once and is always the same for each horizon and each modeling. No tuning are required since the modeling is robust to this choice [6]. Repeating the IBR procedure for each horizon, we get the $\hat{f}_{25}, \cdots, \hat{f}_{72}$ and the 48 predictions are obtained immediately calculating

$$\hat{f}_H(Z_T, \ldots, Z_{T-p})$$

and every day, we update X and Y_H with the new data.

The different days of the week have different patterns (Fig. 3). It seems reasonable to take into account this information and to build forecasting models according to the type of day. We tested two approaches here.

- The first approach consists in adding a new explanatory variable D which corresponds to the forecast day type $\{Monday, \ldots, Sunday\}$ and we estimate the more general model

$$Y_H = f_{day,H}(X) \times \{D = day\} + \varepsilon_{day,H}$$

  So for forecasting a given day in $\{Monday, \ldots, Sunday\}$, we calculate

$$\hat{Z}_{t+H} = \hat{f}_{day,H}(Z_T, \ldots, Z_{T-p}). \tag{5}$$

  This means that we estimate $7 \times 48 = 336$ models. By doing so, we decrease the size of the database by gathering the data by day.
- The second approach splits the week into working days, Saturday and Sunday (see [2]). In order to keep learning database with the same pattern, when the

autoregression model used only implies the previous day, instead of using seven different scenarios as done in the previous approach, we only use four: Monday, Tuesday-Friday, Saturday and Sunday. When the autoregression model used implies the two previous days, we use five scenarios Monday, Tuesday, Wednesday-Friday, Saturday and Sunday.

Now, let discuss the size of the explanatory variables. Previous analysis conduct us to take in memory the 2 previous days so 96 lagged variables. Such model will be denoted in the following IBR96. In order to reduce the number of variables but still keeping 2 days of history, we could work with hourly data and use only 48 lagged variables. The corresponding model will be denoted IBR48h. If one wishes to use only one day of history, one could run IBR48 (built on measurements every half hour) or IBR24h (built on measurements every hour).

4 Results

Given data measured every half hour from January 2006 to December 2008, we are able to forecast year 2009. Several forecasting methods have been considered:

1. The RTE SARIMA method (the current RTE method),
2. Our 4 IBR models: IBR96, IBR48h, IBR48 and IBR24h (see previous section),
3. The simplest type of aggregative estimator [10] where we average all the five previous predictions (Agg.)
4. Functional Nonparametric approach (NP) described in [2] with the same choices: epanechnikov kernel, bandwidth selection adapted from [3] and L^2 distance,
5. The naive approach which uses the previous day (of the database) as a prediction,
6. Multivariate Adaptive Regression Splines (MARS) estimated with **mda** package [11],
7. Nonlinear Additive Autoregressive models (NAAR) of order p estimated with **mgcv** package [24]. The order p is chosen using a forward selection with AIC criterion based on the past values. The variables are the lagged variable (built on measurements every half hour),
8. Nonlinear Additive Autoregressive models (NAARh) of order p estimated with **mgcv** package [24]. The order p is chosen using a forward selection with AIC criterion on the past values. The variables are the lagged variable (built on measurements every hour).

Forecasting specific days or periods are beyond the scope of that paper and RTE decides to delete 32 days from the year 2009: National holidays, Christmas period. Thus, all the errors are calculated on 333 days. We compute two types of errors, the first one on the corrected load curve and the second one on the global French load curve where the correction belonging to the climatic and tariff models are added to

the forecasts. We compute the Mean Absolute Percentage Error (MAPE) over 2009

$$MAPE = \frac{1}{333 \times 48} \sum_{t=1}^{333} \sum_{H=25}^{72} \frac{|Z_{T+H} - \hat{Z}_{T+H}|}{Z_{T+H}} \times 100.$$

The TSO is also interested in considering large forecasting errors (in absolute value). To have an insight on these kind of errors, we compute the percentage of absolute errors greater than 2,000 MW among all the forecasts.

$$PE = \frac{1}{333 \times 48} \sum_{t=1}^{333} \sum_{H=25}^{72} \mathbb{1}\{|Z_{T+H} - \hat{Z}_{T+H}| > 2{,}000\} \times 100,$$

where $\mathbb{1}$ is the indicator function.

We summarize in Table 1 the forecast results on (i) the corrected time series (denoted Cor.) and on (ii) the French load curve (denoted Glob.). Forecastings for the last one are obtained by adding to the forecasts of the corrected time series, the predicted values given by the complex nonlinear parametric regression model (used to take into account the different tariffs and the climate) used with the true values of the exogenous variables. These *ex post* forecasts only take into account the errors due to the model and not those due to the temperature forecasts for example. We only present here the results obtained by using a different model for each day. Using only three models (working day, Saturday and Sunday) are for all the methods proposed here much less accurate.

Working with the first approach (one model for each day of the week) leads to better results than the second approach with aggregated working days. This remark is valid for all modeling type: IBR, non-parametric approach, nonlinear additive autoregressive models. Thus, even if the size of the sample decreases the accuracy proposed by each modeling is better. As the second approach is not well suited to our problem, we will focus on the first one in In what follows.

It is worth noting that independently of the size of integrated lagged variables IBR provides better results than the SARIMA model and also better than other competitors. Integrating two days of history on a half an hour basis, which had sense for practitioner, give the best results for IBR. Integrating one day on a half an hour basis or two days on a hour basis give approximately the same results which are better than one day on a hour basis. The same is true for the corrected series, the global series and the number of forecast errors greater than 2,000 MW. The nonlinear additive autoregressive models perform better when the variables are chosen on an hour basis (compared to half an hour basis). This can be explained by the fact that the selected p is usually in the same order of magnitude for NAAR and NAARh and thus the NAARh can take into account more elements from the past (by dropping the half hourly information).

It is interesting to note that the classical mean of the five first modeling (Agg.) is doing well and this is a possible way for the future, although the forecaster has to maintain numerous models. In the following, we will only compare SARIMA

Table 1 Percentage errors (MAPE) for year 2009 on the corrected series (Cor.) and on the global one (Glob.), and percentage of forecast errors greater than 2,000 MW (PE)

Modeling by day											
Method	SARIMA	IBR96	IBR48	IBR48h	IBR24h	Agg.	NP	Naive	MARS	NAAR	NAARh
MAPE (glob.)	1.12	0.98	1.00	1.00	1.05	0.92	1.35	1.39	1.61	1.47	1.24
MAPE (cor.)	1.34	1.16	1.18	1.19	1.24	1.09	1.65	1.67	1.94	1.71	1.45
PE	4.53	2.79	3.14	3.18	3.50	2.76	7.99	8.16	9.73	7.45	5.38
Modeling by type of day (working days-Saturday-Sunday)											
MAPE (glob.)	–	1.96	2.26	1.89	2.49	1.62	1.63	4.27	2.60	2.82	2.78
MAPE (cor.)	–	2.29	2.62	2.17	2.88	1.87	1.97	4.83	3.07	3.19	3.16
PE	–	14.51	19.72	13.77	20.89	11.21	12.48	29.47	21.81	24.81	24.00

Table 2 Daily percentage errors (MAPE) for the corrected series and the global one

Series	Model	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Corrected	SARIMA	1.22	1.19	1.34	1.61	1.42	1.26	1.37
Corrected	IBR96	1.13	1.11	1.16	1.12	1.02	1.21	1.37
Global	SARIMA	1.03	1.01	1.12	1.32	1.15	1.04	1.14
Global	IBR96	0.97	0.92	1.00	0.97	0.85	0.98	1.14

Table 3 Monthly percentage errors (MAPE) for the corrected series and the global one

Series	Model	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Corrected	SARIMA	4.96	1.48	1.97	1.28	1.26	0.91	0.45	0.05	0.53	1.40	1.31	1.32
Corrected	IBR96	2.68	1.37	1.54	1.49	1.15	0.85	0.49	0.18	0.46	1.36	1.45	1.38
Global	SARIMA	3.53	1.08	1.60	1.16	1.22	0.89	0.44	0.05	0.53	1.28	1.13	0.98
Global	IBR96	1.76	0.99	1.26	1.35	1.11	0.83	0.48	0.19	0.45	1.25	1.23	1.04

Table 4 Peak percentage errors (MAPE) for the corrected series and the global one

Series	Model	10:30	11:00	11:30	18:00	18:30	19:00	19:30	20:00
Corrected	SARIMA	1.17	1.18	1.19	1.61	1.63	1.60	1.65	1.74
Corrected	IBR96	1.04	1.03	0.99	1.41	1.34	1.34	1.41	1.55
Total	SARIMA	0.98	0.99	1.00	1.32	1.35	1.33	1.39	1.47
Total	IBR96	0.89	0.88	0.85	1.19	1.13	1.13	1.21	1.33

model and IBR96: we will focus on daily errors (Table 2), monthly errors (Table 3) and peak hours errors (Table 4).

A more refined analysis of forecasting errors can be done by considering the error for each day of the week or for each hour of the day. When considering errors for each day of the week, it can be seen in Table 2 that IBR96 is uniformly better for each type of day but the errors obtained are extremely similar for the weekends. This can be easily explained by the low level of consumption of these days, and their regularity.

When considering errors for each hour of the day (or saying differently errors by horizon of forecasts), Fig. 4 shows that the MAPE of IBR96 is uniformly better for each hour. When zooming on the peaks of consumption (in the morning and in the evening), Table 4 confirms that IBR96 is better. When focusing on the month, the classical SARIMA approach does better forecasts for April, July, August, November and December. It seems that this modeling allows better forecasts for low consumption months (such as July) and leads to very bad forecasts for very high consumption months such as January.

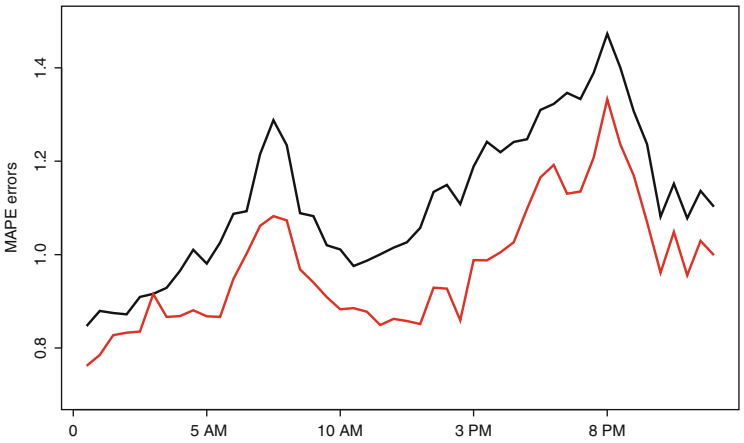


Fig. 4 Hourly percentage errors (MAPE) for the corrected series (SARIMA: *black*, IBR96: *red*)

5 Conclusion

Cornillon et al. [5] propose a new smoothing method, IBR, that has the desirable property of being simple and adaptive, which suggests that it may be used to perform fully nonparametric forecasting procedure. This paper compares this new method with the current state-of-the-art forecasting model at RTE which is used every day, and different competitors. This study on real French consumption of electricity (from January 2006 to December 2009) shows that IBR can lead to significant improvement over both the current RTE model and other competitors such as nonparametric functional approach or nonlinear additive autoregressive approach.

These improvements are not only on the global Mean Absolute Percentage Error over 2009 but also when MAPE are measured over high consumption date (consumption above 2,000 MW). When the MAPE are monthly detailed, some months show that the current RTE model outperforms IBR model. This remarks leads to a possible improvement by partitioning months in two category (or more): the low level consumption months and the high level consumption months. This qualitative variable *month* with two levels *high* and *low* could be used to partition the database in *month* $\times$ *day* instead of *day* only. To apply this remark with more than two category, one needs more observations or less memory in order to have enough data to make robust forecast.

Finally, from the TSO's point of view, this new modeling could be extended to other problems, like wind and solar generation forecasting. To insert renewable generation into the network according to the security of the grid, RTE needs appropriate tools to provide accurate forecasts and supervision.

Acknowledgements The authors would like to thank the editors and the two anonymous referees for their valuable comments which helped in improving the paper.

References

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.
2. Aneiros, G., Vilar, J. M., Cao, R., & Muñoz San Roque, A. (2013) Functional prediction for the residual demand in electricity spot markets. *IEEE Transactions on Power Systems, 28*(4), 4201–4208.
3. Antoniadis, A., Paparoditis, E., & Sapatinas, T. (2008). Bandwidth selection for functional time series prediction. *Statistics and Probability Letters, 79*, 733–740.
4. Chakhchoukh, Y. (2010). A new robust estimation method for ARMA models. *IEEE Transactions on Signal Processing, 58*(7), 3512–3522.
5. Cornillon, P. A., Hengartner, N., & Matzner-Løber, E. (2014, to appear). Recursive bias estimation for multivariate regression. *ESAIM/probability and statistics, 18*, 483–502.
6. Cornillon, P. A., Hengartner, N., Jégou, N., & Matzner-Løber, E. (2013). Iterative bias reduction: A comparative study. *Statistics and Computing, 23*, 777–791.
7. Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik, 31*, 377–403.
8. Cugliari, J. (2011). *Prévision non-paramétrique de processus à valeurs fonctionnelles. Application à la prévision de la consommation d'électricité*. Phd thesis, University Paris Sud 11.
9. Fan, S., & Hyndamn, R. J. (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE transactions on power systems, 27*(1), 134–141.
10. Goude, Y. (2008). *Mélange de prédicteurs et application à la prévision de la consommation électrique*. Phd thesis, University Paris Sud 11.
11. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.
12. Hengartner, N., Wegkamp, M., & Matzner-Løber, E. (2002). Bandwidth selection for local linear regression smoothers. *Journal of the Royal Statistical Society: Series B, 64*, 1–14.
13. Hurvich, C., Simonoff, G., & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using and improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B, 60*, 271–294.
14. Lefieux, V. (2007). *Modèles semi-paramétriques appliqués à la prévision des séries temporelles: cas de la consommation d'électricité*. Phd thesis, Rennes.
15. Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics, 15*, 958–975.
16. Martin, M. M. (1999). Filtrage de Kalman d'une série temporelle saisonnière. Application à la prévision de consommation d'électricité. *Revue de Statistique Appliquée, 47*(4), 69–86.
17. Meslier, F. (1976). *Contribution à l'analyse des séries chronologiques et application à la mise au point de modèles de prévision à court terme relatifs à la demande journalière relevée à Paris Monsouris*. Phd thesis, University Paris Sud 9.
18. Pierrot, A., & Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. In *Proceedings of ISAP power*, Hersonissos, Greece.
19. Poggi, J. M. (1994). Prévision non-paramétrique de la consommation électrique. *Revue de Statistique Appliquée, 42*(4), 83–98.
20. Reiss, P., & Ogden, R. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B, 71*, 505–523.
21. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.
22. Taylor, J. W., & McSharry, P. E. (2007). A new robust estimation method for ARMA models. *IEEE Transactions on Power Systems, 22*(4), 2213–2219.
23. Vilar, J. M., Cao, R., & Aneiros, G. (2012). Forecasting next-day electricity demand and price using nonparametric functional methods. *Electrical Power and Energy Systems, 39*, 48–55.
24. Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton: Chapman & Hall/CRC.

Forecasting Electricity Consumption by Aggregating Experts; How to Design a Good Set of Experts

Pierre Gaillard and Yannig Goude

Abstract Short-term electricity forecasting has been studied for years at EDF and different forecasting models were developed from various fields of statistics or machine learning (functional data analysis, time series, non-parametric regression, boosting, bagging). We are interested in the forecasting of France's daily electricity load consumption based on these different approaches. We investigate in this empirical study how to use them to improve prediction accuracy. First, we show how combining members of the original set of forecasts can lead to a significant improvement. Second, we explore how to build various and heterogeneous forecasts from these models and analyze how we can aggregate them to get even better predictions.

1 Introduction

Electricity consumption forecasting is a crucial matter for electricity providers like EDF to maintain the equilibrium between production and demand. Overestimating the consumption leads to overproduction, which has a negative environmental impact and implies unnecessary loss of benefits for the company. On the other hand, underestimating the consumption may cause a shortage of energy and black outs. In the past years EDF R&D has therefore developed several competitive forecasting models achieving around 1.4 % error in MAPE (the average of percentage errors, see (2) for a formal definition) at the daily horizon. However the electrical scene in France is constantly evolving (nuclear power, electric cars, air conditioning are developing for instance) and the opening of the electricity market induces potential

P. Gaillard (✉)
EDF R&D, 1 av du Général de Gaulle, Clamart, France

GREGHEC, CNRS, Jouy-en-Josas, France
e-mail: pierre@gaillard.me

Y. Goude
EDF R&D, 1 av du Général de Gaulle, Clamart, France
e-mail: yannig.goude@edf.fr

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_6

customer losses. Therefore the historical models have to be regularly reconsidered and challenged. As daily forecasts are the main inputs for optimizing the production units we consider in this paper the goal of improving short-term (daily) forecasting of France's electricity consumption.

As the historical French electricity provider, EDF has investigated the issue of load forecasting for years and developed models from a wide range of statistical or machine learning methods. Among many, we consider in this study three approaches presented below. They were chosen for two main reasons. First, they have a good forecasting accuracy. Second, they are derived from quite different statistical frameworks, which results in a sort of heterogeneity. The first model is a non-parametric model based on regularized regression on spline basis (see Wood [28]). It will be referred to next as the generalized additive model (GAM). This model has performed well on France's load consumption signal (see Pierrot and Goude [25]), on EDF portfolio data (see Wood et al. [29]) and was proven to be a good competitor on US data (see Nedellec et al. [24]). The second model is based on curve linear regression (CLR) via dimension reduction. It is introduced and applied to electricity consumption forecasting in Cho et al. [10, 11]. The third and last model, kernel wavelet functional (KWF), is detailed in Antoniadis et al. [2–4]. It combines clustering functional data and detection of similar patterns in functional processes based on a wavelet distance. These three approaches are based on extremely different insights and we expect it can induce different behaviors that an aggregation algorithm can take advantage of in some online fashion. The GAM model captures non-linear relationships between electricity load and the different covariates driving it (temperature, fare effects...) and provides smooth estimates of these transfer functions without any transformation of the original data. The CLR model performs a data-driven dimension reduction as well as a data transformation so that the relationship between the transformed data is linear and can be captured by simple multivariate regression models. The KWF approach is non-parametric and does not use any exogenous variable but the past consumption. It is particularly robust to special days (bank holidays, holiday seasons) and meteorological forecasts errors. In the GAM setting, observations (half-hourly electricity load and covariates) are considered as finite dimensional whereas in the CLR and the KWF approaches, daily electricity load is the realization of a functional process.

As we have at our disposal three forecasting models, a straightforward question is how to combine them to produce accurate forecasts. The art of combining forecasts has been extensively studied for the past four decades (see the review of Clemen [12]) and the empirical literature is voluminous. However, few real-world empirical studies consider the framework of individual sequences to design the aggregation rules. Some of them include for instance climate prediction [23], air-quality prediction [21, 22], quantile prediction of daily call volumes entering call center [6], or electricity consumption [13]. The vast majority of these studies focuses however on the aggregation rules and how to weight the experts. Little consideration goes into designing the set of experts to include in the combination. Aiolfi et al. in their technical report [1] studied the construction of a varied enough set of experts by considering the combination of linear autoregressive models with

non-linear models (logistic smooth transition autoregressive and neural networks). They however did not consider the same aggregation rules as we do: because of the small length of their time series, none of their rules had time to learn the weights and the best results were obtained using uniform aggregation scheme.

We now describe the methodology followed in this study. We aim first at designing a set of base forecasting methods (henceforth referred to as experts) by using the three models described above. We show how an aggregation rule that sequentially outputs forecasts of the electricity consumption for the next instances can significantly improve upon these experts. The aggregation rules and the framework of prediction with expert advice is detailed in Sect. 2. Then, we propose different strategies to design a larger set of experts from the three initial experts and give a detailed analysis of the corresponding combined forecasts.

2 Sequential Aggregation of Experts

The content of this section reviews the framework of sequential prediction with expert advice, a setting which received considerable attention in the past 20 years (see the monograph by Cesa-Bianchi and Lugosi [9]). It considers an online learning scenario in which a forecaster has to guess element by element future values of an observed time series. To form its prediction it receives and combines before each instance the opinions of a finite set of experts. This framework makes possible to consider several stochastic models with extremely different assumptions in a single approach. To do so, it adopts the deterministic and robust point of view of the literature of individual sequences. It is thus particularly adapted to our application.

2.1 Mathematical Context

We now present the mathematical setting of prediction with expert advice. We suppose that at each time instance $t = 1, \ldots, T$ the next outcome y_t of a sequence of observations $y_1, \ldots, y_T$, like half-hourly electricity consumptions, is to be predicted. We assume that the observations are all bounded by some positive constant B, so that $y_t \in [0, B]$. Before each time instance t, a finite number K of experts provide forecasts $\boldsymbol{x}_t = (x_{1,t}, \ldots, x_{K,t}) \in [0, B]^K$ of the next observation y_t. A forecaster is then asked to form its own prediction with knowledge of the past observations $y_1^{t-1} = y_1, \ldots, y_{t-1}$ and of the past expert advice $\boldsymbol{x}_1^t = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$. Let denote by $\cdot$ the inner product in $\mathbb{R}^K$. Formally the forecaster forms a mixture $\hat{\boldsymbol{p}}_t = (\hat{p}_{1,t}, \ldots, \hat{p}_{K,t}) \in \mathbb{R}^K$ and predicts $\hat{y}_t = \hat{\boldsymbol{p}}_t \cdot \boldsymbol{x}_t = \sum_{k=1}^K p_{k,t} x_{k,t}$ by linearly combining the predictions of the experts.

The accuracy of a prediction x proposed by an expert or by the aggregation rule at time instance t for the outcome y_t is measured through a convex loss function ℓ_t. In this paper, we consider the special case of the square loss $\ell_t(x) = (y_t - x)^2$. The

analysis can however be easily extended to any convex loss function. On instance t, expert k suffers loss $\ell_t(x_{k,t}) = (y_t - x_{k,t})^2$ and the aggregation rule incurs loss $\ell_t(\hat{y}_t) = (y_t - \hat{y}_t)^2$. The goal of the forecaster is to design aggregation rules (that is, applications $\mathscr{A} : (\boldsymbol{x}_1^t, y_1^{t-1}) \mapsto \hat{\boldsymbol{p}}_t$) with small average error. The latter can be decomposed as

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 \quad \triangleq \quad \inf_{\boldsymbol{q}\in S}\left\{\frac{1}{T}\sum_{t=1}^{T}(y_t - \boldsymbol{q}\cdot\boldsymbol{x}_t)^2\right\} + R_T\,, \tag{1}$$

where S is some closed and bounded subset of $\mathbb{R}^K$; and this defines the regret R_T. As we explain next this decomposition highlights the well-known trade-off between approximation error and estimation error. Because these two terms add up to the error incurred by the aggregation rule they act as two opposing forces.

The first term in (1) is the error encountered by the best constant weight vector chosen in hindsight in a closed and bounded set $S \subset \mathbb{R}^K$. This best mixture is called an *oracle*. Its performance is the target that the aggregation rule intends to reach and is thus used as a benchmark value to be compared to the performance of an aggregation rule. Several oracles can be defined according to the set S the aggregation rule aims at competing with. We can list several oracles: the *best expert* oracle suffers $\min_{k=1,\dots,K}\sum_{t=1}^{T}(y_t - x_{k,t})^2$; the *best convex weight vector* corresponds to the best element in $S = \Delta_K \triangleq \{\boldsymbol{q} \in \mathbb{R}_+^K \,:\, \sum_i q_i = 1\}$; and finally the *best linear* oracle is defined by $S = B_K(r)$ the ball of radius r in $\mathbb{R}^K$. The larger the set S we aim at competing with, the smaller the first term in (1) is, but the harder it is for the aggregation rule to remain competitive. The second term grows in general. This approximation error is closely related to the expert forecasts. It decreases with increasing heterogeneity of the expert set.

The second term R_T is the estimation error. It evaluates the ability of the aggregation rule to retrieve online the oracle, i.e., the best possible mixture. If the aggregation rule is well designed, R_T will vanish to 0 as the length T of the experiment grows to infinity.

We assume in this paper that we have an efficient aggregation rule and we focus on reducing the approximation error; indeed many efficient aggregation rules are already well-known—see Sect. 2.2, but the approximation error is often left out of the debate.

2.2 Aggregation Rules

Experiments are performed by considering four different aggregation rules: the exponentially weighted average forecaster (EWA), the fixed share forecaster (FS), the ridge regression forecaster (Ridge), and the polynomially weighted average forecaster with multiple learning rates (ML-Poly). EWA, FS, and Ridge are described in the book of Cesa-Bianchi and Lugosi [9] for constant values of their

learning parameters. Devaine et al. [13] already applied EWA and FS to short-term load forecasting. They suggested in Sect. 2.4 an empirical tuning of the learning parameters which comes with no theoretical guarantees but works empirically well. It consists of optimally choosing the learning parameters on adaptive finite grids. Except for ML-Poly which already comes with its own learning parameter calibration rule, the parameters are tuned online following the method of Devaine et al. [13].

The exponentially weighted average forecaster (EWA) is an online convex aggregation rule introduced in learning theory by Littlestone and Warmuth [20] and by Vovk [27]. At time instance t, it assigns to expert k the weight

$$\hat{p}_{k,t} = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(x_{k,s})}}{\sum_{i=1}^{K} e^{-\eta \sum_{s=1}^{t-1} \ell_s(x_{i,s})}},$$

which is exponentially small in the cumulative loss suffered so far by the expert. When the learning parameter η is properly tuned, it has a small average regret $R_T = O\left(1/\sqrt{T}\right)$ with respect to the best fixed expert oracle—see Cesa-Bianchi and Lugosi [9].

The fixed share forecaster (FS) is due to Herbster and Warmuth [18]. It has the property to compete not only with the best fixed expert but with the best sequence of experts that may change a small number of times. It is particularly interesting when dealing with non stationary environments, in which the best expert should regularly be reconsidered. The fixed share forecaster considers a learning parameter η as well as a mixing parameter $\alpha \in [0, 1]$ that evaluates the number of changes in the oracle sequence of experts we are competing with.

We now provide a short mathematical description of the fixed share aggregation rule. The initial weight distribution is uniform $\hat{p}_1 = (1/K, \ldots, 1/K)$. Then, at each instance t, the weights are updated twice. First, a *loss update* takes into account the new loss incurred by each expert,

$$\hat{v}_{k,t} = \frac{\hat{p}_{k,t-1} e^{-\eta \sum_{s=1}^{t-1} \ell_s(x_{k,s})}}{\sum_{i=1}^{K} \hat{p}_{i,t-1} e^{-\eta \sum_{s=1}^{t-1} \ell_s(x_{i,s})}}.$$

Second a *mixing-update* ensures that each expert gets a minimal weight α/K by assigning

$$\hat{p}_{k,t} = (1 - \alpha)\hat{v}_{k,t} + \alpha/K.$$

This update captures the possibility that the best expert may have switched at time instance t. The fixed share forecaster was proven to have nice theoretical properties and vanishing average regret R_T with respect to sequences of experts with few shifts.

Algorithm 1: The polynomially weighted average forecaster with multiple learning rates (ML-Poly)

Initialization: $\boldsymbol{p}_1 = (1/K, \ldots, 1/K)$ and $\boldsymbol{R}_0 = (0, \ldots, 0)$
For each instance $t = 1, 2, \ldots, T$

0. pick the learning rates

$$\eta_{k,t-1} = 1/\left(1 + \sum_{s=1}^{t-1} \big(\ell_s(\hat{y}_s) - \ell_s(x_{k,s})\big)^2\right)$$

1. form the mixture $\hat{\boldsymbol{p}}_t$ defined component-wise by

$$\hat{p}_{k,t} = \eta_{k,t-1} \big(R_{k,t-1}\big)_+ \;/\; \boldsymbol{\eta}_{t-1} \cdot (\boldsymbol{R}_{t-1})_+$$

   where $\boldsymbol{x}_+$ denotes the vector of non-negative parts of the components of $\boldsymbol{x}$
2. output prediction $\hat{y}_t = \widehat{\boldsymbol{p}}_t \cdot \boldsymbol{x}_t$
3. for each expert k update the regret

$$R_{k,t} = R_{k,t-1} + \ell_t(\hat{y}_t) - \ell_t(x_{k,t})$$

For more details about the fixed share aggregation rule the reader is referred to Cesa-Bianchi and Lugosi [9, Section 5.2].

The polynomially weighted average forecaster with multiple learning rates (ML-Poly) is obtained via a version of the polynomially weighted average forecaster detailed in Cesa-Bianchi and Lugosi [8], see also Cesa-Bianchi and Lugosi [9, Section 2.1]. The multiple learning rate version is due to Gaillard et al. [17] whose implementation is recalled in Algorithm 1. Gaillard et al. [17] proved the regret bound $R_T = O\big(1/\sqrt{T}\big)$ with respect to the best fixed expert. ML-Poly is particularly interesting since despite the theoretical tuning of the learning parameters, it achieves as good performance as the other ones. It runs also much faster than the empirical tuning described by Devaine et al. [13] and used for the other rules which needs to run as many times the aggregation rule as the size of the parameter grid.

The ridge regression forecaster (Ridge) is presented in Algorithm 2. It was introduced in a stochastic setting by Hoerl and Kennard [19]. It forms at each instance the linear combination of experts minimizing a L_2-regularized least-square criterion on past data. It was first studied in the context of prediction with expert advice by Azoury and Warmuth [5] and Vovk [26] and was proved to enjoy nice theoretical properties, namely a regret bound $R_T = o(1)$ as $T \to \infty$ with respect to the best linear oracle. Once again, the learning parameter λ of the ridge regression aggregation rule has to be calibrated online. This tuning can be done using the methodology detailed in Devaine et al. [13, Section 2.4].

Ridge forms linear mixtures. The weights may be negative and not sum to one, while the other three aggregation rules restrict themselves to convex combination

Algorithm 2: The ridge regression forecaster (Ridge)

Parameter: $\lambda > 0$
Initialization: $\hat{\boldsymbol{p}}_0 = (1/K, \ldots, 1/K)$
For each instance $t = 1, 2, \ldots, T$

1. form the mixture $\hat{\boldsymbol{p}}_t$ defined by

$$\hat{\boldsymbol{p}}_t = \underset{\boldsymbol{u} \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ \sum_{s=1}^{t-1} (y_s - \boldsymbol{u} \cdot \boldsymbol{x}_s)^2 + \lambda \left\| \boldsymbol{u} - \boldsymbol{p}_0 \right\|_2^2 \right\}$$

2. output prediction $\hat{y}_t = \hat{\boldsymbol{p}}_t \cdot \boldsymbol{x}_t$

of experts. In other words they only propose weight vectors $\hat{\boldsymbol{p}}_t \in \Delta_K$ where $\Delta_K = \{\boldsymbol{x} \in \mathbb{R}_+^K \ : \ \sum_i x_i = 1\}$. While linear aggregation rules might have more flexibility to detect correlation between experts and therefore often reach better performance, convex aggregation offers easy interpretation and safe predictions. Indeed convex weight vectors only assign non-negative weights to experts and their predictions always lie in the convex hull of experts predictions. Thus if all the experts are known to perform well, the aggregation rule will do so as well.

The gradient trick In the versions described above, EWA, FS, and ML-Poly compete only with the best fixed expert oracle. In Eq. (1) they cannot per se ensure vanishing average regret R_T with respect to the best fixed convex combination (i.e., $S = \Delta_K$). But it exists a standard reduction from the problem of competing with the best convex combination oracle to the goal of competing with the best fixed expert. This reduction is a well-known trick in the literature of individual sequences and is known as the *gradient trick*. The theoretical proof of this reduction is beyond the scope of this empirical research and is detailed in Cesa-bianchi and Lugosi [9, Section 2.5].

We only provide a brief description of the gradient trick. For each time instance t, we denote by $f_t : \boldsymbol{p} \in \Delta_K \mapsto \ell_t(\boldsymbol{p} \cdot \boldsymbol{x}_t) \in \mathbb{R}_+$ the function which evaluates the losses incurred by the weight vectors at time instance t. When the loss functions ℓ_t are convex and (sub)differentiable, the functions f_t are convex and (sub)differentiable over Δ_K. That is the case for instance for the square loss. We denote by ∇f_t the (sub)gradient function of f_t. The gradient trick relies then in not directly running the aggregation rule with the loss functions ℓ_t but with modified gradient loss functions $\tilde{f}_t : \boldsymbol{p} \in \Delta_K \mapsto \nabla f_t(\hat{\boldsymbol{p}}_t) \cdot \boldsymbol{p}$. In other words, the aggregation rules are run the same way by replacing the loss $\ell_t(\hat{y}_t)$ incurred by the algorithm by $\tilde{f}_t(\hat{\boldsymbol{p}}_t)$ and the loss $\ell_t(x_{k,t})$ suffered by expert k by $\tilde{f}_t(\delta_{k,t})$, where $\delta_k \in \Delta_K$ is the Dirac mass on k. Experiments of the next section are run using the gradient trick.

3 Experiments

We now describe the data we are dealing with and how we intend to build new experts from the three forecasting models described in the introduction. We then report the results obtained by mixing the different sets of experts as well as the performance of three reference oracles (best experts, best convex combination, best linear combination). As explained in Sect. 2 the performance of these oracles corresponds to the one aggregation rules hope to reach. Remember that the fixed share aggregation rule does not only compete with the best fixed convex combination but has a more ambitious goal. It aims at coming close to the performance of the best sequence of convex combinations that vary slowly enough. The results obtained by this more complex oracle will however not be reported in this research and we will only compare the performance of the fixed share aggregation rule to the best fixed convex combination of experts.

3.1 Presentation of the Data Set

We consider an electricity forecasting data set which corresponds to an updated version of the one analyzed by Devaine et al. [13]. It contains half-hourly measurements of the total electricity consumption of the EDF market in France from January 1, 2008 to June 15, 2012, together with several covariates, including temperature, cloud cover, wind, etc. Our goal is to forecast the consumption every day at 12:00 for the next 24 h; that is, for the next 48 time instances.

Atypical days are excluded from the data set. They correspond to public holidays as well as the days before and after them. Besides, the data set is cut into two subsets. A training set of 1,452 days from January 1, 2008 to August 31, 2011 is used to build the forecasting methods. The performance of the methods is then measured using the testing set of 244 days between September 1, 2011 to June 15, 2012. Prediction accuracy is measured in megawatts (MW) by the root mean squared error (RMSE)

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t-\hat{y}_t)^2}$$

and by the absolute percentage of error (MAPE)

$$\frac{1}{T}\sum_{t=1}^{T}\frac{|y_t-\hat{y}_t|}{y_t}. \tag{2}$$

Operational forecasting purposes require the predictions to be made simultaneously at 12:00 for the next 24 h (or equivalently for the next 48 half-hourly time

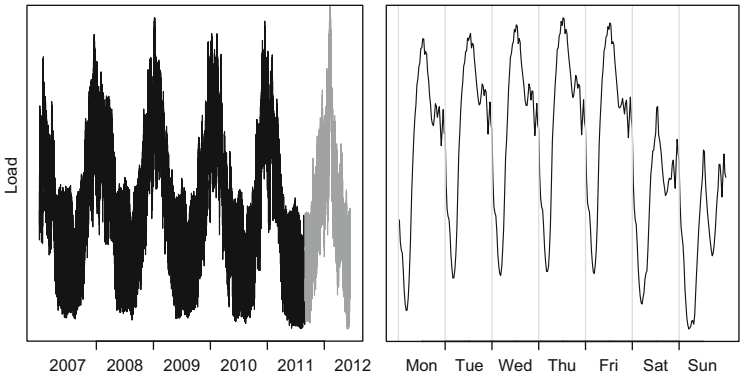


Fig. 1 (*left*) The observed half-hourly electricity consumptions between January 1, 2008 to June 15, 2012. An overall trend as well as a yearly seasonality can be pointed out in the data. The electrical heating in winter has a major impact in France on the electricity consumption. Approximately the last year is used to test the methods. (*right*) The observed half-hourly electricity consumptions during a typical week. A weekly pattern can be observed with a reduction of consumption during the week-end

instances) (Fig. 1). Aggregation rules can be adapted to this constraint via a generic extension detailed in Devaine et al. [13, Section 5.3].

3.2 Combining the Three Initial Models

From each of the three forecasting models described in the introduction, one expert is obtained: one from the generalized additive model (GAM), one from the curve linear regression (CLR) and a last one from the kernel approach based on wavelets (KWF). The experts are trained using the total training set from January 1, 2008 to August 31, 2011 described in the previous section. We calibrate the methods as presented in [4, 11, 25]. This starting set of three experts is denoted in the rest of the paper by E_0.

Table 1 reports the performance obtained by mixing the three experts in E_0. It describes also the reference results of the corresponding benchmark oracles: the best expert in E_0, the best convex combination and the best linear combination. The best convex combination and the best linear combination obtain similar results with RMSEs of 629 MW. Due to confidentiality constraints, we cannot provide detailed characteristics of the observed electricity consumptions. The relative performance of the methods can be enjoyed by noting that MAPEs are around 1 %. A significant improvement in performance can be noted in comparison to the best expert which obtains 744 MW. This motivates the necessity of mixing these models whose forecasts bring different information.

Table 1 Performance of oracles and aggregation rules using the set of experts E_0: GAM, CLR, and KWF

Oracles and aggregation rules	RMSE (MW)	MAPE (%)
Best expert	744	1.29
Best convex combination	629	1.06
Best linear combination	629	1.06
EWA	624	1.07
FS	625	1.05
ML-Poly	626	1.05
Ridge	638	1.06

EWA, FS, and ML-Poly are designed to compete with the best convex combination of experts while Ridge aims at approaching the performance of the best linear combination. The latter suffer RMSEs between 624 and 638 MW, which corresponds to reductions of the RMSE of approximatively 15 % compared to the best expert RMSE.

To quantify if our improvements are significant, we computed the dispersion of the errors among time instances of the aggregation rules and of the oracles—see technical report from Gaillard et al. [16, Section 1.2] for details. The dispersion is measured by the 95 % standard error

$$\hat{s}_t = \sqrt{\frac{\frac{1}{T}\sum_{t=1}^{T}\left((y_t-\hat{y}_t)^2 - \frac{1}{T}\sum_{t=1}^{T}(y_t-\hat{y}_t)^2\right)^2}{\frac{4}{T}\sum_{t=1}^{T}(y_t-\hat{y}_t)^2}},$$

that is, the half-width of the 95 % symmetric confidence interval of the error around the RMSEs reported in Tables 1–6. The 95 % standard error of the RMSEs are around 10 MW while the 95 % standard error of the MAPE are approximatively 0.02 %. Hence any reduction of the RMSE of more than 10 MW can be considered significant in the following.

Figure 2 reports the time evolution of the weights formed by ML-Poly and Ridge. The weight vectors created by Ridge converge but that is not obvious with ML-Poly. Stability is beneficial in an industrial context where weights have to be interpreted and understood by human beings. The weights formed by EWA and FS behave similarly to the ones of ML-Poly and are thus not reported here.

In the next section we will investigate how more experts can be designed based on these three models in order to improve further the predictions (Figs. 3 and 4).

3.3 Creating New Experts

We aim now at reducing the approximation error in Eq. (1), i.e., at improving the performance of the oracles, by adding new experts to our initial set E_0. If the new experts are not different enough from the base ones, the approximation term will

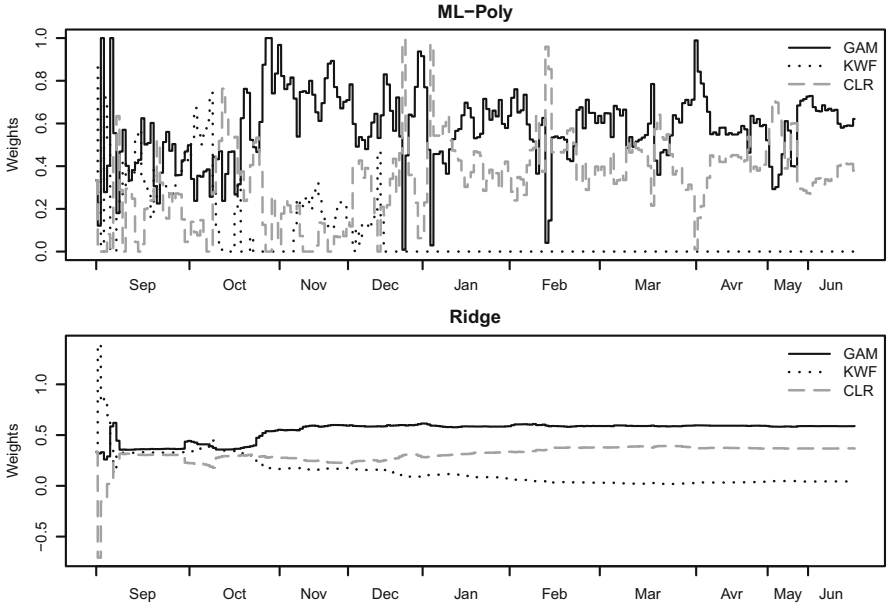


Fig. 2 Time evolution of the weight vectors formed by ML-Poly (*top*) and Ridge (*bottom*). We remark that the weights assigned by ML-Poly are always non-negative and sum to 1. Ridge can form negative weights

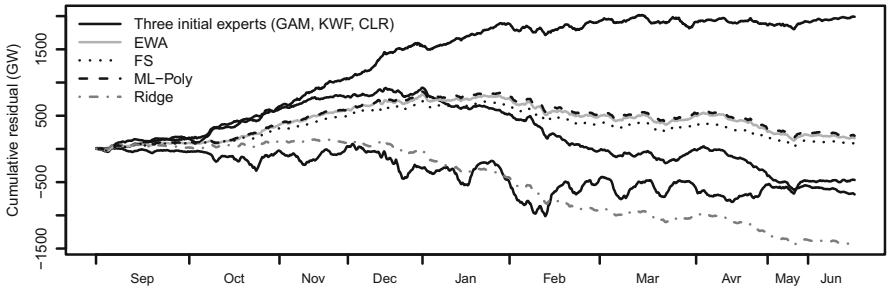


Fig. 3 Time evolution of cumulative residual of the three experts in E_0 and of the considered aggregation rules. The aggregation rules have smaller gradient in comparison to the experts. Besides it can be noticed that Ridge behaves very differently when compared to the other aggregation rules

not decrease; and worse, the right-most term in (1), the sequential estimation error, may increase, as the aggregation rule will have to face more experts. Note that none of the newly constructed experts will significantly outperform the performance of the best expert in E_0, which achieves a RMSE of 744 MW and a MAPE of 1.29 %. The benchmark performance of the best expert oracle thus remains the same for all considered extended sets of experts in this study.

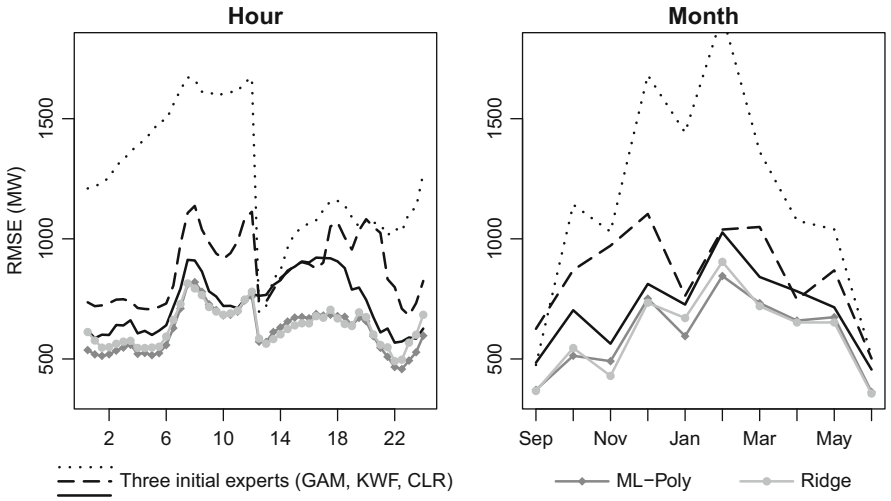


Fig. 4 Hourly and monthly RMSE of the first three experts and two aggregation rules described in Table 6. Because they obtain similar results to the ML-Poly aggregation rule, the EWA and the fixed share aggregation rules are not reported here

3.3.1 Bagging

The first method that we investigate is inspired from bagging, a machine learning method that combines bootstrapping with aggregating. It was introduced by Breiman [7] in order to improve the stability and the accuracy of a forecasting model. As most averaging methods it is known to reduce the variance and to avoid over-fitting. We aim at creating new experts by bootstrapping and at averaging online the newly constructed set of experts by running the aggregation rules.

Given a forecasting model, a bootstrapped expert is obtained by estimating the model on a random training strict subset S_0' (that does not include the whole training set S_0 of $n = 1{,}452$ days). The training set S_0' is generated by sampling n days from S_0 uniformly and with replacement. As the sampling is performed with replacement, some days may be present multiple times in S_0'. Breiman [7] pointed out that it leaves out $e^{-1} \approx 37\,\%$ of the days.

The bootstrap procedure is repeated 20 times using each of the three models at hand: GAM, CLR, and KWF. We name E_1 the set of 60 new experts, thus created. In our experiments we used 20 bootstrapped replicates of each model. This does not mean that more or fewer replicates would have led to worse performance. We wanted to add enough replicates to get sufficient variety but in the other hand we did not want to have too many bootstrapped experts in comparison to the experts we will build in the following sections. We tested several values and 20 expert replicates for each model seemed to be a reasonable trade off.

The performance of aggregation rules and oracles on $E_0 \cup E_1$ is reported in Table 2. The best linear combination oracle achieves a RMSE of 571 MW, which

Table 2 Performance of oracles and aggregation rules using the set of experts $E_0 \cup E_1$: GAM, CLR, KWF as well as the 60 bootstrapped experts

Oracles and aggregation rules	RMSE (MW)	MAPE (%)
Best convex combination	601	1.01
Best linear combination	571	0.99
EWA	614	1.01
FS	619	1.03
ML-Poly	612	1.02
Ridge	629	1.05

is a slightly better performance than the one of the best convex combination oracle, that equals 601 MW. This can be explained by two facts. First, the new experts might be biased. As their weights do not need to sum to one, linear mixtures correct better such bias. Second, as many experts are built using the same method, there are important correlations between them that can be better modeled using negative weights. However Ridge seems to have a hard time estimating the linear oracle and the performance is not much improved compared to Table 1. The empirical gain is about 10 MW for all aggregation rules. The improvement is thus not really significant.

3.3.2 Specialization

We start this section with the intuition that we need variety in our set of experts. We try to reuse the idea of bootstrapping to create new experts by modifying the training set. However, instead of sampling days uniformly in the training set E_0, we aim at assigning weights to training days with the goal to maximize the variety among themselves. To do so, we choose weights according to the values of the corresponding covariates (temperature, nebulosity, wind, type of day, ...). *Specialized experts* are created this way to some specific scenarios like heatwave, cold spell, sunny days or cloudy days. Hopefully if we choose different enough scenarios, these experts may catch different effects in the consumption that we might combine by aggregating them.

We now describe how to design such new experts. We suppose that we have at our disposal a forecasting model such that, during the training of the model, we can assign different weights to the elements of the training data. This is the case for GAM, CLR, and KWF for example. We assume that we also have access to an exogenous variable $Z \in [0, 1]$ like the temperature or the nebulosity which was normalized in $[0, 1]$. Given this model and this exogenous variable Z, we build two specialized experts: the first one by assigning to the day d the weight $(1 - Z_d)^2$, the second one with the choice Z_d^2. We thus get one expert focusing on high values of Z, and another one focusing on low values. The form of these weights was set empirically but we might want to replace it by many other forms. For instance, we had first looked at weights in $\{0, 1\}$ so as to select days according to a threshold on Z. However this led to unstable experts and poor performance. We

Table 3 Performance of oracles and aggregation rules using the set of experts $E_0 \cup E_2$: GAM, CLR, KWF as well as the 24 specialized experts

Oracles and aggregation rules	RMSE (MW)	MAPE (%)
Best convex combination	604	1.02
Best linear combination	582	0.99
EWA	609	1.01
FS	610	1.02
ML-Poly	602	1.00
Ridge	613	1.01

chose four covariables all based on temperature scenarios: the average, maximum, and minimum temperature of the day, and the variation of temperature with the previous day. We thus got 8 ($=$ 4 scenarios $\times$ 2 experts: hight and low) specialized experts by using each of the three models: GAM, CLR, and KWF. We call E_2 this set of 24 ($=$ 8 experts $\times$ 3 methods) experts. The performance obtained by mixing the experts in $E_0 \cup E_2$ is reported in Table 3. We observe a better performance for all aggregation rules with respect to bagging although we consider fewer additional experts.

Note that we showed the interest of specialized experts when they are combined with initial experts. The individual performance of specialized experts is often poor. They do not necessarily perform better than initial experts even when they are evaluated only on the data they should be specialized to.

3.3.3 Temporal Double-Scale Model

Now we study another way of constructing new experts by considering a temporal two-scale model. We follow the methodology detailed in Nedellec et al. [24] of the team TOLOLO for the "Kagle Global Energy Forecasting Competition 2012: Load Forecasting".

To forecast the short-term load with the canonical generalized additive model (GAM), the electricity consumption is usually explained by a single model including all the covariates (meteorological, and calendar ones) together with the recent consumption. The consumption Y_t is here decomposed into two parts: a medium-term part Y_t^{mt} including meteorological and calendar effects and a short-term part Y_t^{st} containing what could not be captured in large temporal scales, $Y_t = Y_t^{mt} + Y_t^{st}$. The short-term part Y_t^{st} basically consists of capturing local effects like extreme weather, network reconfigurations and so on. The modeling approach is thus divided into two estimation steps. First, we fit a mid-term generalized additive model including the meteorological and calendar covariates only. Second, we perform a residual analysis and we correct online the forecasts by using the observed consumptions of the last 30 days. This short-term readjusting is done by fitting another generalized additive model on the residuals.

The set containing this new expert is called E_3 and the performance obtained by combining this new expert with the three experts in E_0 is reported in Table 4. We

Table 4 Performance of oracles and aggregation rules using the set of experts $E_0 \cup E_3$: only four experts

Oracles and aggregation rules	RMSE (MW)	MAPE (%)
Best convex combination	596	1.00
Best linear combination	595	1.00
EWA	601	1.01
FS	599	1.00
ML-Poly	605	1.01
Ridge	605	1.00

observe RMSEs around 600 MW for all aggregation rules, which is a significant improvement considering that we add only one expert. The extension to other methods, like CLR and KWF, of this new way of creating experts is left for future work.

3.3.4 Boosting

In this section we investigate a final method to create new experts. We take now inspiration from boosting methods, like the AdaBoost algorithm of Freund and Schapire [15], that aims at correcting the mistakes of weak learners (or experts). The experts constructed in this section will be referred to as *boosted experts*.

Suppose that we have an expert that at an instance t of the training data estimates the consumption y_t by x_t. We want to build another expert predicting x'_t. Then reminding that our final aim is to aggregate well these predictions, it is irrelevant wether the second expert does not predict well y_t as soon as it counterbalances the error made by the original expert x_t. Improving the performance of the best convex combination should indeed only improve the prediction of the mixture. We can thus try to build the second expert so that the constant mixture $\gamma x_t + (1-\gamma)x'_t$ performs well for some $\gamma \in [0,1]$. This can be done by training the second experts not directly on the observed consumption y_t but on the modified one $y'_t = (y_t - \gamma x_t)/(1-\gamma)$. We can create several new experts by considering different values for $\gamma \in [0,1]$. Small values might lead to experts too similar from the original one, while larger values may create unstable experts.

We create 45 ($= 5 \times 3 \times 3$) new experts by using $\gamma \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, each of the three initial experts in E_0 are used as the original expert x_t and each of the three models (GAM, CLR, and KWF) are used to create the modified experts x'_t. We denote by E_4 the set of 45 experts thus constructed.

We report in Table 5 the performance obtained by mixing experts in $E_0 \cup E_4$. We did not consider $\gamma < 0.5$ because the created experts were too similar to the original ones. Considering all $\gamma \in \{0.1, \ldots, 0.9\}$ does not affect the results (neither improve nor worsen them). The step size 0.1 of the grid was chosen arbitrarily and the investigation of different values is left for future research. The best convex combination oracle achieves a RMSE of 528 MW and the best linear combination oracle suffers a RMSE of 543 MW. The performance of *EWA* and *FS* is

Table 5 Performance of oracles and aggregation rules using the set of experts $E_0 \cup E_4$: GAM, CLR, KWF as well as the 45 boosted experts

Oracles and aggregation rules	RMSE (MW)	MAPE (%)
Best convex combination	543	0.93
Best linear combination	528	0.92
EWA	609	0.99
FS	609	0.99
ML-Poly	588	1.00
Ridge	578	0.98

Table 6 Performance of oracles and aggregation rules using the full set of experts $E_0 \cup E_1 \cup E_2 \cup E_4 \cup E_3$: all the 133 constructed experts

Oracles and aggregation rules	RMSE (MW)	MAPE (%)
Best convex combination	521	0.95
Best linear combination	479	0.84
EWA	578	0.95
FS	581	0.95
ML-Poly	565	0.95
Ridge	557	0.95

not much improved compared to previous experiments. They both incur RMSEs of 609 MW. But ML-Poly and Ridge suffer *rmse*s under 580 MW, which is a significant improvement.

3.3.5 Combining the Full Set of Experts

Table 6 reports the performance obtained by mixing all the experts created in the previous sections. We have now 133 experts at our disposal: 3 experts from in the starting set E_0, 60 bootstrapped experts in E_1, 24 specialized experts in E_2, 45 boosted experts in E_4 and 1 temporal two-scale model in E_3. The best linear combination and the best convex combination perform better. But at the same time it is harder to compete with them. Thus while the performance of aggregation rules is improved, the gap between oracles and aggregation rules is increased as well.

Ridge suffers in Table 6 a RMSE of 557 MW while it got 638 MW when mixing only the three experts in E_0 (see Table 1). The several refinement of the set of experts thus reduced its RMSE by approximatively 13 %. Similarly, the errors of EWA and FS were improved by about 7 % while ML-Poly got a 10 % reduction.

Figure 5 provides the RMSEs according to the number of experts aggregated with ML-Poly and Ridge. The experts included in the mixture were chosen by induction on the number of experts by following a forward approach. The induction was initialized with the expert which performed the best (744 MW). Suppose we had a set of K experts, the expert $K + 1$ was the one among the remaining experts that got the best results when it was mixed with the K experts using the considered rule. The procedure was stopped when all the 133 experts were used in the aggregation. The symbols in the figures represent the category (bootstrapped, specialized, boosting, etc.) of the last added expert.

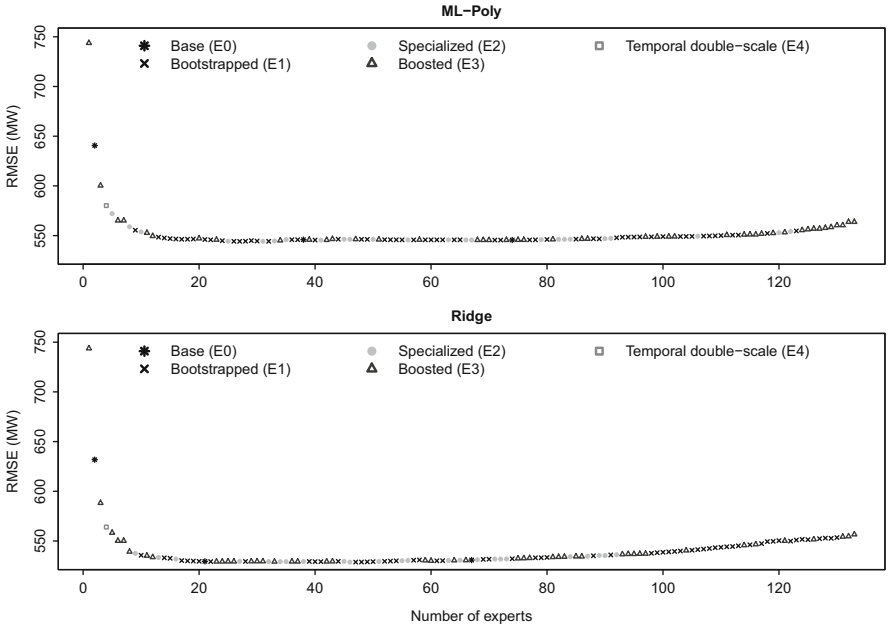


Fig. 5 Evolution of the performance according to the number of aggregated experts with ML-Poly (*top*) and Ridge (*bottom*)

Figure 5 shows the usual trade-off between having enough experts and overfitting. If we could select a good subset of experts to include in the mixture we could reduce the RMSE under the 530 MW bar by using Ridge (and approximatively under 545 MW by using ML-Poly). A suitable number of experts seems to lie between 15 and 90 experts. In future work, a pruning step, that would remove the less important experts before combining the forecasts of the remaining ones, might thus be a good option. Eban et al. [14] investigated in the framework of prediction of individual sequences a setting with many experts and few prediction instances. They remarked that trimming the worst experts often improves performance and suggested a procedure to do so online.

Note that the weights formed by ML-Poly and Ridge were different enough in Fig. 2. The aggregation rules might thus capture different information and we may thus try to combine them in a second layer. The simple uniform average of the forecasts of these two rules incurs a RMSE of 541 MW, while using one of the fancier sequential aggregation rules for the second layer gets us around 548 MW.

Figure 6 plots the hourly and monthly RMSEs of the two best aggregation rules and the RMSEs of the benchmark oracles described in Table 6. It shows that the aggregation rules always outperform in average the best single expert at all 48 half-hours of the day and at all 10 months of the testing set. In addition, we note a significant improvement of the performance at 12:30. This can be explained by the update of the weights, which occurs at noon. The best expert oracle, which is built

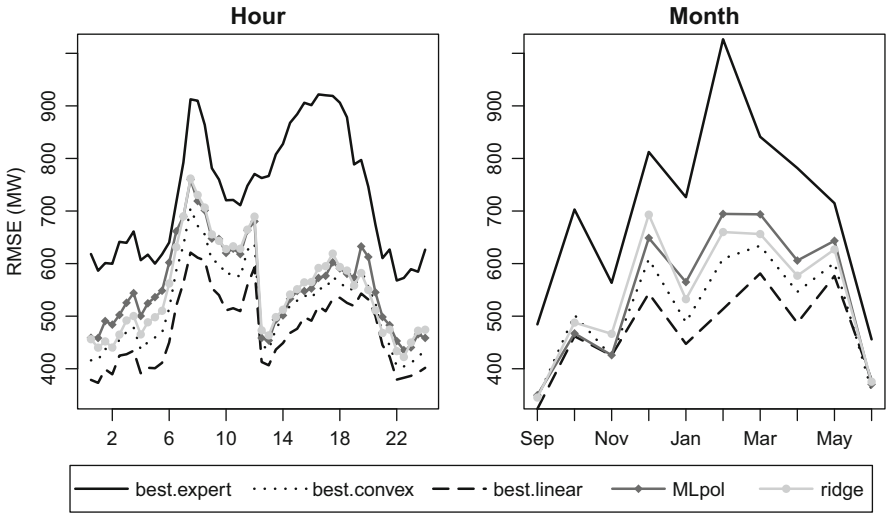


Fig. 6 Hourly and monthly RMSEs of the three benchmark oracles and of ML-Poly and Ridge described in Table 6

with a version of GAM, does not favor any hour of the day. The figure with monthly averaged RMSEs shows that aggregation rules do not only focus in improving forecasts when the task is easy. The best expert oracle is indeed outperformed every month, including November or February, which are month that are notoriously difficult to predict. Second, it illustrates that aggregation rules have a short learning period. They indeed encounter almost no regret during September and October with respect to all oracles although they just started to learn on September 1.

4 Conclusion

We presented in this paper an extensive application of aggregation rules from the literature of individual sequences to short-term electrical consumption forecasting. We focused on building an efficient set of experts from three initial ones, where the efficiency is viewed in terms of what these new experts bring to the combined forecasts. In other terms, we assumed that we had an efficient aggregation rule and focused more on reducing the approximation error, that is, the first term in (1). We noticed that despite the vast literature on combining forecasts (including empirical studies) rare papers dealt with this important topic. We proposed different strategies to generate experts from the three initial approaches: KWF, GAM, and CLR. We then quantified the gains in terms of forecast accuracy of the combined forecasts on the test set (about 10 month of half-hourly data). A summary of our results is presented in Fig. 7 for the two best aggregation rules: ML-Poly and Ridge.

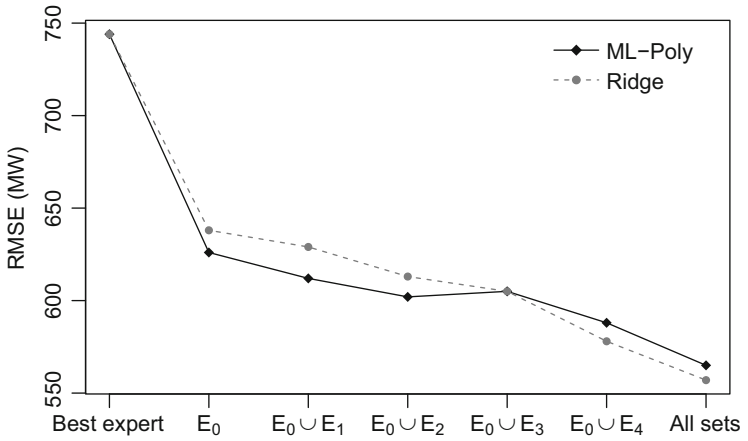


Fig. 7 RMSEs suffered by combining experts in E_0, $E_0 \cup E_1, \ldots, E_0 \cup E_4$ by using ML-Poly and Ridge. The performance of the best expert in E_0, and the final performance obtained by mixing all the experts in $E_0 \cup \cdots \cup E_4$ (referred to as 'All sets') are also reported

Combining all the experts that we generated with four different strategies, we achieved a 25 % gain over the best expert (around 200 MW in RMSE), which is a significant gain considering that the three original experts had already been refined and worked extremely well (they are not week learners as in classical boosting). This gain can be decomposed into two parts: roughly half of it comes from combining three heterogeneous initial experts, the other half is due to the construction of new experts. Among the four proposed strategies, our boosting trick and what we call specialized experts bring the most important improvements. We believe that these strategies could be applied to other forecasting problems and there is still some work to derive theoretical guarantees for these tricks. We also observe that aggregating rules are quite robust to adding new experts, and it is clear in Fig. 5 that combining forecasts does not suffer much from over fitting. Nevertheless, these results suggest that there is a way for improving the aggregation rules accuracy by adding a pruning step that could select the best set of experts in some online fashion.

Acknowledgements We thank the anonymous reviewers, the editors, and Gilles Stoltz for their valuable comments and feedback.

References

1. Aiolfi, M., Capistrán, C., & Timmermann, A. (2010). *Forecast combinations* (Working Papers 2010-04). Banco de México. http://EconPapers.repec.org/RePEc:bdm:wpaper:2010-04.
2. Antoniadis, A., Brossat, X., Cugliari, J., & Poggi, J. (2012). Prévision d'un processus à valeurs fonctionnelles en présence de non stationnarités. Application à la consommation d'électricité. *Journal de la Société FranÃČÂğaise de Statistique, 153*(2), 52–78.
3. Antoniadis, A., Brossat, X., Cugliari, J., & Poggi, J. (2013). Clustering functional data using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing, 11*(1), 1–30.
4. Antoniadis, A., Paparoditis, E., & Sapatinas, T. (2006). A functional wavelet–kernel approach for time series prediction. *Journal of the Royal Statistical Society: Series B, 68*(5), 837–857.
5. Azoury, K. S., & Warmuth, M. K. (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning, 43*(3), 211–246.
6. Biau, G., & Patra, B. (2011). Sequential quantile prediction of time series. *IEEE Transactions on Information Theory, 57*(3), 1664–1674.
7. Breiman, L. (1996). Bagging predictor. *Machine Learning, 24*(2), 123–140.
8. Cesa-Bianchi, N., & Lugosi, G. (2003). Potential-based algorithms in on-line prediction and game theory. *Machine Learning, 51*(3), 239–261.
9. Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge/ New York: Cambridge University Press.
10. Cho, H., Goude, Y., Brossat, X., & Yao, Q. (2013). Modeling and forecasting daily electricity load curves: A hybrid approach. *Journal of the American Statistical Association, 108*, 7–21.
11. Cho, H., Goude, Y., Brossat, X., & Yao, Q. (2014, to appear). Modeling and forecasting daily electricity load using curve linear regression. In *Lecture notes in statistics 217: Modeling and stochastic learning for forecasting in high dimension*, 35–52.
12. Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559–583.
13. Devaine, M., Gaillard, P., Goude, Y., & Stoltz, G. (2013). Forecasting electricity consumption by aggregating specialized experts. *Machine Learning, 90*(2), 231–260.
14. Eban, E., Birnbaum, A., Shalev-Shwartz, S., & Globerson, A. (2012). Learning the experts for online sequence prediction. In *Proceedings of ICML*, Edinburgh.
15. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an applicationto boosting. *Journal of Computer and System Sciences, 55*, 119–139.
16. Gaillard, P., Goude, Y., & Stoltz, G. (2011). *A further look at the forecasting of the electricity consumption by aggregation of specialized experts* (Technical report). pierre.gaillard.me/doc/GaGoSt-report.pdf.
17. Gaillard, P., Stoltz, G., & van Erven, T. (2014). A second-order bound with excess losses. ArXiv:1402.2044.
18. Herbster, M., & Warmuth, M. K. (1998). Tracking the best expert. *Machine Learning, 32*(2), 151–178.
19. Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67.
20. Littlestone, N., & Warmuth, M. K. (1994). The weighted majority algorithm. *Information and Computation, 108*(2), 212–261.
21. Mallet, V. (2010). Ensemble forecast of analyses: Coupling data assimilation and sequential aggregation. *Journal of Geophysical Research, 115*(D24303), 1–10.
22. Mallet, V., Stoltz, G., & Mauricette, B. (2009). Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research, 114*(D05307), 1–13.
23. Monteleoni, C., Schmidt, G. A., Saroha, S., & Asplund, E. (2011). Tracking climate models. *Statistical Analysis and Data Mining, 4*(4), 372–392.

24. Nedellec, R., Cugliari, J., & Goude, Y. (2014). Gefcom2012: Electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting, 30*(2), 375–381.
25. Pierrot, A., & Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. In: *Proceedings of ISAP power*, Hersonisos, Greece (pp. 593–600).
26. Vovk, V. (2001). Competitive on-line statistics. *International Statistical Review, 69*(2), 213–248.
27. Vovk, V. G. (1990). Aggregating strategies. In *Proceedings of the Third Workshop on Computational Learning Theory*, Rochester (pp. 371–386).
28. Wood, S. (2006). *Generalized additive models: An introduction with R*. Boca Raton: Chapman and Hall/CRC.
29. Wood, S., Goude, Y., & Shaw, S. (2015). Generalized additive models for large datasets. *Journal of Royal Statistical Society, Series C, 64*(1), 139–155.

Flexible and Dynamic Modeling of Dependencies via Copulas

Irène Gijbels, Klaus Herrmann, and Dominik Sznajder

Abstract In this chapter we first review recent developments in the use of copulas for studying dependence structures between variables. We discuss and illustrate the concepts of unconditional and conditional copulas and association measures, in a bivariate setting. Statistical inference for conditional and unconditional copulas is discussed, in various modeling settings. Modeling the dynamics in a dependence structure between time series is of particular interest. For this we present a semiparametric approach using local polynomial approximation for the dynamic time parameter function. Throughout the chapter we provide some illustrative examples. The use of the proposed dynamical modeling approach is demonstrated in the analysis and forecast of wind speed data.

1 Introduction

When the aim is to model the dependence structure between d random variables, denoted by $Y_1, \ldots, Y_d$, we can distinguish between several approaches. In a regression approach, one is interested in how a variable of primary interest, say Y_d, and called the response variable, is influenced on average by $Y_1, \ldots, Y_{d-1}$, the covariates. A general regression model is of the form

$$Y_d = g(Y_1, \ldots, Y_{d-1}) + \varepsilon \ ,$$

where $g : \mathbb{R}^{d-1} \to \mathbb{R}$ is a $(d-1)$-dimensional function of the covariates, and where the error term ε has conditional mean $E(\varepsilon | Y_1, \ldots, Y_{d-1})$ equal to zero. Consequently, the conditional mean function of Y_d given the covariates $Y_1 = y_1, \ldots, Y_{d-1} = y_{d-1}$, with $(y_1, \ldots, y_{d-1}) \in \mathbb{R}^{d-1}$ equals $E(Y_d | Y_1 = y_1, \ldots, Y_{d-1} = y_{d-1}) = g(y_1, \ldots, y_{d-1})$. For the mean regression

I. Gijbels (✉) • K. Herrmann • D. Sznajder
Department of Mathematics and Leuven Statistics Research Centre (LStat), KU Leuven, Celestijnenlaan 200B, Box 2400, B-3001 Leuven (Heverlee), Belgium
e-mail: Irene.Gijbels@wis.kuleuven.be; Klaus.Herrmann@wis.kuleuven.be; Dominik.Sznajder@wis.kuleuven.be

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_7

function g one can either assume some known parametric form, or leaving its form fully unspecified (and to be determined from the data), in respectively parametric and nonparametric regression. An alternative is a semiparametric modeling in which the influence of some covariates might be modeled parametrically, whereas the influence on average of other covariates on Y_d might be described via an unknown function (a nonparametric functional part). In a general approach, the dependencies between the various components in the random vector $(Y_1, \ldots, Y_d)$ are fully described by the joint distribution of $(Y_1, \ldots, Y_d)$, i.e. by the d-variate cumulative distribution function of $(Y_1, \ldots, Y_d)$ denoted by $P\{Y_1 \leq y_1, \ldots, Y_d \leq y_d\}$. From this one can calculate for example the conditional mean function $E(Y_d | Y_1 = y_1, \ldots, Y_{d-1} = y_{d-1})$.

Denote the marginal cumulative distribution function of Y_j, by F_j, for $j = 1, \ldots, d$. According to Sklar's Theorem [34] there exists a copula function C defined on $[0, 1]^d$ such that

$$P\{Y_1 \leq y_1, \ldots, Y_d \leq y_d\} = C(F_1(y_1), \ldots, F_d(y_d)) \qquad \forall (y_1, \ldots, y_d) \in \mathbb{R}^d .$$

In case the marginal distribution functions, $F_1, \ldots, F_d$, are continuous, the copula function C is unique. See [28]. The copula function *couples* the joint distribution function to its univariate margins $F_1, \ldots, F_d$. The dependence structure between the components of $(Y_1, \ldots, Y_d)$ is fully characterized by the copula function C.

Note that a copula function is nothing but a joint distribution function on $[0, 1]^d$ with uniform margins. Based on a joint distribution function, we can study conditional distribution functions derived from it, as well as characteristics of these (e.g. moments, medians, ...). Translated into the copula context this leads to various concepts for describing dependence structures.

The aim of this chapter is to first provide a review of copula modeling concepts, and statistical inference for them in various settings (parametric, semiparametric and nonparametric). This is done in a setting of independent data in Sects. 2 and 3. For simplicity of presentation, we restrict throughout the chapter to the setting of bivariate copulas (the case $d = 2$). In Sect. 4 we turn to the dynamical modeling of the dependence between time series, extending existing approaches of local polynomial fitting to this setting. We conclude the chapter by an illustration of the use of the method in a practical forecasting application in Sect. 5.

2 Global Dependencies and Unconditional Copulas

2.1 Population Concepts

Consider two random variables Y_1 and Y_2, with joint distribution function H, and continuous marginal distributions functions F_1 and F_2 respectively. There then exists

a (unique) bivariate copula function C, such that

$$H(y_1, y_2) = P\{Y_1 \le y_1, Y_2 \le y_2\} = C\left(F_1(y_1), F_2(y_2)\right) \qquad (y_1, y_2) \in \mathbb{R}^2 \,. \tag{1}$$

It is common practice to measure the strength of the relationship between Y_1 and Y_2 via a so-called association measure. There are various statistical association measures. Among the most well-known are: Pearson's correlation coefficient, Kendall's tau, Spearman's rho, Gini's coefficient, and Blomqvist's beta. See [4, 16, 17, 27] and [25], among others. Pearson's correlation coefficient, defined as $\mathrm{Cov}(Y_1, Y_2)/\sqrt{\mathrm{Var}(Y_1)\,\mathrm{Var}(Y_2)}$, only exists if the second order moments of both margins Y_1 and Y_2, exist, and equals ± 1 in case Y_2 is a (prefect) linear transform of Y_1. Gini's coefficient and Blomqvist's beta, are often used in economical sciences (for example, as a measure of inequality of income or wealth). Several well-known association measures can be expressed as functionals of the copula function. Denote by (Y_1', Y_2') and (Y_1'', Y_2''), two independent copies of (Y_1, Y_2). For the following association measures we give their definitions followed by an expression in terms of the copula function (for some measures alternative expressions in terms of C exist).

- Kendall's tau:

$$\begin{aligned} \tau_{Y_1,Y_2} &= P\left\{(Y_1 - Y_1')(Y_2 - Y_2') > 0\right\} - P\left\{(Y_1 - Y_1')(Y_2 - Y_2') < 0\right\} \\ &= 4 \iint_{[0,1]^2} C(u_1, u_2) dC(u_1, u_2) - 1 \,. \end{aligned} \tag{2}$$

- Spearman's rho:

$$\begin{aligned} \rho_{Y_1,Y_2} &= 3\left[P\left\{(Y_1 - Y_1')(Y_2 - Y_2'') > 0\right\} - P\left\{(Y_1 - Y_1')(Y_2 - Y_2'') < 0\right\}\right] \\ &= 12 \iint_{[0,1]^2} C(u_1, u_2) du_1 du_2 - 3. \end{aligned}$$

- Gini's coefficient:

$$\begin{aligned} \gamma_{Y_1,Y_2} &= 2E\left(|F_1(Y_1) + F_2(Y_2) - 1| - |F_1(Y_1) - F_2(Y_2)|\right) \\ &= 2 \iint_{[0,1]^2} \left(|u_1 + u_2 - 1| - |u_1 - u_2|\right) dC(u_1, u_2). \end{aligned}$$

- Blomqvist's beta:

$$\beta_{Y_1,Y_2} = 2P\left\{(Y_1 - F_1^{-1}(0.5))(Y_2 - F_2^{-1}(0.5)) > 0\right\} - 1 = 4C\left(\frac{1}{2}, \frac{1}{2}\right) - 1 \,,$$

where $F_j^{-1}(0.5)$ is the median of the margin F_j, for $j = 1, 2$.

See [28] for an overview of association measures, their basic properties, and their interrelationships.

When talking about unraveling the dependence structure between Y_1 and Y_2, it is transparent from (1) that informations about the copula function C as well as about the two margins F_1 and F_2 are of importance. We illustrate the impact of these elements with some examples in Sect. 2.2.

2.2 Illustration: Examples

The copula function and the marginal distributions together determine the joint distribution function (see (1)), and consequently all of the population characteristics. This is reflected in, among others, the typical observed scatter plots and the different values for the association measures.

As an illustration we consider the following examples:

Example 1.

$$Y_1 \sim N(0,4), \qquad Y_2 \sim \text{Exp}(2)$$
$$C(u_1,u_2) = \left(\max\left(u_1^{-\theta} + u_2^{-\theta} - 1, 0\right)\right)^{-\frac{1}{\theta}}, \qquad \theta = 1\,,$$

where the copula belongs to the Clayton copula family.

Example 2.

$$Y_1 \sim \text{Exp}(2), \qquad Y_2 \sim \text{Beta}(1,4)$$

and with the same copula as in Example 1.

Example 3.

$$Y_1 \sim \text{Student}(5,3), \qquad Y_2 \sim \text{Ex}(0.2)$$
$$C(u_1,u_2) = \exp\left(-\left((-\log u_1)^\theta + (-\log u_2)^\theta\right)^{\frac{1}{\theta}}\right), \qquad \theta = 3\,,$$

where Student(ν,μ) is a noncentral Student distribution with ν degrees of freedom and noncentrality parameter μ, and where the copula belongs to the Gumbel copula family.

Example 4.

$$Y_1 \sim \text{Exp}(2), \qquad Y_2 \sim \text{Beta}(1,4)$$

and the same copula as in Example 3.

In Fig. 1 we present a typical sample $((Y_{11}, Y_{21}), \ldots, (Y_{1n}, Y_{2n}))$ from (Y_1, Y_2) (with n the sample size) from the above models (left column of pictures) together with their pseudo-observations, defined as $(F_1(Y_{1i}), F_2(Y_{2i}))$, for $i = 1, \ldots, n$, (right column of pictures). Due to the probability integral transformation, the pseudo-observations $F_j(Y_{ji})$ are uniformly distributed (for each $j = 1, 2$). For Examples 1 and 2 the marginal distribution functions are different, but the dependence structure (the copula) is the same. In the left panels of rows one and two of Fig. 1 we depict scatter plots based on random samples of sizes 700 and 400 from, respectively, Examples 1 and 2. The scatter plots look quite different for the samples from the two models, but notice the similarity between the plots for the pseudo-observations (right panels). We can observe a mild positive dependence everywhere with a higher concentration in the lower tails (the lower left corner of the plots). In Example 3, both the margins and the copula are different, and the scatter plots based on a sample of size $n = 700$ from that model looks very different from the previous examples. Here, there is clearly more positive dependence visible, and we also notice heavier right tail characteristics. When comparing the plots for samples from Examples 3 and 4 (rows three and four in Fig. 1) we can observe similarities in the scatter plots of the pseudo-observations (the right panels), due to the fact that these examples share the same underlying copula. Furthermore, looking at scatter plots of typical observations from Examples 2 and 4, we can see the impact of changing a copula while keeping the same marginal distributions.

In Table 1 we present the values of some association measures for the four examples. Note the equality of the measures for Examples 1 and 2 on the one hand, and for Examples 3 and 4, on the other hand, since these examples share the same copula (i.e. have the same dependence structure). The values of the association measures only depend on the underlying copula function and not on the marginal distributions.

2.3 *Statistical Inference*

Suppose now that $((Y_{11}, Y_{21}), \ldots, (Y_{1n}, Y_{2n}))$ is a sample of size n of independent observations from (Y_1, Y_2), and the interest is in estimating the copula function in (1). Once an estimator for the copula function is available, the way is open to obtain estimates for association measures that can be expressed as functionals of the copula, as those for example listed in Sect. 2.1.

According to available information on either the copula and/or the margins we distinguish between different situations in the modeling aspects. In the fully parametric setting the copula function is assumed to be known, up to some parameters, and the same for the distribution of the margins. Other settings are listed in Table 2. We briefly review statistical inference under each of these settings.

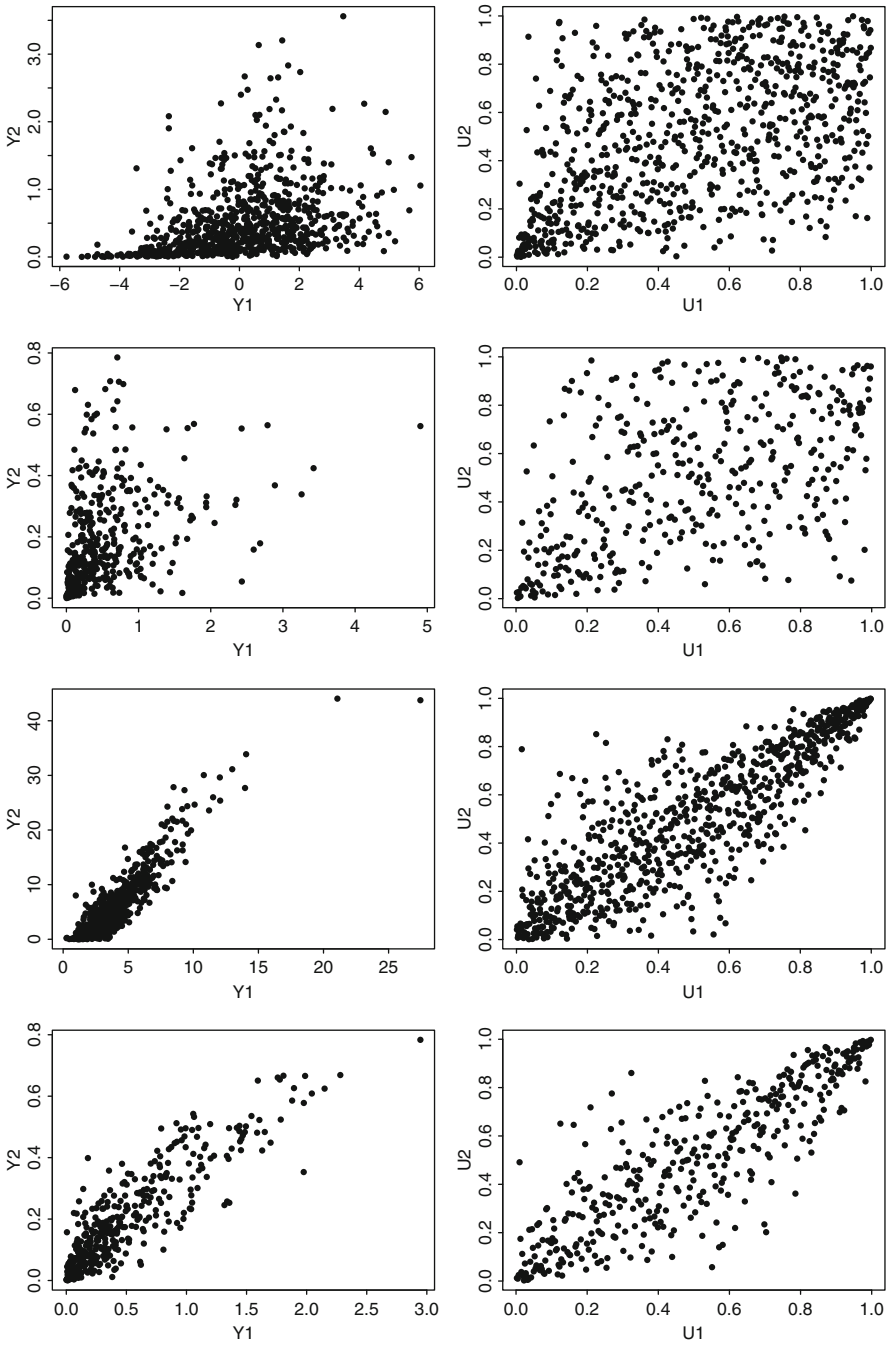


Fig. 1 Typical samples for Examples 1–4 (*top row* to *bottom row*); *left columns*: scatter plot of the observations, *right columns*: scatter plots of the pseudo-observations

Table 1 Association measures for Examples 1–4

Example	Kendall's tau	Spearman's rho	Gini's index	Blomqvist's beta
1 & 2	0.333	0.478	0.382	0.333
3 & 4	0.667	0.849	0.725	0.670

Table 2 Situations for statistical inference

Copula	Margins	Approach
Parametric	Parametric	Fully parametric
Parametric	Nonparametric	Semiparametric
Nonparametric	Parametric	Semiparametric
Nonparametric	Nonparametric	Fully nonparametric

2.3.1 Fully Parametric Approach

In a fully parametric approach one starts by assuming a specific parametric model for the copula function as well as for the margins. More formally, suppose that the copula $C(\cdot,\cdot) = C(\cdot,\cdot;\theta_C)$, and that $F_j(\cdot) = F(\cdot;\theta_j)$, for $j = 1, 2$, where $\boldsymbol{\theta}_C$, $\boldsymbol{\theta}_j$, for $j = 1, 2$ are the respective parameter vectors, taking values in parameter spaces Θ_C, Θ_1 and Θ_2 respectively. These parameters spaces can have nonempty intersections, in other words, the parameter vectors $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_j$ can have common elements.

Assume for simplicity that the density of the copula function exists, i.e. the second order partial derivative of the copula function exists

$$c(u_1, u_2; \boldsymbol{\theta}_C) = \frac{\partial^2 C(u_1, u_2; \boldsymbol{\theta}_C)}{\partial u_1 \partial u_2} \qquad \forall (u_1, u_2) \in [0, 1]^2 \, .$$

If in addition the corresponding densities f_j of F_j, for $j = 1, 2$, exist, then the joint density of (Y_1, Y_2) is given by (see (1))

$$h(y_1, y_2) = c\left(F_1(y_1), F_2(y_2)\right) f_1(y_1) f_2(y_2) \qquad \forall (y_1, y_2) \in \mathbb{R}^2 \, .$$

Keeping this in mind, the logarithm of the likelihood function then equals

$$\ell_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_C) = \sum_{i=1}^{n} \log\left(c(F_1(Y_{1i}; \boldsymbol{\theta}_1), F_2(Y_{2i}; \boldsymbol{\theta}_2); \boldsymbol{\theta}_C) f_1(Y_{1i}; \boldsymbol{\theta}_1) f_1(Y_{2i}; \boldsymbol{\theta}_2)\right) \, , \tag{3}$$

which needs to be maximized with respect to $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_C)$. Denote this maximizer by $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_C)$.

An estimator for, for example, the associated Kendall's tau is then obtained via (2)

$$\hat{\tau}_{Y_1, Y_2} = 4 \iint_{[0,1]^2} C(u_1, u_2; \hat{\boldsymbol{\theta}}_C) dC(u_1, u_2; \hat{\boldsymbol{\theta}}_C) - 1 \, .$$

2.3.2 Semiparametric Approach

Suppose now that the margins F_1 and F_2 cannot be parametrized, but are fully unknown. For the copula function, on the contrary, we still believe that a parametric model $C(\cdot, \cdot; \boldsymbol{\theta}_C)$ is a reasonable assumption. In this case, we thus need to estimate the margins from the available data. The log-likelihood in (3), where the margins F_1 and F_2 are unknown, is replaced by the pseudo log-likelihood

$$\ell_n(\theta_C) = \sum_{i=1}^{n} \log\left(c(F_{1n}(Y_{1i}), F_{2n}(Y_{2i}); \theta_C)\right) \ , \tag{4}$$

where the unknown margins F_1 and F_2 are replaced by the estimates

$$F_{1n}(y_1) = \frac{1}{n+1} \sum_{i=1}^{n} I\{Y_{1i} \leq y_1\} \qquad F_{2n}(y_2) = \frac{1}{n+1} \sum_{i=1}^{n} I\{Y_{2i} \leq y_2\} \ ,$$

with $I\{y \in A\}$ the indicator function on a set A, i.e. $I\{y \in A\} = 1$, if $y \in A$ and $I\{y \in A\} = 0$, if $y \notin A$. In the empirical estimates, it is recommended to use the modified factor $(n+1)^{-1}$ instead of the usual factor n^{-1}, because by using $(n+1)^{-1}$ the values $F_{jn}(Y_{ji})$ are in the set $\{\frac{1}{n+1}, \ldots, \frac{n}{n+1}\}$ instead of in the set $\{\frac{1}{n}, \ldots, \frac{n-1}{n}, 1\}$, and hence by using this modified factor, one stays away from both boundary points, 0 as well as 1. Note that this modification has no effect on the (asymptotic) properties of the resulting estimates. See, for example, [15].

The pseudo log-likelihood estimate $\hat{\boldsymbol{\theta}}_C$ of θ_C is then obtained by maximizing the pseudo log-likelihood in (4) with respect to $\boldsymbol{\theta}_C$.

See [33] for a study on semiparametric efficient estimation in case of Gaussian copulas with unknown margins.

2.3.3 Fully Nonparametric Approach

We now turn to the fully nonparametric approach where one can neither for C nor for the margins (F_1 and F_2) propose an appropriate parametric model. Hence C, F_1 and F_2 are fully unknown.

Nonparametric estimation of a copula goes back to the early seventies. In the paper [10] the empirical copula estimator was introduced and studied. The basic idea behind this estimator is very simple. As can be seen from (1), $C(\cdot, \cdot)$ is in fact nothing else but the joint cumulative distribution function after the margins have been transformed, via the probability integral transformation

$$U_1 = F_1(Y_1) \qquad \text{and} \qquad U_2 = F_2(Y_2) \ .$$

In other words, $C(\cdot, \cdot)$ is the joint cumulative distribution function of (U_1, U_2). If independent observations $((U_{11}, U_{21}), \cdots, (U_{1n}, U_{2n}))$ from (U_1, U_2) would be

available, then the cumulative distribution function could be estimated via the usual bivariate empirical distribution function $\frac{1}{n}\sum_{i=1}^{n} I\{U_{1i} \le u_1, U_{2i} \le u_2\}$. Since we have no observations from $F_1(Y_1)$ and $F_2(Y_2)$ we simply replace the unobserved $U_{ji} = F_j(Y_{ji})$ by a pseudo-observation, its 'slightly modified' rank $F_{nj}(Y_{ji})$ in the original sample, and obtain the empirical copula estimator

$$C_n(u_1, u_2) = \frac{1}{n}\sum_{i=1}^{n} I\left\{\tilde{U}_{1i} \le u_1, \tilde{U}_{2i} \le u_2\right\} \qquad \tilde{U}_{1i} = F_{1n}(Y_{1i}) \quad \tilde{U}_{2i} = F_{2n}(Y_{2i})\ . \tag{5}$$

This estimator was also studied further in [14] and [32], among others.

Obviously, the estimator in (5) is a step function, which might not be very desirable, when the function $C(\cdot, \cdot)$ is continuous or even differentiable.

Nonparametric estimation methods that lead to smooth estimators for the copula C have been derived. Among these are kernel estimators. See, for example, [6, 18] and [29]. Kernel estimators of the copula C are essentially obtained by replacing the non-smooth indicator function $I\{\tilde{U}_{1i} \le u_1, \tilde{U}_{2i} \le u_2\} = I\{\tilde{U}_{1i} \le u_1\}\, I\{\tilde{U}_{2i} \le u_2\}$ in (5) by a smooth kernel function. The non-smooth function $I\{\tilde{U}_{ji} \le u_j\}$ is replaced by the smooth function $K\left(\frac{u_j - \tilde{U}_{ji}}{h_n}\right)$, where $K(y) = \int_{-1}^{y} k(t)dt$ is the kernel distribution function associated with the kernel k, a symmetric density function, with support the interval $[-1, 1]$, and $h_n > 0$ is a bandwidth parameter. The bandwidth parameter determines the size of the neighbourhood in which the jump in the indicator function is 'smoothed out'. An example of a kernel function k is the Epanechnikov kernel $k(x) = \frac{3}{4}(1 - x^2)I\{|x| \le 1\}$. As an illustration we depict in Fig. 2 the indicator function $I\{0.6 \le u\}$ as well as a smooth version of it, namely $K\left(\frac{u-0.6}{h_n}\right)$, with K based on the Epanechnikov kernel, for two different values of the bandwidth h_n. The larger the bandwidth, the larger the neighbourhood over which the jump is 'smeared out'.

Such a simple replacement of the indicator part by a smooth part, leads to the kernel estimator

$$\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{u_1 - \tilde{U}_{1i}}{h_n}\right) K\left(\frac{u_2 - \tilde{U}_{2i}}{h_n}\right)\ .$$

Since the copula C is defined on the interval $[0, 1]^2$ (a compact support) special attention however is needed to obtain a kernel estimator that shows the same nice asymptotic properties at the boundaries of $[0, 1]^2$ as in the interior of that support. The aim is to obtain kernel estimators for which the convergence rate is the same on the interior of the square $[0, 1]^2$ as well as on the edges of it. This can be done for example, by using a reflection type of method, which consists of reflecting each pseudo-observation $(\tilde{U}_{1i}, \tilde{U}_{2i})$ with respect to all four corner points, and all four edges of the interval $[0, 1]^2$, resulting into an augmented data set of size $9n$, from

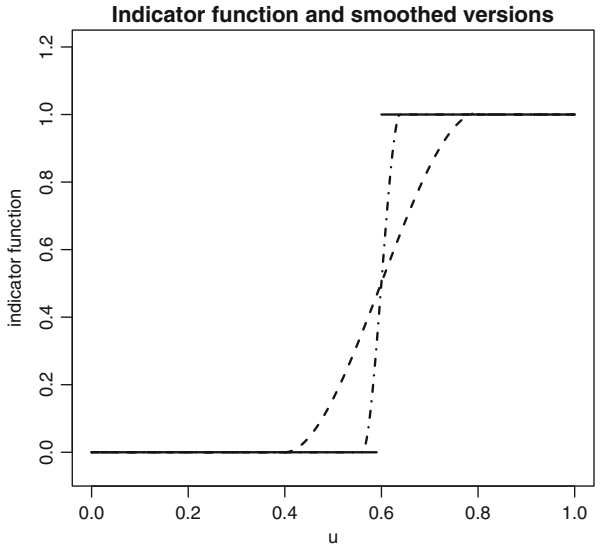


Fig. 2 The indicator function $I\{0.6 \leq u\}$ and its smooth version $K\left(\frac{u-0.6}{h_n}\right)$, using the Epanechnikov kernel, and bandwidths $h_n = 0.04$ (*dashed-dotted curve*) and $h_n = 0.20$ (*dashed curve*)

which the kernel estimator is defined. See Fig. 3 for an illustration of a data point and the eight points resulting from reflections of the given point with respect to all corners and edges of the unit square $[0, 1]^2$.

This leads to the kernel Mirror-Reflection type estimator introduced and studied in [18]:

$$\hat{C}_n^{\mathrm{MR}}(u_1, u_2) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{9} \left[K\left(\frac{u_1 - \tilde{U}_{1i}^{(\ell)}}{h_n} \right) - K\left(\frac{-\tilde{U}_{1i}^{(\ell)}}{h_n} \right) \right] \times \left[K\left(\frac{u_2 - \tilde{U}_{2i}^{(\ell)}}{h_n} \right) - K\left(\frac{-\tilde{U}_{2i}^{(\ell)}}{h_n} \right) \right],$$

where

$$\{(\tilde{U}_{1i}^{(\ell)}, \tilde{U}_{2i}^{(\ell)}),\ i = 1, \ldots, n,\ \ell = 1, \ldots, 9\}$$
$$= \{(\pm\tilde{U}_{1i}, \pm\tilde{U}_{2i}), (\pm\tilde{U}_{1i}, 2 - \tilde{U}_{2i}), (2 - \tilde{U}_{1i}, \pm\tilde{U}_{2i}), (2 - \tilde{U}_{1i}, 2 - \tilde{U}_{2i}),\ i = 1, \ldots, n\}.$$

and K is the integral of the considered kernel k, as mentioned above.

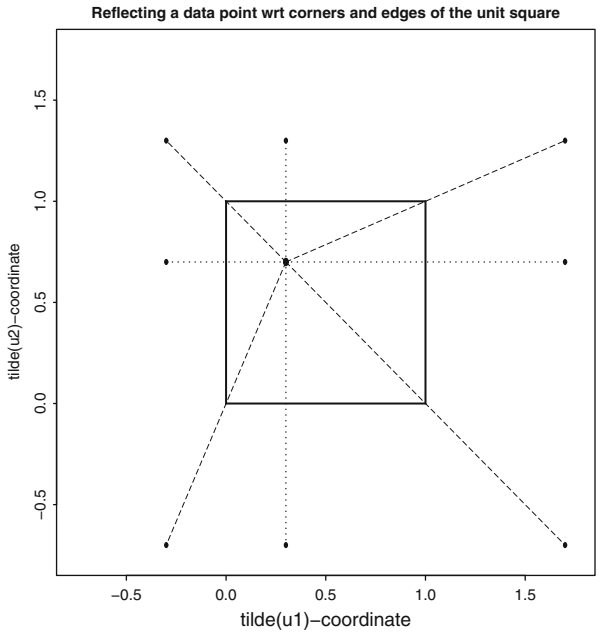


Fig. 3 A data point (indicated by a "•" in the unit square), and the reflected points (indicated by a "•") with respect to all corners (see the *dashed lines*) and edges (see the *dotted lines*) of the unit square

An alternative approach to deal with the boundary issue is by using local linear fitting, and its implicit boundary kernel. This was done by [6], and resulted in the kernel Local Linear estimator:

$$\hat{C}_n^{\mathrm{LL}}(u_1,u_2) = \frac{1}{n}\sum_{i=1}^{n} K_{u_1,h_n}\left(\frac{u_1-\tilde{U}_{1i}}{h_n}\right) K_{u_2,h_n}\left(\frac{u_2-\tilde{U}_{2i}}{h_n}\right),$$

where K_{u,h_n} is the integral of the modified boundary kernel

$$k_{u,h}(x) = \frac{k(x)\,(a_2(u,h)-a_1(u,h)x)}{a_0(u,h)a_2(u,h)-a_1^2(u,h)}\, I\{\frac{u-1}{h} < x < \frac{u}{h}\},$$

where

$$a_\ell(u,h) = \int_{\frac{u-1}{h}}^{\frac{u}{h}} t^\ell k(t)dt \quad \text{for } \ell = 0,1,2\,.$$

Note that nonparametric methods such as the above kernel methods, involve the choice of a bandwidth parameter. This issue is not discussed here. See for example [29] (Section 3.2 in that paper), and references therein, for some discussion on

bandwidth choice. One could also use other smoothing methods than the kernel method. We do not discuss these.

3 Local Dependencies and Conditional Copulas

3.1 Population Concepts

Suppose now that the interest is in the relationship between the random variables Y_1 and Y_2 but that both random variables are possibly influenced by another random variable, say X. A first interest is then in studying the conditional dependence between Y_1 and Y_2 given a specific value for X, say $X = x$. Instead of simply looking at the joint distribution between Y_1 and Y_2, as in Sect. 2, we now focus on the joint distribution of (Y_1, Y_2) *conditionally* upon $X = x$:

$$H_x(y_1, y_2) = P\{Y_1 \leq y_1, Y_2 \leq y_2 \,|X = x\} \,.$$

Applying Sklar's theorem to this conditional joint distribution function results into

$$H_x(y_1, y_2) = C_x\left(F_{1x}(y_1), F_{2x}(y_2)\right) \qquad (y_1, y_2) \in \mathbb{R}^2 \,, \tag{6}$$

where

$$F_{1x}(y_1) = P\{Y_1 \leq y_1 \,|X = x\} \qquad \text{and} \qquad F_{2x}(y_2) = P\{Y_2 \leq y_2 \,|X = x\} \,,$$

denote the marginal cumulative distributions functions of Y_1 and Y_2, respectively, *conditionally* upon $X = x$. The main difference between (1) and (6) is that the copula function C_x changes with the fixed value of X ($X = x$), as well as the margins F_{jx} (for $j = 1, 2$). We refer to C_x as the *conditional copula* function. This notion was first considered by [30] in the specific context of modeling the dynamics of exchange rates, where the conditioning variable is related to time. See also Sect. 4.

Analogously to the case of the (unconditional) copula C in Sect. 2, the strength of the dependence relationship between Y_1 and Y_2, but now *conditionally* upon the given value of $X = x$, can be measured using an association measure. For simplicity of presentation, we just focus on the Kendall's tau association measure. Denote by (Y_1', Y_2', X') an independent copy of (Y_1, Y_2, X). Then the *conditional* Kendall's tau *function* is defined as

$$\begin{aligned} \tau(x) &= P\left\{(Y_1 - Y_1')(Y_2 - Y_2') > 0 \,\middle|\, X = X' = x\right\} \\ &\qquad - P\left\{(Y_1 - Y_1')(Y_2 - Y_2') < 0 \,\middle|\, X = X' = x\right\} \\ &= 4 \iint_{[0,1]^2} C_x(u_1, u_2) dC_x(u_1, u_2) - 1 \,. \end{aligned} \tag{7}$$

For not making the notation too involved, we dropped the superscript $\{Y_1, Y_2\}$ to indicate that we are interested in the dependence structure between Y_1 and Y_2

(conditionally upon $X = x$). See for example [20], for illustrations and examples of use of conditional association measures.

Note from (6) that the conditional dependence of Y_1 and Y_2 (conditionally given $X = x$) may be different for different values taken by X, i.e. the dependence structure, as well as its strength, may change with the value taken by the third variable X. This possible change in the strength of the relationship is reflected in the fact that Kendall's tau is now a function of x.

It is often noticed that in applications, one uses the following simplification: the dependence structure itself, captured by the copula function, does not change with the specific value that X takes, and the dependence on x only comes in via the conditional margins, i.e.

$$H_x(y_1, y_2) = C\left(F_{1x}(y_1), F_{2x}(y_2)\right) \qquad (y_1, y_2) \in \mathbb{R}^2 \ .$$

In, for example, the literature on C-vine and D-vine copulas this assumption is inherently present. See [24] and [3], among others. See also the chapter (and its discussion) by [36], in which the dependence structure is assumed to stay constant in time.

3.2 Illustration: Examples

We now illustrate the concepts of Sect. 3.1 with some examples. A first example, Example 5, is an extension of Examples 1 and 2 of Sect. 2.2: instead of a bivariate Clayton copula we start from a three-variate Frank copula. In a second example, Example 6, we modify Example 3 of Sect. 2.2 by allowing the parameter of the copula and the marginal distribution of the second component to change with the random variable X. More precisely, the examples are as follows.

Example 5.

$$Y_1 \sim N(0,4), \qquad Y_2 \sim \text{Exp}(2), \qquad X \sim \text{Beta}(1,4)$$

$$C(u_1,u_2,u_3) = -\frac{1}{\theta}\log\left(1 + \frac{\left(e^{-\theta u_1}-1\right)\left(e^{-\theta u_2}-1\right)\left(e^{-\theta u_3}-1\right)}{\left(e^{-\theta}-1\right)^2}\right), \qquad \theta = 3 \ .$$

Example 6.

$$Y_1 \sim \text{Student}(5,3), \qquad Y_2|X \sim \text{Exp}(0.2(10X+1)), \qquad X \sim \text{U}(0,1)$$

$$C_x(u_1,u_2) = \exp\left(-\left((-\log u_1)^{\theta(x)} + (-\log u_2)^{\theta(x)}\right)^{\frac{1}{\theta(x)}}\right)$$

$$\theta(x) = 2\sin(2\Pi x) + 3 \ ,$$

with X independent from (Y_1, Y_2).

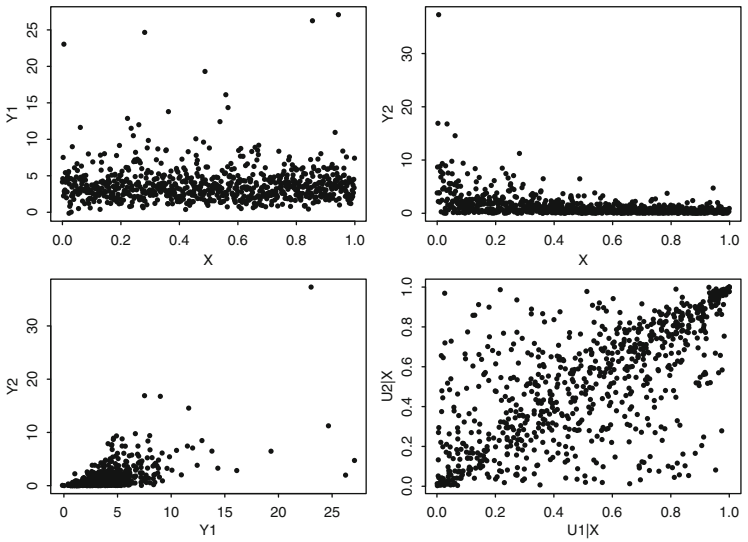


Fig. 4 Scatter plots based on a typical sample of size $n = 800$ from Example 6

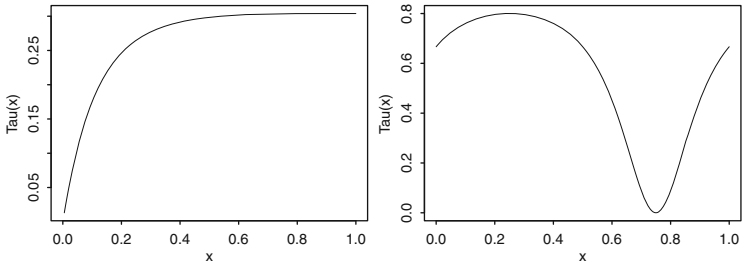


Fig. 5 Conditional Kendall's tau function for Examples 5 (*left panel*) and 6 (*right panel*)

In Fig. 4 we present the pairwise scatter plots for a typical sample of size $n = 800$ from Example 6, revealing the independence between Y_1 and X, but the dependence between Y_2 and X. In the bottom right panel of Fig. 4 we plot $F_{1X_i}(Y_{1i})$ versus $F_{2X_i}(Y_{2i})$, for each $i = 1, \ldots, n$. Each of these $F_{jX_i}(Y_{ji})$ should be (close to) uniformly distributed. In Fig. 5 we plot the conditional Kendall's tau function for Examples 5 and 6. Note that in Example 5 there is a mild positive dependence between Y_1 and Y_2, conditionally upon $X = x$, but that the dependence increases with x. In Example 6 the dependence switches from very positive via independence back to strongly positive dependence.

3.3 Statistical Inference

Suppose now that $((Y_{11}, Y_{21}, X_1), \ldots, (Y_{1n}, Y_{2n}, X_n))$ is a sample of size n of independent observations from (Y_1, Y_2, X). The interest is in estimating the conditional copula function C_x. As in Sect. 2, one can distinguish between different modeling settings, depending on what is known on possible appropriate parametric forms for C_x on the one hand and F_{jx} on the other hand. For convenience of the reader, we discuss similar modeling settings as in Sect. 2, but in a different order.

3.3.1 Fully Parametric Approach

In a fully parametric approach, we model $C_x(\cdot, \cdot)$ via $C(\cdot, \cdot; \theta_C(x))$ where $\theta_C(x)$ is a known parametric function of x, for example a polynomial of degree p: $\theta_C(x) = \theta_{C,1} + \theta_{C,2}x + \ldots \theta_{C,p}x^p$. We denote the corresponding parameter vector by $\boldsymbol{\theta}_C = (\theta_{C,1}, \ldots, \theta_{C,p})$. Similarly, the conditional margins can be modeled via

$$F_{jx}(\cdot) = F_j(\cdot; \theta_j(x)) \qquad \text{with a parametric function } \theta_j(x) \ .$$

For example, the functions $\theta_j(x)$ could be polynomial functions or any other given parametric functional form. Denote the resulting parameter vectors by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively. For example, if $\theta_1(x)$ is a cubic function of x, then the dimension of $\boldsymbol{\theta}_1$ is 4.

With these parametrizations, we are again in a setting that is quite similar to that of Sect. 2.3.1. Indeed, considering the second order partial derivative of $C_x(u_1, u_2)$ with respect to its arguments, we obtain

$$c_x(u_1, u_2) = \frac{\partial^2 C(u_1, u_2; \theta_C(x))}{\partial u_1 \partial u_2} \qquad \forall (u_1, u_2) \in [0, 1]^2 \ ,$$

which due to the structure can be written as $c(u_1, u_2; \theta_C(x))$. Analogously denote the marginal densities by $f_1(\cdot; \theta_1(x))$ and $f_2(\cdot; \theta_2(x))$.

A data point (Y_{1i}, Y_{2i}, X_i) contributes to the likelihood with the factor

$$c\left(F_1(Y_{1i}; \theta_1(X_i)), F_2(Y_{2i}; \theta_2(X_i)); \theta_C(X_i)\right) f_1(Y_{1i}; \theta_1(X_i)) f_2(Y_{2i}; \theta_2(X_i)) \ .$$

Finally, we get to the logarithm of the likelihood function

$$\begin{aligned} &\tilde{\ell}_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_C) \\ &= \sum_{i=1}^{n} \log\left(c\left(F_1(Y_{1i}; \theta_1(X_i)), F_2(Y_{2i}; \theta_2(X_i)); \theta_C(X_i)\right) f_1(Y_{1i}; \theta_1(X_i)) f_2(Y_{2i}; \theta_2(X_i))\right) \end{aligned}$$

$$= \sum_{i=1}^{n} \log \left(c \left(F_1(Y_{1i}; \theta_1(X_i)), F_2(Y_{2i}; \theta_2(X_i)); \theta_C(X_i) \right) \right)$$

$$+ \sum_{i=1}^{n} \log \left(f_1(Y_{1i}; \theta_1(X_i)) f_2(Y_{2i}; \theta_2(X_i)) \right) \ , \tag{8}$$

which needs to be maximized with respect to $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_C)$. From that point on we proceed as in Sect. 2.3.1. Denote the maximizer of (8) by $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_C)$.

An estimator for, for example, the associated conditional Kendall's tau function is then obtained by substituting $C_x(\cdot, \cdot) = C(\cdot, \cdot; \theta_C(x))$ in (7) by its estimator $C(\cdot, \cdot; \hat{\theta}_C(x))$, where $\hat{\theta}_C(x)$ is obtained by replacing the parameter vector $\boldsymbol{\theta}_C$ in the parametric form of $\theta_C(x)$ by its maximum likelihood estimator $\hat{\theta}_C$. For example, in case $\theta_C(x) = \theta_{C,1} + \theta_{C,2}x + \ldots \theta_{C,p}x^p$, this is $\hat{\theta}_C(x) = \hat{\theta}_{C,1} + \hat{\theta}_{C,2}x + \ldots \hat{\theta}_{C,p}x^p$ based on $\hat{\boldsymbol{\theta}}_C = (\hat{\theta}_{C,1}, \ldots, \hat{\theta}_{C,p})$. So, the estimator for the conditional Kendall's tau is then

$$\hat{\tau}(x) = 4 \iint_{[0,1]^2} C(u_1, u_2; \hat{\theta}_C(x)) dC(u_1, u_2; \hat{\theta}_C(x)) - 1 \ .$$

3.3.2 Fully Nonparametric Approach

An alternative expression for (6) is

$$C_x(u_1, u_2) = H_x \left(F_{1x}^{-1}(u_1), F_{2x}^{-1}(u_2) \right) \qquad (u_1, u_2) \in [0, 1]^2 \ , \tag{9}$$

where $F_{jx}^{-1}(\cdot)$ denotes the quantile function corresponding to $F_{jx}(\cdot)$, for $j = 1, 2$.

From (9) it is transparent that we need to find estimators for the conditional joint cumulative distribution function $H_x(\cdot, \cdot)$ as well as for the (quantiles of the) conditional margins $F_{1x}(\cdot)$ and $F_{2x}(\cdot)$. Since these are conditional quantities, some smoothing in the domain of X is needed. Nonparametric estimation of a conditional distribution function using kernel methods has been well-studied in the literature. See for example [23] and [38], among others. A general estimator is obtained by 'smearing out' the mass n^{-1} that is in the expression for a bivariate (unconditional) empirical distribution function, in the covariate domain, using a weight function:

$$\hat{H}_x(y_1, y_2) = \sum_{i=1}^{n} w_{ni}(x, b_n) I \{ Y_{1i} \leq y_1, Y_{2i} \leq y_2 \} \ , \tag{10}$$

with $w_{ni}(x, b_n) \geq 0$, a sequence of weights that 'smooths' over the covariate space. Herein $b_n > 0$ is a sequence of bandwidths. Since $H_x(y_1, y_2)$ is a distribution function, the weights need to tend to 1 when y_1 and y_2 tend to infinity. This is achieved by ensuring that the weights are such that $\sum_{i=1}^{n} w_{ni}(x, b_n) = 1$ (either

exactly or asymptotically, as $n \to \infty$). There are many scenario's of appropriate weight functions available in the literature. The simplest set of weights is given by the Nadaraya-Watson type of weights defined as

$$w_{ni}(x, b_n) = \frac{k_{b_n}(X_i - x)}{\sum_{j=1}^{n} k_{b_n}(X_j - x)} ,$$

with $k_{b_n}(\cdot) = \frac{1}{b_n} k(\cdot/b_n)$ a rescaled version of $k(\cdot)$. Alternative scenario's of weights are local linear weights, Gasser-Müller weights, etc.

From (10) we obtain estimators for the conditional marginal distribution functions $F_{jx}(\cdot)$ by simply letting the other argument in the estimated joint cumulative distribution function tend to infinity:

$$\hat{F}_{jx}(y) = \sum_{i=1}^{n} w_{ni}(x, b_{jn}) I\left\{Y_{ji} \le y\right\} \qquad j = 1, 2 , \tag{11}$$

where the bandwidth sequences $b_{1n} > 0$ and $b_{2n} > 0$ (for estimation of the conditional margins) do not need to be the same and/or do not need to be the same as this for the joint estimation. For practical simplicity one can take $b_n = b_{1n} = b_{2n}$.

Asymptotic properties for kernel type estimators of conditional distributions functions have been established in [35, 37] and [39], among others. For a recent contribution in the area, see [38].

Remark that in (10) and (11) one can again replace the indicator function by a smooth function, if differentiability properties of the resulting estimators are of importance.

From the estimators $\hat{F}_{jx}(\cdot)$ in (11), we obtain estimators for the quantile functions $F_{jx}^{-1}(\cdot)$. From the estimators for $H_x(\cdot, \cdot)$ and $F_{jx}^{-1}(\cdot)$ one then derives an estimator for $C_x(\cdot, \cdot)$ by replacing in (9) the former quantities by their estimators. In the literature these and improved estimators are studied, also in more complex frameworks (of multivariate or functional covariates). See, for example, [39] and [19].

3.3.3 Semiparametric Approach

There are at least a few semiparametric approaches, depending on the particular modeling setting. We just discuss some major approaches.

Firstly, assume that the conditional margins are fully known, i.e. $F_{jx}(\cdot)$, for $j = 1, 2$, are fully known for all x in the domain of X. Suppose that the conditional copula function $C_x(\cdot, \cdot)$ depends on x through a parameter function $\theta_C(x)$, i.e. $C_x(\cdot, \cdot) = C(\cdot, \cdot; \theta_C(x))$, but contrary to Sect. 3.3.1 the function $\theta_C(x)$ is fully unknown. This setting has been studied by [22] and [2], among others. In the sequel we drop the subscript C in θ_C and $\theta_C(x)$ for simplifying the notation.

The parametric copula family $C(\cdot, \cdot; \theta)$ that serves as a starting point here (and in fact also in Sect. 3.3.1, and before) of course has some restrictions on the parameter

space Θ for θ. For example, for a Gaussian copula: $\theta \in (-1, 1)$; for a Clayton copula $\theta \in (0, \infty)$. Such restrictions on the parameter space of the parametric copula $C(\cdot, \cdot; \theta)$ should in fact not be ignored. The same holds when looking at a conditional copula modeled via $C_x(u_1, u_2) = C(u_1, u_2; \theta(x))$, with corresponding copula density $c(u_1, u_2; \theta(x))$.

Since the function $\theta(x)$ is fully unknown, we are going to approximate this function locally, i.e. in the neighbourhood of x by, for example, a polynomial of degree p say. But since a polynomial takes on values in $\mathbb{R}$, we need to take care of the restrictions on the parameter space Θ of the parametric copula family $C(\cdot, \cdot; \theta)$. One therefore transforms the function $\theta(x)$, which takes on values in Θ, via a given transformation $\psi(\cdot)$, into the function

$$\eta(x) = \psi(\theta(x)) \ ,$$

which takes on value in $\mathbb{R}$.

In the sequel, we assume that the inverse transformation $\psi^{-1}(\cdot)$ exists, such that we can obtain $\theta(\cdot)$ from $\eta(\cdot)$:

$$\theta(x) = \psi^{-1}(\eta(x)) \ , \tag{12}$$

which takes values in Θ.

Consider now independent observations $((Y_{11}, Y_{21}, X_1), \ldots, (Y_{1n}, Y_{2n}, X_n))$ from (Y_1, Y_2, X). A data point (Y_{1i}, Y_{2i}, X_i) then contributes to the (pseudo) log-likelihood with

$$\log c\left(F_{1X_i}(Y_{1i}), F_{2X_i}(Y_{2i}); \theta(X_i)\right) = \log c\left(F_{1X_i}(Y_{1i}), F_{2X_i}(Y_{2i}); \psi^{-1}\left(\eta(X_i)\right)\right) \ . \tag{13}$$

For a data point X_i in a neighbourhood of x, we then can approximate $\eta(X_i)$ using a Taylor expansion, by

$$\begin{aligned} \eta(X_i) &\approx \eta(x) + \eta'(x)(X_i - x) + \cdots + \frac{\eta^{(p)}(x)(X_i - x)^p}{p!} \\ &\equiv \beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p \ , \end{aligned}$$

where we denoted

$$\beta_r = \beta_r(x) = \frac{\eta^{(r)}(x)}{r!} \qquad r = 0, \ldots, p \ .$$

If X_i is near x, then the contribution in (13) to the (pseudo) log-likelihood, can be approximated by

$$\begin{aligned} &\log c\left(F_{1X_i}(Y_{1i}), F_{2X_i}(Y_{2i}); \theta(X_i)\right) \\ &\quad \approx \log c\left(F_{1X_i}(Y_{1i}), F_{2X_i}(Y_{2i}); \psi^{-1}\left(\beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p\right)\right) \ . \end{aligned}$$

This approximation is only valid for X_i near x, and this is taken care off by multiplying this contribution in the log-likelihood by a weight factor $k_{h_n}(\cdot) = \frac{1}{h_n} k(\frac{\cdot}{h_n})$, with $k(\cdot)$ as before, and $h_n > 0$ a bandwidth parameter.

This leads to the local log-likelihood

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log c \left(F_{1X_i}(Y_{1i}), F_{2X_i}(Y_{2i}); \psi^{-1} \left\{\beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p\right\}\right) \times k_{h_n}(X_i - x) \,, \tag{14}$$

which is a localized version of the (pseudo) log-likelihood function in the parametric setting. Maximization of this local log-likelihood with respect to $\boldsymbol{\beta}$ leads to the estimated vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$, and hence, in particular, an estimator for $\eta(x)$ is $\hat{\beta}_0$. From (12) an estimator for $\theta(x)$ is

$$\hat{\theta}(x) = \psi^{-1}(\hat{\beta}_0) \,.$$

By maximizing the local log-likelihood (14) in a grid of x-values, one obtains estimates of the unknown parameter function $\theta(\cdot)$ in a grid of points.

We next turn to the setting where also the conditional margins are fully unknown, i.e. $F_{jx}(\cdot)$, for $j = 1, 2$, are fully unknown for all x in the domain of X. For X_i in a neighbourhood of x, we then can replace the contribution

$$\log c \left(F_{1X_i}(Y_{1i}), F_{2X_i}(Y_{2i}); \psi^{-1} \left\{\beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p\right\}\right) k_{h_n}(X_i - x)$$

in the log-likelihood function by

$$\log c \left(F_{1x}(Y_{1i}), F_{2x}(Y_{2i}); \psi^{-1} \left\{\beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p\right\}\right) k_{h_n}(X_i - x) \,,$$

and next we substitute the unknown quantities $F_{1x}(Y_{1i})$ and $F_{2x}(Y_{2i})$ by nonparametric estimators, such as these provided in (11). This then leads to the local log-likelihood

$$\tilde{\ell}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log c \left(\hat{F}_{1x}(Y_{1i}), \hat{F}_{2x}(Y_{2i}); \psi^{-1} \left\{\beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p\right\}\right) \times k_{h_n}(X_i - x) \,,$$

which needs to be maximized with respect to $\boldsymbol{\beta}$.

This estimation method is called a local polynomial maximum pseudo log-likelihood estimation method. Properties of the resulting estimator have been studied in [1]. That paper also contains a brief discussion on some practical bandwidth selection methods, including a rule-of-thumb type of bandwidth selector and a cross-validation procedure. For a general treatment of the use of local polynomial

modeling in a maximum likelihood framework, see for example [13]. For a general discussion on the choice of the degree of the polynomial approximation see [12].

4 Dynamics of a Dependence Structure and Copulas

When introducing copulas to the modeling of time series data different approaches are possible. Following, for example, [9] copulas can be used to model the inter-temporal dependence within one time series by specifying the transition probabilities in a Markov process. See [8] for recent developments and further references for such settings. In this section we focus on a different approach.

4.1 Dynamical Modeling of a Dependence Structure

Alternatively copulas can be used to model the spatial dependence of a bivariate stochastic process $\mathbf{Y}_t = (Y_{1,t}, Y_{2,t})$, $t \in \mathbb{Z}$. As time series analysis is naturally formulated conditionally upon the history of the process we revert to the conditional copula concept, now enlarging the conditioning from one variable as in Sect. 3 to the entire past of the process. This is done in a mathematical rigorously way by conditioning upon the sigma-algebras generated by the past of the time series. We opted for a more layman's term presentation here. As first introduced by [30] such a setting allows time dependent variation in the joint distribution of $(Y_{1,t}, Y_{2,t})$ conditionally upon $\mathbf{W}_t = (\mathbf{Y}_{t-k})_{k>0}$ via

$$P\{Y_{1,t} \le y_1, Y_{2,t} \le y_2 | \mathbf{W}_t = \mathbf{w}_t\} = C_t(F_{1,t}(y_1), F_{2,t}(y_2)) \,, \tag{15}$$

where $F_{j,t}(y_j) = P\{Y_{j,t} \le y_j | \mathbf{W}_t = \mathbf{w}_t\}, j = 1, 2$, and $C_t(\cdot, \cdot) = C(\cdot, \cdot | \mathbf{W}_t = \mathbf{w}_t)$ is the conditional copula implied by Sklar's Theorem.

The conditional modeling in (15) can be readily combined with, for example, a GARCH(r, s) error structure for the two involved time series $\{Y_{1,t}\}$ and $\{Y_{2,t}\}$. See [5], for example, for a standard reference to Generalized Autoregressive Conditional Heteroskedasticity (GARCH) type of modeling. More precisely, we model the marginal time series as

$$Y_{j,t} = \mu_j + \varepsilon_{j,t}, \qquad \text{where} \qquad \varepsilon_{j,t} = \sigma_{j,t}\eta_{j,t} \,, \tag{16}$$

$$\sigma_{j,t}^2 = \alpha_j + \sum_{\ell=1}^{r} \beta_{j,\ell}\sigma_{j,t-\ell}^2 + \sum_{m=1}^{s} \gamma_{j,m}\varepsilon_{j,t-m}^2 \,, \tag{17}$$

where $\alpha_j, \beta_{j,\ell}, \gamma_{j,m} \geq 0$, $\mu_j \in \mathbb{R}$, $r, s \in \mathbb{N}_0$ and $(\eta_{j,t})_{t\in\mathbb{Z}}$ is white noise with zero mean and unit variance. GARCH models are designed to account for the time varying and clustering volatility of shocks observed frequently, but not exclusively, in financial time series. This is accomplished by relating the time t variance of $\varepsilon_{j,t}$ to the lagged r realized variances as well as to the lagged s realized shocks, where it is noteworthy that the dependence on the lagged shocks makes the variance itself stochastic (random). Combining (15) and (16) the marginal time series can now be fused into a joint model by specifying its conditional joint distribution

$$\mathbf{Y}_t|\mathbf{W}_t = \mathbf{w}_t \sim C_t(F_{1,t}, F_{2,t}),$$

where the conditional mean and variance of $F_{j,t}$, for $j = 1, 2$, are only determined by information from the past (up to time point $t-1$) and are given as μ_j and $\sigma^2_{j,t}$. This framework combines autocorrelated shocks with a flexible modeling of the conditional joint distribution. A detailed review of copula models in economic time series can be found in [31]. For the purpose of this chapter, we focus on applying a semiparametric approach as discussed in Sect. 3.3.3 to the time series framework, with the difference that in the semiparametric approach described here, we model the marginal time series via parametric GARCH models. For simplicity, let $t \in \mathbb{N}_0$, and denote by $(Y_{1,t}, Y_{2,t})_{t=1}^T$ the available sample of size T ($T \in \mathbb{N}_0$). Furthermore, denote the observed (standardized) time points by t/T, so that all observational points t/T are in the interval $[0, 1]$. The conditional copula is chosen to be time dependent through an unknown parameter function $\theta_C(t^*)$ for $t^* \in [0, 1]$. To obey restrictions in the parameter space we again consider a suitable one-to-one transformation ψ such that $\eta(t^*) = \psi(\theta_C(t^*)) \in \mathbb{R}$ and recover θ_C via $\theta_C(t^*) = \psi^{-1}(\eta(t^*))$. For a sample $(Y_{1,t}, Y_{2,t})_{t=1}^T$ we can write the log-likelihood of the overall joint density by successive conditioning in terms of the contributions of the bivariate densities $\mathbf{Y}_t|\mathbf{W}_t = \mathbf{w}_t$ to the log-likelihood as

$$\begin{aligned}\ell_T &= \sum_{t=1}^{T} \log\left(c\left(F_{1,t}(Y_{1,t}), F_{2,t}(Y_{2,t})\right); \theta_C(t/T)\right) + \sum_{j=1}^{2}\sum_{t=1}^{T} \log\left(f_{j,t}(Y_{j,t})\right) \\ &= \ell_{T,C} + \ell_{T,1} + \ell_{T,2}\,.\end{aligned}$$

See also (8) in Sects. 3.3.1 and 3.3.3. Following the so-called inference of margins approach (a two-steps procedure), see [26], maximizing ℓ_T can be accomplished by first separately maximizing $\ell_{T,1}$ and $\ell_{T,2}$, under our semiparametric setting by a standard parametric maximum likelihood estimation method, and then maximizing $\ell_{T,C}$ taking estimates of the previous step into account by replacing the distribution functions $F_{j,t}$ by their respective estimates $\hat{F}_{j,t}$ (for $j = 1, 2$). In order to maximize $\ell_{T,C}$ we extend the local constant fitting approach of [22] to the local polynomial approach discussed in Sect. 3.3.2. Asymptotic normality of the resulting estimator in case of local constant fitting can be found in [22]. The study of the asymptotic properties of the local polynomial dynamic copula estimator, presented in this

section, is part of future research. Consider a fixed point $t^* \in [0,1]$, and denote by $k_{h_n}(\cdot)$ a rescaled kernel with bandwidth h_n, as before. The local log-likelihood for the problem considered is then given by

$$\begin{aligned}
&\ell_{T,C}(\boldsymbol{\beta}) \\
&= \sum_{t=1}^{T} \log c \left(\hat{F}_{1,t}(Y_{1,t}), \hat{F}_{2,t}(Y_{2,t}); \psi^{-1} \right. \\
&\left. \times \left\{ \beta_0 + \beta_1 \left(\frac{t}{T} - t^* \right) + \cdots + \beta_p \left(\frac{t}{T} - t^* \right)^p \right\} \right) \times k_{h_n} \left(\frac{t}{T} - t^* \right) , \qquad (18)
\end{aligned}$$

which needs to be maximized with respect to $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$.

4.2 Illustrative Example

We illustrate the presented methodology by simulating from a bivariate GARCH(1,1) model, where the conditional marginal distributions are set to be Normal distributions, and the conditional copula is assumed to be a student t copula where the degrees of freedom are fixed to 4. The remaining free copula parameter is a time varying parameter function $\theta_C(t) = 2\sin(0.95\pi(2B(t;2,3)-1)/6)$, with $B(t;a,b)$ the cumulative distribution function of the Beta distribution. See [11] for a reference on (student) t copulas. The remaining parameters are set to $\mu_1 = \mu_2 = 0$, $\alpha_1 = \alpha_2 = 0.1$ $\beta_{1,1} = \beta_{2,1} = 0.4$ and $\gamma_{1,1} = \gamma_{2,1} = 0.4$.

Figure 6 show the first and second marginal time series for a simulated trajectory of the process, highlighting the volatility clustering inherent in the process. In Fig. 7 we show simulated scatter plots of the unconditional distribution at three different time points. As expected the time varying conditional dependence structure carries

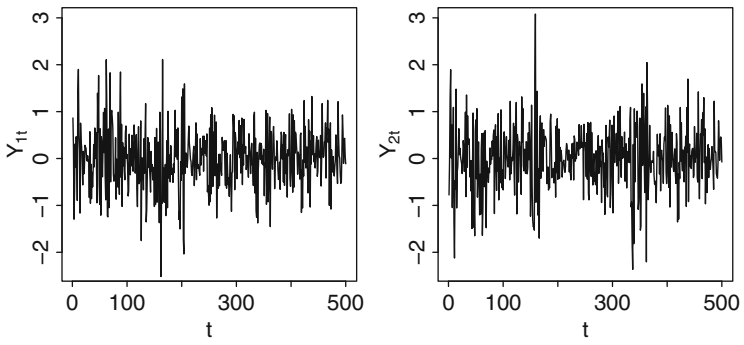


Fig. 6 Marginal time series $Y_{1,t}$ (*left*) and $Y_{2,t}$ (*right*) of a simulated bivariate copula-GARCH(1,1) model, $t = 1, \ldots, 500$

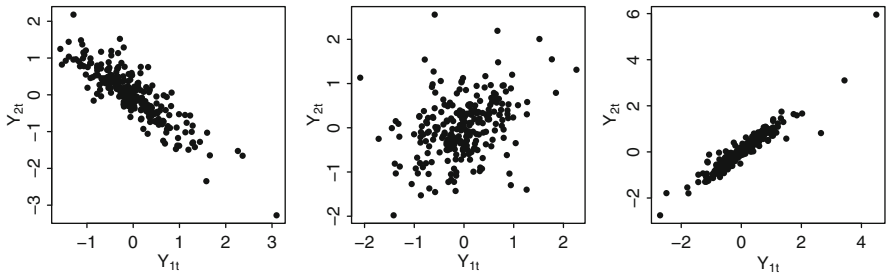


Fig. 7 Simulated scatter plot of $(Y_{1,50}, Y_{2,50})$ (*left panel*); $(Y_{1,250}, Y_{2,250})$ (*middle panel*) and $(Y_{1,450}, Y_{2,450})$ (*right panel*). The scatter plots are based on 250 independently simulated trajectories

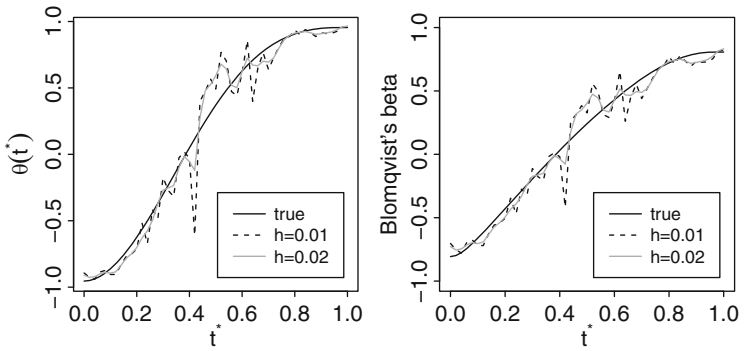


Fig. 8 True (*solid curve*) and estimated curve using bandwidths $h_n = 0.01$ (*dashed*) and $h_n = 0.02$ (*dotted*). *Left panel*: true and estimated $\theta(t^*)$. *Right panel*: true and estimated conditional Blomqvist's beta. Calculations are based on a simulated sample of size $T = 500$

over to the distribution of $(Y_{1,t}, Y_{2,t})$, displaying a negative dependence at early time stages, and gradually switching to a positive dependence later on (moving from the left panel to the right panel).

To illustrate the local log-likelihood estimation of $\theta(\cdot)$ we first obtain estimates $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_{1,1}$, $\hat{\beta}_{2,1}$, $\hat{\gamma}_{1,1}$ and $\hat{\gamma}_{2,1}$ by fitting a GARCH(1,1) model to each of the marginal time series individually. From the estimates we can then recover, for $j = 1, 2$, the conditional variances $\hat{\sigma}^2_{j,t}$ to find $\hat{F}_{j,t}(Y_{j,t}) = \Phi((Y_{j,t} - \hat{\mu}_j)/\hat{\sigma}_{j,t})$, where Φ denotes the standard normal distribution function. To perform the local log-likelihood estimation in (18) we settle for a local approximation with a polynomial of degree $p = 1$, i.e. performing local linear fitting. For a fixed point $t^* \in [0, 1]$ we then maximize $\ell_{T,C}(\beta_0, \beta_1)$ as given in (18), where we choose $\psi^{-1} : \mathbb{R} \to (-1, 1)$, with $\psi^{-1}(x) = \tanh(x)$.

In the left panel of Fig. 8 we show the true and estimated $\theta(t^*)$ when using different bandwidths h_n in the estimation procedure. In the right panel of Fig. 8 we also plot the true and estimated conditional Blomqvist's beta as a function of time, using the same bandwidths as for the estimation of $\theta(\cdot)$.

5 Dynamic Modeling via Copulas: Application in Forecasting

We consider wind speed data obtained by weather stations in Kennewick (southern Washington) and Vansycle/Butler Grade (north-eastern Oregon), in the USA. The raw data[1] consist of average wind speeds (in miles per hour = mph) for intervals of 5 min. For the analysis here we restrict to the period between April 1 and July 1, 2013. A more detailed description and analysis of these data can be found in [21]. In this section, we illustrate how the discussed methods can be used for estimation and for forecasting of wind speeds.

To fit the time series data we extend the model described in (16) and (17) to include an autoregressive component, leading to an AR(q)-GARCH(r, s) model, where (16) is replaced by (for $j = 1, 2$)

$$Y_{j,t} = \mu_j + \sum_{\ell=1}^{q} \phi_{j,\ell} Y_{j,t-\ell} + \varepsilon_{j,t} \,, \qquad \text{where} \qquad \varepsilon_{j,t} = \sigma_{j,t}\eta_{j,t} \,, \tag{19}$$

and (17) is kept. Herein $\phi_{j,\ell} \in \mathbb{R}$, $q \in \mathbb{N}_0$. As in Sect. 4 the conditional marginal distributions are Normal, and the marginal time series are coupled by a time dependent copula to form a bivariate model.

As wind speed forecasts with a two-hour forecast horizon are needed to meet practical demands (see [21]), we transform the raw data into the hourly averages, shown in Fig. 9, and denote the sample by $(Y_{1,t}, Y_{2,t})_{t=1}^{2,184}$. In the next step we fit an AR(1)-GARCH(1, 1) model, described by (19) and (17), to the marginal time series in an one hour rolling window type fashion as follows. The first estimates

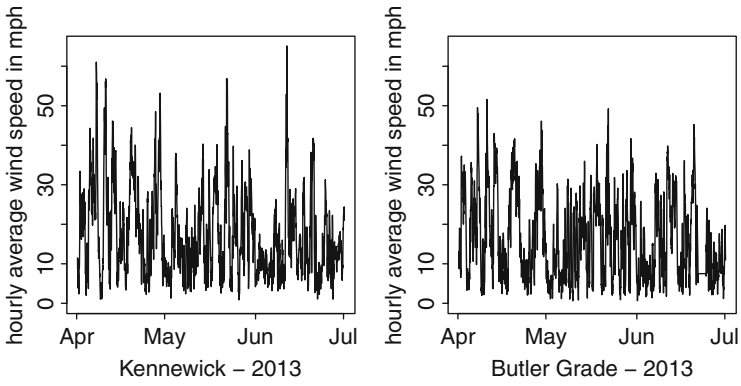


Fig. 9 Hourly averages of wind speed (mph) at Kennewick (*left*) and at Butler Grade (*right*) from April 1 to July 1, 2013

[1] Datasets can be obtained from the web site of the Bonneville Power Administration under http://transmission.bpa.gov/Business/Operations/Wind/MetData.aspx.

$(\hat{\mu}_j, \hat{\phi}_{j,1}, \hat{\alpha}_j, \hat{\beta}_{j,1}, \hat{\gamma}_{j,1})$, $j = 1, 2$, are obtained by fitting the AR-GARCH model to $(Y_{j,t})_{t=1}^{744}$. The second estimates are then based on the shifted data $(Y_{j,t})_{t=2}^{745}$ and so forth. We repeat this process 240 times, yielding estimates covering a span of 10 days. By keeping the number of observations fixed at 744 all estimates are effectively based on data of the last respective 31 days. The so obtained estimates are plotted against the window shift in Figs. 10 and 11. While the mean and AR coefficients (Fig. 10) are very comparable between both sites, the GARCH parameters show a differing pattern: while the baseline variance and lagged realized shock coefficients are generally higher in Kennewick than in Butler Grade (left and right panels of Fig. 11), the situation is reversed considering the dependence on lagged realized variances. See the middle panel of Fig. 11. The two weather stations are on different altitudes, the difference being around 130 m. This could may be explain the differences noticed.

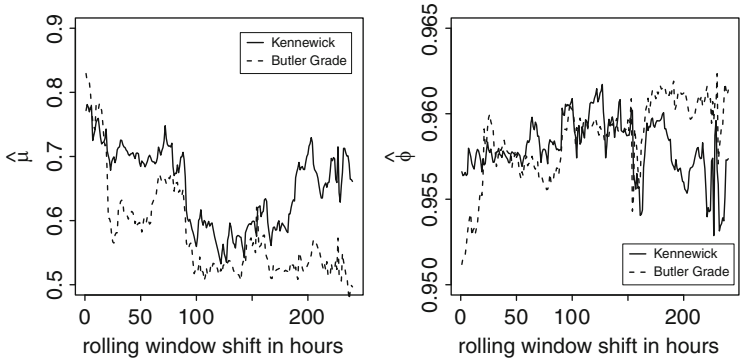


Fig. 10 Estimates of μ_j (*left panel*) and of $\phi_{j,1}$ (*right panel*), for $j = 1, 2$, for the 240 rolling windows, based on 744 observations each

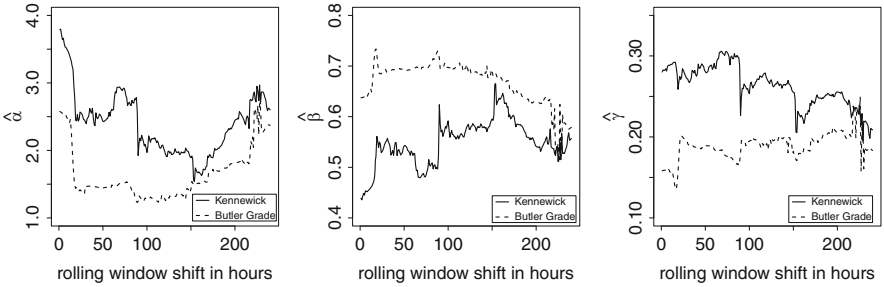


Fig. 11 Estimates of $\alpha_{j,1}$ (*left panel*), of $\beta_{j,1}$ (*middle panel*) and of $\gamma_{j,1}$ (*right panel*), for $j = 1, 2$, for the 240 rolling windows, based on 744 observations each

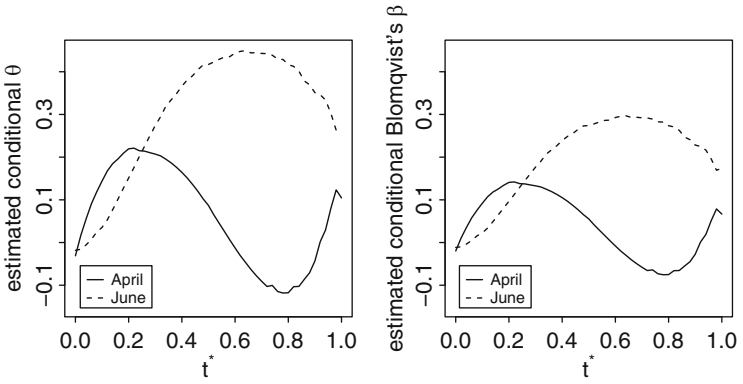


Fig. 12 Estimates of $\theta(t^*)$ (*left*) and the conditional Blomqvist's beta (*right*), by local-linear fitting with $h = 0.2$, for a student t copula with 4 degrees of freedom based on the observations $(Y_{1,t}, Y_{2,t})_{t=1}^{744}$ (for the April period) and on the observations $(Y_{1,t}, Y_{2,t})_{t=1,441}^{2,184}$ (for the June period). *Solid line*: April period; *Dotted line*: June period

The conditional dependence structure between the marginal time series is modeled by a student t copula with 4 fixed degrees of freedom, where the remaining parameter θ is allowed to vary as a smooth function of time, as explained in Sect. 4. In Fig. 12 (left panel) we show the estimated conditional copula parameter function for observations $(Y_{j,t})_{t=1}^{744}$ (the solid curve) using the previously estimated AR-GARCH parameters, and applying local linear fitting with bandwidth $h = 0.2$ (see Sect. 4). We repeat the procedure also based on the very last 744 observations $(Y_{j,t})_{t=1,441}^{2,184}$ (i.e. the June period) and show the results in the left panel of Fig. 12 (the dotted curve). We also present the corresponding results for the estimated conditional Blomqvist's beta in the right panel of Fig. 12. As can be seen, the dependence structure (between the observations from the two stations) varies within the periods (April and June), but also seems to be different for the two periods examined (early spring and summer period).

Turning towards forecasting we compute for each set of rolling window estimates the conditional copula parameter at $t^* = 1$, i.e. $\hat{\theta}(744)$ for the first window, $\hat{\theta}(745)$ for the second and so on. This yields successive estimates $(\hat{\beta}_0, \hat{\beta}_1)$ that we use to predict the one hour ahead forecast and, respectively, the two hours ahead forecast of $\theta(t)$ by $\tilde{\theta}(T + k) = \psi^{-1}(\hat{\beta}_0 + k\hat{\beta}_1/744)$, with $k = 1$, respectively $k = 2$, where ψ^{-1} is the link function as in Sect. 4 and T denotes the last time point of the respective rolling window sample. The obtained 2 h ahead forecast of θ is shown in the left panel of Fig. 13.

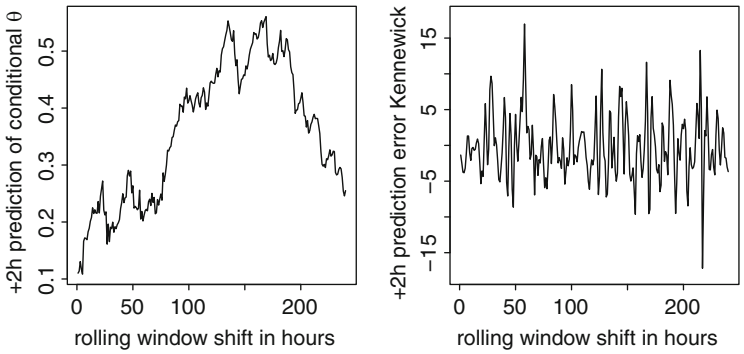


Fig. 13 *Left*: Two hours ahead prediction of $\theta(\cdot)$ for the rolling windows, based on 744 observations. *Right*: Realized minus two hours ahead predicted wind speeds in Kennewick for the rolling windows

Concerning the marginal time series we use the point estimates of the parameters as one and two hours ahead forecasts of the parameters. This allows to compute point forecasts of the wind speed by (see (19))

$$\tilde{Y}_{j,T+1} = \mathrm{E}\left[Y_{j,T+1}|Y_{j,T-\ell}, \ell \geq 0\right] = \mu_j + \phi_{j,1}\, Y_{j,T}$$
$$\tilde{Y}_{j,T+2} = \mathrm{E}\left[Y_{j,T+2}|Y_{j,T-\ell}, \ell \geq 0\right] = \mu_j + \phi_{j,1}\, \left(\mu_j + \phi_{j,1} Y_{j,T}\right)$$

for $j = 1, 2$.

To assess the forecast quality we compute the square root of the mean squared error of the predicted to the realized values for the 240 forecasts, denoted by RMSE. For one hours ahead forecasts we obtain: RMSE = 2.9485 for the Kennewick station, and RMSE = 3.0026 for the Butler Grade station. For the two hours ahead forecasts these are: RMSE = 4.4725 for Kennewick and RMSE = 4.8190 for Butler Grade. The right panel of Fig. 13 depicts the differences of realized to predicted two hours ahead forecasts at Kennewick.

Having predictions of the marginal time series, as well as the conditional dependence between them allows to go beyond point forecasting and to predict their joint behaviour. In Fig. 14 we show contours of the predicted two hours ahead joint distribution for the first rolling window, and then 5.5 days later, in respectively the left and right side panels. As shown by the figure, not only the mean of the distribution, represented by the wind speed point forecasts, but also the shape changes as implied by the prediction of $\theta(\cdot)$.

To further visualize the impact of the time varying association we forecast the probability of $Y_{1,T+2}$ and $Y_{2,T+2}$ staying jointly below their conditional $\alpha \times 100\,\%$ quantiles. Within a copula framework this probability equals $C(\alpha, \alpha; \theta(T + 2))$. See [7] for an application in economics. For example based on the first rolling

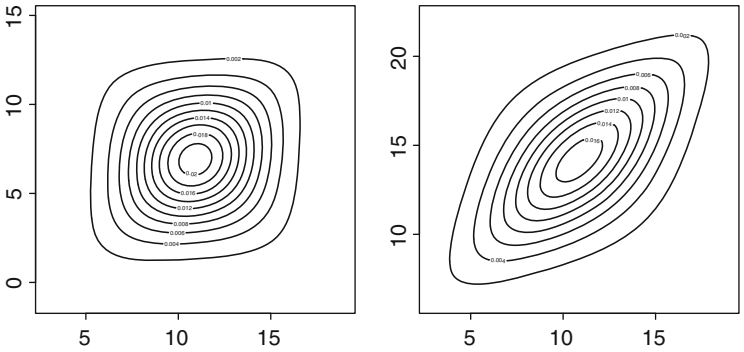


Fig. 14 Two hours ahead prediction of the conditional joint density of $(Y_{1,746}, Y_{2,746})$ (*left*) and the conditional joint density of $(Y_{1,877}, Y_{2,877})$ (*right*) based on the first rolling window, respectively on rolling window 132

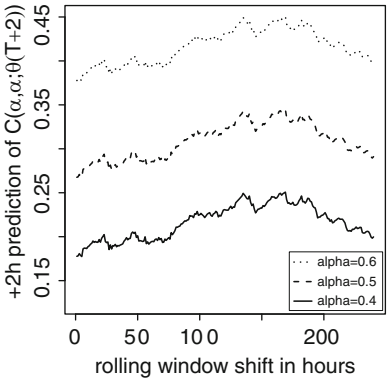


Fig. 15 Two hours ahead prediction of $C(\alpha, \alpha; \theta(T+2))$ for the rolling windows, based on 744 observations

window we compute the conditional two hours ahead 40 % quantiles as 10.1487 (for Kennewick) and 6.1633 (for Butler Grade). This yields a prediction of $P\{Y_{1,746} \leq 10.1487, Y_{2,746} \leq 6.1633\} = 0.1777$. Figure 15 shows the obtained predictions for different values of α.

Acknowledgements The authors thank the organizers of the "Second workshop on Industry Practices for Forecasting" (WIPFOR 2013) for a very simulating meeting. This research is supported by IAP Research Network P7/06 of the Belgian State (Belgian Science Policy), and the project GOA/12/014 of the KU Leuven Research Fund. The third author is Postdoctoral Fellow of the Research Foundation – Flanders, and acknowledges support from the foundation.

References

1. Abegaz, F., Gijbels, I., & Veraverbeke, N. (2012). Semiparametric estimation of conditional copulas. *Journal of Multivariate Analysis, Special Issue on "Copula Modeling and Dependence", 110*, 43–73.
2. Acar, E. F., Craiu, R. V., & Yao, F. (2011). Dependence calibration in conditional copulas: A nonparametric approach. *Biometrics, 67*, 445–453.
3. Acar, E. F., Genest, C., & Nešlehová, J. (2012). Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis, Special Issue on "Copula Modeling and Dependence", 110*, 74–90.
4. Blomqvist, N. (1950). On a measure of dependence between two random variables. *The Annals of Mathematical Statistics, 21*, 593–600.
5. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*, 307–327.
6. Chen, S. C., & Huang, T.-M. (2007). Nonparametric estimation of copula functions for dependence modelling. *The Canadian Journal of Statistics, 35*, 265–282.
7. Cherubini, U., & Luciano, E. (2001). Value-at-risk trade-off and capital allocation with copulas. *Economic Notes, 30*, 235–256.
8. Cherubini, U., Mulinacci, S., Gobbi, F., & Romagnoli S. (2011). *Dynamic copula methods in finance*. New York: Wiley.
9. Darsow, W. F., Nguyen, B., & Olsen, E. T. (1992). Copulas and Markov processes. *Illinois Journal of Mathematics, 36*, 600–642.
10. Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. *Académie Royale de Belgique, Bulletin de la Classe des Sciences, 5e Série, 65*, 274–292.
11. Demarta, S., & McNeil, A. J. (2005). The t copula and related copulas. *International Statistical Review, 73*, 111–129.
12. Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.
13. Fan, J., Farmen, M., & Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society, Series B, 60*, 591–608.
14. Fermanian, J.-D., Radulović, D., & Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli, 10*, 847–860.
15. Genest, C., Ghoudi, K., & Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika, 82*, 543–552.
16. Gini, C. (1909). Concentration and dependency ratios (in Italian). English translation in *Rivista di Politica Economica, 87*, 769–789 (1997).
17. Gini, C. (1912). Variability and mutability (in Italian, 156 p.). Bologna: C. Cuppini. (Reprinted in E. Pizetti & T. Salvemini (Eds.), *Memorie di metodologica statistica*. Rome: Libreria Eredi Virgilio Veschi (1955))
18. Gijbels, I., & Mielniczuk, J. (1990). Estimating the density of a copula function. *Communications in Statistics – Theory and Methods, 19*, 445–464.
19. Gijbels, I., Omelka, M., & Veraverbeke, N. (2012). Multivariate and functional covariates and conditional copulas. *Electronic Journal of Statistics, 6*, 1273–1306.
20. Gijbels, I., Veraverbeke, N., & Omelka, M. (2011). Conditional copulas, association measures and their applications. *Computational Statistics & Data Analysis, 55*, 1919–1932.
21. Gneiting, T., Larson, K., Westrick, K., Genton, M., & Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching-space-time method. *Journal of the American Statistical Society, 101*, 968–979.
22. Hafner, C., & Reznikova, O. (2010). Efficient estimation of a semiparametric dynamic copula model. *Computational Statistics & Data Analysis, 54*, 2609–2627.
23. Hall, P., Wolff, R. C. L., & Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association, 94*, 154–163.

24. Hobæk Haff, I., Aas, K., & Frigessi, A. (2010). On the simplified pair-copula construction – Simply useful or too simplistic? *Journal of Multivariate Analysis, 101*, 1296–1310.
25. Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.). New York: Wiley.
26. Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis, 94*, 401–419.
27. Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association, 53*, 814–861.
28. Nelsen, R. B. (2006). *An introduction to copulas* (Lecture notes in statistics, 2nd ed.). New York: Springer.
29. Omelka, M., Gijbels, I., & Veraverbeke, N. (2009). Improved kernel estimation of copulas: Weak convergence and goodness-of-fit testing. *The Annals of Statistics, 37*, 3023–3058.
30. Patton, A. (2006). Modeling asymmetric exchange rate dependence. *International Economical Review, 47*, 527–556.
31. Patton, A. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis, 110*, 4–18.
32. Segers, J. (2012). Weak convergence of empirical copula processes under nonrestrictive smoothness assumptions. *Bernoulli, 18*, 764–782.
33. Segers, J., van den Akker, R., & Werker, B. J. M. (2014). Semiparametric gaussian copula models: Geometry and efificent rank-based estimation. *Annals of Statistics, 42*(5), 1911–1940.
34. Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de L'Institut de Statistique de L'Université de Paris, 8*, 229–231.
35. Stute, W. (1986). On almost sure convergence of conditional empirical distribution functions. *The Annals of Statistics, 14*, 891–901.
36. Tatsu, J., Pinson, P., Madsen, H. (2015). Space-time trajectories of wind power generation: Parametrized precision matrices under a Gaussian copula approach. *Lecture Notes in Statistics 217: Modeling and Stochastic Learning for Forecasting in High Dimension*, 267–296.
37. Van Keilegom, I., & Veraverbeke, N. (1997). Estimation and bootstrap with censored data in fixed design nonparametric regression. *The Annals of the Institute of Statistical Mathematics, 49*, 467–491.
38. Veraverbeke, N., Gijbels, I., & Omelka, M. (2014). Pre-adjusted nonparametric estimation of a conditional distribution function. *Journal of the Royal Statistical Society, Series B, 76*, 399–438.
39. Veraverbeke, N., Omelka, M., & Gijbels, I. (2011). Estimation of a conditional copula and association measures. *The Scandinavian Journal of Statistics, 38*, 766–780.

Online Residential Demand Reduction Estimation Through Control Group Selection

Leslie Hatton, Philippe Charpentier, and Eric Matzner-Løber

Abstract Demand response levers, as tariff incentive or direct load control on residential electrical appliances, are potential solutions to efficiently manage peak consumption and aid in grid security. The major objective is to estimate the consumption that would have been used in the absence of demand reduction: the baseline. This is an important issue to enhance demand response for electricity markets and to allow the grid operators to efficiently manage the grid. For these reasons, baseline estimation methods have to satisfy the following operational objectives: highly accurate, computationally efficient, cost-effective and flexible to the demand response customer turnover. In general, methods using available data from the control group give the best results, but current control group methods do not satisfy the aforementioned operational objectives. Having a real control group is highly costly because it requires to meter thousands customers who will not be used in the demand response offer. So there is a need to find new methods to select a control group. The advancement of smart meters can now provide a wealth of data to construct this group. This paper proposes the use of individual smart meter loads to select a control group. The method satisfies the aforementioned operational objectives since the selected control group is adaptable in operations even to demand response customers changes. The methodology developed is based on a selection algorithm and constraint regression approach. These new methods have been successfully tested in an online environment.

L. Hatton (✉)
EDF R&D, Dept. ICAME and agrocampus Rennes, Rennes, France
e-mail: hatton@agrocampus-ouest.fr

P. Charpentier
EDF R&D, Dept. ICAME, 92140 Clamart, France
e-mail: philippe.charpentier@edf.fr

E. Matzner-Løber
Universite Rennes 2, 35000 Rennes, France
e-mail: eml@uhb.fr

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_8

1 Introduction

In many countries, electricity demand is increasing globally every year but the peak demand is increasing much faster. Classically, peak demand occurs on hot summer afternoons, as for example in North America due to air conditioner, or in cold winter evenings, as for example in France due to electric heating (Fig. 1) when customers return back home and start household appliances (cooking, information technology devices), heating and lighting.

Demand response (DR) is a mechanism involving the demand side in the electric grid management. It consists in appealing to the customer so that he adjusts his electric consumption in order to meet economic issues (increasing prices) or grid security (disequilibrium between supply and demand). This customer's action is then integrated on the whole operations aiming at balancing the grid at every time. The consumption modulations mainly consist in reducing, shifting or shedding the consumption on a given time period. There are two kinds of programs inciting customers to adapt their consumption: *tariff incentive* and *direct load control*.

Tariff incentive was the first tool used to shift or reduce the customers' consumption. Electricité De France (EDF) is a pioneer in this kind of tariffs since it introduced the peak and off-peak hours rate in 1965 to shift the consumption on the daily off-peak periods, generally the midday and night hours. To control the seasonal variations and particularly the winter peak demand, the EJP rate ("Effacement Jour de Pointe") was introduced in the eighties. The EJP customers

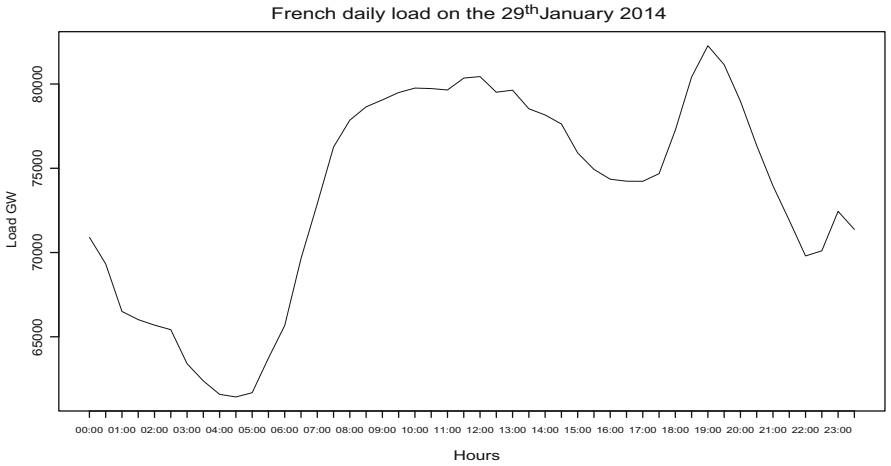


Fig. 1 Daily French electricity consumption. When professional activities end and customers return back home, the consumption highly increases. On this winter day, the peak demand is obviously perceptible at 7 pm

pay 0.74 times the basic electricity price during the year except for 22 critical winter days determined according to weather forecasts, where they pay 3.63 times the basic price. Then, in the middle of the nineties, the TEMPO rate was introduced and defines three kinds of days: blue, white and red. During 300 blue (respectively 43 white and 22 red) days, the customers pay 0.66 (resp. 0.93 and 3.7) times the basic. EJP and TEMPO rates allow to smooth the daily consumption. They enable the customer to be active towards his consumption by inciting him to shift his peak consumption. Customers create demand reduction capacities by substantially reducing their electricity consumption.

Direct load control consists in interrupting the consumption of an electric appliance on a given time period. Currently, the direct load control mainly concerns industrial customers which have available significant demand reduction capacities. Indeed, they can momentarily interrupt their process or call on generators.

DR is a flexible mean to adjust demand according to supply. Using tariff incentive or direct load control, the obtained result is a reduction or an increase of the demand. Some companies aim at inciting their customers to increase their consumption when the renewable energy production is high. We only focus here on situations where the supplied capacity is a reduction, a suppression or a substitution of the customer demand leading to a demand reduction. As residential electric heating and air conditioner are flexible usages and partly responsible for the peak demand, many countries around the world experimented in the last decade new DR programs on residential customers: dynamic pricing [4] or direct load control [5].

To enhance the residential demand reduction on electricity markets, it has to be evaluated. It is impossible to measure the reduction, the only available measure is the consumption obtained with the DR program. To quantify the curtailment, one has to estimate the consumption which would have been used in the absence of the DR program, called the baseline. The curtailment is obtained by difference between the baseline and the metered load during the DR event (Fig. 2).

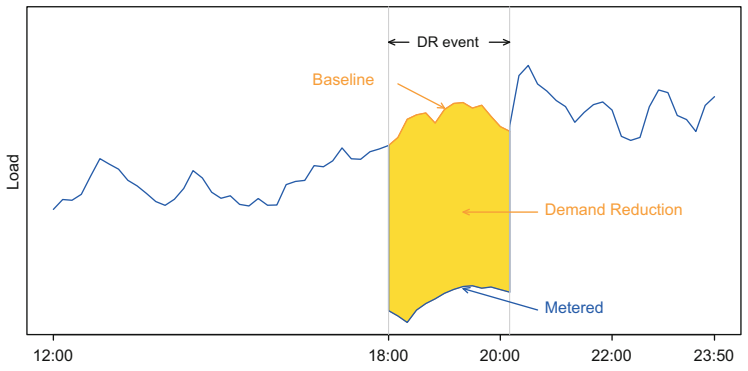


Fig. 2 Representation of the average load curve of a DR customers group with a DR event between 6 and 8 pm. The metered load is the *dark line* and the estimated baseline the *light line*. The demand reduction is the difference between these two loads, corresponding to the *coloured area*

Baseline estimators have to satisfy different objectives. On one hand, they have to be operational, accurate (in order to remunerate the customers), cost-effective, optimal and flexible with time and with the DR customers group size. Indeed, when DR offers will roll-out on the residential market, the DR group size will vary with time because of the customers entrance or leaving. This is already the case in the experiments. Moreover, as the number of customers participating in experiments is evolving, methods have to be able to estimate demand reduction on thousands customers. This will be the case in the Smart Electric Lyon[1] (SEL) project where 25,000 customers will experiment DR. On the other hand, as demand reductions can be enhanced on the electricity markets, they could allow to maintain the grid balanced during peak consumption. Consequently, for the grid operators who supervise the balance online and in order to rapidly react, it could be interesting to control the demand reduction effectiveness and to quantify it online. Baseline estimations are then expected to be evaluated in real time.

Many baseline methods are issued from experiments. In general, methods using available data from control group give the best results [1]. When no control group was specified during the experiment, marginal calibration method [2] can be employed to select one from individuals available in a database. This method consists in calculating weights from individual characteristics. Variables as address, tariff plan, subscript power, are classical informations collected. To be accurate, the method requires a set of individual characteristics having an important impact on the consumption (house and heating surface, occupants number and their presence, details on electric appliances, ...). Collecting these additional data on thousands customers is a costly process. Moreover, the weights are calculated from invariant individual characteristics while the consumption is highly weather dependent and fluctuates from one day to the next. These weights being fix, the control group does not evolve and the method is not flexible for a daily estimation.

Operators have residential portfolio database with individual characteristics (such as address, tariff plan ...) and possibly individual loads. With the roll-out of smart meters, individual loads will be available. Electricity loads largely reflect the customers behaviours and it is possible to use this information to select a control group. In addition, individual covariates can be used to restrict the geographic area.

In this paper, we propose to build control group based on loads' shape matching. It consists in selecting individual loads without demand reduction from the datamart such that the distance between their average load and that of the DR group is

[1]The Smart Electric Lyon project, launched in 2012, is an ADEME project, the French energy control and environment agency. The program aims at offering technical solutions and testing tariff incentives on thousands residential and services customers. http://www.smart-electric-lyon.fr/

minimal. This method is exposed in Sect. 2. The Sect. 3, illustrated by a real application, describes how this solution is operational and allows to estimate the demand reduction online with thousands customers. We conclude in Sect. 4.

2 Control Group Selection Methods

Before introducing the methods, let us consider the following notations. The individual load curve which is a time series is $P_j(t)$. Even so, it is much proper to index it through the day and the hour of the day: $P_j(d,h)$. So, if the entire day is considered, we will simplify the notation by $P_j(d)$. We evaluate the average load curve of the DR program by

$$P^{DR}(d,h) = \frac{1}{n}\sum_{j=1}^{n} P_j^{DR}(d,h)$$

where $P_j^{DR}(d,h)$ is the load curve of the individual j among the n customers participating in the DR program (example on Fig. 3).

The operator managing the DR program has obviously others customers and by the way their loads recorded in a datamart (eventually the same datamart containing the DR loads). All these customers are potentially useful for being used as "control". The objective is to select individuals from the datamart having load curve denoted

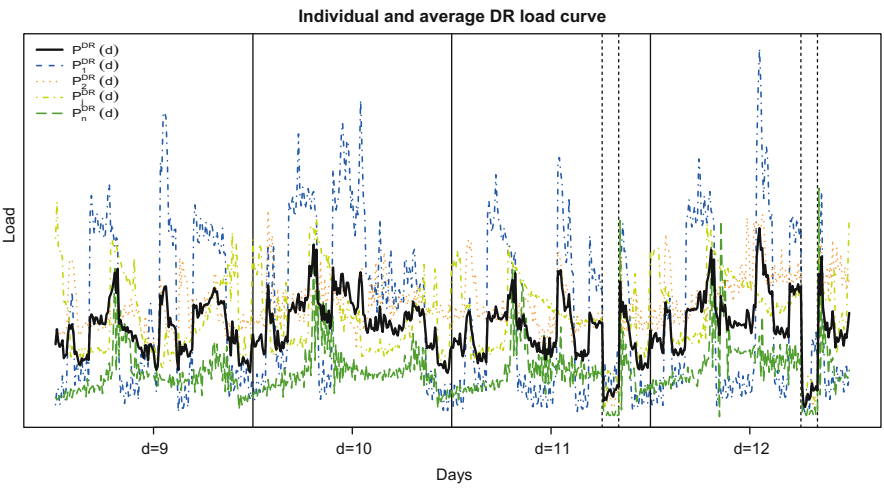


Fig. 3 The individual 4-days DR load curves are represented by different *dotted* and *dashed lines* and show different variabilities. The *bold solid line* characterizes the average load curve of the DR group. Demand reduction events occur between 6 and 8 pm on the days 11 and 12

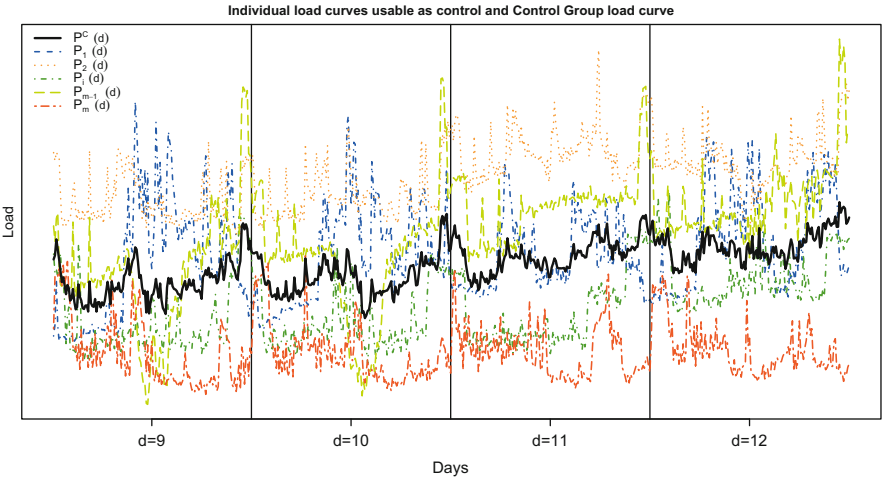


Fig. 4 Some individual non-DR load curves are represented by the different *dotted* and *dashed lines*. The *bold solid line* characterizes the control group load curve obtained by averaging all the individual load curves

$P_i(d,h)$ such that the average curve of the selected individuals

$$P^C(d,h) = \frac{1}{\#selected} \sum_{i \in selected} P_i(d,h)$$

is the "closest" to $P^{DR}(d,h)$. We will discuss the notion of closeness later on.

The data are collected everyday at midnight, making available the complete day, we then consider the whole day and simplify the notations. So, the individual load curve of the n individuals belonging to the DR program is $P_j^{DR}(d)$ and the average load curve $P^{DR}(d) = \frac{1}{n}\sum_{j=1}^{n} P_j^{DR}(d)$. The m individual load curves in the datamart which could be partially used as "control" are noted $P_i(d)$. We aim at constructing control group load curve $P^C(d)$ to estimate the baseline on event days (example on Fig. 4).

Suppose that the curtailment occurs at 6 pm of day d we would like to select individuals whom the average load curve is "similar" to the load curve of the DR group which could be written as:

$$\underset{\beta_1,\cdots,\beta_m}{\operatorname{argmin}} \left\| P^{DR}(\underline{d}) - \sum_{i=1}^{m} \beta_i P_i(\underline{d}) \right\|^2 . \tag{1}$$

The coefficients β_i assigned to each individual load curve have to be estimated.

When considering the day before as historic, $\underline{d} = \{d-1\}$, and if it is a week, $\underline{d} = \{d-1, \cdots, d-7\}$, where $\underline{d}$ are non event days. To estimate the unknown β_i, we propose constraint regression methods and a sequential algorithm.

2.1 Constraint Regressions

To estimate the baseline from a control group, we aim at selecting individual load curves. We consider in this subsection the DR group's load curve $P^{DR}(\underline{d})$ as the response variable Y and the individual load curves $P_i(\underline{d})$ as the explanatory variables $X_1, \cdots, X_m$. We have to estimate the model $Y = X\beta + \epsilon$ where ϵ is an error term. However, the length of Y depends to $\underline{d} \times 144$. If $\underline{d} = 7$ days, $\underline{d} \times 144 = 1{,}008$ which is highly probable to be lesser than m. In this case, we face the rank deficiency. The linear system of the regression model is under-determined and there is no unique solution. The dimension has to be reduced by selecting the most important variables.

For the sake of interpretation we want to keep the load curves as explanatory variables excluding projection methods as Principal Component Regression (PCR) and Partial Least Squares (PLS). An other dimension reduction method consists in regularizing the regression model by penalizing the l^p norm of the coefficients. When $p = 2$ it is the ridge regression and the Lasso regression when $p = 1$. Each method is then used to select the control group.

2.1.1 Ridge Regression

Ridge regression [6] penalizes the l^2 coefficients' norm. Applied to the control group selection, the minimization problem is:

$$\hat{\beta}^R(\lambda) = \underset{\beta \in \mathbb{R}^m}{\operatorname{argmin}} \left\| P^{DR}(\underline{d}) - \sum_{i=1}^{m} \beta_i P_i(\underline{d}) \right\|^2 \quad \text{wrt} \quad \sum_{i=1}^{m} \beta_i^2 \leq \lambda\,.$$

The average load curve of the control group is: $P^C(\underline{d}) = \sum_{i=1}^{m} \hat{\beta}_i^R(\lambda) P_i(\underline{d})$.

Geometrically, it consists in constraining the least squares coefficients to belong to a λ radius ball. This leads to shrink the least squares estimator. The tuning parameter λ is estimated by cross-validation, selecting the value which minimizes the mean squared error (MSE) between the estimation and the real value. For m large, each coefficient will be close to zero.

2.1.2 Lasso Regression

Lasso regression [7] penalizes the l^1 coefficients' norm. Applied to the control group selection, the minimization problem is:

$$\hat{\beta}^L(\tau) = \underset{\beta \in \mathbb{R}^m}{\operatorname{argmin}} \left\| P^{DR}(\underline{d}) - \sum_{i=1}^{m} \beta_i P_i(\underline{d}) \right\|^2 \quad \text{wrt} \quad \sum_{i=1}^{m} |\beta_i| \leq \tau.$$

The average load curve of the control group is: $P^C(\underline{d}) = \sum_{i=1}^{m} \hat{\beta}_i^L(\tau) P_i(\underline{d})$.

Geometrically, it consists in constraining the least squares coefficients to belong to a diamond-shaped convex polytope (l^1 constraint expression in $\mathbb{R}^m$). Coefficients are then shrunk and some of them equal zero. Thus, Lasso regression proceeds in variables selection (here curves selection). The method allows to select the most relevant curves to build the control group. There are many methods to find the Lasso solution, the Least Angle Regression (LARS) algorithm developed by [3] is the most used. This algorithm proceeds in forward selection of variables mostly correlated with the current residuals. The tuning parameter τ is estimated by cross-validation, selecting the value which minimizes the MSE. Through Lasso regression, we get a sparse solution of the coefficients vector $\hat{\beta}^L(\tau)$.

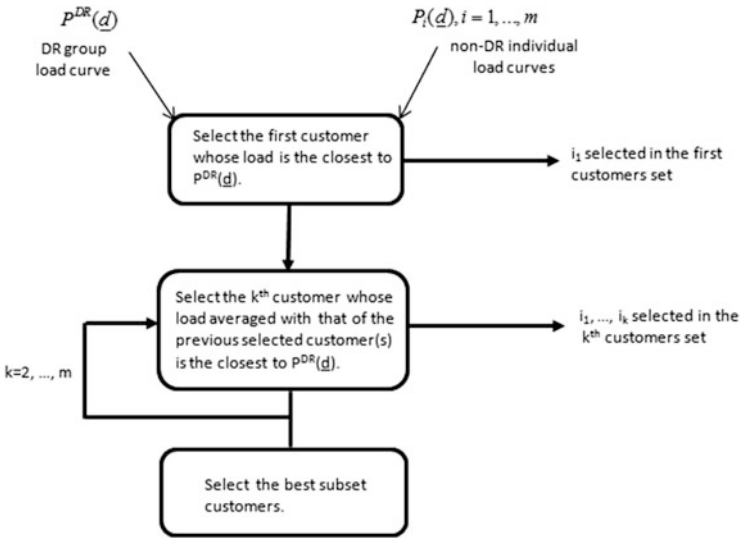
Remark Let us note that the number of non zero coefficients (the number of selected load curves) could not be larger than the length of Y (here $\underline{d} \times 144$) [3].

2.2 Algorithm

To solve problem (1) with coefficients β_i valued in $\{0, 1\}$ (absence/presence), we have to evaluate all the possible control groups with m customers which is of complexity $2^m - 1$. In order to reduce this number and like forward selection methods, we minimize

$$\| P^{DR}(\underline{d}) - \frac{1}{\#selected} \sum_{i \in selected} P_i(\underline{d}) \|_p$$

by sequentially selecting the individual loads $P_i(\underline{d}), i = 1, \cdots, m$, to construct the control group. The algorithm developed here aims at selecting the individual load curves, such that its average minimises the L^p norm with the DR group load curve. Before introducing in details the algorithm, let us present a flow diagram:



Going into details, the algorithm is:

- Initialization:
  Choose the length of the history $\underline{d}$ and evaluate:

$$S_{1,i} = \left\| P^{DR}(\underline{d}) - P_i(\underline{d}) \right\|_p$$

$$i_1 = \underset{i=1,\ldots,m}{\operatorname{argmin}} \{S_{1,i}\}, \quad P_{(1)}(\underline{d}) = P_{i_1}(\underline{d}), \quad \hat{S}_1 = \left\| P^{DR}(\underline{d}) - P_{(1)}(\underline{d}) \right\|_p$$

- Loop on $k = 2, \ldots, m$:

$$i_k = \underset{i \in \{1,\ldots,m\} \setminus \{i_1,\ldots,i_{k-1}\}}{\operatorname{argmin}} \left\| P^{DR}(\underline{d}) - \frac{1}{k} \left[\sum_{l=1}^{k-1} P_{(l)}(\underline{d}) + P_i(\underline{d}) \right] \right\|_p$$

$$\hat{S}_k = \left\| P^{DR}(\underline{d}) - \frac{1}{k} \left[\sum_{l=1}^{k} P_{(l)}(\underline{d}) \right] \right\|_p$$

- Select the customers set $\{i_1, \cdots, i_k\}$ minimizing the norm $\hat{S}_k$.

The average load curve of the control group is $P^C(\underline{d}) = \frac{1}{\#selected} \sum_{i \in selected} P_i(\underline{d})$.

This algorithm is easy to implement. As the constraint regression methods, it aims at selecting individual load curves to build the control group by attributing a weight one if the curve is selected and zero if not. The algorithm can be applied in a forward or backward way. By modifying it, this is also possible to select identical loads several times.

Remark Unlike Lasso regression, the number of selected load curves is not related to the length of Y (here $\underline{d} \times 144$).

3 An Operational and Online Solution

3.1 Operational Baseline Estimation

There are two steps to estimate the baseline. First, one has to choose the historical period $\underline{d}$ to estimate the coefficients. Second, to evaluate the control group load curve, the estimated coefficients are applied on the event day d. Historical periods could be $\underline{d} = \{d-1\}$, $\underline{d} = \{d-1, d-2\}, \ldots, \underline{d} = \{d-1, \cdots, d-7\}$. The control group selection is done for each event day. The control group adapts to the $P^{DR}(d)$ which is changing because of weather conditions but also because individuals in the program could evolve with the customer turnover. In the case of consecutive event days, the selection is done on the $\underline{d}$ period without demand reduction event prior to the first event day of the consecutive set. Then the selected control group is common over all those consecutive event days but the baseline curve varies for each day.

To illustrate the methodology, we apply the algorithm and the constrained regressions to the baseline estimation of a DR group of three hundred customers of whom electric heating is cut off between 6 and 8 pm during 20 winter days. For confidential reasons, we can not mentioned the exact number, but we have thousands non-DR load curves available in the datamart to select the control group. As the DR customers, these control customers have the peak and off-peak hours rate. They are distributed throughout the France and around 6 % of them are sharing the same area as the DR customers.

To evaluate the accuracy baseline, we estimate for each day and hours h the error $err(d,h) = P^{C}(d,h) - P^{DR}(d,h)$ for the hours h between 6 am and 6 pm corresponding to the longer daily period without DR event. We evaluate the MAPE(d) of the errors on each event day:

$$MAPE(d) = \frac{1}{\#h} \sum_{h} \frac{|err(d,h)|}{P^{DR}(d,h)} \times 100$$

and to compare the methods' variability on the 20 event days, we estimate the standard deviation of the MAPE(d) denoted $\hat{\sigma}_d$.

We are also interested in analysing the methods regarding to the energy used during the day. In order to do so, we estimate the MPE(d) for h between 6 am and 6 pm:

$$MPE(d) = \frac{1}{\sum_h P^{DR}(d,h)} \sum_{h} err(d,h) \times 100.$$

Table 1 MAPE (standard deviation) and MPE evaluated between 6 am and 6 pm for the algorithm and constrained regression methods applied on the 20 event days

	Algo L^1	Algo L^2	Ridge	Lasso
MAPE	12.06	11.83	14.44	16.36
$\hat{\sigma}_d$	8.74	8.50	9.07	10.52
MPE	1.19	0.59	2.55	4.95

Through MPE, positive and negative errors are balanced. It indicates if methods over or underestimate the true value. A MPE close to zero means that positive errors counterbalance negative errors (or inversely) and that, in energy, the estimations are accurate.

Table 1 presents the MAPE, its standard deviation and the MPE averaged on the 20 DR days. The best accuracy of the baseline estimation is obtained with the algorithm which also presents the smallest variability. A previous study showed that the accuracy is largely improved (by half) if the geographic region of non-DR customers is restricted only to the DR customers region. That makes sense that customers from different French regions have not the same consumption or the same load shape. In fact, the main reason is that the off-peak hours are placed according to the geographic regions. This is considerable because off-peak hours could occur at midday or during the afternoon, modifying then the non-DR individual loads shape. However, limiting the geographic area of non-DR individuals considerably reduces the currently available control loads number which is beyond the scope of this paper. The future smart meters roll-out or the shortly available individual loads from the SEL project will provide thousands load curves belonging to different French areas. Regarding the MPE, we observe that the algorithm used with the L^2 distance provides, on average, balanced estimations while regression methods overestimate the daily load curve, by 5 % for the Lasso regression. This overestimation induces an overestimation of the demand reduction.

Apart from this, these baseline methods allow to estimate the entire daily load without using data on the hours prior the event. This means that the methods adapt to any DR formats, even if there are several DR events in the day. Moreover, estimating accurately the whole daily load curve enables to quantify the anticipation or the load report after the DR.

3.2 Online Demand Reduction Estimation

The demand reduction estimation is essential for electric stakeholders since it is possible to enhance demand reduction on the electric markets. The demand reduction estimation needs to be estimated on real time. The proposed solution

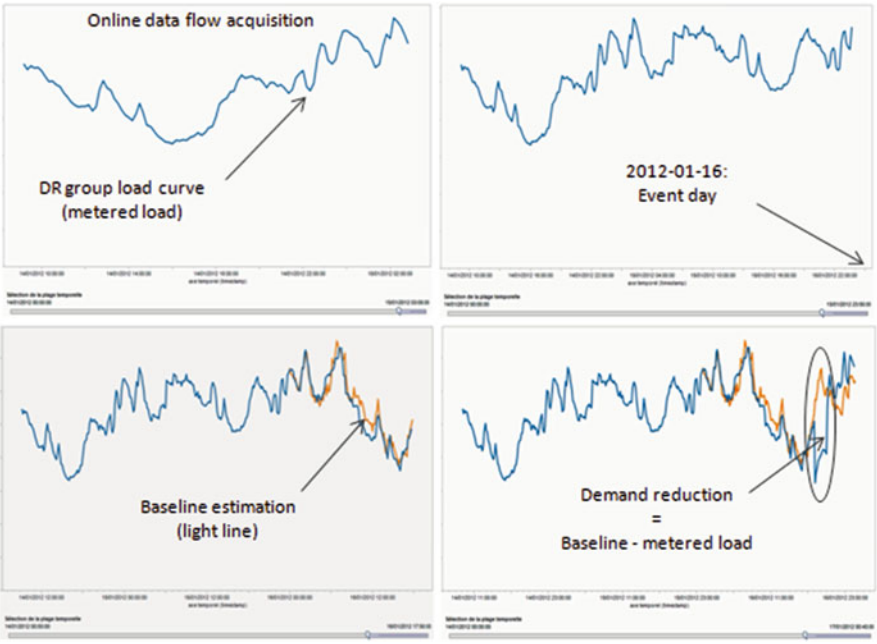


Fig. 5 Online demand reduction estimation: DR group load curve acquisition (*top left*), event day (*top right*), baseline estimation (*bottom left*) and demand reduction estimation (*bottom right*)

consists in a data flow module linked with a statistical software. Data acquisition and results production are realized as following:

1. **Input data treatment:**
   Data flow could provide from different sources (files, databases or network ports). Each data source is processed online and grouped together in one data flow. This flow contains historical data necessary to estimate coefficients of the problem (1).
2. **Control group estimation:**
   The module communicates the data flow to the statistical software via a server, a communication network interface. Statistical programs are executed and results returned to the module.
3. **Output data visualization:**
   Data are saved and exploited to view online results. It is possible to visualize the online data flow acquisition and the estimated baseline or demand reduction produced (Fig. 5).

This innovative process brings a real time solution to all electrical stakeholders in order to act rapidly in the interest of the electric grid. It is largely able to process lots of input and output DR and non-DR loads at the same time. It is then a reliable solution to evaluate the demand reduction on thousands customers.

4 Conclusion

In this article, baseline methods relied on new control group selection methods were presented. Control group selection methods are currently using individual characteristics. To provide an adequate solution for an operational use, we based the selection on the loads' shape, containing many more informations than the observable variables. Thus, the weather component can be easily caught and included in the control group selection. The algorithm and constrained regression methods we proposed here comply with the expected objectives:

1. Methods estimate the entire daily load curve without using the DR group load on the event day. This resultant particularity allows to quantify demand reductions on short, long or various event periods. Estimating the whole daily load curve enables to quantify the anticipation or the load report after the demand reduction.
2. Methods, only using individual load curves, are cost-effective. As future roll-out of smart meters will provide many loads in different French regions, it could be possible to improve the baseline accuracy.
3. The new control group selection methods are flexible and adapt easily to the customer turnover in operational conditions.
4. To control and secure the grid balance, one requires to visualize and quantify the DR online. The solution brought totally meets this requirement and is able to doing it on thousands customers.

For the above reasons, our new control group selection methods are efficient for an operational use. They are currently applied in some EDF's experiments, particularly the algorithm which is easy to understand and implement. When the SEL project will provide thousands loads with demand reduction, these new methods will be operational to quantify the demand reduction.

Acknowledgements The authors would like to thank the editor and two anonymous referees for their valuable comments which helped in improving the paper.

The online DR estimation solution is the result of a common work realized with Benoît Grossin, EDF R&D, Dept. ICAME, we thank him for the accomplished work.

References

1. Bode, J. L., Sullivan, M. J., Berghman, D., & Eto, J. H. (2013). Incorporating residential AC load control into ancillary service markets: Measurement and settlement. *Energy Policy, 56*, 175–185.
2. Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87*(418), 376–382.
3. Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle regression. *The Annal of Statistics, 32*, 407–499.
4. Faruqui, A., & Sergici, S. (2010). Household response to dynamic pricing of electricity: A survey of 15 experiments. *Journal of regulatory Economics, 38*, 193–225.

5. Frontier Economics and Sustainability First, Department of Energy and Climate Change. (2012). *Demand Side Response in the domestic sector – A literature review of major trials*. UK Department of Energy and Climate Change the report is available at: http://www.frontier-economics.com/documents/2013/10/frontier-report-demand-side-response-in-the-domestic-sector.pdf
6. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67.
7. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267–288.

Forecasting Intra Day Load Curves Using Sparse Functional Regression

Mathilde Mougeot, Dominique Picard, Vincent Lefieux, and Laurence Maillard-Teyssier

Abstract In this paper we provide a prediction method, *the prediction box*, based on a sparse learning process elaborated on very high dimensional information, which will be able to include new – potentially high dimensional – influential variables and adapt to different contexts of prediction. We elaborate and test this method in the setting of predicting the national French intra day load curve, over a period of time of 7 years on a large data basis including daily French electrical consumptions as well as many meteorological inputs, calendar statements and functional dictionaries. The prediction box incorporates a huge contextual information coming from the past, organizes it in a manageable way through the construction of a *smart* encyclopedia of *scenarios*, provides experts elaborating strategies of prediction by comparing the day at hand to referring scenarios extracted from the encyclopedia, and then harmonizes the different experts. More precisely, the prediction box is built using successive learning procedures: elaboration of a data base of historical scenarios organized on a high dimensional and functional learning of the intra day load curves, construction of expert forecasters using a retrieval information task among the scenarios, final aggregation of the experts. The results on the national French intra day load curves strongly show the benefits of using a sparse functional model to forecast the electricity consumption. They also appear to meet quite well with the business knowledge of consumption forecasters and even shed new lights on the domain.

M. Mougeot (✉) • D. Picard
Université Paris-Diderot, CNRS-LPMA, UFR de Mathématiques, 75013 Paris, France
e-mail: mougeot@math.univ-paris-diderot.fr; picard@math.univ-paris-diderot.fr

V. Lefieux
RTE-EPT & UPMC-ISUP, 92919 La Défense Cedex, France
e-mail: vincent.lefieux@rte-france.com

L. Maillard-Teyssier
RTE-R&D-I, 78000 Versailles, France
e-mail: laurence.maillard-teyssier@rte-france.com

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_9

1 Prediction Box: Forecasting the Electrical Consumption

This paper is the result of a cooperation between industrial and academic research. RTE, the French electricity transmission system operator, is responsible for operating, maintaining and developing the high and extra high voltage network. RTE is required to guarantee the security of supply, so anticipating French electricity demand helps to ensure the balance between generation and consumption at all times, and directly influences the reliability of the power system.

Demand forecasts are carried out for several different timeframes: for the long-term, in the form of the Generation Adequacy Report or network development studies, for the medium-term (annual, monthly and weekly forecasts) and lastly on a day-ahead basis.

From a short term point of view, electricity demand fluctuates depending on cycles (annual, weekly and daily), on temperature and cloud cover (to take into account variations in the outside temperature affecting the use of heating equipment in winter and air-conditioning in summer), and on other factors such as economic activity (e.g. holiday periods), demand response offers or daylight saving time changes. Note that the French load curve is very sensitive to temperatures, it contributes to half of the European thermo-sensitivity.

Today RTE uses a complex nonlinear parametric regression model with around one thousand coefficients estimated twice a year, and also a SARIMA model. These decision-making tools cannot predict exceptional events which may disrupt the demand profile (heavy snow fall, sporting events, strikes), and final day-ahead forecasts are provided by the French national dispatchers (see: http://www.rte-france.com/en/sustainable-development/eco2mix/electricity-demand).

If this process currently provides good forecasts, the context of the smart grids and the energy transition will lead to more variability in the load curve. Moreover, the aim is not only to obtain a low mean error but also to avoid big forecast errors which have a direct influence on the reliability of the power system.

Taking into account new explanatory variables (e.g. wind, new tariffs, electricity prices), economic uncertainties (e.g. economic crisis), new innovative heating systems (e.g. heat pumps) requires to work with more adaptive and dynamic models. Many models and approaches have already been considered, from the robust SARIMA [5, 13] to the semi-parametric model MAVE [10] or functional regression using wavelets [2, 3].

Referring to this context, in this paper, we address the following program:

(i) Construct a prediction method, *the prediction box*, based on a sparse learning process including high dimensional information, which will be able to include new – potentially also high dimensional – influential variables or to adapt to different settings of prediction in terms of time ahead (one day ahead, 48-h ahead, medium-term) or geographical context (European, or at the opposite regional or even more local) and eventually to more general situations of forecasting.

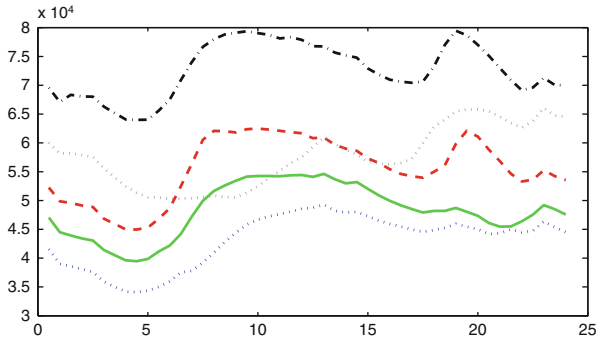


Fig. 1 Intraday load curves for various days. 2010-02-03 winter: *black dashed dot line*, 2010-05-21 spring: *red dashed line*, 2009-10-23 autumn: *green solid line*, 2010-08-19 summer: *blue dot line*, 2010-01-01 public day: *gray dot line*

(ii) Elaborate and test this method in the setting of predicting the national French intraday load curve, over a period of time of 7 years on a large data basis including daily electrical consumption as well as many meteorological inputs, calendar statements and functional dictionaries, described in the next subsection.

Figure 1 illustrates the observed dissimilarity between some daily consumption signals. For instance electrical consumption is mostly higher and more *active* in winter than in summer and is characterized by two large peaks of consumption. However some winter public days may show weak consumption with unusual pattern, and spring days can reach high consumption with also specific features characterized by a unique peak of consumption.

A crucial step in the forecasting process is the modeling. It is commonly admitted that many variables are influential for the prediction in this context. On the other hand, it is well known that a model relying on a small number of well chosen predictors is more robust and efficient than a model without variable pre-selection. The challenge here is then for each day to produce a small number of predictors, after considering all the variables which can be potentially significant.

The *prediction box* will provide three drawers of unequal sizes using different learning procedures at each different scale:

(a) The first drawer contains a *smart* encyclopedia of *scenarios* coming from the observed past. A smart encyclopedia is a very large but very well organized structure. For each day of the past sample, the encyclopedia provides the *background* of the day measured by a large (but manageable) number of significant explanatory variables. It also contains, associated to the *background* of the day, a sparse approximation of the consumption of the day (using much fewer explanatory variables than the number of initial variables).
(b) The second drawer contains a bunch of experts. Each of them provides a strategy of prediction for the load curve after consulting the referred encyclopedia. Each expert essentially bases its strategy on comparing the day at hand to referring scenarios extracted from the backgrounds of the past. This step allows to find a

day in the past which is closest (according to the expert) to the day at hand. The prediction then uses the sparse approximation of this closest day.

(c) The final action of the box will be to harmonize the experts using an aggregation process.

In the following section we describe the data basis. Section 3 is devoted to the construction of the encyclopedia. This section is crucial and will especially describe the choice of variables for the backgrounds as well as the sparse approximation. Section 4 details the experts, their respective performances and the aggregation process. The last section is devoted to analyze the results of the prediction box as well as the perspectives of the method.

2 The Data Basis

2.1 Electrical Consumption

The French national electrical consumption has been recorded every half hour from January 1st, 2003 to August 31st, 2010 and stored in a database. We focus our study on daily recording signals. For this period of time, the global consumption signal is split into $N_{obs} = 2{,}800$ sub signals $(Y_1, \ldots, Y_t, \ldots, Y_{N_{obs}})$. $Y_t \in R^n$ defines the intra day load curve for the tth day of size $n = 48$. These intra day signals will constitute our data basis to approximate and then to forecast the daily consumption.

Figure 2 shows a week of electrical consumption defined by seven successive intra day curves.

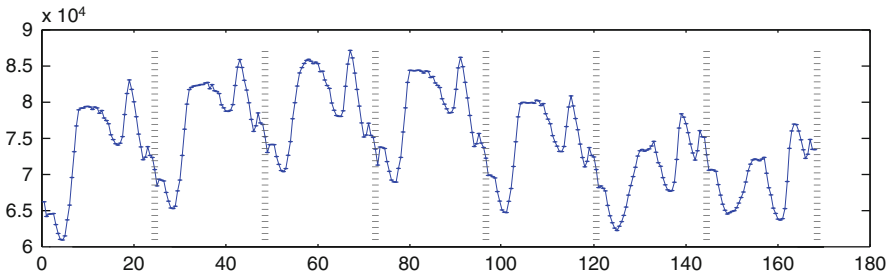


Fig. 2 Electrical consumption week from Monday January 25th to Sunday January 31st 2010 regrouping seven successive intra day load curves of size $n = 48$

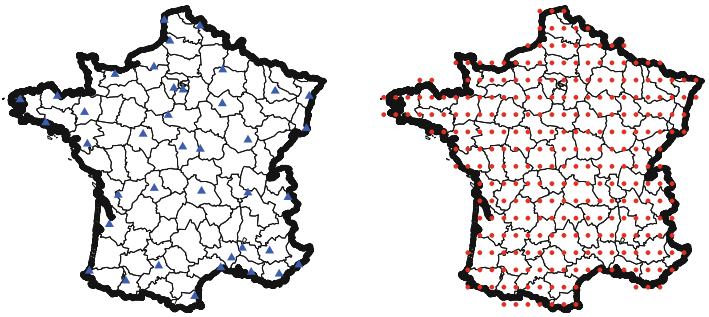


Fig. 3 Temperature and cloud cover (*left*) measurement stations. Wind strength network available points (*right*)

2.2 Meteorological Inputs

For this study, available meteorological inputs are recorded each half hour on the same period of time, from January 1st, 2003 to August 31st, 2010:

Temperature: T^k for $k = 1, \ldots, 39$ denotes the temperatures measured in 39 weather stations scattered all over the French territory as indicated in Fig. 3 (left).
Cloud Cover: N^k for $k = 1, \ldots, 39$ is an indicator of the cloud cover which is also measured in the same 39 weather stations. The cloud cover is a fraction between 0 (free of clouds sky) and 80 tenths of octas (completely clouded sky). Cloud cover are nowadays built on satellite based observations on the same meteorological stations than temperature.
Wind: $W^{k'}$ for $k' = 1, \ldots, 293$ denotes the 100 m wind speed analyses available at the 293 network points scattered all over the French territory (see Fig. 3, right).

As the electrical consumption, the meteorological inputs (temperature, cloud Cover and wind) are sampled each half hour and available for the same time period. Every day, for each meteorological input and weather spot, a $n = 48$ signal can be extracted as shown in Fig. 4.

T_t^k (resp. N_t^k, $W_t^{k'}$) denotes the daily temperature (reps. cloud cover, wind) for day t, $1 \leq t \leq 2{,}800$, and station k with $1 \leq k \leq 39$ or network point k' with $1 \leq k' \leq 293$.

Figure 4 illustrates the variability of weather conditions in France. Temperature, cloud cover and wind signals are chosen in three different meteorological spots localized in West (Brest), North (Lille) and South (Marseille) of France. In Marseille, large variations of temperature can be observed during this day, with a stationary high cloud cover and an increasing wind. On the opposite during the same day, stationary temperature and wind can be observed in Brest with a decreasing cloud cover. All these meteorological factors are known to have an impact on the electrical regional consumption which has been established for the French PACA and Bretagne regions for instance.

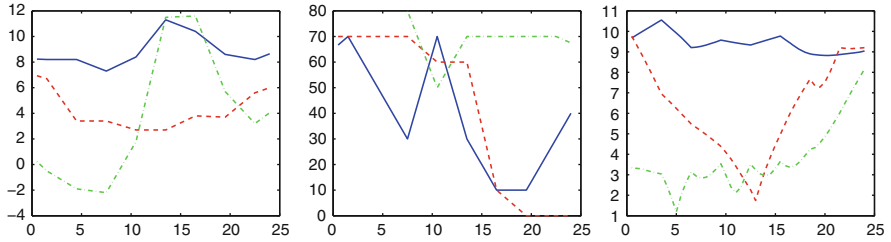


Fig. 4 Temperature (*left*), cloud cover (*middle*) and wind (*right*) intra day signal for the 3rd February 2010 in Brest (*blue line*), Lille (*red line*) and Marseille (*green*) cities

In this study, a total of 371 ($= 2\times39 + 293$) raw meteorological variables are available. Both types of data (surface weather points and grid points) are used which constitutes a new way of integrating meteorologic data into a load curve modeling.

2.3 Calendar Statements

As illustrated in Figs. 1 and 2, the intra day load curves are quite different, depending on the day and on the season. Qualitative variables are introduced in order to characterize days and seasons.

Variable D takes seven modalities characterizing the type of day {1:Monday, . . ., 7:Sunday }.

As in [22], C is a qualitative variable, taking one of the 20 modalities depending on the 5 groups of days (Monday, Friday, Saturday, Sunday and the others) subdivided by the four seasons Winter, Spring, Summer and Autumn.

D_t and C_t describe the modalities of both variables for day t.

3 Building the *Smart* Encyclopedia

Electrical consumption, meteorological inputs and calendar statements contributes to the elaboration of the *smart encyclopedia*, recording the historical data base.

The encyclopedia proposed in this work has been considered as a *smart* encyclopedia since part of its elements are built using a learning process. In particular, the encyclopedia contains sparse approximations of the intra day load curves which can be considered as a 'clever' and significant representation of the consumption signals; the choice of the dictionary being a key point.

Parts of the construction of the *encyclopedia* use different steps interesting by themselves which are described in [17]. We just recall here the principal ones and refer to [17] for more details.

3.1 Patterns of Consumption

It is well known that intra day load curves can be explained using two types of variables: on one hand specific patterns of consumption (also called endogenous variables), on the other hand meteorological variables (also called exogenous variables) [6, 18, 19]. Patterns of consumption are usually built using calendar information and it is important to address the problem of determining which calendar information will be used to best represent the curves, with the serious issue that some choices can lead to representation which are highly correlated with meteorological variables and very often will disappear in sparse representations.

In most applications, in order to integrate calendar information, the set of days is split on deterministic statements. Taylor [22] uses a partition of size 20 already mentioned in Sect. 2.3. To study the Spanish consumption, [14] uses Kohonen maps to build adaptive groups of consumption.

Our point of view is slightly different and consists in providing typical *patterns of consumption* using a three step pre-processing: we first provide a sparse modeling of the profiles of daily electrical consumption as functions of the time. Then, using the sparse representation of each load curve, clusters of consumption are defined. A final interpretation of the clusters yields *typical* profiles, the *group centroids* or it patterns of consumption that will be time variables entering into the final model, in addition to meteorological variables. More precisely,

1. The first task is defined by the compression of the intraday load curves Y_t using a nonparametric regression on a dictionary of functions of the time variable, with the help of the sparse algorithm LOLA described in Table 1 [12, 15, 17].

   In other words, the intraday signals Y_t are treated as functions of the time [20] and sparsely represented in a dictionary, which has to be well adapted to produce a full reduction of the problem. A combination of Fourier basis and Haar basis has been chosen as dictionary for this task.

Table 1 Description of LOLA Algorithm

Step 1: Selection by thresholding
A first thresholding procedure allows to reduce the dimensionality of the problem in a rather crude way by a simple inspection of the empirical correlations between the signal and each element of the dictionary.
This first threshold is data driven and is chosen adaptively [15].
Step 2: OLS
Ordinary Least Square method is then used on the linear sub-model obtained by considering the variables retained after the first step.
Step 3: Denoising by thresholding
A second thresholding is performed on the estimators of the parameters of the sub-model.
This second step is more refined and corresponds to a denoising phase of the algorithm.
As the first one, the second threshold is also data driven and is chosen adaptively [15].

Table 2 Groups, 1 . . . 8, are defined using a calendar interpretation of clusters from Monday (day 1) to Sunday (day 7) and from January (month 1) to December (month 12) computed form January 1st to August 31st [17]

	Months											
Days	1	2	3	4	5	6	7	8	9	10	11	12
1	7	8	5	3	3	3	3	1	3	3	5	7
2	7	8	5	3	3	3	3	1	3	3	5	7
3	7	8	5	3	3	3	3	1	3	3	5	7
4	7	8	5	3	3	3	3	1	3	3	5	7
5	7	8	5	3	3	3	3	1	3	3	5	7
6	6	8	4	4	2	2	2	2	2	2	4	6
7	6	6	4	4	2	2	2	2	2	2	4	6

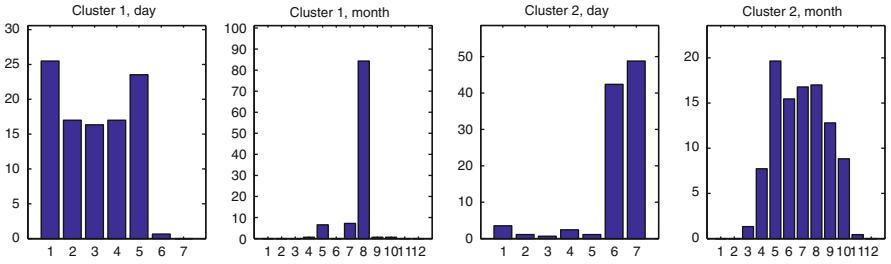


Fig. 5 Illustration of the calendar repartition of days and months for the two clusters 1, 2

2. The second task consists in clustering the previous sparse representations of the signals. The K-means algorithm is here chosen for its simplicity. An additional investigation [17] shows that the clustering end up with 8 different groups, as described in the following Table 2.
3. The last step defines the *group centroid* variables or *patterns of consumption* as the mean signal inside each group. This yields eight signals $S_1, \ldots, S_8$ summarizing the typical behavior of each cluster.

An analysis is then performed in order to retrieve a correspondence between clusters and calendar statements. For an illustration of the different calendar statements extracted between the clusters, Fig. 5 provides, as an example, the occurrences observed between days and months, for two clusters. For cluster 1, a large majority of load curves are week-days (1–5) for the month of August (8). For cluster 2, a large majority of load curves are week-ends (6–7) for months from May to October (5–10). It should be noted that, at this stage, no specific treatments have been done for bank holidays or other special days. Interpretation for all clusters is available in [17].

The *pattern variable* G is a variable taking eight (functional) modalities and assigning each day t to the center of the group where it belongs: $S_{g(t)}$, where $g(t)$ is simply the labeling function of the calendar situation of t, according to Table 2.

During our preliminary study an additional pattern variable emerged and appeared to be as well a key endogenous variable, defined by the intra day load curve, Y_{t-7}, recorded 1 week before. This signal can be considered as a general

trend for Y_t and provides also indirectly some calendar information related to the type of day (Sunday, Monday, ...) and seasons.

The endogenous variables are then defined by these two patterns and will represent the consumption signal as a function of the time, denoted in the sequel by $P_t = [G_t\, Y_{t-7}]$.

3.2 Meteorological Variables

In this study, the target variable is the French national electrical consumption, which is impacted by all the meteorological conditions of the French territory, but, of course, with different contributions regarding each region. Inside each group of temperatures $\{T^k, 1 \leq k \leq 39\}$, cloud cover $\{N^k, 1 \leq k \leq 39\}$ or winds $\{W^{k'}, 1 \leq k' \leq 293\}$, variables appear to be highly correlated and show strong scale effect. A first non linear pre-processing is applied to build meteorological *indicators*, both to sum up the information as well to reduce the redundancy between the meteorological variables as already observed in [9] for instance. The new following standard indicator variables are then computed.

For each label $U \in \{T, N, W\}$, we introduce four non linear transformations of the meteorological inputs computed each half hour for the set of $n_U = 39$ stations for (T, N) and for the set of $n_U = 293$ network points for W:

- $U^{min} = \min(U^1, \ldots, U^{n_U})$
- $U^{max} = \max(U^1, \ldots, U^{n_U})$
- $U^{med} = \text{median}(U^1, \ldots, U^{n_U})$
- $U^{std} = \sqrt{\text{Var}(U^1, \ldots, U^{n_U})}$

These standard indicators provide a non linear sum-up of the variations and sizes of temperature, cloud cover or wind all over the French territory. Hence, $12 = 4 \times 3$ indicators are computed, half-hourly sampled, and stored from January 1st 2003 to August 31st 2010.

With a slight abuse of notation, for each label $U \in \{T, N, W\}$, we denote now by $U = [U^{min}, U^{max}, U^{med}, U^{std}]$ the reduced meteorological variable, where U^{min}, U^{max}, U^{med}, U^{std} are the standard indicators defined previously.

Finally, piling up all the days, a total of 12 indicators defines the meteorological process $M = [T\, N\, W]$ over the time. The meteorological conditions for day t are defined then by $M_t = [T_t N_t W_t]$, which is a 48×12 matrix.

Even if at this step of reading, the forecasting methodology has not been yet presented, we should say a brief remark in order to justify, at this stage, the choice of these four indicators. During this project, different methods have been investigated to forecast intra day load curves. One of them investigated to model the intra day load curves with a high dimensional model using all 371 meteorological available variables ($371 = 273 + 2^*39$). As the fit of the intraday load curves was extremely accurate, the performance of the forecast was quite poor. This is explained by the fact that the global consumption, as we have already said, is actually composed

of various regional consumptions. In this study, we did not have access to the regional consumptions which are known additionally to show quite high variability in space and in time. For a forecasting point of view, the method which sums up the meteorological data using four indicators performs the best at this stage.

3.3 Sparse Approximation of Intraday Load Curves

For each day t, we model the daily electrical consumption signal Y_t in a linear way using the following equation

$$Y_t = Z_t \beta_t + u_t \tag{1}$$

where the unknown parameter β_t (so depending on the day t) belongs to R^p with $p = 14$ and where the variable $Z_t = [P_t\ M_t] = [G_t\ Y_{t-7}\ T_t\ N_t\ W_t]$ is the concatenation of the pattern variables and meteorological variables previously described. G_t is the pattern variable previously defined. G_t takes eight modalities (Table 2) depending on days (variable D_t) and months of the year. The size of Z_t is $(n \times p) = (48 \times 14)$.

β_t is estimated using the LOLA algorithm especially chosen to produce a sparse representation: $\hat{\beta}_t = LOLA(Z_t, Y_t)$.

Note that LOLA is an algorithm providing good sparse approximation in very high dimension (see [17] in the case of the intra day load curve) and very accurate selection properties in medium high dimension (see [16]), which is the case here ($p = 14$). The adjusted electrical consumption is then:

$$\hat{Y}_t = Z_t \hat{\beta}_t \tag{2}$$

To evaluate the quality of this preliminary fit, we report here the *MAPE* and the *RMSE* errors, computed.

We observe that the selected covariables offer a quite high sparsity representation (Table 3): in average, $S = 2.5$ non zero coefficients are used to approximate the 365 intra-day load signals, with an average MAPE error of 1.2 % (median 1 %).

Figure 6 shows the sparse approximations $\hat{Y}_t = Z_t \hat{\beta}_t$ of Y_t for 4 days belonging to different seasons. For each graph, the number of selected coefficients with LOLA

Table 3 Statistical indicators for the sparsity, the MAPE and the RSME between the daily signal Y_t and its fit signal $\hat{Y}_t$, computed from September 1st 2009 to August 31st 2010. Groups are computed using previous years of available data

Statistical indicator	Sparsity	MAPE (%)	RMSE
Mean	2.49	1.24	833
Median	2.00	1.05	695
Std	0.81	0.79	531

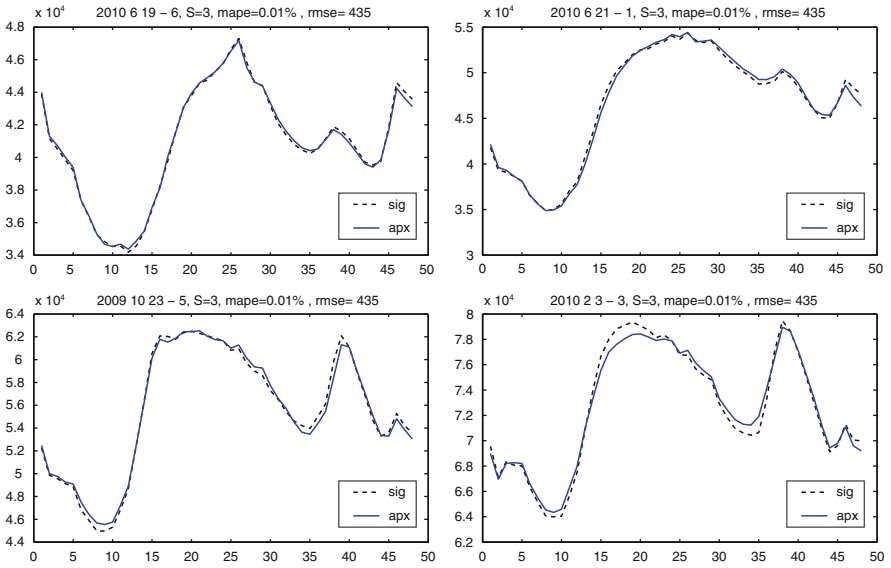


Fig. 6 Spring, Summer, Autumn and Winter intra day signals (*dashed line*) are here approximated using at most 3 coefficients selected by LOLA algorithm (*solid line*)

algorithm equals at most 3. Hence, we achieved here one step in our box: well approximating intra day consumption signals for various shapes and sizes using only few coefficients.

3.4 Smart Encyclopedia Contents

To summarize, the raw temporal data of electrical consumption as well as the meteorological inputs, available on the given historical period of 7 years and an half of data, have been daily processed as described in the previous subsections to produce the *Encyclopedia*, $\mathscr{E}$, providing for each day t, $1 \leq t \leq 2{,}800$:

- The daily electrical consumption Y_t,
- A qualitative description of t, given by calendar statements: D_t, C_t,
- A qualitative description of t, defined after an adaptive clustering on the data: G_t,
- The meteorological indicators over the French territory $M_t = [T_t\, N_t\, W_t]$,
- The estimated coefficient $\hat{\beta}_t$,
- The approximation of the daily consumption $\hat{Y}_t = Z_t \hat{\beta}_t$.

4 Intra Day Forecasting

We are now interested in the one day ahead forecast of the intraday load curve Y_t. Consumption and meteorological variables are then supposed to be known until day $t-1$, and we want to propose a forecast of the intraday load curve for the next day t. In this case, it is obviously not possible to use for the forecast, the approximated coefficient $\hat{\beta}_t$, since its computation involves the knowledge of the electrical consumption Y_t we precisely want to forecast.

To forecast the intra day load curve of the next day, called $\tilde{Y}_t$, we refer again to the previous linear model and write:

$$\tilde{Y}_t = Z_t \, \tilde{\beta}_t$$

- The matrix $Z_t = [P_t \; M_t]$ is known at t.

  $P_t = [G_t \; Y_{t-7}]$ defines the pattern of day t. Using the calendar interpretation of the clusters described in Table 2, the centroid of G_t is known as Y_{t-7} which is the intra day load curve, one week ahead and can be easily computed.

- The meteorological variables M_t are here supposed to be known. In real applications, these variables will be provided by Meteo France, the French company for weather prediction.
- The main issue here is to provide $\tilde{\beta}_t$.

Our approach will be to choose a "good candidate" for $\tilde{\beta}_t$, among the set of already estimated coefficients $\hat{\beta}_u$ with $u < t$. This strategy is motivated by the fact that the linear model introduced in Eq. (1) appears to be quite a good model to approximate the intra day load curve. Moreover, this model is sparse and thus relies only on a small number of coefficients.

In the forecast problem there is a typical balance to find between the need for increasing the number of coefficients to better approximate (bias correction), and the fact that each added coefficient increases the variability of the forecast, which strongly justifies the use of a sparse methodology.

4.1 The Experts

The forecasting begins with an information retrieval task. As explained before, at this stage, we will produce a variety of experts. Each expert essentially bases its approach on comparing the day t at hand to referring scenarios extracted from the backgrounds of the past i.e. finding a day t^* in the past which is closest (according to this expert) to the day t.

The motivation is that, for one day, similar causes of weather or calendar conditions or identical groups of consumption should provide similar effects and then a similar electrical consumption.

In order to retrieve t^*, different *strategies* s are introduced. Each strategy, s, is a function defined from $\mathscr{T}$ to $\mathscr{T}$ such that for any $t \in \mathscr{T}$, $s(t) = t^* < t$, where $\mathscr{T}$ denotes the set of indices of the different days. A forecasting *Expert* is then simply associated to a strategy s and provides a forecast of the intra day load signal of the next day t by plug-in the approximated coefficients $\hat{\beta}_{s(t)}$ calculated at day $s(t)$ chosen by strategy s:

$$\tilde{Y}_t^s = Z_t \hat{\beta}_{s(t)}$$

How to choose the experts? Many factors are known to have a potential impact on the electrical consumption and the next paragraph provides the 17 strategies introduced here to potentially forecast the intra-day load curves. Of course, much more strategies can be included but, for a sake a clarity we detail here some of the simplest as well as most efficient ones (Table 4).

Time-lags: Studies of historical intra day load curves typically show that the day before as well as the day one week before are significantly influential,[6, 22]. Consequently, strategy *Week* recalls the approximated coefficients of the same

Table 4 The forecasting experts

Strategies	Time lags impact
Yday	$t-1$
Week	$t-7$
Strategies	Meteorological scenarios, $s(t) = t^* = ArgMin_u sup \|\|.\|\|$
T	$[T_u - T_t]$
T^{med}	$[T_u^{med} - T_t^{med}]$
T^{med}/N	$[T_u^{med} - T_t^{med}]$ with $\|\|N_u^{med} - N_t^{med}\|\|_1 / \|\|N_t^{med}\|\|_1 < 2\,\%$
T^{med}/W	$[T_u^{med} - T_t^{med}]$ with $\|\|W_u^{med} - W_t^{med}\|\|_1 / \|\|W_t^{med}\|\|_1 < 2\,\%$
T	$[T_u - T_t]$
T/G	$[T_u - T_t]$ with $G_u = G_t$
T/D	$[T_u - T_t]$ with $D_u = D_t$
T/C	$[T_u - T_t]$ with $C_u = C_t$
N	$[N_u - N_t]$
N/G	$[N_u - N_t]$ with $G_u = G_t$
N/D	$[N_u^{med} - N_t^{med}]$ with $D_u = D_t$
N/C	$[N_u^{med} - N_t^{med}]$ with $C_u = C_t$
W	$[W_u - W_t]$
W/G	$[W_u - W_t]$ with $G_u = G_t$
W/D	$[W_u^{med} - W_t^{med}]$ with $D_u = D_t$
W/C	$[W_u^{med} - W_t^{med}]$ with $C_u = C_t$

day, one week before, $s(t) = t - 7$ and strategy *Yday* involves the "yesterday" approximated coefficients, $s(t) = t - 1$ (which is known to provide useful information from Tuesday to Wednesday) [21].

Meteorological scenarios (MS): Temperature is commonly admitted to be an important factor in France as, in winter, 80 % of the French heating comes from electrical devices. So called *windchill temperature*, which is a more complex phenomenon depending both on temperature, wind and cloud cover has also an impact on electrical consumption.

The strategies we provide in this domain will be contextual and retrieve in the past a day corresponding to the nearest neighbor, regarding meteorological intra day signals (temperature, wind and/or cloud cover). Different metrics have been investigated but the sup distance seems to be especially suitable. The distance is measured taking into account the median signal or all the indicators (min, max, med, std) of meteorological variables introduced in Sect. 3.2.

Strategy T (resp. T^{med}) refers to the day having the closest temperature indicators (resp. the closest median temperature) among all the days in the past. Strategy N (resp. W) refers to the day having the closest cloud cover (resp. wind) indicator.

Cloud cover and wind may have special effects, so we consider strategies that specifically single out their effect.

Temperature constrained by cloud cover and wind: Strategy T^{med}/N refers to the day having the closest median temperature given cloud cover. More precisely, this means that we begin by selecting the days in the past which have a cloud cover signal in a small vicinity of 'today', and among these ones we choose the day with the closest T^{med}.

In the same way, strategy T^{med}/W refers to the day having the closest median temperature given the wind.

Impact of meteorological factors on electrical consumption also depends on day type (week days, week end or public day).

MS constrained by groups: Clustering methods applied on historical consumption data have exhibit specific groups of consumption [17]. Strategy T/G refer to the day having the closest temperature indicator given the group of t. Strategy N/G (resp. W/G) refers to the day having the closest cloud cover (resp. wind) indicator given the group of day t.

MS of the day constrained by the type of day: [1] shows the importance of the type of day. Strategy T/D refers to the day having the closest indicators (min, max, med, sd), given the kind of day of t Strategy N/D (resp. W/D) refers to the day having the closest cloud cover (resp. wind) indicator (min, max, med, sd) given the kind of day of t.

MS of the day constrained by a calendar group: [21] introduced five calendar groups to forecast (see Sect. 2.3). Strategy T/C refers to the day having the closest temperature indicators(min, max, med, std) for days belonging to the same calendar group as t. Strategy N/C (resp. W/C) refers to the day having the closest cloud

cover (resp. wind) indicators (min, max, med, std) for days belonging to the same calendar group as Y_t.

Here a meaningful discussion could be engaged on the choice of the experts. For instance, experts could integrate information about "special events" such as announced strike, big soccer games... We did not include such experts in the present study, since we observed in our data, that these events are at the same time rare and with high variable responses, in such a way that the learning process on them did not seem to show clear effects. Consequently we preferred to postpone this delicate point to a further study.

4.2 Smoothing Parameter

In this approach, the approximated intra day load is computed day by day and up to now, no continuity assumptions have been introduced between two consecutive days. In this perspective, several approaches can be used. For sake of simplicity, we chose to present here the simplest one.

In order to address this problem of introducing a regularity constraint between days, a parameter called δ_t^s reflecting the possible lag between day $t-1$ and prediction at day t is introduced and maintain in a *security zone*: $\tilde{Y}_t^s \leftarrow \tilde{Y}_t^s + \delta_t^s$

In this application, where the forecast is computed for 24 h, we simply choose $\delta_t^s = Y_t(1) - \tilde{Y}_t^s(1)$, and maintained it to be zero. $Y_t(1)$ (resp. $\tilde{Y}_t^s(1)$) is the value of the first point ($00:30'$) of the load curve (predicted load curve using strategie s) for day t. It means that the forecast is actually beginning each day at $00:30'$ for the next 23 h $30'$. When this method is used in other contexts (for a 36 h forecast for instance), a more refined δ_t^s parameter can be computed.

4.3 Performances of the Various Experts

Table 5 presents the forecast performances, for the $K = 17$ experts considered previously, for 1 year of data, from September 1st 2009 to August 31st 2010. The historical set of data to retrieve the $\hat{\beta}_t^*$, contains 3 years of data, from September 1st 2006 to August 31st 2009.

For sake of comparison, an additional *naive* expert is introduced which forecasts the daily electrical consumption of day t by simply using the intra day raw consumption signal of the previous day: $\tilde{Y}_t = Y_{t-1}$.

An approximation expert (called "Apx") is also introduced. The oracle expert cannot be used in practice, but gives a benchmark of the method: in this case $\tilde{\beta}_t = \hat{\beta}_t$.

The difference between these two performances measures the gain that can be expected with respect to a crude prediction.

Table 5 MAPE performances for each expert in forecasting 1 year of data from September 1st 2009 to August 31st 2010

Names	Average	Median	Std
Naive	0.0634	0.0415	0.0514
Apx	0.0170	0.0145	0.0145
The experts:			
Yday	0.0300	0.0231	0.0231
Week	0.0293	0.0236	0.0236
T^{med}	0.0315	0.0261	0.0261
T^{med}/W	0.0351	0.0252	0.0252
T^{med}/N	0.0320	0.0257	0.0257
T	0.0311	0.0238	0.0238
T/G	0.0310	0.0232	0.0232
T/D	0.0321	0.0262	0.0262
T/C	0.0295	0.0249	0.0249
N	0.0406	0.0293	0.0293
N/G	0.0282	0.0210	0.0210
N/D	0.0284	0.0220	0.0220
N/C	0.0287	0.0220	0.0220
W	0.0384	0.0294	0.0294
W/G	0.0309	0.0241	0.0241
W/D	0.0381	0.0305	0.0305
W/C	0.0317	0.0256	0.0256

4.3.1 Detailed Performances of the Experts

The naive approach shows a 6.3 % MAPE error. Time lag strategies (Yday, Week) behave well compared to the overall strategies (MAPE average of 3.0 % or 2.9 %). Forecast results are significantly improved by plug-in the sparse estimated coefficients computed the day before instead of taking the raw intra day load curve of the previous day. In order to stress the importance of variable selection, we have computed ordinary Least Square regression (OLS), for theses two time lag strategies. Compared to the LOLA algorithm actually used, no sparsity constraint is introduced in the least square method. For the same period of analyze, the OLS approach shows a MAPE average of 3.98 % for the Yday strategy and of 4.15 % for the Week strategy. With no variables selection, the MAPE average error increases of approximatively 30 %. Similar deteriorations of performances are observed using OLS instead of LOLA for the other strategies.

These results strongly show the benefits of using a sparse functional model to forecast the electrical consumption.

Using LOLA algorithm, strategies associated to the closest cloud cover conditions with constraints (N/G, N/D, N/C), behave especially well, compared for instance, to strategies only based on temperature (T), cloud cover (N) or wind (W). The strategy retrieving the day relying on cloud cover given group information

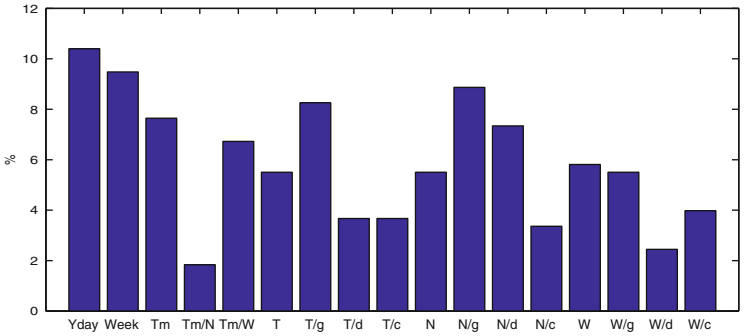


Fig. 7 Frequencies for the expert which perform at best computed for 1 year of data from September 1st 2009 to August 31st 2010

provides, in average, the best MAPE error (mean: 2.82 %; median: 2.10 %). Hence, these results strongly suggest that cloud cover seem to have a important impact in the electricity consumption and then the forecast.

Figure 7 provides for each strategy, the frequency of hits, i.e. the frequency of days over 1 year when the strategy performs at best.

The different strategies seem quite competitive and we observe that all of them performs at best at least 2 % of the days. Strategies based on finding the closest day regarding temperature, cloud cover given the group of day performs in general well (N/G: 9 % of hits), as it was already observed in Table 5. Time dependent strategies, Yday or Week seem also quite competitive with hit frequencies equal to 10 % or 9 %.

These quantitative forecasting results seem to reflect quite well the opinion of the human experts of the discipline.

4.4 Aggregation of Experts

If we were able to find each day the best strategy to apply regarding the MAPE error, the oracle MAPE average error over 1 year would be equal to 1.44 % (standard deviation 0.74 %), as presented in Table 6. This is similar to the approximation MAPE error (Table 4, Apx: 1.70 %) and quite a good performance for these prediction experts which purpose are, as explained before, more adaptation to high dimensional information.

As seen in the previous part, the experts perform successively well depending on days, or meteorological issues. But no one among them achieves the best performance most of the time. There is an obvious need to combining them.

Table 6 MAPE performances in % for aggregation method

Names	Average	Median	Std
Oracle	1.44	1.29	0.74
Exponential weights	2.25	1.92	1.24

In the recent years, many interesting theoretical results as well as practical simulations have been obtained using aggregation and especially exponential penalization: see [4, 7, 8, 11, 23]. These techniques will be for us a good source of inspiration. However, a crucial problem then is to find a weighting, learning the performances of each expert and optimizing them. In this context of prediction, this is quite a challenging issue which can give rise to very sophisticated procedures.

For sake of simplicity we present here a very understandable and manageable one, which only records the approximation properties of each expert and penalizes those with poor approximation results. More precisely, let us recall that $\mathscr{M}$ is the set of strategies introduced above, and $\hat{Y}^s$ the expert forecast computed with the strategy s.

The aggregated expert is a weighted sum of all the forecast consumptions provided by the different experts:

$$\hat{Y}_t = \frac{\sum_{s\in\mathscr{M}} w_t^s \tilde{Y}_t^s}{\sum_{s\in\mathscr{M}} w_t^s}$$

where w_t^s are positive weights depending on the day t and the strategy s.

As explained above, our procedure penalizes by putting small weights, on the strategies which were not able to well approximate the signal at $s(t)$: e.g. the weights w_t^s depend in an exponential way on the l_2 error of $\|Y_{s(t)} - \hat{Y}_{s(t)}||_2^2$:

$$w_t^s = exp(-\|Y_{s(t)} - \hat{Y}_{s(t)}\|_2^2/\theta)$$

$\theta > 0$ is a standard tuning parameter (also called temperature parameter with reference to statistical physics). In the performances presented in the Table 6, this parameter was chosen using cross validation on the past. Using aggregation with exponential weights, we observe that the MAPE decreases to 2.25 % in average and to 1.92 % in median, with a standard deviation of 1.20 %. This error is much smaller than the different errors computed for each individual experts, presented in Table 4 showing the benefits of the different contributions.

Figures 8 and 9 give a graphical illustration of forecast for two different weeks chosen in winter and spring. We observe that forecasts are more accurate during spring periods than winter periods. In Fig. 9, local maxima seem to be overestimated, while local minima are underestimated.

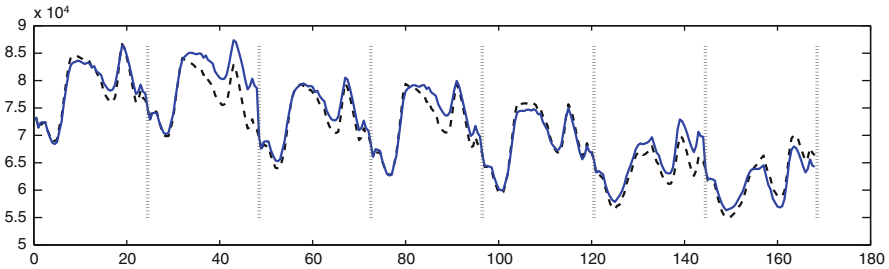


Fig. 8 Forecast (*solid blue line*) and observed (*dashed dark line*) electrical consumption for a winter week from Monday February 1st to Sunday January 7th 2010

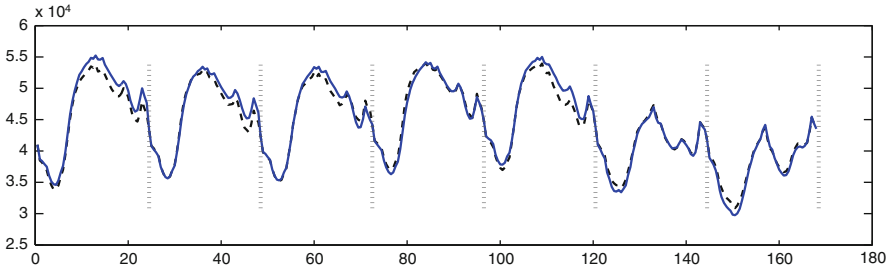


Fig. 9 Forecast (*solid blue line*) and observed (*dashed dark line*) electrical consumption for a spring week from Monday June 14th to Sunday June 21st 2010

5 Conclusion and Perspectives

The method described above will be implemented in a short-term consumption forecasting platform, aggregating various models, which is currently tested at RTE.

This collaboration between academics and industrials provides results which, although coming from automatic statistical methods, happen to agree surprisingly well with the business knowledge of RTE. In fact, they go further and shed new lights:

- The differences between the performances of OLS predictions and the sparse method approach if they were expected from a theoretical point of view, happen to be surprisingly striking (30 % !), strongly motivating for going in this direction in the future.
- The approach for the construction of the experts (searching for *similar days* in the past) which has been established using a mathematical perspective is finally quite close to the strategies of the RTE forecasters. If we refer for instance to the comparison between the expert performances, some results are *common sense* (the lag-expert performances for instance). However, the impressive performances of cloud cover experts were a little more difficult to predict, but already observed by RTE forecasters.

– The competitiveness of these experts, also expected from the RTE forecasters, have been highlighted and quantified. In the near future, more experts will be introduced. As well, the method of aggregation will be diversified according to the feedback of the short term consumption forecasting platform.

Due to forecasting operational needs, different adaptations of the forecasting box will be provided. Particularly, the horizon forecast will be extended to 48 h, or more and the method will be adapted to choose the delivery time of prediction, according to business constraints. In this perspective, the smoothing parameter will be refined to integrate nonparametric regularization methods as well as designed strategies of RTE forecasters.

Acknowledgements The authors thank RTE for the financial support through two industrial contracts, LPMA for hosting our researches, and Karine Tribouley for taking part of an earlier elaboration of this project.

References

1. Antoniadis, A., Brossat, X., Cugliari, J., & Poggi, J. M. (2010). Clustering functional data using wavelets. In *Proceedings of the 19th international conference on computational statistics, COMPSTAT*, Paris, 697–704.
2. Antoniadis, A., Brossat, X., Cugliari, J., & Poggi, J. M. (2012). Prévision d'un processus à valeurs fonctionnelles en présence de non stationnarités. Application à la consommation d'électricité. *Journal de la Société; Française de Statistique, 153*(2), 52–78.
3. Antoniadis, A., Paparoditis, E., & Sapatinas, T. (2006). A functional wavelet–kernel approach for time series prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(5), 837–857.
4. Catoni, O. (2004). *Statistical learning theory and stochastic optimization, volume 1851 of Lecture notes in mathematics*. Berlin: Springer.
5. Chakhchoukh, Y., Panciatici, P., & Bondon, P. (2009). Robust estimation of SARIMA models: Application to short-term load forecasting. In *IEEE workshop on statistical signal processing*, Cardiff
6. Cho, H., Goude, Y., Brossat, X., & Yao, Q. (2013). Modelling and forecasting daily electricity load curves: A hybrid approach. *Journal of the American Statistical Association, 108*(501), 7–21.
7. Dalalyan, A. S., & Tsybakov, A. B. (2008). Aggregation by exponentiel weighting, sharp oracle inequalities and sparsity. In *COLT*, Helsinki (pp. 97–111).
8. Devaine, M., Gaillard, P., Goude, Y., & Stoltz, G. (2012). Forecasting electricity consumption by aggregating specialized experts. *Machine Learning, 90*, 1–30.
9. Fan, S., & Hyndman, R. J. (2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems, 25*(2), 1142–1152.
10. Fan, S., & Hyndman, R. J. (2012). Short-term load forecasting based on a semi-parametric additive approach. *IEEE Transactions on Power Systems, 27*(1), 134–140.
11. Juditsky, A. B., & Nemirovski, A. S. (2000). Functional aggregation for nonparametric regression. *The Annals of Statistics, 28*(3), 681–712.
12. Kerkyacharian, G., Mougeot, M., Picard, D., & Tribouley, K. (2009). Learning out of leaders. In *Multiscale, nonlinear and adaptive approximation* (Lecture notes in computer science). Berlin: Springer.

13. Lefieux, V. (2007). Modèles semi-paramétriques appliqués à la prévision des séries temporelles: cas de la consommation d' électricité.
14. Marin, F. J., Garcia-Lagos, F., & Sandoval, F. (2002). Global model for short term load forecasting using artificial neural networks. *IEE Proceedings – Generation, Transmission, and Distribution, 149*, 121–125.
15. Mougeot, M., Picard, D., & Tribouley, K. (2012). Learning out of leaders: Regression for high dimension. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74*, 475–513.
16. Mougeot, M., Picard, D., & Tribouley, K. (2014). LOL selection in high dimension. *Computational Statistics and Data Analysis, 71*, 743–757.
17. Mougeot, M., Picard, D., Tribouley, K., Lefieux, V., & Maillard-Teyssier, L. (2013). Sparse approximation and fit of intraday load curves in a high dimentional framework. *Advanced in Adaptive Data Analysis, 5*. http://www.worldscientific.com/doi/pdf/10.1142/S1793536913500167.
18. Muñoz, A., Sánchez-Úbeda, E. F., Cruz, A., & Marin, J. (2010). Short-term forecasting in power systems: A guided tour. In *Handbook of power systems II* (pp. 129–160). Berlin/Heidelberg: Springer.
19. Poggi, J. M. (1994). Prévision non paramétrique de la consommation d'électricité. *Revue de Statistique Appliquée, 42*, 83–98.
20. Ramsay, J. O., & Silverman B. W. (2005). *Functional data analysis*. New York: Springer.
21. Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research, 204*(1), 139–152.
22. Taylor, J. W. (2012). Short-term load forecasting with exponentially weighted methods. *IEEE Transactions on Power Systems, 27*, 458–464.
23. Tsybakov, A. B. (2003). Optimal rates of aggregation. In *COLT*, Washington, DC (pp. 303–313).

Modelling and Prediction of Time Series Arising on a Graph

Matthew A. Nunes, Marina I. Knight, and Guy P. Nason

Abstract Time series that arise on a graph or network arises in many scientific fields. In this paper we discuss a method for modelling and prediction of such time series with potentially complex characteristics. The method is based on the lifting scheme first proposed by Sweldens, a multiscale transformation suitable for irregular data with desirable properties. By repeated application of this algorithm we can transform the original network time series data into a simpler, lower dimensional time series object which is easier to forecast. The technique is illustrated with a data set arising from an energy time series application.

1 Introduction

Many multivariate time series encountered in practice will be observed on a network or graph. The network on which these time series is observed is often large, limiting the feasibility of some data analyses in some settings. This article discusses the potential of a particular wavelet-like data transformation, called lifting, to facilitate modelling and forecasting of time series arising on large networks. The motivation for this approach stems from lifting's ability to 'simplify' the series under analysis. The transformation employed by the method naturally handles data on irregular observation domains, such as spatial or network data. As such, time series arising on complex domains can be analysed, and we propose that the technique described will be of use for practitioners in many industrial and scientific fields. Lifting was introduced by Sweldens [7, 8]. We review and utilize the 'lifting one coefficient at

M.A. Nunes (✉)
Department of Mathematics and Statistics, Lancaster University, Lancaster, UK
e-mail: m.nunes@lancaster.ac.uk

M.I. Knight
Department of Mathematics, University of York, York, UK
e-mail: marina.knight@york.ac.uk

G.P. Nason
School of Mathematics, University of Bristol, Bristol, UK
e-mail: g.p.nason@bristol.ac.uk

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_10

a time' (LOCAAT) version of the lifting transformation [3, 4] in Sect. 2, followed by a brief description of how we use this algorithm for network time series analysis in Sect. 3. Section 4 explores the performance of the technique on a toy example involving wind energy time series.

Notation The data under analysis are observations on the nodes of a fully connected graph taken through time. The graph consists of a set of $R \in \mathbb{N}$ nodes, $\mathscr{R} = \{1, \ldots, R\}$, which are connected by a set of edges. Define the edge set $\mathscr{E}_r = \{j \in \mathscr{R} : \text{node } j \text{ is connected to node } r\}$. The entire graph $\mathscr{G} = (\mathscr{R}, \mathscr{E})$ where $\mathscr{E} = \cup_{r \in \mathscr{R}} \mathscr{E}_r$. The data are values of a function defined on the nodes of the graph. Suppose our data set is $\{X_{r,t}\}$ where $r \in \mathscr{R}$ are nodes in $\mathscr{G}$ and $t = 1, \ldots, T$ is time where $T \in \mathbb{N}$. In what follows, we shall also denote this data by $\mathbf{X}_t$ to illustrate its time dependence over the network structure.

2 A Wavelet-Like Transform for Network Data

Our new forecasting technique, described in Sect. 3, makes use of a dimension reduction step via a wavelet-like lifting transformation called LOCAAT. This section gives an overview of the LOCAAT lifting transformation for network data based on the exposition in Jansen et al. [3, 4].

Lifting 'one coefficient at a time' (LOCAAT) Let $X^{\mathscr{R}}$ denote data on a network $\mathscr{G}$ for a fixed timepoint, t. We do not explicitly mention the time index in the description of LOCAAT, but it must be remembered in the background. The lifting algorithm transforms a set of data into a simpler form by iteratively repeating a number of linear combinations of the original data. In particular, lifting a discrete signal consists of three steps: *split, predict* and *update*. These steps can be described as follows:

Let us denote our initial data by $\{c_{n,r}\}_{r=1}^{R} := X^{\mathscr{R}}$.

Split. A node r_n is chosen to be *lifted*.

Predict. The function value at the chosen node r_n is then *predicted* using the function values of its immediate neighbours in the network, $\mathscr{E}_{r_n}$, as regressors. The difference between the true and predicted function value at r_n is then computed:

$$d_{r_n} = c_{n,r_n} - \sum_{j \in \mathscr{E}_{r_n}} w_j^n c_{n,j},$$

for regression weights $\underline{w}^n$. Typically, these weights are chosen according to some measure of association between the lifted node r_n and the neighbour nodes. We denote the difference as d_{r_n} and refer to it as the detail coefficient at the node r_n.

Update. The function values at the neighbouring positions $j \in \mathscr{E}_{r_n}$ are updated by using linear combinations of the detail coefficient obtained in the prediction step.

In other words,

$$c_{n-1,j} := c_{n,j} + \tilde{w}_j^n d_{r_n}, \ \forall j \in \mathscr{E}_{r_n}, j \neq r_n,$$

where $\underline{\tilde{w}^n}$ are weights chosen to keep the energy of the signal constant at each lifting step (see Jansen et al. [4] for more details).

The node r_n is then removed from the network, thus creating a 'coarser' network structure; the edge sets $\mathscr{E}_j$ for each neighbour j of r_n will thus change after the node r_n is lifted. A schematic of the lifting steps 1–3 is shown in Fig. 1.

The three lifting steps are then repeated successively removing a node at each step. After a large number of network nodes have been lifted, the original data is transformed into a set of detail coefficients. The remaining unlifted coefficients are called *scaling coefficients*. The detail coefficients $d^{\mathscr{R}}$ capture the local changes in the data at a particular node in the network. An example of the lifting transform operating on some example wind speed data (see Sect. 4) is shown in Fig. 3. The original data along with its transformed version is shown.

The lifting transform can also be written as a matrix operator, i.e.

$$d^{\mathscr{R}} = \mathscr{W} X^{\mathscr{R}}.$$

The lifting transformation fully takes into account the network structure, and is invertible: the original data can be retrieved using the inverse transformation matrix $X^{\mathscr{R}} = \mathscr{W}^{-1} d^{\mathscr{R}}$.

LOCAAT can be used for regression (function estimation) on a network in a signal plus noise model, see Jansen et al. [3, 4] and Mahadevan et al. [6].

LOCAAT lifting for network time series A benefit of the lifting transformation is its ability to transform data in an efficient manner in order to obtain desirable properties. Suppose we have the time series $\mathbf{X}_t$ on a network $\mathscr{G}$. In our setup, the network can either be real, or can be constructed according to a practitioner's insight into the underlying problem for the series under analysis. Our simplification of the network time series proceeds as follows. For *each* fixed timepoint $t = 1, \ldots, T$, we perform the network LOCAAT lifting algorithm. This results in transformed data $\{Y_{r,t}\}_{t=1}^T$. Notationally, we write this as

$$\mathbf{Y}_t = \mathscr{W} \mathbf{X}_t, \tag{1}$$

where $\mathbf{Y}_t$ is an r-dimensional vector time series and $\mathbf{X}_t$ is the original function $X_{r,t}$ on the graph $\mathscr{G}$ at time $t = 1, \ldots, T$. The transformation of the entire time series object can be inverted by simply performing the inverse lifting transform for each fixed timepoint t via $\mathscr{W}^{-1}$. In this article the network topology is fixed as a function of time. The values of the network function at the nodes may change at each time step, but the number of nodes and how they are connected remains constant.

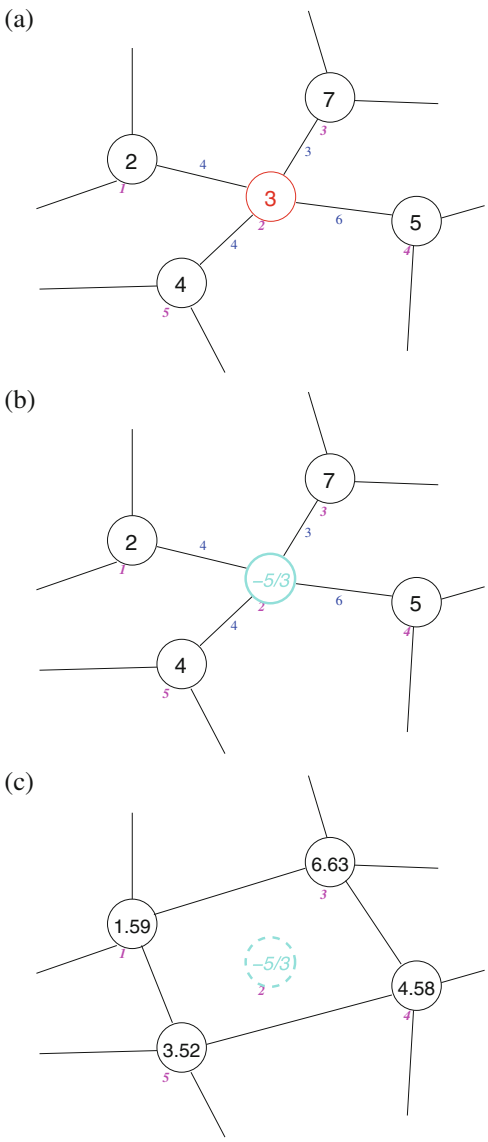


Fig. 1 Example of LOCAAT lifting: values in circles (nodes) represent observed process values X_t. Node indices are shown underneath circles; weights associated to network edges are also shown. (**a**) Region of network before lifting step: node chosen to be lifted (node 2, centre). (**b**) Region of network after *prediction* step: the detail coefficient is produced as a linear combination of neighbouring process values. (**c**) Region of network after *update* step: lifted node is removed. Process values at neighbouring nodes are updated according to their contribution to the detail coefficient in the prediction step; the network is then reconnected

3 Network Time Series Forecasting via LOCAAT Lifting

After LOCAAT lifting, the transformed time series $\mathbf{Y}_t$ can be split into two sets:

$$Y_{r,t} = \begin{cases} s_{r,t} \text{ for } r \in \mathscr{S}(\mathscr{G}), \\ q_{r,t} \text{ for } r \in \mathscr{Q}(\mathscr{G}), \end{cases} \tag{2}$$

where $\mathscr{S}(\mathscr{G})$ are the lifting scaling function coefficients from performing lifting on $\mathscr{G}$, representing the "trend" of the time series object. The LOCAAT wavelet coefficients of the new time series object are denoted by $\mathscr{Q}(\mathscr{G})$. The sets $\mathscr{S}(\mathscr{G})$ and $\mathscr{Q}(\mathscr{G})$ form a disjoint union of the time series across the network.

Suppose we want to forecast the network time series object $\mathbf{X}_t$ at time t for h steps ahead, denoted by $\hat{\mathbf{X}}_{t+h}$. Our strategy for this task is to transform the data via LOCAAT lifting, forecast in the transformation domain, and then invert the transformation including the forecast. Specifically, our forecasting procedure is as follows.

1. Transform the original data via the LOCAAT transformation, described in Sect. 2 to form $\mathbf{Y}_t = \mathscr{W}\mathbf{X}_t$.
2. Construct the wavelet domain forecast $\hat{\mathbf{Y}}_{t+h}$ by
   (a) Extrapolating the scaling coefficient series $s_{r,t}$ in $\mathscr{S}(\mathscr{G})$,
   (b) Modelling the wavelet coefficient time series in $\mathscr{Q}(\mathscr{G})$ as stationary ARMA processes, and use a forecasting method to predict each time series $q_{r,t}$ in $\mathscr{Q}(\mathscr{G})$, e.g. using a Box-Jenkins [1] or exponential smoothing [2] approach.
3. Invert the network time series transform with $\mathscr{W}^{-1}$ to form the time domain forecast for the *original series* as $\hat{\mathbf{X}}_{t+h} = \mathscr{W}^{-1}\hat{\mathbf{Y}}_{t+h}$.

Note that in step 2b above, any wavelet domain model can be chosen according to the context or choice of the practitioner.

The next section demonstrates the potential of this lifting-based time series prediction technique with an application arising from energy time series.

4 Application to Energy Time Series

The data that we analyse with our network time series prediction method are hourly wind speeds collected from a number of UK Metereological Office weather stations [9]. Our data set consisted of $N = 102$ weather stations, and $T = 721$ hourly records.

Note that the wind data are not supplied with a network, but we can construct one from geographical proximity information. In what follows, we obtain a network for the data using a *minimal spanning tree* (MST) construction as suggested in Jansen et al. [4]. A minimal spanning tree is a method of connecting a set of points in space in a network (in our case spatial locations) such that all nodes are connected to at least one other. The network has the added property that the total "weight" along all edges in the constructed tree is minimal amongst all potential networks. As a result, the network contains a low number of edges. For the data described above, these edge weights are distances between the weather stations. The MST thus provides a representation of potential geographic information linking the data at wind speed sites. See Krzanowski and Marriott [5] for more details on the MST construction. The minimal spanning tree that we constructed for the wind data is shown in Fig. 2.

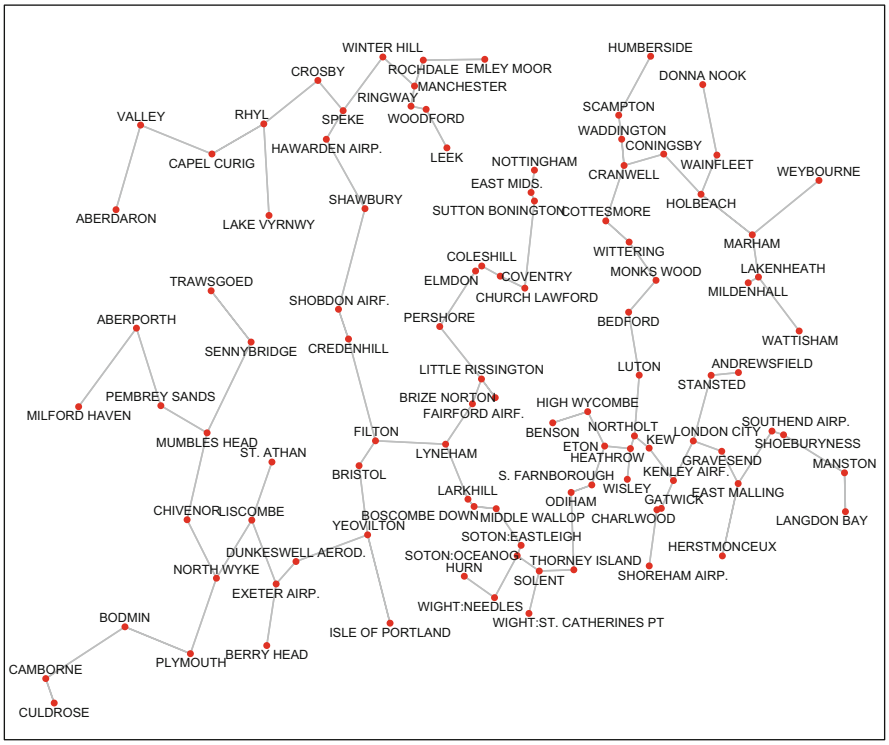


Fig. 2 UK Metereological Office weather stations together with the constructed minimal spanning tree network between stations

We aim to use our proposed method for forecasting the wind speed across the entire network at the last time instant ($t = 721$). Therefore in what follows we shall treat the data as having been observed at the first 720 timepoints. The lifting transformation described in Sect. 2 was applied to the network time series data for each timepoint $t = 1, \ldots, 720$, transforming the original wind data into a lifted time series. More specifically, 100 lifting steps were performed for a particular fixed timepoint, leaving two scaling coefficient series, namely $\mathscr{S}(\mathscr{G}) = \{Filton, Kenley\ Airfield\}$. The original and transformed data for the first time instance are shown in Fig. 3a, b respectively. In this example, the order in which nodes are removed to form lifting coefficients is decided by the default choice in the lifting software. Roughly speaking, points are removed from higher density regions first, progressively moving to lower density regions. This behaviour mimics the usual wavelet methodology of proceeding from fine to coarse scales.

Following this transform, the original network series object is now split into the 100 weather stations associated to the lifting (wavelet) coefficients and the two stations corresponding to the scaling functions; at each node we have a new (lifted) time series (of length 720). This new LOCAAT object was then used in the fore-

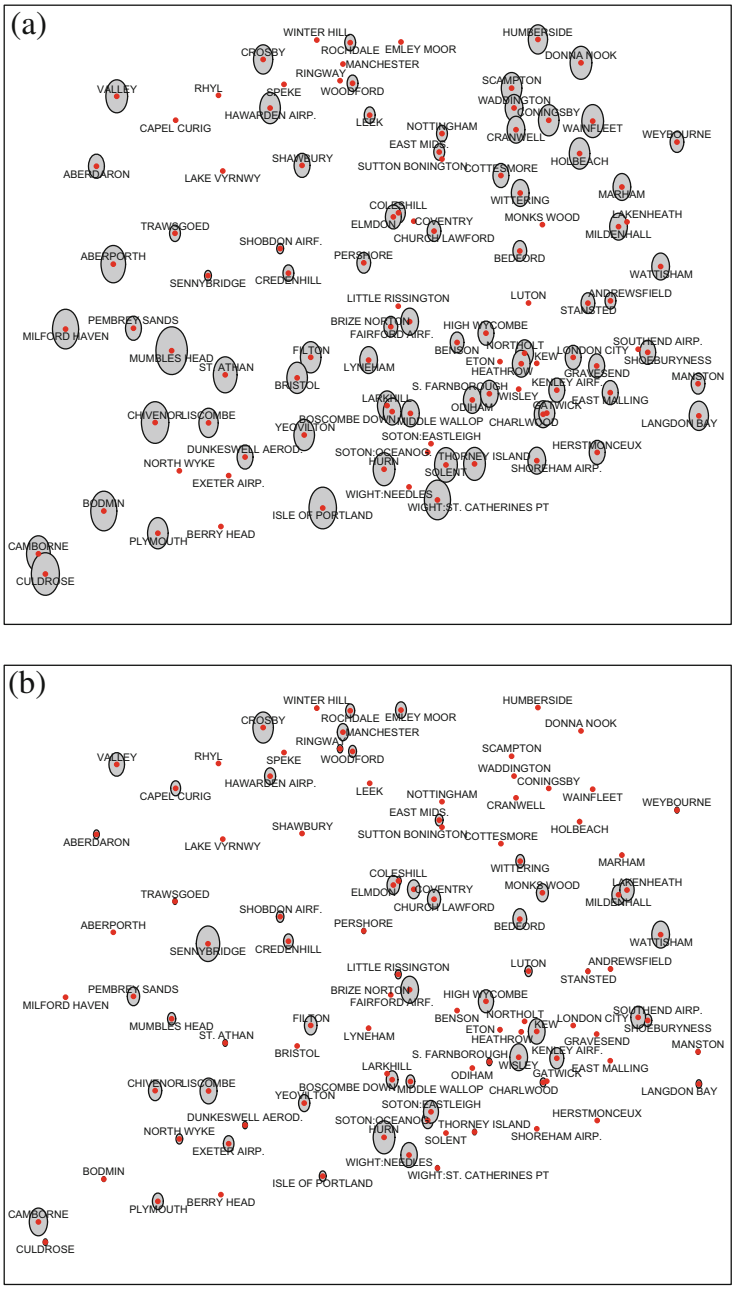


Fig. 3 Example of LOCAAT lifting across the weather station network, before and after LOCAAT algorithm (for timepoint $t = 1$). Process values are transformed into detail coefficients capturing local changes in the data. Radii of circles represent magnitude of series values. (**a**) Wind speed data before lifting for $t = 1$. (**b**) Wind speed lifting coefficients for $t = 1$

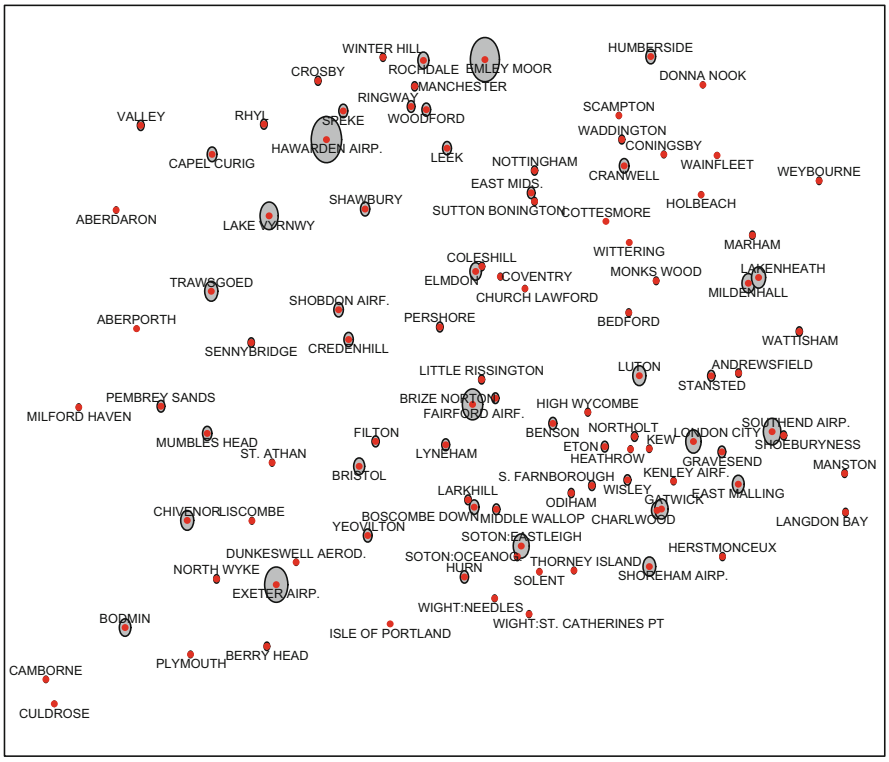


Fig. 4 Prediction error for the weather stations network with the forecasting technique based on the lifting algorithm. Circle diameters represent the squared forecasting error for each station in the network

casting technique outlined in Sect. 3, using $h = 1$ step ahead forecasts. We permit both AR and ARMA modelling of detail coefficient series in the lifting domain through automatic selection of the ARMA model parameters. Figure 4 displays the squared prediction error of the transformation-based forecasting method, i.e. $\{e_r\}_{r=1}^{102} = \{(X_{r,721} - \hat{X}_{r,721})^2\}_{r=1}^{102}$, when using the data up to the previous timepoint, $X^{\mathscr{R}}$, $t = 1, \ldots, 720$.

A comparison with forecasting the data without transformation (using the previous timepoint as a predicted value, as well as time-domain ARMA prediction) indicates that our LOCAAT method reduces the average squared forecasting error $\bar{e} = \frac{1}{102} \sum_{r=1}^{102} e_r$ significantly. A typical example of the one-step ahead prediction of the last datapoint for the three methods (at the Lakenheath weather station) is shown in Fig. 5.

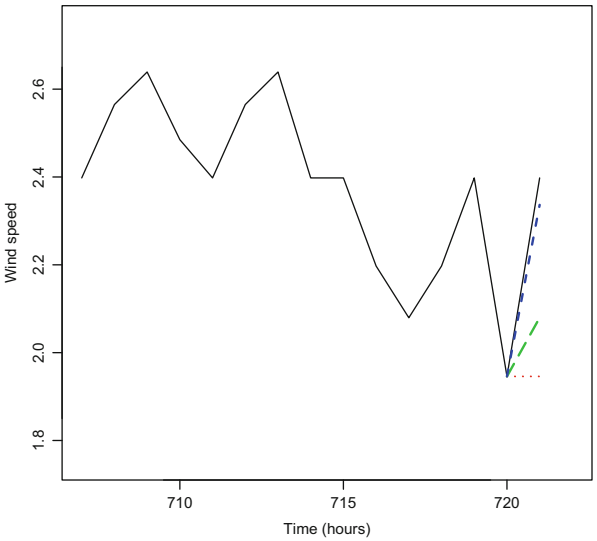


Fig. 5 Example one step ahead forecasts for the wind speed data at Lakenheath. Graph shows true data (*solid*); naive prediction with last datapoint (*dotted*); time domain ARMA forecasting (*long dashed*) and LOCAAT-based prediction (*short dashed*)

5 Summary

This article has described a method for predicting time series data which arise on a graph or network by forward transforming the values on the network at each time point, modelling and forecasting the coefficients in the transform domain and then transforming back to obtain forecasts in the original domain. The results of the method for the practical application to wind speed modelling and forecasting are promising. The aim of our method is to use LOCAAT to gain improved prediction performance in many areas in which data and analysis simplification are of importance, such as geostatistical or complex network data. It is likely that other transforms related to be one presented here could be used to simplify other time series analysis tasks, where the dimension of the problem renders analysis difficult or computationally infeasible.

Whilst in this article we have considered data arising on a fixed graph, we recognise that in many applications, analysis of time series on a dynamic (time-varying) graph is of interest. One change, that we are currently investigating, is for the number of nodes and basic connectivity to remain the same, but for the edge weights to vary over time. This is left as an interesting avenue of further research.

Acknowledgements We would like to thank the organisers of the "2nd Workshop on Industry Practices for Forecasting" (WIPFOR13) for an instructive and stimulating meeting. The authors would like to gratefully acknowledge the UK Met Office and British Atmospheric Data Centre (BADC) for access to the wind speed data.

References

1. Anderson, O. D. (1976). *Time series analysis and forecasting: The Box-Jenkins approach*. London: Butterworths.
2. Gardner, E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting, 4*(1), 1–28.
3. Jansen, M., Nason, G. P., & Silverman, B. W. (2001). Scattered data smoothing by empirical Bayesian shrinkage of second generation wavelet coefficients. In M. Unser & A. Aldroubi (Eds.), *Wavelet applications in signal and image processing IX* (Vol. 4478, pp. 87–97). SPIE.
4. Jansen, M., Nason, G. P., & Silverman, B. W. (2009). Multiscale methods for data on graphs and irregular multidimensional situations. *Journal of the Royal Statistical Society: Series B, 71*(1), 97–125.
5. Krzanowski, W. J., & Marriott, F. H. C. (1995). *Multivariate analysis part 2: Classification, covariance structures and repeated measurements* (Kendall's library of statistics, Vol. 2). London: Edward Arnold.
6. Mahadevan, N., Nason, G., & Munro, A. (2008). Multi-dimensional network function estimation. In *IEEE international conference on communications*, Beijing.
7. Sweldens, W. (1996). The lifting scheme: A custom-design construction of biorthogonal wavelets. *Applied and Computational Harmonic Analysis, 3*(2), 186–200.
8. Sweldens, W. (1998). The lifting scheme: A construction of second generation wavelets. *SIAM Journal on Mathematical Analysis, 29*(2), 511–546.
9. UK Meteorological Office. (2013). *Met Office integrated data archive system (MIDAS) land and marine surface stations data (1853-current)*. Harwell/Oxford: British Atmospheric Data Centre.

Massive-Scale Simulation of Electrical Load in Smart Grids Using Generalized Additive Models

Pascal Pompey, Alexis Bondu, Yannig Goude, and Mathieu Sinn

Abstract The emergence of Smart Grids is posing a wide range of challenges for electric utility companies and network operators: Integration of non-dispatchable power from renewable energy sources (e.g., photovoltaics, hydro and wind), fundamental changes in the way energy is consumed (e.g., due to dynamic pricing, demand response and novel electric appliances), and more active operations of the networks to increase efficiency and reliability. A key in managing these challenges is the ability to forecast network loads at low levels of locality, e.g., counties, cities, or neighbourhoods. Accurate load forecasts improve the efficiency of supply as they help utilities to reduce operating reserves, act more efficiently in the electricity markets, and provide more effective demand-response measures. In order to prepare for the Smart Grid era, there is a need for a scalable simulation environment which allows utilities to develop and validate their forecasting methodology under various what-if-scenarios. This paper presents a massive-scale simulation platform which emulates electrical load in an entire electrical network, from Smart Meters at individual households, over low- to medium-voltage network assets, up to the national level. The platform supports the simulation of changes in the customer portfolio and the consumers' behavior, installment of new distributed generation capacity at any network level, and dynamic reconfigurations of the network. The paper explains the underlying statistical modeling approach based on Generalized Additive Models, outlines the system architecture, and presents a number of realistic use cases that were generated using this platform.

P. Pompey (✉) • M. Sinn
IBM Research, Damastown Industrial Estate, Dublin 15, Ireland
e-mail: papompey@ie.ibm.com; mathsinn@ie.ibm.com

A. Bondu • Y. Goude
EDF R&D, 1 Avenue du Général De Gaulle, 92140 Clamart, France
e-mail: alexis.bondu@edf.fr; yannig.goude@edf.fr

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_11

1 Introduction

The French electrical grid is currently being fundamentally modernized by deploying Information and Communication Technology at a massive scale. The emerging "*Smart Grid*" is designed to meet multiple objectives: (i) optimizing the control of the grid and the quality of the electricity supply, despite the fact that power generation is becoming more decentralized; (ii) scheduling the production of energy while taking into account the uncertainty related to renewable energy sources (e.g., photovoltaics, hydro and wind); (iii) coordinating and shaping the energy demand to flatten consumption peaks and limit their impact on the networks and on the electricity markets.

"*Smart Meters*" constitute the fundamental building block of the Smart Grid architecture. Within the next few years, these digital meters are expected to be installed at all French households.[1] Smart Meters record the individual power consumptions in real time, and send this information to a data center through a communication network. The expected volume of Smart Meter data (in France: 35 millions signals sampled every 30 min) poses a significant challenge for utility companies. In France, one year of Smart Meter data amounts to more than 600 billion data points, which is equivalent to 4.4 Terabytes.[2] Electricité de France (EDF), the main French provider of electricity, needs to anticipate managing such amounts of data in terms of **storage**, **querying** and **data analysis** capabilities. Currently, only a small subset of the 35 million Smart Meters has already been deployed, mostly through pilot studies in specific geographic areas. In order to prepare for the full deployment and test different types of distributed data management systems, EDF needs to simulate consumption data for individual households at a massive scale.

Previous studies on massive-scale processing of electrical load time series have been carried out using the Hadoop framework [8]. Also the data storage and querying aspects have been investigated in this context. The present paper describes a platform for **more realistic** simulations of electricity consumption in order to validate forecasting approaches at different levels of the electrical grid. The platform also supports the generation of **what-if-scenarios** to foresee the impact of changes in electricity usage on the quality of the forecasts. Note that electricity consumption data at the level of individual households have several distinctive features: (i) the overall number of time series is very large; (ii) the diversity of individual behavior induces a wide variety of shapes; (iii) the volatility of these time series is very high; (iv) the sum of these time series is a smooth time series with cyclical patterns. The upper time series in Fig. 1 shows the total consumption in France during 1 week, and the lower time series gives an example of an individual consumption time series during the same period of time. As can be seen, the characteristics of the load profiles at these different aggregation levels are very different.

[1] More details are available at http://www.erdfdistribution.fr/linky/

[2] Assuming that each data point requires 8 bytes memory.

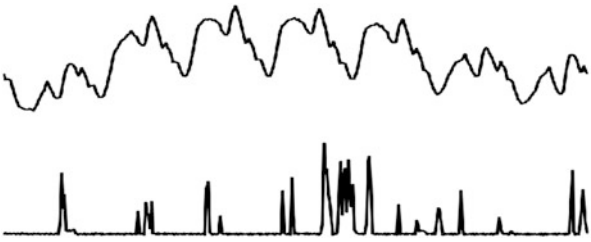


Fig. 1 Example of an individual consumption signal during 1 week *(lower time series)*, in comparison with the sum of individual consumption signals during the same period of time *(upper time series)*

This paper is organized as follows. After a review of related work on the simulation of electrical networks in Sect. 2, Sect. 3 introduces the statistical modeling approach for simulating and forecasting electrical load based on Generalized Additive Models. Section 4 describes the architecture of the simulation platform. Use cases demonstrating applications of the simulation platform are presented in Sect. 5. Finally, Sect. 6 proposes a benchmark method to evaluate how realistic are the simulations generated by the platform at different aggregation levels. Section 7 concludes with an outlook on directions for future work.

2 Related Work

There exists a wide body of literature and software tools for simulating electrical networks. Most of these tools focus on physical properties of the grid (e.g., power flows, voltage drops), typically under steady-state conditions and for a limited part of the network (e.g., transmission or distribution), and with a great level of detail in modeling the physical assets of the grid (lines, transformers, etc.). The purpose of the simulation platform presented in this paper is to emulate **statistical** properties of electrical load. In this context, bottom-up and top-down approaches have been proposed in the literature (see [19] for a detailed review). Bottom-up methods start by modeling the usage of individual electrical appliances (e.g., by a Multi-Agent System) and then compute the aggregated load, e.g., at the household or neighborhood level. While those approaches yield detailed and realistic simulations at a high temporal resolution, they are computationally expensive, require considerable modeling effort, and typically rely on assumptions about the usage of appliances that are difficult to justify empirically. Typically, bottom-up methods are used for loads only at low-level aggregations, e.g., to simulate Microgrids.

Top-down methods start by modeling aggregated load curves which are then iteratively disaggregated using statistical methods to obtain the consumption at lower levels. The main advantage of this approach is that a variety of models can be used to accurately represent features of aggregated load, and usually high-quality data for fitting those models is available at the top aggregation levels. However, top-down approaches often fail to reproduce distinctive features of disaggregated loads,

e.g., the volatility of loads at lower aggregation levels, and the localized effects of meteorological and socio-economic variables.

The simulation platform presented in this paper is designed to emulate loads throughout the entire electrical network (from individual households over low- to medium-voltage network assets up to the transmission and national level) for a country the size of France, over multiple years and under various what-if-scenarios. To the best of the authors' knowledge, there exists no previous solution for simulation studies of this scale. Another special feature of the platform presented in this paper is the modeling approach based on Generalized Additive Models, which will be discussed in the following section. As will be shown in Sect. 6, while this approach does not capture all the distinctive features of loads at individual households, it reflects well the characteristics of aggregates of 70 households or more. Hence, it can be argued that the modeling approach proposed in this paper offers a good compromise between top-down and bottom-up methods.

3 Generalized Additive Models

3.1 Background

Generalized Additive Models (GAMs) are a class of semi-parametric regression models introduced in [12] and [13]. Originally, the learning of GAMs was done using the backfitting algorithm, but recently more efficient methodologies have been introduced, among them boosting procedures (see [3]) and penalized regression methods (see [22]). GAMs have been successfully applied to electrical load forecasting at different geographical scales and network aggregation levels. For example, [18] uses GAMs to forecast the French load at the national level, achieving a Mean Absolute Percentage Error (MAPE) of less than 2 %. Ba et al. [1] studies the same data set and proposes an online learning algorithm for GAMs which is shown to further improve the forecasting accuracy. Fan and Hyndman [9] applies GAMs to regional data in the National Electricity Market of Australia, [16] shows results on data from a US utility company, and [11] demonstrates forecasting at the substation level in France. Experiments in Sect. 6 of the present paper suggest that GAMs are applicable to small aggregates of down to 70 households.

GAMs have properties which make them useful both for simulation and forecasting: They are able to capture complex non-linear relationships (e.g., between electrical load and temperature), and their estimation and prediction are straightforward. Another interesting feature of GAMs is their simplicity due to their additive structure, which makes them easy to use and understand by practitioners. This property is of particular importance in the simulation context, because it allows domain experts to design specific what-if-scenarios.

Mathematically, GAMs have the following form:

$$y_i = f_1(x_{1,i}) + f_2(x_{2,i}) + \ldots + f_p(x_{p,i}) + \varepsilon_i$$

where y_i is a univariate response variable (here the electrical load), $x_{q,i}$ are the covariates that shape y_i (e.g., meteorological conditions, the time of day, the day of week, etc.). ε_i denotes the model error at time i, also called "*noise*" in this paper. The simulation platform presented in this paper supports different types of noise: White noise, Autoregressive noise, and Heteroscedastic noise where the variance of ε_i at time i could depend on the covariates $x_{q,i}$. The functions f_q, called "*transfer functions*" in this paper, are centered around 0 to achieve model identifiability and represented using splines (in particular, they can be non-linear). A penalization term in the model estimation enforces smoothness of the transfer functions. More specifically, using the spline representation each transfer function can be written as follows:

$$f_q(x) = \sum_{j=1}^{k_q} \beta_{q,j} b_j^q(x)$$

where k_q is the dimension of the spline basis, and $b_j^q(x)$ are the corresponding basis functions (e.g., cubic B-splines) with the spline coefficients $\beta_{q,j}$. In order to estimate the spline coefficients of all the transfer functions while enforcing smoothness, the following objective is minimized:

$$\sum_{i=1}^{n} (y_i - \sum_{q=1}^{p} f_q(x_i))^2 + \sum_{q=1}^{p} \lambda_q \int \|f_q''(x)\|^2 dx.$$

Here $\Lambda = (\lambda_1, \ldots, \lambda_p)$ is a vector of penalty parameters controlling the degree of smoothness of each transfer function (the higher λ_q, the smoother f_q). This parameter is optimized through a model selection criterion, e.g., see the methodology in [21] and [23] which minimizes the Generalized Cross Validation criterion proposed in [7]. For practical computations in this paper, the implementation in the R package `mgcv` (see [20] and [22]) is used.

3.2 Load and Wind Farm Modeling

This subsection provides examples of GAMs which will be used in Sect. 5 to configure different use cases running on the simulation platform. The data set used for learning the load models was compiled by the Irish Commission for Energy Regulation (CER) in a Smart Metering trial (see the reports [5] and [6]). The data were collected half-hourly for every meter participating in the trial from

July 14th, 2009, to December 31st, 2010. In this paper, meters with missing values or replications were discarded; the resulting cleaned data set consisted of 4,623 m (residential customers and small-to-medium enterprises), each with 48 half-hourly meter readings per day over 536 days. For simplicity, days corresponding to daylight savings were dropped: October 25th, 2009, March 28th and October 31st, 2010. As the location of the individual meters is anonymized for confidentiality reasons, the weather data from the Dublin airport (downloaded from wunderground.com) were used as the meteorological covariates in the load models.

As part of the CER Smart Metering trial, one out of five different tariff classes was offered to each residential household. For the experiments in this paper, the load of households using the same tariff was aggregated, and one GAM per class was estimated. Figure 2 shows 2 weeks of data for each of the five classes. The GAM learned for each class is given by

$$y_i = \sum_{k=1}^{7} s_k(\text{TimeOfDay}_i) I_{\text{WeekDay}_i=k} + s(\text{Temperature}_i) + s(\text{TimeOfYear}_i) + \varepsilon_i \quad (1)$$

where y_i is the electrical load, TimeOfDay_i is the time of day (ranging from 0 to 47, corresponding to the half-hourly measurements at 0:30, 1:00, ..., 24:00), WeekDay_i is the day of week (1 = Sunday, 2 = Monday, ..., 7 = Saturday), Temperature_i is the temperature at the Dublin airport, and TimeOfYear_i is the time in the year (ranging between 0 on January 1st and 1 on December 31st). Note that $I_{\text{WeekDay}_i=k}$ denotes the indicator function which evaluates to 1 if $\text{WeekDay}_i = k$, and to 0, otherwise. Hence, the model includes a transfer function depending on the time of day which is specific for each week day. The transfer functions are represented using cubic B-splines, and cyclic splines for the TimeOfYear effect which enforces continuity between December 31st and January 1st. In the simulations, the noise term ε_i is sampled from a normal distribution with zero mean and a standard deviation proportional to 1 % of the signal, i.e., as explained in the previous subsection, the variance also depends on the model covariates (here: TimeOfDay_i, WeekDay_i, Temperature_i and TimeOfYear_i).

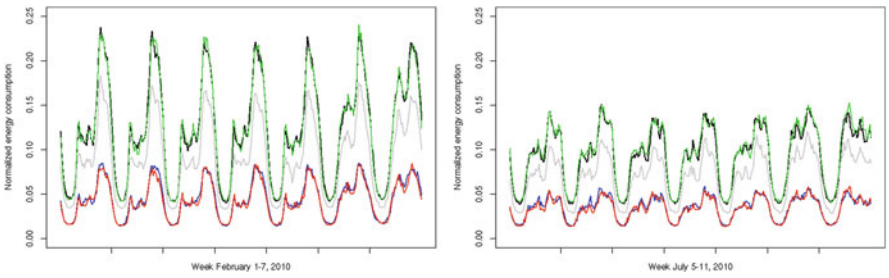


Fig. 2 Irish CER data set: Electricity consumption of residential customers signed up to five different tariff classes (represented by the curves in different colors)

For the learning of a wind farm model, a public data set from the wind power forecasting track of the GEFcom competition (see [16]) was used. In the experiments of this paper, this model was standardized and, in order to simulate wind farms of different sizes, scaled to the desired level. The GAM model is given by

$$y_i = \sum_{k=1}^{12} s_k(\text{WindSpeed}_i) I_{\text{WindDirection}_i = k} + \varepsilon_i \tag{2}$$

where y_i is the wind power, WindSpeed_i is the wind speed, and WindDirection_i the wind direction (1 = N, 2 = NNE, 3 = NE, ..., 12 = NNW). Note that the wind direction was discretized into 12 sectors (instead of using a bivariate transfer function) for parsimony reasons. In the simulations, the noise term ε_i is sampled from a normal distribution with zero mean and a standard deviation proportional to 5 % of the signal (to simulate higher uncertainty of production data), i.e., the variance again also depends on the model covariates (here: WindSpeed_i and WindDirection_i).

4 Simulator Platform Architecture

This section describes the architecture and design of the Smart Grid simulation platform, with particular emphasis on the modeling of the electrical network, the representation of load at the Meter level, and design considerations related to the scalability of the platform.

4.1 Network Modeling

Simulating the load at each level of an electrical grid requires a model of the network. The simulation platform presented in this paper models the initial network structure, and dynamic changes (e.g., reconfiguration events) applied to it. The initial **network structure** is a tree of depth six, with the nodes – from the lowest level to the root – representing Meters, Low-Voltage Stations (LVS), High-Voltage Stations (HVS), Source Substations, Regional Agencies, and the National Level. An example of a subtree, up to the Regional Agency level, is shown in Fig. 3. The numbers on the right hand side correspond to the number of nodes per network level for a country the size of France. Note that the tree structure only allows for the representation of radial networks; modeling meshed networks is a direction for future work.

To ensure the resilience and security, numerous backup lines exist in real electrical networks that enable to redirect the electrical flow from one element to

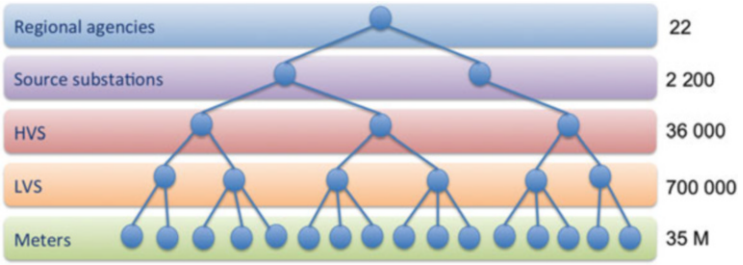


Fig. 3 Tree-based representation of the network structure grid with the approximate number of nodes per network level for a country the size of France

another. The simulation platform can take into account **dynamic reconfigurations** where, at a given point in time, a leaf node or an internal node (with its subtree) connect to a different parent node at the upper level. This can also be used to emulate mobile network elements like electric vehicles which might change their connection point to the grid depending on their location. In most networks, backup lines only exist between few, but not all the nodes. The simulation platform is able to enforce "can connect/cannot connect" constraints to ensure that dynamic reconfigurations only connect network elements that are physically linked with each other.

The **aggregated load** at any internal node in the network structure is obtained by simply taking the sum of the loads from all children elements in the tree. Electricity production (e.g., from distributed renewable energy sources) can also be taken into account and modeled as negative load. The simulation platform supports separate aggregation of load, production and net load (i.e., the difference between load and production); moreover, load can be aggregated separately for different customer classes. Note that the simulation platform does not model physical properties and only aggregates active powers. In particular, line losses are neglected, and there is no calculation of currents, voltages and other physical quantities in the network.

4.2 Representation of Load at the Meter Level

The simulation platform uses two attributes for characterizing load at the Meter level in the network: The statistical model which is used for simulating the load at a particular Meter, and the geographical location of the Meter. Typically, the simulation model is chosen from a set of "*customer classes*", e.g., representing the behavior of customers signed up to different tariffs as shown in Sect. 3.2.[3] Similarly, also simulation models for energy production (e.g., from wind farms)

[3]It is important to note that the GAMs learned on aggregated load data do not really represent load at the individual Meter level, but more an "average consumer". As will be shown in Sect. 6, GAMs fit well for aggregates of 70 households or more. The purpose for using GAMs, nevertheless, at

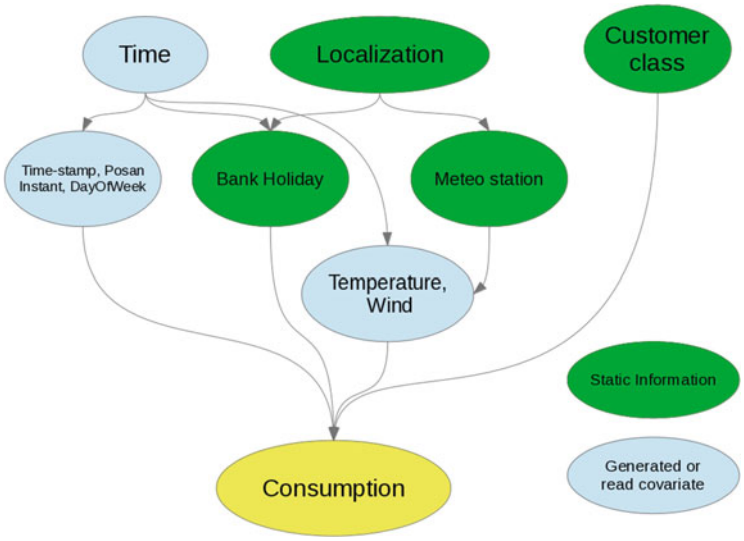


Fig. 4 Bayesian Belief Network representing the dependencies among the simulation models and covariates based on localization, time and customer type information

can be deployed at the Meter level. The geographic location of the Meter allows the simulation platform to retrieve the relevant covariates for the simulation model, e.g., the temperature data from the nearest weather station.

By taking into account the location of Meters, the simulation platform can represent complex **spatial correlations** among the simulated time series. In particular, by using the meteorological information from the nearest weather station, nearby Meters will use similar covariates in their simulation models. Another way to induce correlations between Meters is via the customer class, i.e., the type of model that is used for simulation. Figure 4 shows a Bayesian Belief Network representing the dependencies among the simulation models and covariates based on localization, time and customer type information.

Finally, the simulation platform allows for changes in the simulation model assigned to a particular Meter at given points in time. This capability can be used to represent consumers changing their behavior (e.g., due to dynamic pricing or the usage of novel electrical appliances such as electric vehicles or heat pumps), to model changes in the customer portfolio of an energy supplier, and to simulate installment of new wind farms and solar systems. Use cases illustrating this capability are described in Sects. 5.1 and 5.2.

---

the Meter level, is to represent shifts in the customer portfolio and changes in the consumers' behaviors, as will be explained at the end of this subsection.

4.3 Scalability Aspects

An important aspect in the design of the architecture of the simulation platform was scalability to enable massive-scale simulations of extended time periods much faster than realtime (e.g., simulate one year of half-hourly data from 35 millions Meters in less than 30 h). A key paradigm to achieve scalability was to use **parallel processing for streaming data**. Streams processing is a computational model designed for handling large amounts of data flows in a parallel and distributed manner. The rationale is similar to assembly-lines for manufacturing: each data element goes through different processing units, is processed and then forwarded to the next unit. Storage of the processed elements is avoided throughout the processing pipeline and performed only for the finished end product of the computations. IBM InfoSphere Streams [14] is a computing platform designed to enable high-performance, parallel and distributed processing of data streams. The challenge in designing a streaming application is to carefully design the processing line to take maximal advantage of distributed computing resources while keeping the volume of communication among these resources at a reasonable level.

A full description of the design is beyond the scope of this paper. The most important consideration was that, in the simulation platform, most of the data volume is generated at the lowest levels of the network (the Mete' and LVS levels in Fig. 3). In the case of very large networks, this requires to heavily distribute the computation at those levels. Also the volume of communication between network elements at those low levels is significant (in particular, when aggregating loads from the Meter to the LVS level), which requires to fuse Streams operators into single processing elements in order to avoid impractical communication overhead.

Scalability results from experiments with the simulation platform are shown in Fig. 5. The horizontal axis shows the number of parallel processing elements used

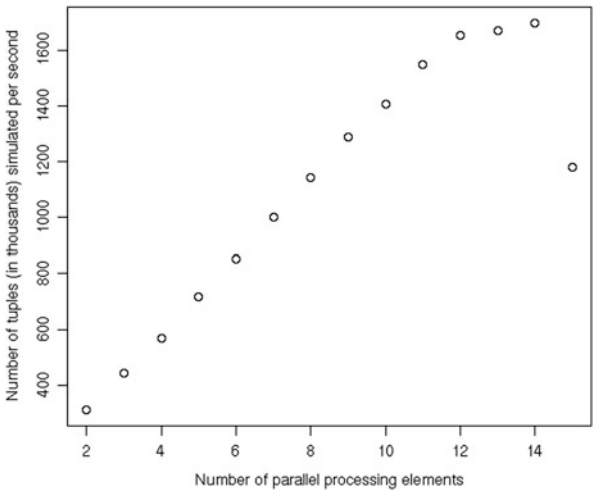


Fig. 5 Scalability of the simulation platform: The horizontal axis shows the number of parallel processing elements used in the simulation, the vertical axis the number of simulated data points per second. As can be seen, the platform scales almost perfectly linearly until the number of parallel processing elements reaches the number of physical CPUs (which was 12 in this experiment)

in the simulation, the vertical axis the number of simulated data points per second. As can be seen, the platform scales almost perfectly linearly until the number of parallel processing elements reaches the number of physical Central Processing Units (CPUs) which was 12 in this experiment. Approximately 140,000 data points can be simulated per CPU in one second. Based on this experiment, it can be estimated that 40 cores are sufficient to simulate one year of half-hourly data from 35 millions Meters (corresponding to 613.2 billion data points) in approximately 30 h.

5 Use Cases

This section presents three different use cases generated with the simulation platform presented in this paper, each of them addressing a specific challenge for utility companies from the emerging Smart Grids.

5.1 Forecasting a Time-Varying Portfolio

The first use case studies the impact of losses and gains of customers in a utility company's portfolio on the aggregated consumption. It is motivated by the deregulation and competition in retail electricity markets which will allow customers to change their electricity provider. Another goal of this use case is to illustrate the effectiveness of the online learning algorithm for GAMs introduced in [1] to forecast the aggregated consumption.

To simulate the changes in the portfolio, the five different customer classes learned from the Irish CER data set (see Sect. 3.2) are used. Two different kinds of changes are simulated in this use case: abrupt and gradual changes. Let $P_t = (p_{t,k})_{k=1,\ldots,5}$ denote the proportion of customers in the portfolio belonging to each class at a given time t. An abrupt change occurs at time t_0 if there is a significant difference between P_{t_0} and P_{t_0+1}. A gradual change is a linear transition of P_t between two points in time t_0 and t_1. Losses and gains of customers can be simulated by introducing a sixth "void" class which represents zero consumption, and simulating customers switching from/to this class to/from any of the five tariff options in the portfolio.

Figure 6 shows an example: Here, a portfolio of residential customers was simulated, uniformly distributed over the five tariff classes, with a loss of 20 % of the customers over the course of two years. The black line in the left plot shows a simulated abrupt change, while the blue line depicts a gradual, linear loss over the two years. The right plot illustrates the performance of forecasting algorithms in the gradual loss scenario. Here the black line shows the actual loads, the blue line shows

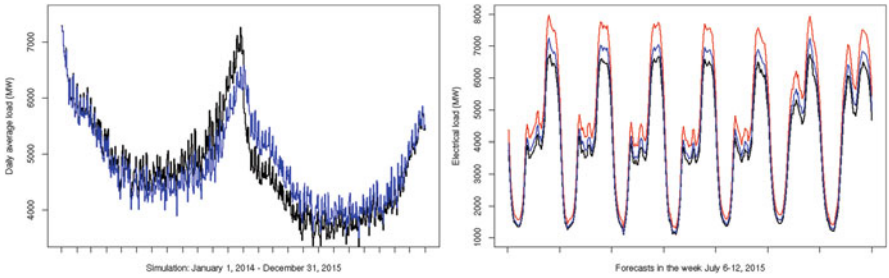


Fig. 6 *Left*: Simulation of a customer portfolio with a loss of 20 % of the customers over two years. *The black line* shows an abrupt loss after the first year, the *blue line* a gradual, linear loss over the two years. *Right*: Performance of forecasting algorithms in the gradual loss scenario. Here the *black line* shows the actual loads, the *blue line* shows the forecasts obtained by a GAM with online learning, and the *red line* the forecasts obtained by a GAM without online learning. As can be seen, the online learning is able to track some of the losses, resulting in a higher forecasting accuracy

the forecasts obtained by a GAM with online learning (using the algorithm proposed in [1]), and the red line the forecasts obtained by a GAM without online learning. As can be seen, the online learning is able to track some of the losses, resulting in a higher forecasting accuracy than the non-adaptive method. More generally, this example shows the usefulness of the simulation platform for comparing the performance of forecasting algorithms under different what-if-scenarios.

5.2 Impact of Wind Power Generation on the Distribution Grid

Managing the injection of power from renewable energy sources into the electrical grid, particularly wind power, raises high levels of concern for utility companies. Electricity providers and network operators need to optimize their production and grid management, respectively, to cope with those intermittent energy sources. Due to the high variability of wind power and its localized properties, simulations are an important tool for making decisions in this context.

Figure 7 shows examples of the simulations generated by the platform. The blue curves represent actual loads, generated using the same models as in the previous use case. The green curves show the simulated amount of wind power injected into the distribution network. For the simulation of wind power, the GAM introduced at the end of Sect. 3.2 was used. The difference between the two curves (i.e., the net load) is shown by the red curves. The plot on the left-hand-side displays a detail of 1 week, while the right plot shows the evolution over one year with an increase of 20 % in wind power capacity, corresponding to the installment and connection

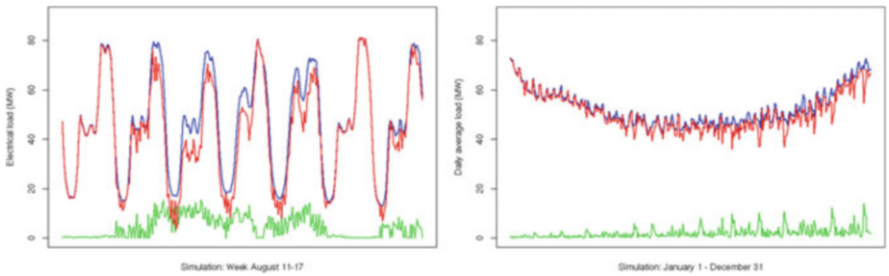


Fig. 7 Simulation of actual loads (*blue*), power from distributed wind farms (*green*), and the resulting difference, i.e., the net loads (*red*). The *left graph* shows a detail of 1 week, the *right graph* the daily averages over one year, with a simulated 20 % increase in wind power capacity

of new wind farms to the grid.[4] Note that, similarly, the simulation platform also supports the simulation of distributed power generation from photovoltaic systems.

5.3 Network Reconfigurations

In the last use case, the effect of network reconfiguration events is simulated. Such events, where loads are transferred over alternative lines or to different substations, become increasingly important in the operation of distribution networks where the trend is towards a more active management of the grid in order to increase the efficiency while coping with the challenges, e.g., due to power injections from distributed renewable energy sources. In this paper, only reconfigurations between the LVS and HVS network levels (see Sect. 4.1) are considered, where an LVS node connects to a different HVS parent node. In general, however, the simulation platform can represent reconfiguration events at any level in the network. Interestingly, the same logic can be applied to simulate electric vehicles (nodes at the Meters level) connecting to different charging stations (nodes at the LVS level), e.g., related to changes in location. Note that the platform presented in this paper can read reconfiguration events either from static files (e.g., generated by the user based on statistical assumptions and/or historical data), or dynamically receive them via a web server interface.

Figure 8 shows an example. The graph on the left shows how network entities and their current status (load, outside temperature etc.) are displayed on a map. The same interface can be used to dynamically introduce reconfiguration events by selecting a new HVS parent node for a particular LVS node. Typically, the new

[4]The installment of new wind power capacity can be represented by network nodes which, at specified time points, change their simulation model from a "void" GAM (producing zero values) to a GAM model simulating wind farm output. Compare with the remark at the end of Sect. 4.2.

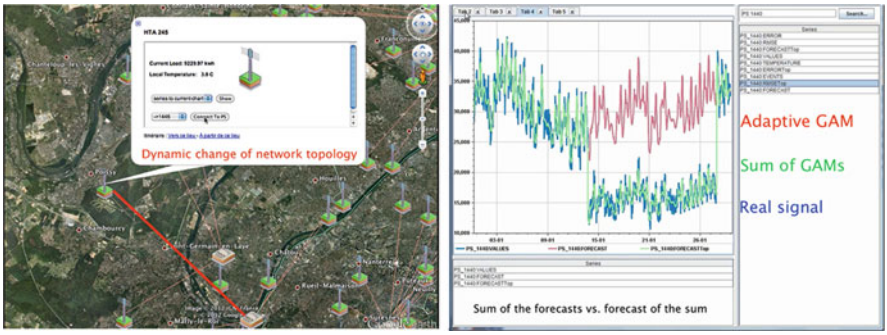


Fig. 8 Dynamic reconfiguration events simulated by the platform. The *left picture* displays LVS and HVS network elements on a map. By using the menu options in the *white balloon*, the user can manually connect the LVS element to a new parent at the HVS level. Changes in the connectivity will be reflected by the *red lines* displayed on the map. The *right hand side* shows how reconfiguration events impact the load signal and forecasts at the parent node. Here the *blue line* represents the actual load, the *red line* the forecasts at the parent node, and the *green line* the sum of forecasts from all children nodes. After approximately half of the displayed time period, one of the children is connected to a different parent node at the HVS level, resulting in a significant decrease in load (*blue line*). While the forecasts at the parent node (*red line*) are unable to quickly adapt to this change, taking the sum of forecasts from all children nodes (*green line*) reflects the actual configuration. Shortly before the end of the displayed time period, the children node is reconnected to its original parent, hence the load goes back to the original level

parent node is chosen from a list of candidates to which physical connections exist. The blue curve in the right graph shows the load at an HVS node. As can be seen, there is a significant load decrease after 2 weeks, which is due to a child of this node connecting to a different parent at the HVS level. After 2 weeks, the child reconnects to its original parent, and the load reaches the previous level. The red curve shows the load forecasts for the HVS node using an adaptive GAM model. While these adaptive models are very effective in tracking long-term trends and changes (see Sect. 5.1), they are not capable to adapt to such sudden shifts. The green curve represents the load forecasts obtained by taking the sum of the load forecasts for the children of this HVS node. Clearly, this approach is favorable in the presence of reconfiguration events.

6 Statistical Evaluation

The goal of this section is to evaluate how realistic are the load simulations generated by the platform, both at an aggregated and at the individual Meter level. Most approaches in the literature for this purpose use statistical hypothesis tests to assess whether the simulated and the real data have the same distribution. For

instance, in [15] a Mann-Whitney U test is used to test the similarity of the real and simulated data distributions. In [17], different statistics of the simulated and real data set are compared to assess how realistic the simulations are. The evaluation protocol in this paper is based on a classifier which aims at discriminating real and simulated data. The more difficult it is for the classifier to distinguish these data, the more realistic the simulations are. This approach is motivated by previous work which combines supervised and unsupervised approaches in order to evaluate the quality of the unsupervised task. For instance, the cascade evaluation [4] enriches a supervised dataset with the *cluster id* of each example. Then the *cluster id* is used by a classifier as an additional explicative variable. The cascade evaluation estimates the quality of the unsupervised task by measuring the improvement of the classifier when the *cluster id* is used. Another example is the use of a classifier to detect changes in the distribution of a data stream [2]. In this approach, two time windows are used to capture the "*current*" and the "*normal*" behavior of the observed system, respectively. Changes are quantified by the ability of the classifier to discriminate the both classes.

6.1 Experimental Protocol

The goal of the first experiment in this section is to assess the accuracy of GAMs depending on the size of the groups over which the load is aggregated. Same as in Sect. 3.2, the data set for this experiment is the Irish CER Smart Metering trial, and the GAM is given by Eq. (1). For aggregation sizes between $k = 10$ and $k = 800$, a random sample of k meters is drawn from the CER data set and then aggregated into a single time series. A GAM is learned on the first 70 % of this time series, then the model's Mean Squared Error (MSE) and standard deviation of the error (Sd) is calculated on the remaining 30 % of the time series. Overall, this procedure is repeated $n = 1{,}000$ times for each aggregation size k, and the average MSE and Std are computed for each k.

The results of this experiment are shown in Fig. 9. As to be expected, the models become more accurate (i.e., the MSE decreases) with increasing sample sizes, essentially illustrating the Law of Large Numbers which states that aggregating independent random variables following the same distribution yields stabilized variables around the mean value. Noteworthy is the inflection point in the Sd curve around the sample size $k = 70$: Beyond this point, the standard deviation of the model errors is slightly increasing. Similarly, the decrease in the MSE beyond this point is much less pronounced. A possible explanation is that the distributions of the individual meter signals are not identical, therefore, if too many signals are aggregated, information specific to some meters is lost while the benefit of aggregation to reduce noise does not compensate that loss of information. Therefore, the variance of a model learned on that sample will increase.

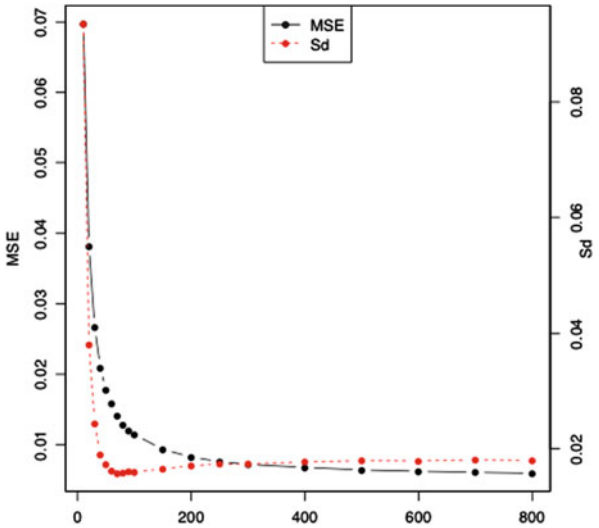


Fig. 9 Accuracy of GAM depending on the size of randomly aggregated groups of meters, measured in terms of Mean Squared Error (MSE) and Standard deviation of the error (Sd)

This experiment suggests two directions how to improve the quality of the simulations. First, the GAM approach is effective for simulating aggregations of 70 (or more) households, but not suitable for smaller sizes. Hence, for those low-level aggregations, other modeling approaches will be required. Second, blindly aggregating meters can lead to an information loss and an increase of variance of the error. Therefore, clustering meters into similar classes could improve the modeling accuracy.

Next, the effectiveness of this clustering approach using the k-means algorithm with the Euclidean distance is investigated. The clustering of the meters is used to build a **generative model** which is obtained by learning different GAMs for the aggregation of meters from each cluster. The k-means algorithm is parameterized in two different ways:

1. **Naive setting:** The number of clusters is arbitrarily fixed at $k = 10$. The corresponding generative model is used as a base line.
2. **Taking into account GAM performance:** Using the results from Fig. 9, an aggregation size of 70 m per cluster is found to be optimal, because it yields a good performance in terms of the MSE and the minimal standard deviation of errors. Correspondingly, the number of clusters is fixed at $k = 60$, leading to an average group size of 70 m (n.b.: the total number of residential meters in the data set is approximately 4,000).

For both settings, the k-means algorithm is applied to one year of half-hourly meter data.

6.2 Evaluation Protocol

The generative model obtained from the k-means clustering is first evaluated on simulations of the aggregated consumption. In particular, the sum of the 4,000 simulated individual meter signals is compared with the sum of the 4,000 real signals from the CER data set over the same time period. The Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE) are calculated as evaluation criteria.

In order to assess how realistic the simulated individual meter signals are, a classifier for discriminating real and simulated signals is used. The classification task is defined as follows: the data set consists of 8,000 time series (4,000 simulated and 4,000 real ones), each described by 336 numerical explicative variables (denoted by v_i), corresponding to 48 data points per day over 7 days. The target class variable c is equal to "0" for the simulated signals, and equal to "1" for the real ones. The data set is split into two disjoint parts: 70 % of the data are used for training, and 30 % for testing. For the classification, a simple Naive Bayes classifier is used. In particular, the range of each explicative variable v_i is discretized into 10 intervals, such that the numbers of training observations lying in each interval are equal. The conditional probabilities $P(v_i|c)$ for $i = 1, 2, \ldots, 336$ are estimated by the corresponding sample frequencies, and then $P(c|v_1 \ldots v_{336})$ is computed by applying Bayes' rule. The classifier is evaluated by using the Area Under Curve (AUC) metric [10]. Recall that a perfect classifier reaches an AUC equal to 1, and a random classifier an AUC equal to 0.5.

6.3 Results

Table 1 reports the RMSE and MAPE of the GAMs for the two different numbers of clusters $k = 10$ and $k = 60$. These two metrics assess the ability of the simulator to fit aggregated individual load signals. In both cases, the value of k has an insignificant impact on the RMSE and the MAPE. Note that a MAPE of 10 % is relatively high, however, it needs to be taken into account that the GAMs were learned on small aggregates and not at the national level.

Table 1 also reports the AUC score of the Naive Bayes classifier for the generative models with $k = 10$ and $k = 60$. In both cases the classifier is able to separate almost perfectly the simulated signals from the real signals, which underlines the difficulty of building a realistic simulator for individual load signals. This result can be intuitively explained by the fact that the GAMs in this experiment were learned on aggregated loads, which are much smoother than the individual signals. The Gaussian noise added to the simulated signals fails to exactly reproduce the characteristics of individual consumption signals. Alternative approaches will be discussed in the conclusions of this paper. Nevertheless, a significant drop in the classifier accuracy from AUC 0.927 for $k = 10$ to 0.806 for $k = 60$ can be observed.

Table 1 Comparative evaluation of generative models based on $k = 10$ and $k = 60$ clusters. The MAPE and RMSE measure how accurately the models are fitting the real data, while the AUC indicates how difficult it is for a classifier to distinguish between real and simulated data (hence, how realistic the simulations are). Note the drop in the classifier accuracy from AUC 0.927 for $k = 10$ to 0.806 for $k = 60$, which indicates that optimizing the granularity of the generative model can significantly improve the authenticity of simulations

	Criterion	$k = 10$	$k = 60$
How accurate?	Mean Absolute Percentage Error (MAPE)	10.81 %	10.73 %
	Root Mean Squared Error (RMSE)	283.21	283.05
How realistic?	Area Under Curve (AUC)	0.927	0.806

This means that using the clustering of consumer signals in the generative model can significantly improve the authenticity of the simulated signals.

7 Conclusion

In this paper, a platform for massive-scale simulation of electrical load in Smart Grids has been presented. The paper has provided details on the underlying statistical methodology, based on Generalized Additive Models (GAMs), and explained the architecture of the platform, with particular emphasis on scalability aspects. Experiments have shown the scalability and computational power of the platform, which is able to simulate one year of half-hourly load data for the entire electrical network in a country the size of France. The paper has presented three different use cases generated by the simulation platform, illustrating the value of the platform for power system engineers, statisticians and econometricians to study various what-if-scenarios, e.g., related to dynamic reconfigurations of the electrical network, changes in the customer portfolio and consumers' behavior, and increasing capacity of distributed renewable energy sources such as solar and wind.

In an evaluation study, the paper has shown that GAMs provide realistic simulations for aggregated load signals of at least 70 individual households. However, it has been demonstrated that novel modeling approaches are needed for simulating lower-level aggregates. Possible ideas for future research in this direction are: (i) using point processes (e.g., non-homogeneous Poisson); (ii) taking into account ancillary information (e.g., higher-resolution meteorological data and socio-economic indicators); (iii) considering GAMs with random effects and spatio-temporal correlations.

References

1. Ba, A., Sinn, M., Goude, Y., & Pompey, P. (2012). Adaptive learning of smoothing functions: Application to electricity load forecasting. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25, pp. 2519–2527). Curran Associates, Inc.
2. Bondu, A., & Boullé, M. (2011). A supervised approach for change detection in data streams. In *IJCNN (International joint conference on neural networks)*, San Jose (pp. 519–526). IEEE.
3. Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science, 22*, 477–522.
4. Candillier, L., Tellier, I., Torre, F., & Bousquet, O. (2006). Cascade evaluation of clustering algorithms. In J. Fürnkranz, T. Scheffer, & M. Spiliopoulou (Eds.), *17th European conference on machine learning (ECML'2006)*, Berlin (Volume LNAI 4212 of LNCS, pp. 574–581). Springer.
5. Commission for Energy Regulation. (2011). *Electricity smart metering customer behavior trials findings report* (Technical report). Commission for Energy Regulation, Dublin.
6. Commission for Energy Regulation. (2011). *Results of electricity cost-benefit analysis, customer behavior trials and technology trials* (Technical report). Commission for Energy Regulation, Dublin.
7. Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimated the correct degree of smoothing by the method of general cross validation. *Numerische Mathematik, 31*, 377–403.
8. dos Santos, L. D. P., Picard, M. L., da Silva, A. G., Worms, D., Jacquin, B., & Bernard, C. (2012). Massive smart meter data storage and processing on top of Hadoop. In *International workshop on end-to-end management of big data, VLDB (International conference on very large data bases)*, Istanbul.
9. Fan, S., & Hyndman, R. J. (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems, 27*(1), 134–141.
10. Fawcett, T. (2003). *ROC graphs: Notes and practical considerations for data mining researchers* (Technical report HPL-2003-4). HP Labs.
11. Goude, Y., Nedellec, R., & Kong, N. (2013, to appear). Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE Transactions on Smart Grid, 5*(1), 440–446.
12. Hastie, T., & Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science, 1*, 297–318.
13. Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. Boca Raton: Chapman & Hall/CRC.
14. International Technical Support Organization. (2013). Addressing data volume, velocity, and variety with IBM InfoSphere streams V3.0. http://www.redbooks.ibm.com/redbooks/pdfs/sg248108.pdf. March 2013.
15. Muratori, M., Roberts, M., Sioshansi, R., Marano, V., & Rizzoni, G. (2013). A highly resolved modeling technique to simulate residential power demand. *Applied Energy, 107(C)*, 465–473.
16. Nedellec, R., Cugliari, J., & Goude, Y. (2014, to appear). Electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting, 30*(2), 375–381.
17. Paatero, J. V., & Lund, P. D. (2006). A model for generating household electricity load profiles. *International Journal of Energy Research, 30*(5), 273–290.
18. Pierrot, A., & Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. In *Proceedings of ISAP power*, Hersonissos (pp. 593–600).
19. Swan, L., & Ugursal, V. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews, 13*(8), 1819–1835.
20. Wood, S. (2001). mgcv: GAMs and generalized ridge regression for R. *R News, 1*(2), 20–25.

21. Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association, 99*, 673–686.
22. Wood, S. (2006). *Generalized additive models, an introduction with R*. Boca Raton: Chapman and Hall/CRC.
23. Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society Series (B), 73*(1), 3–36.

Spot Volatility Estimation for High-Frequency Data: Adaptive Estimation in Practice

Till Sabel, Johannes Schmidt-Hieber, and Axel Munk

Abstract We develop further the spot volatility estimator introduced in Hoffmann et al. (Ann Inst H Poincaré (B) Probab Stat 48(4):1186–1216, 2012) from a practical point of view and make it applicable to the analysis of high-frequency financial data. In a first part, we adjust the estimator substantially in order to achieve good finite sample performance and to overcome difficulties arising from violations of the additive microstructure noise model (e.g. jumps, rounding errors). These modifications are justified by simulations. The second part is devoted to investigate the behavior of volatility in response to macroeconomic events. We give evidence that the spot volatility of Euro-BUND futures is considerably higher during press conferences of the European Central Bank. As an outlook, we present an estimator for the spot covolatility of two different prices.

1 Introduction

Semimartingales provide a natural class for modeling arbitrage-free log price processes (cf. [23, 24]). In this context, estimation of the volatility and its surrogates such as integrated volatility or higher moments is inevitable for many purposes as for example hedging or option pricing. Under a semimartingale assumption, estimation of the volatility can be done using realized quadratic variation techniques (cf. for example [31]). During the last decades, however, technical progress of trading platforms allowed to trade and to record data on very high frequencies. On these fine

T. Sabel (✉) • A. Munk
Department of Mathematics and Computer Science, Institute for Mathematical Stochastics, Georg-August-University Göttingen, Goldschmidtstraße 7, 37077 Göttingen, Germany
e-mail: tsabel@uni-goettingen.de; munk@math.uni-goettingen.de

J. Schmidt-Hieber
University of Leiden, Niels Bohrweg 1, 2333 CA, Leiden
e-mail: schmidthieberaj@math.leidenuniv.nl

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_12

scales, microstructure effects due to market frictions have to be taken into account (for an overview of such market frictions cf. [34] and [43]). Following Zhou [55], these are often modelled by an additive noise process in the literature. Incorporating microstructure noise, our observations are given by

$$Y_{i,n} = X_{i/n} + \epsilon_{i,n}, \quad i = 1, \ldots, n \tag{1}$$

where the (latent) log price process X is considered to be a continuous Itô semimartingale, that is $dX_t = \sigma_t dW_t +$"drift", with W a Brownian motion. The quantity of interest, the spot volatility $s \rightsquigarrow \sigma_s^2$ (which is sometimes referred to as squared spot volatility), has to satisfy some regularity conditions, in order to make everything well-defined. Adding the noise process $(\epsilon_{i,n})$ accounts for microstructure effects.

Microstructure noise leads to severe difficulties for estimation: As the noise is generally rougher than the original log price process X, methods based on increments of the data become inconsistent as the resulting estimators are first order dominated by noise. For example, the realized quadratic variation does not converge to the integrated volatility as the sample size increases (cf. [8]). Rather, it tends to infinity (cf. [55]). See also Ait-Sahalia and Yu [2] for a comprehensive empirical analysis of the noise level of different NYSE stocks.

Beginning with the work of Ait-Sahalia et al. [3] and Zhang et al. [54], various sophisticated regularization methods have been developed in order to estimate the integrated volatility under microstructure noise, cf. Zhang [52], Fan and Wang [30], and Barndorff-Nielsen et al. [9], to name just a few. Of particular interest in this work is the pre-average technique proposed in Podolskij and Vetter [46] and Jacod et al. [38].

These methods target on integrated volatility, that is the spot volatility integrated over a fixed time interval. Spot volatility estimation, that is pathwise reconstruction of the function $s \rightsquigarrow \sigma_s^2$ itself, is more difficult and therefore much less studied, since naive numerical differentiation of the integrated volatility estimators is not consistent without sophisticated additional regularization. To obtain consistent estimators, one needs to combine tools from nonparametric statistics and stochastic analysis. In Munk and Schmidt-Hieber [44], an estimator of the spot volatility was proposed, which is based on a Fourier series expansion of σ^2. Although this estimator could be shown to be asymptotically rate-optimal in Sobolev ellipsoids and hence is a first step towards a rigorous approach to spot volatility estimation, it suffers from various drawbacks. First, it obeys Gibb's effects which are well- known for Fourier estimators given non-smooth signals. Secondly, it requires knowledge of the smoothness of the underlying spot volatility, which is unknown in practice. To overcome these issues, Hoffmann et al. [37] introduced a wavelet estimator of σ^2. This estimator fully adapts to the smoothness of the underlying function and is rate-optimal over Besov classes. However, notice that Hoffmann et al. [37] deals with the abstract estimation theory in model (1) without making a particular connection

to finance. We fill this gap in the current paper by specifically tuning the estimator for application to stock market data, while at the same time keeping the procedure purely data-driven and adaptive. In the following, we refer to the modified estimator as *Adaptive Spot Volatility Estimator (ASVE)*.

The key idea of the estimation method is to exploit the different smoothness properties of the semimartingale and the noise part: In a first step, we compute weighted local averages over data blocks of size $c\sqrt{n}$, for a constant $c > 0$ independent of n. We show that the squared averages can be thought of as being observations in a regression type experiment. This is essentially the pre-averaging trick presented in Jacod et al. [38] and Podolskij and Vetter [46]. On one hand, local averaging reduces the impact of the noise (by a CLT type argument), while at the same time, the semimartingale part is (up to some small bias) not affected due to its a.s. Hölder continuity. On the other hand, treating the squared averages as new observations results in a reduction of the sample size from n to $c^{-1}\sqrt{n}$. Therefore, pre-averaging acts here as a denoising technique. In a second step, the pre-averaged data are decomposed via discrete wavelet transform and a robust thresholding procedure is applied. A detailed explanation concerning the construction of ASVE is given in Sect. 2.

Let us summarize in the following the main difficulties that we address in order to make the estimator applicable to real financial data.

1. *Thresholding:* One of the main challenges is to find a suitable and robust wavelet thresholding method. We argue in Sect. 2.4 that rewriting the initial model via the pre-average transform yields, as outlined above, a regression model with errors following approximately a centered χ^2_1-distribution. Furthermore, the errors are dependent and heteroscedastic causing severe difficulties for wavelet estimation. Therefore, a crucial point in our method is the choice of the thresholding procedure. We address this problem in Sect. 2.5.
2. *Parameter tuning:* ASVE requires to pick a bandwidth and a weight function. The specific choice will heavily influence the finite sample performance and even the asymptotic variance. In Sect. 3.1, we propose a method to chose these values based on an explicit computation of the asymptotic variance in a toy model. In a second part, the finite sample performance for these choices is studied in simulations.
3. *Model violations:* Given real data, model violations often occur. These include rounding errors, which is a non-additive microstructure effect as well as various types of jumps (cf. [1, 5]). In Sect. 4.2, we show that rounding has almost no impact on the performance of the estimator, while the presence of jumps is indeed a very delicate problem. In order to eliminate jumps in the price, we propose in Sect. 3.2 a specific pre-processing of the data.
4. *Trading times:* We have to deal with data recorded at non-equidistant time points. One possibility to 'convert' data into the equispaced framework of model (1) is to subsample the process, that is to sample for example every 10th second.

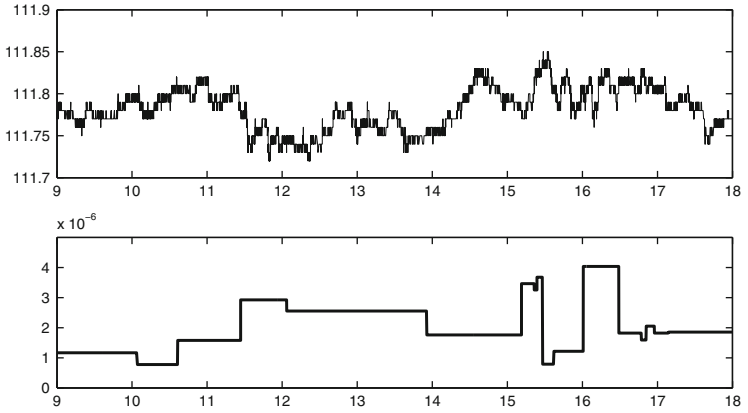


Fig. 1 Application to real data. *Upper panel*: Price of FGBL data on June, 4th 2007. *Lower panel*: Adaptive spot volatility estimator (ASVE)

In Sect. 5, we propose another method by defining different time schemes. Especially, we distinguish between real time and tick time and clarify their connection.

While Sect. 3 is devoted to calibration of ASVE especially focussing on the issues mentioned above, in Sect. 4, we evaluate ASVE by numerical simulations. This includes a stability analysis regarding model violations and different types of microstructure noise.

As an illustrating example for a real data application, Fig. 1 shows Euro-BUND (FGBL) prices for June 4th, 2007 together with the reconstructed volatility. Notice that ASVE appears to be locally constant. This is due to the specific wavelets which are the building blocks of this estimator. Note further, that ASVE is still quite regular, while spot volatility is commonly assumed to have no finite total variation. This relies on the fact that microstructure noise induces additional ill-posedness to the problem which leads to relatively slow convergence for any estimator (cf. [47]). Therefore, only key features of the spot volatility can be reconstructed, while fine details cannot be recovered by any method.

In Sect. 6, a more extensive investigation of real data is done concerning the reaction of spot volatility in answer to macroeconomic announcements: We study characteristics of the volatility of FGBL prices during the monthly ECB press conference on key interest rates. We observe that the spot volatility as well as the volatility of the volatility is higher during these conferences.

Finally in Sect. 7, we discuss extensions of ASVE to spot covolatility estimation.

The proposed estimator is implemented within the Matlab based Spotvol toolbox, available at http://www.stochastik.math.uni-goettingen.de/SpotvolToolbox.

2 The Adaptive Spot Volatility Estimator (ASVE)

2.1 Wavelet Estimation

A common tool for adaptive, nonparametric function estimation is wavelet thresholding (cf. for example [25] and [26], for some early references). Assume our signal, say f, is a function in $L^2[0,1]$. Then, for given scaling function φ and corresponding wavelet ψ, the function f can be decomposed into

$$f = \sum_k \langle f, \varphi_{j_0,k} \rangle \varphi_{j_0,k} + \sum_{j=j_0}^{\infty} \sum_k \langle f, \psi_{j,k} \rangle \psi_{j,k}, \quad j_0 \in \mathbb{N}, \tag{2}$$

where the sum converges in $L^2[0,1]$. Here, we denote $\langle f, g \rangle = \int_0^1 f(x)g(x)dx$, $\varphi_{j,k}(\cdot) = 2^{j/2}\varphi(2^j \cdot -k)$, and $\psi_{j,k}(\cdot) = 2^{j/2}\psi(2^j \cdot -k)$. The scaling and wavelet coefficients are $\langle f, \varphi_{j_0,k} \rangle$ and $\langle f, \psi_{j,k} \rangle$, respectively. See Daubechies [22] and Cohen [18] for an introduction to wavelets, Cohen et al. [19] for wavelets on $[0,1]$, and Wassermann [51] for a reference to wavelets in statistics.

Suppose that we have estimators for scaling and wavelet coefficients, denoted by $\widehat{\langle f, \varphi_{j_0,k} \rangle}$ and $\widehat{\langle f, \psi_{j,k} \rangle}$, respectively. A thresholding estimator for f is given by

$$\hat{f} = \sum_k \widehat{\langle f, \varphi_{j_0,k} \rangle} \varphi_{j_0,k} + \sum_{j=j_0}^{j_1} \sum_{k \in \mathbb{Z}} \mathscr{T}\big(\widehat{\langle f, \psi_{j,k} \rangle}\big) \psi_{j,k}, \tag{3}$$

for some thresholding procedure $\mathscr{T}$. Traditional choices for $\mathscr{T}$ include hard thresholding ($\mathscr{T}_{HT}(x) = x\mathbf{1}_{\{|x|>t^*\}}$) and soft thresholding ($\mathscr{T}_{ST}(x) = (x - t^*)\mathbf{1}_{\{x>t^*\}} + (x + t^*)\mathbf{1}_{\{x<-t^*\}}$), both for some threshold level t^*. The idea of term-by-term thresholding is to keep large coefficients while discarding small ones for which one cannot be sure that they contain significant information about the true signal.

Even though coefficientwise thresholding has many appealing theoretical properties, it nevertheless might lead to unstable reconstructions if applied to real data. Robustification of wavelet thresholding is typically based on variations of the following idea. Assume for the moment that ψ is the Haar wavelet, which has compact support on $[0,1]$. Then, $\langle f, \psi_{j,k} \rangle$ depends only on f restricted to the interval $[2^{-j}k, 2^{-j}(k+1)]$. If the absolute value of the estimate of $\langle f, \psi_{j,k} \rangle$ is large, while the absolute values of the estimates of nearby coefficients are small, then it is likely that this is due to an outlier and hence the wavelet coefficient should be discarded as well.

There are two types of methods for detecting such situations. Tree-structured wavelet thresholding using the hierarchical pattern of multiresolution analysis (cf. for example [7]) and block thresholding methods, which are based on neighboring coefficients for fixed level j. For our problem, SURE block thresholding (cf. [14]) turns out to work well. For more details, we refer to Sect. 2.5 as well as Sect. 4.

2.2 Model

Consider the process X defined via $dX_t = \sigma_t dW_t + b_t dt$ and $X_0 = 0$ on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$, where W denotes a standard Brownian motion. The processes σ and b are assumed to be $\mathscr{F}_t$-adapted and càdlàg. We will always suppose that σ and b are chosen in such a way that a unique weak solution of the SDE above exists.

Recall (1), that is we observe

$$Y_{i,n} = X_{i/n} + \epsilon_{i,n}, \quad i = 1, \ldots, n. \tag{4}$$

While X should be interpreted as the true, uncorrupted log price process, the noise process $(\epsilon_{i,n})$ models the microstructure effects. We allow for inhomogeneous variation in the noise, that is

$$\epsilon_{i,n} = \tau(\tfrac{i}{n}, X_{i/n})\eta_{i,n}, \tag{5}$$

where $(\eta_{i,n})_i$ is an i.i.d. sequence of random variables independent of X. Notice that the noise level may depend on the price itself. For identifiability, we assume further that $(\eta_{i,n})_i$ is centered and second moment normalized, that is $\mathbb{E}\eta_{i,n}^2 = 1$ for $i = 1, \ldots, n$.

To summarize, $Y_{i,n}$ is the observed log price, which is the sum of the latent true log price process X at time point i/n under additional microstructure noise $\epsilon_{i,n}$.

While the drift b is of only minor importance for high-frequency data, the process σ^2 is the key quantity in this model as it drives the fluctuation and variation behavior of the process. Although under debate, the additive microstructure noise model (4) is commonly believed to perform very well in practice, as it is able to reproduce many stylized facts found in empirical financial data. Moreover, to the best of our knowledge, it is the only model incorporating microstructure noise for which a theory of pathwise estimation of the volatility exists.

2.3 Pre-averaging and Estimation of Series Coefficients

The key step behind the construction of ASVE is a transformation of the data, which allows to rewrite the original problem as a nonparametric regression problem. This transformation is based on the pre-averaging method as introduced in Jacod et al. [38] and Podolskij and Vetter [46]. Since then, pre-averaging became an important tool to tackle estimation under microstructure. For an extension of pre-averaging to data measured on an endogenous time grid, cf. Li et al. [41]. Recently, the practical performance of these methods in estimation of integrated volatility was investigated in Hautsch and Podolskij [35].

In a first step, let us introduce a class of suitable weight functions (cf. [37], Definition 3.1).

Definition 1 (Pre-average function) A piecewise Lipschitz continuous function $\lambda : [0, 2] \to \mathbb{R}$ satisfying $\lambda(t) = -\lambda(2 - t)$, for all $t \in [0, 1]$ and

$$\Big(2\int_0^1 \Big(\int_0^s \lambda(u)du\Big)^2 ds\Big)^{1/2} = 1 \tag{6}$$

is called a (normalized) pre-average function.

Notice that whenever we have a function $\tilde{\lambda}$ satisfying all assumptions of the previous definition except (6), then by dividing $\tilde{\lambda}$ through the l.h.s. of (6), we obtain a proper pre-average function. Next, we define local averages using weights generated from pre-average functions.

Define $m = n/\lfloor n^{1/2}/c \rfloor$ for some fixed $c > 0$. Notice that $m = c\sqrt{n} + O(1)$ and that m divides n. The divisibility property allows to get rid of some discretization errors later. For $i = 2, \dots, m$, set

$$\overline{Y}_{i,m}(\lambda) := \frac{m}{n} \sum_{\frac{j}{n} \in [\frac{i-2}{m}, \frac{i}{m}]} \lambda\big(m\tfrac{j}{n} - (i-2)\big) Y_{j,n}. \tag{7}$$

Further, let us introduce

$$\mathfrak{b}(\lambda, Y.)_{i,m} := \frac{m^2}{2n^2} \sum_{\frac{j}{n} \in [\frac{i-2}{m}, \frac{i}{m}]} \lambda^2\big(m\tfrac{j}{n} - (i-2)\big)\big(Y_{j,n} - Y_{j-1,n}\big)^2$$

which plays the role of a bias correction. For any L^2-function g, the estimator of the scalar product $\langle \sigma^2, g \rangle$ is given by its empirical version applied to the bias-corrected squares $\overline{Y}^2_{i,m}$ via

$$\widehat{\langle \sigma^2, g \rangle} := \sum_{i=2}^{m} g\big(\tfrac{i-1}{m}\big)\big[\, \overline{Y}^2_{i,m} - \mathfrak{b}(\lambda, Y.)_{i,m}\big] = \frac{1}{m}\sum_{i=2}^{m} g\big(\tfrac{i-1}{m}\big) Z_{i,m}, \tag{8}$$

where

$$Z_{i,m} := m\big[\, \overline{Y}^2_{i,m} - \mathfrak{b}(\lambda, Y.)_{i,m}\big]. \tag{9}$$

Definition 2 The random variables $Z_{i,m}$, $i = 1, \dots, m$, are called *pre-averaged values*.

As we will show below, the pre-averaged values can be interpreted as observations coming from a nonparametric regression experiment with the spot volatility being the regression function. For $g \in \{\varphi_{j_0,k}, \psi_{j,k}\}$, we obtain estimates for the

scaling/wavelet coefficients $\langle \sigma^2, \varphi_{j_0,k} \rangle$ and $\langle \sigma^2, \psi_{j,k} \rangle$, respectively. In practice, fast computations of these coefficients can be performed using a discrete wavelet transform (DWT).

2.4 A Heuristic Explanation

In this part, we will present the main idea underlying the construction of the estimator, which is to think of the pre-averaged values $(Z_{i,m})_i$ as coming from a nonparametric regression problem. First, note that for $i = 2, \dots, m$,

$$\overline{Y}_{i,m}(\lambda) \approx \int_{\frac{i-2}{m}}^{\frac{i}{m}} m\lambda\big(ms - (i-2)\big) X_s ds + \xi_{i,m}$$

with

$$\xi_{i,m} = \frac{m}{n} \sum_{\frac{j}{n} \in [\frac{i-2}{m}, \frac{i}{m}]} \lambda\big(m\tfrac{j}{n} - (i-2)\big)\epsilon_{j,n}.$$

Now, let $\Lambda(u) = -\int_0^u \lambda(v)dv \mathbb{I}_{[0,2]}(u)$. By Definition 1, $\Lambda(0) = \Lambda(2) = 0$. Hence, $\Lambda'(ms - (i-2)) = m\lambda(ms - (i-2))$ and using partial integration

$$\overline{Y}_{i,m} \approx \int_{\frac{i-2}{m}}^{\frac{i}{m}} \Lambda\big(ms - (i-2)\big) dX_s + \xi_{i,m}.$$

It is easy to verify that $\xi_{i,m} = O_p(\sqrt{m/n})$ and $\mathbb{E}\xi_{i,m}^2 = \mathbb{E}\mathfrak{b}(\lambda, \epsilon.)_{i,m} \approx \mathbb{E}\mathfrak{b}(\lambda, Y.)_{i,m}$. For the diffusion term, $\int_{(i-2)/m}^{i/m} \Lambda(ms-(i-2))dX_s = O_p(m^{-1/2})$ and by Itô's formula there exists $U_{i,m}$, such that $\mathbb{E}U_{i,m} = 0$, $U_{i,m} = O_P(m^{-1})$, and

$$\Big(\int_{\frac{i-2}{m}}^{\frac{i}{m}} \Lambda\big(ms - (i-2)\big) dX_s \Big)^2 = \int_{\frac{i-2}{m}}^{\frac{i}{m}} \Lambda^2\big(ms - (i-2)\big)\sigma_s^2 ds + U_{i,m} \tag{10}$$

$$\approx \tfrac{1}{m}\sigma^2_{(i-1)/m} + U_{i,m}, \tag{11}$$

using the definition of a pre-average function for the last step. Recall (9). Then, $\mathbb{E}[Z_{i,m} - \sigma^2_{(i-1)/m}] \approx 0$ and $Z_{i,m} - \sigma^2_{(i-1)/m} = O_P\Big(1 + \frac{m}{n^{1/2}} + \frac{m^2}{n}\Big) = O_P(1)$, since $m = c\sqrt{n} + O(1)$. To summarize,

$$Z_{i,m} = \sigma^2_{(i-1)/m} + \tilde{\epsilon}_{i,m}, \quad i = 2, \dots, m, \tag{12}$$

with $\mathbb{E}\tilde{\epsilon}_{i,m} \approx 0$ and $\tilde{\epsilon}_{i,m} = O_P(1)$. Hence, we may interpret $(Z_{i,m})_{i=2,\dots,m}$ as a random vector generated from a regression problem with regression function σ^2 and additive (dependent) noise $\tilde{\epsilon}_{i,m}$.

Let us conclude this section with the following remarks.

- Notice that the estimator of $\widehat{\langle \sigma^2, g\rangle}$ in (8) is just the empirical version of the scalar product $\langle \sigma^2, g\rangle$ in the regression model (12).
- By some CLT argument, the distribution of $\overline{Y}_{i,m}$ as defined in (7), will converge to a Gaussian law. But since we are considering the squares of $\overline{Y}_{i,m}$ in (9), the noise process in (12) will not be Gaussian. Rather, one can think of the $\widetilde{\epsilon}_{i,m}$'s as centered χ^2_1 random variables.
- The variance of $\tilde{\epsilon}_{i,m}$ (which is here approximately the second moment) is (up to some remainder terms) a quadratic function in $\sigma_{i/n}$ and $\tau(i/n, X_{i/n})$. Therefore, the regression problem (12) is strongly heteroscedastic. This point is separately addressed in Sect. 2.5.
- Rewriting the original problem as regression model, as outlined above, reduces the effective number of observation from n to m and thus to the order $n^{1/2}$. This implies that if we can estimate a quantity in the regression model (for example pointwise estimation of the regression function σ^2) with rate m^{-s}, given m observation, we obtain the rate of convergence $n^{-s/2}$ in the original model (4). Therefore, we always lose a factor $1/2$ in the exponent of the rate of convergence. It is well-known that this is inherent to spot volatility estimation under microstructure noise. As proved in Munk and Schmidt-Hieber [44], Reiß[47] for various situations, these rates are optimal.

2.5 *Thresholding and Construction of ASVE*

Having the estimates of the wavelet coefficients at hand, let us outline the thresholding procedure. The proposed method extends SURE block thresholding as introduced in Cai and Zhou [14] to heteroscedastic problems.

In order to formulate the thresholding estimator define, for a vector v, Stein's unbiased risk estimate (SURE) as

$$\mathrm{SURE}(v, \lambda, L) = L + \frac{\lambda^2 - 2\lambda(L-2)}{\|v\|_2^2}\mathbb{I}_{\{\|v\|_2^2 > \lambda\}} + (\|v\|_2^2 - 2L)\mathbb{I}_{\{\|v\|_2^2 \le \lambda\}}.$$

First, we start with SURE block thresholding for homoscedastic data. For convenience, set $\hat{d}_{j,k} = \widehat{\langle \sigma^2, \psi_{j,k}\rangle}$.

IN: $j_0, j_1, (\hat{d}_{j,k})_{j_0 \le j \le j_1, k}$

(A) For every fixed resolution level $j_0 \le j \le j_1$ define D_j as the set of wavelet dilations $\{k : k \in \mathbb{Z},\ [0,1] \cap supp\psi_{j,k} \neq \varnothing\}$. Denote by T_j the mean of the random variables $\{(\hat{d}_{j,k})^2 - 1 : k \in D_j\}$ and consider the threshold $\gamma(u) = u^{-1/2}\log_2^{3/2}(u)$.

(B) For any given vector $v \in \mathbb{R}^d$ and positive integer L define the qth block (of length L) as $v^{(q,L)} = (v_{(q-1)L+1}, \ldots, v_{qL\wedge d})$, $q \le d/L$. Let $d = |D_j|$. In particular, denote by $(\hat{d}_{j,k})_{k\in D_j}^{(q,L)}$ the qth block of length L of the vector $(\hat{d}_{j,k})_{k\in D_j}$ and define

$$(\lambda^\star, L^\star) = \arg \min_{\substack{1\le L\le d^{1/2} \\ (L-2)\vee 0\le\lambda\le 2L\log d}} \sum_{q=1}^{\lfloor d/L\rfloor} \mathrm{SURE}\big((\hat{d}_{j,k})_{k\in D_j}^{(q,L)}, \lambda, L\big),$$

where $\lfloor . \rfloor$ is the floor function.

(C) For every $k \in D_j$, the block thresholded (and standardized) wavelet coefficient is given by

$$\mathscr{T}(\hat{d}_{j,k}) = \begin{cases} (1 - (2\log d)\, \hat{d}_{j,k}^{-2})_+ \, \hat{d}_{j,k}, & \text{if } T_j \le \gamma(d), \\ \big(1 - \lambda^\star \big\| (\hat{d}_{j,\ell})_{\ell\in D_j}^{(q(k),L^\star)} \big\|_2^{-2}\big)_+ \, \hat{d}_{j,k}, & \text{if } T_j > \gamma(d), \end{cases}$$

with $q(k)$ the (unique) block of length $L^\star$ including k.

OUT: $\mathscr{T}(\hat{d}_{j,k})_{j_0\le j\le j_1,k}$.

SURE block thresholding optimizes levelwise over the block size L and the shrinkage parameter λ in step (B). However, it is well-known that this method does not yield good reconstructions in the case where only a few large wavelet coefficients are present. In order to circumvent these problems, in step (C), soft shrinkage is applied if T_j is small.

As an additional difficulty, we have to deal with errors in (12), that are heteroscedastic with unknown variance. Therefore, we normalize the wavelet coefficients by its standard deviation in a first step, that is for sets $I_{j,k}$, chosen below, define the empirical standard deviation on $I_{j,k}$ by

$$\hat{s}_{j,k} := \Big[\frac{1}{|I_{j,k}| - 1} \sum_{\frac{i}{m}\in I_{j,k}} \Big(Z_{i,m} - \tfrac{1}{|I_{j,k}|} \sum_{\frac{i}{m}\in I_{j,k}} Z_{i,m}\Big)^2\Big]^{1/2} \tag{13}$$

and the *standardized wavelet coefficients* by $\tilde{d}_{j,k} := \hat{d}_{j,k}/\hat{s}_{j,k}$. Now, we run the SURE algorithm applied to $(\tilde{d}_{j,k})_{j_0\le j\le j_1,k}$ instead of $(\hat{d}_{j,k})_{j_0\le j\le j_1,k}$. In a final step we need to invert the standardization. Thus, the thresholded wavelet coefficients are given by $(\hat{s}_{j,k}\mathscr{T}(\tilde{d}_{j,k}))_{j_0\le j\le j_1,k}$. Together with the (truncated) series expansion (2), we have

Definition 3 ASVE is defined by

$$\hat{\sigma}^2(t) = \sum_k \widehat{\langle \sigma^2, \varphi_{j_0,k}\rangle} \varphi_{j_0,k}(t) + \sum_{j=j_0}^{j_1} \sum_{k\in D_j} \hat{s}_{j,k}\mathscr{T}(\tilde{d}_{j,k})\psi_{j,k}(t), \quad t \in [0,1].$$

For estimation of the standard deviations $\hat{s}_{j,k}$, one would instead of (13) rather prefer a robust estimate based on the median (cf. [14], p. 566) or to use variance

stabilizing transformations. Since the error variables $\overline{\epsilon}_{i,m}$ in (12) do not follow a certain prespecified distribution, these approaches are not easily applicable here. Therefore, we rely on (13) and robustify our estimates by the choice of $I_{j,k}$, as described in the next paragraph:

We pick some $j_I, j_0 \le j_I \le j_1$. If $j \le j_I$, we define $I_{j,k}$ as the support of $\psi_{j,k}$. For high resolution levels $j > j_I$, we enlarge the support of $\psi_{j,k}$ such that the length of $I_{j,k}$ never falls below 2^{-j_I}. This guarantees some minimal robustness of the method.

Block thresholding uses the normality of the wavelet coefficients at various places. Thus, to ensure good performance, we need to check whether the distribution of the estimated wavelet coefficients follow approximately a Gaussian law. This is not obvious, because, as we argued in Sect. 2.4, the errors in the regression model (12) behave like centered χ^2_1 random variables. However, since the estimator def est fonct lin is a weighted average of the observations, we indeed find 'almost' Gaussian wavelet coefficients in simulations. Thus, we do not need to include a further correction to account for the non-Gaussianity. Notice that these issues are closely linked to nonparametric variance estimation (cf. [13]).

3 Calibration and Robustness

3.1 Optimal Tuning Parameters

In this section we propose empirical rules for choosing some variables in the ASVE procedure. Notice that the method requires to pick a pre-average function λ and a constant $c > 0$ defining the number of blocks m. By computing the asymptotic variance of ASVE in a simplified model, we derive some insight which pre-average functions might work well. In particular, this shows that λ and c should be chosen dependent on each other, that is $c = c(\lambda)$. In a second step, we study the finite sample performance of these choices for simulated data.

We start with investigating different choices for λ and $c = c(\lambda)$ in a simplified version of model (4) for which the leading term of the mean squared error can be calculated explicitly.

Lemma 1 *Work in model* (4) *with constant* σ, τ *and* $\eta_{i,n} \sim \mathcal{N}(0,1)$ *i.i.d. Then,*

$$\mathrm{MSE}(\widehat{\langle \sigma^2, 1\rangle}) = \tfrac{4}{c}\Big(\int_0^1 \sigma^2 \Lambda(u)\Lambda(1-u) - (\tau c)^2 \lambda(u)\lambda(1-u)du\Big)^2 n^{-1/2} + \tfrac{2}{c}\Big(\sigma^2 + 2(\tau c)^2 \|\lambda\|^2_{L^2[0,1]}\Big)^2 n^{-1/2} + o(n^{-1/2}).$$

A proof of this lemma can be found in Appendix A. Given a pre-average function λ, it allows us to compute the corresponding optimal constant $c^\star$ by minimizing the asymptotic MSE. In general $c^\star$ is a multiple of the signal-to-noise ratio (SNR), that is $c^\star = \text{const.} \times \frac{\sigma}{\tau}$, where the constant depends on λ. In Table 1, the value

Table 1 Different choices for pre-average functions, the optimal tuning parameter $c^\star$ as well as the asymptotic constant of the MSE for estimation of the integrated volatility

i	$\lambda_i(s) =$	$c^\star\tau/\sigma \approx$	$\lim_n n^{1/2}(\tau\sigma^3)^{-1}$ $\cdot$MSE $\approx$
1	$\frac{\pi}{2}\cos(\frac{\pi}{2}s)$	0.49	10.21
2	$\frac{3\pi}{2}\cos(\frac{3\pi}{2}s)$	0.17	31.36
3	$\sqrt{\frac{3}{2}}(\mathbb{I}_{[0,1)}(s) - \mathbb{I}_{(1,2]}(s))$	0.35	10.74
4	$\frac{\pi}{\sqrt{3}}\sin(\pi s)$	0.30	12.52
5	$\frac{2\pi}{\sqrt{3}}\sin(2\pi s)$	0.19	24.35
6	$\frac{3\sqrt{5}}{2}(1-s)^3$	0.47	20.41
7	$\frac{\sqrt{91}}{2}(1-s)^5$	0.38	20.36

of this constant for different pre-average functions and the leading term for the corresponding MSE are derived.

It is well-known (cf. [15, 32, 33]) that the bound $\mathrm{MSE}(\widehat{\langle\sigma^2,1\rangle}) = 8\tau\sigma^3 n^{-1/2}(1+o(1))$ is asymptotically sharp in minimax sense. However, this minimum cannot be achieved within the class of estimators introduced in Sect. 2. Using calculus of variations, we find that the best possible choice for the simplified model introduced above is $\lambda(\cdot) = \pi\cos(\cdot\pi/2)/2$. According to Table 1, the corresponding MSE is $10.21\tau\sigma^3 n^{-1/2}(1+o(1))$ achieving the optimal variance $8\tau\sigma^3 n^{-1/2}(1+o(1))$ up to a factor 1.27.

Computation of $c^\star$ requires knowledge of the SNR, that is σ/τ. As this is unknown, we suggest to estimate the SNR in a first step from the data via

$$\widehat{\mathrm{SNR}} = \left(\frac{\widetilde{\langle\sigma^2,1\rangle}}{\widehat{\langle\tau^2,1\rangle}}\right)^{1/2}, \tag{14}$$

with rescaled quadratic variation $\widehat{\langle\tau^2,1\rangle} = (2n)^{-1}\sum_{i=2}^n (Y_{i,n} - Y_{i-1,n})^2$ and

$$\widetilde{\langle\sigma^2,1\rangle} := \sum_{i=2}^{\tilde{m}} \left(\overline{Y}^2_{i,\tilde{m}} - \mathfrak{b}(\lambda, Y.)_{i,\tilde{m}}\right), \quad \text{with} \quad \tilde{m} = \lfloor n^{1/2}\rfloor$$

as preliminary estimator of $\langle\sigma^2,1\rangle$. It is easy to show that $\widehat{\langle\tau^2,1\rangle}$ is $n^{1/2}$-consistent for estimation of the integrated noise level $\langle\tau^2,1\rangle$ and since we are interested in data sets with sample size $n \sim 10^5$, we may directly divide by $\widehat{\langle\tau^2,1\rangle}$ in (14) without any further regularization.

In the second part of this section, we study the finite sample performance for different pre-average functions. As Table 1 suggests, the MSE deteriorates if the number of oscillations of λ increases. Therefore, we choose the functions $\lambda_1(\cdot) := \pi\cos(\cdot\pi/2)/2$ (the optimal pre-average function in the simplified model),

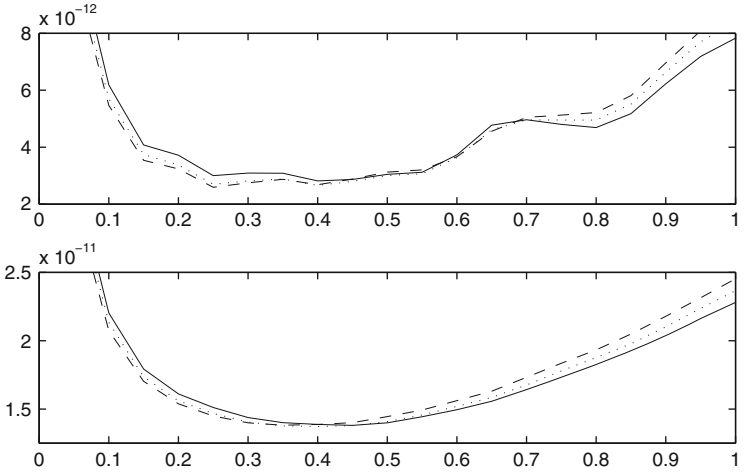


Fig. 2 Empirical MISE for 10,000 repetitions and data with constant $\sigma^2 \equiv 10^{-5}$ (*upper panel*) and data from the Heston model (cf. (16) and (18), *lower panel*). In each panel, the *x*-axis refers to different choices of the optimal constant $c^\star$ and the three curves represent different pre-average functions λ_i (λ_1: *solid line*, λ_3: *dotted line*, λ_4: *dashed line*)

$\lambda_3(\cdot) := (\frac{3}{2})^{1/2}(\mathbb{I}_{[0,1)} - \mathbb{I}_{(1,2]})$ (the pre-average function used in [35]), and $\lambda_4(\cdot) := \pi \sin(\cdot\pi)/3^{1/2}$ as possible candidates.

Figure 2 displays the results of the simulation study. In both panels, we choose $n = 15{,}000$, SNR $= 20$ with constant τ and standard Gaussian white noise. Both display the empirical mean integrated squared error

$$MISE = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \int_0^1 (\hat{\sigma}_i^2(s) - \sigma_i^2(s))^2 ds \tag{15}$$

based on 10,000 repetitions for $\lambda \in \{\lambda_1, \lambda_3, \lambda_4\}$ and different choices of the multiplicative constant c (x-axis). In the upper panel, the data are generated with constant σ. In the lower panel, we simulate the latent log price X according to the Heston stochastic volatility model

$$dX_t = -\tfrac{1}{2}\sigma_t^2 dt + \sigma_t dW_t, \tag{16}$$

$$d\sigma_t^2 = \kappa\big(\theta - \sigma_t^2\big)dt + \epsilon\sigma_t d\tilde{W}_t. \tag{17}$$

In this model, the Brownian motions W and $\tilde{W}$ are correlated, that is $dW_t d\tilde{W}_t = \rho dt$ with $\rho \in [-1, 1]$. It is not difficult to verify that X is indeed a continuous semimartingale. The Heston model is commonly believed to describe stock market data quite well. It only depends on a few parameters which have a clear financial

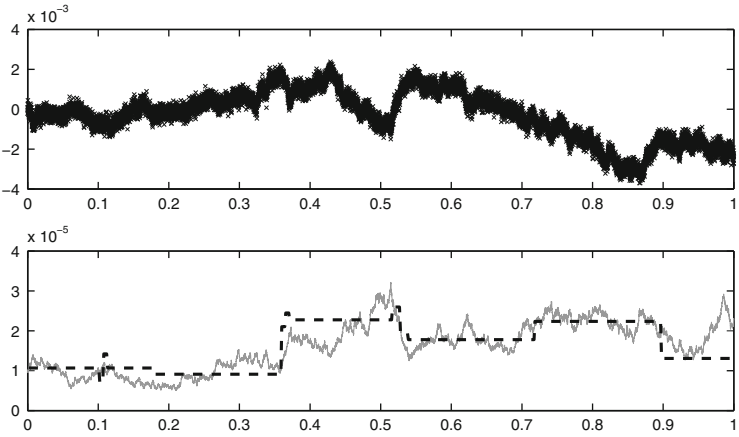


Fig. 3 Simulated data (*Panel 1*) coming from the Heston model with parameter as in (18) for $n = 15{,}000$ and true SNR ≈ 15–20, the true spot volatility function (*solid line*, *Panel 2*) and ASVE (*dashed line*, *Panel 2*)

interpretation allowing in particular for leverage effects ($\rho < 0$). For real data, estimates of the parameters in the Heston model have been carried out in different settings (see for instance Table 5.1 in [49]). For our simulations, we set

$$\rho = -2/3, \theta = 10^{-5}, \kappa = 4, \epsilon = \sqrt{\kappa\theta}. \tag{18}$$

For these parameters, the spot volatility σ^2 typically takes values in $[2{\cdot}10^{-6}, 5{\cdot}10^{-5}]$, see also Fig. 3.

From our simulation study, we find that in the Heston model, there is essentially no difference between the three candidate functions as long as $c^\star$ is chosen appropriately. However, λ_4 seems to produce the best estimators in terms of MISE, when the volatility function is constant. This is surprising, since from an asymptotic point of view, λ_1 is preferable. Our explanation is that non-asymptotically the boundary behavior of the pre-average function matters. Note that in contrast to λ_i, $i = 1, 3$, the function λ_4 vanishes at 0 and 2 and hence downweights observation at the end of the pre-average intervals $((i-2)/m, i/m]$.

Observe that the curves in the lower panel in Fig. 2 are smoother than the ones in the upper panel. We explain this by the fact that the SNR is constant for deterministic σ^2 and varies in the Heston model. Thus, the randomness of the volatility has a smoothing effect and discretization effects become visible in the first case only.

In Fig. 3, we illustrate the procedure for $\lambda = \lambda_4$ and $c = c^\star \cdot \widehat{\mathrm{SNR}}$. Here, X follows again the Heston model with parameters given in (18). Observe that the stylized nature of the reconstruction only reflects the main features of σ^2.

3.2 Jump Detection

Note that our theoretical considerations are based on model (4), that is assuming a continuous Itô semimartingale as log price process corrupted by additive noise. However, the continuity assumption in the model is often too strict in reality, since for example micro- or macroeconomic announcements may cause jumps in the price. The presence of such jumps is discussed in Ait-Sahalia and Jacod [1], Bollerslev and Todorov [12], and the references therein.

The most natural way to include a jump component into the model is to allow for non-continuous semimartingales. Estimation of the integrated volatility under microstructure noise and jumps has been considered for instance in Podolskij and Vetter [46]. Eliminating jumps turns out to be much easier than taking microstructure noise into account.

In order to correct for jumps, we adopt a rather practical point of view here. In fact, looking at financial data, relevant jumps seem to occur very irregularly. Occasionally, there are isolated jumps and quite rarely, jumps clustered over very short time intervals appear (cf. Fig. 4). Therefore, our aim in this section is a hands-on approach to detect and to remove possible jumps as a pre-processing of the data.

As usual, we model jumps as a càdlàg jump process $(J_t)_t$. If jumps are present, ASVE will reconstruct the pointwise sum of the spot volatility plus the jump process $t \mapsto (J_t - J_{t-})^2$, where J_{t-} denotes the left limit of J at time point t. Note that $(J_t - J_{t-})^2$ is either zero or produces a spike depending on whether there is a jump at time point t (cf. Fig. 5, Panel 1). In order to separate spot volatility and jump part, we apply the following method:

Let $m_1 = \lfloor n^{3/4} \rfloor$ and λ be a pre-average function. For $r = \frac{n}{m_1}, \ldots, n - \frac{n}{m_1}$, define

$$Q_r := \frac{m_1}{n} \sum_{j=r-\frac{n}{m_1}}^{r+\frac{n}{m_1}} \lambda\big(1 + (j-r)\frac{m_1}{n}\big) Y_{j,n}. \tag{19}$$

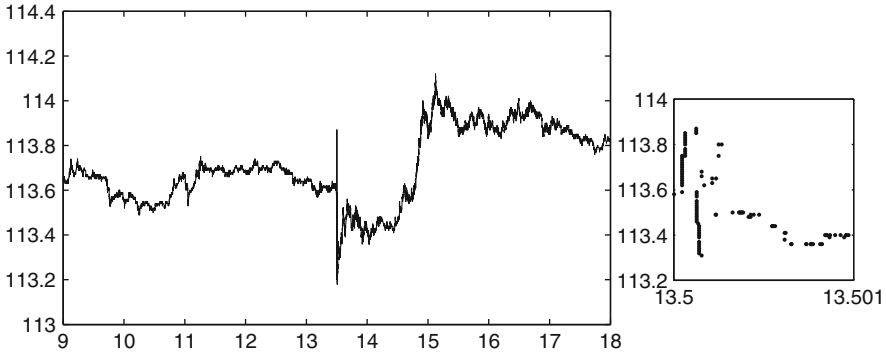


Fig. 4 FGBL data of November 2nd, 2007 and magnification of a small time interval around 1.30 p.m., where multiple consecutive jumps of the process occur

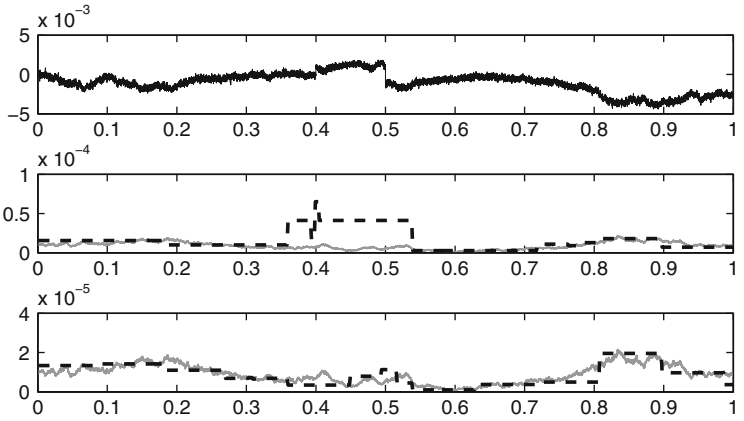


Fig. 5 Simulated data (*Panel 1*) coming from the Heston model with parameter choices given in (18) for $n = 15{,}000$ and true SNR $\approx$ 15–20 with two additional jumps at 0.4 and 0.5, the true spot volatility function (*gray, solid line*, *Panel 2* and *3*) and ASVE neglecting the presence of jumps (*dashed line*, *Panel 2*) and automatically finding and correcting the jumps (*dashed line*, *Panel 3*)

If there is no jump in $[r - \frac{n}{m_1}, r + \frac{n}{m_1}]$, then $Q_r = O_P(n^{-1/8})$ (following the heuristic explanation in Sect. 2.4). Under the alternative, that is, the process jumps with height Δ_r at r/n, we obtain $Q_r = O_P(\Delta_r)$. Note that by some CLT argument, Q_r is approximately Gaussian distributed. Therefore, we may apply a procedure mimicking a local t-test:

1. We partition the set $\{Q_r : r = \frac{n}{m_1}, \ldots, n - \frac{n}{m_1}\}$ into blocks of length $n^{1/2}$.
2. For each of these blocks, we compute the mean $\hat{\mu}$ and the standard deviation $\widehat{sd}$.
3. For each Q_r in a block, we compare $(Q_r - \hat{\mu})/\widehat{sd}$ with a fixed threshold t. Here, simulations show that $t = 2.81$ performs well.

Afterwards, we reject those pre-averaged value $Z_{i,m}$, whose support intersects the support of one of the Q_r's rejected by the procedure. Those rejected values are replaced by the average of the nearest neighbors which are not rejected.

This procedure ensures that isolated jumps are detected. In real data, however, there are occassionally consecutive jumps within a short period of time (cf. FGBL data of November 2nd, 2007 in Fig. 4 as an example). This may result in acceptance of the hypothesis that there is no jump, since a single jump might be not high enough in comparison to the estimated variance of Q_r. However, it is high enough to disrupt the performance of ASVE severely. To overcome this problem, we introduce a second test based on comparing increments of the observations directly which is more suitable to detect jump clusters.

From our data sets, we find that the level of the microstructure noise, that is τ, remains almost constant over a day. Thus, to explain the test, we might assume that τ is constant. Then,

$$Y_{i,n} - Y_{i-1,n} = \tau(\eta_{i,n} - \eta_{i-1,n}) + O_P(n^{-1/2}) \approx \tau(\eta_{i,n} - \eta_{i-1,n}),$$

if there is no jump. Secondly, we observe that the distribution of the noise is well-concentrated around zero. Thus, from a practical perspective, it is justified to assume that the tails of the microstructure noise are not heavier than that of a Gaussian random variable. If $(\eta_{i,n})$ would be i.i.d. standard normal, then using Corollary 2.1 in Li and Shao [40], we find the following extreme value behavior:

$$\lim_{n\to\infty} \mathbb{P}(\max_{i=2,\dots,n} (\eta_{i,n} - \eta_{i-1,n})^2 \leq 4\tau^2 \log n) = 1.$$

Consequently, we identify the difference $Y_{i,n} - Y_{i-1,n}$ as due to a jump, if the squared increment exceeds $4\hat{\tau}^2 \log n$, where $\hat{\tau}^2 = (2n)^{-1} \sum_{i=2}^{n} (Y_{i,n} - Y_{i-1,n})^2$ is an estimator for τ^2. Note that the latter procedure is much less powerful for isolated jumps than the first one, since it cannot detect jumps of size $o_P(\log n)$.

To illustrate the results, Fig. 5 displays simulated data corrupted by two additional jumps at 0.4 and 0.5. ASVE without jump correction (Panel 2) incorporates a bump at the positions of the jumps. In contrast, pre-processing the data in a first step as outlined in this section yields a stable reconstruction (Panel 3).

A simulation study regarding the jump detection procedure is given in Sect. 4.2.

4 Simulations

4.1 Stability

To test the stability of ASVE, we simulate data for sample size $n = 15{,}000$ and X following the Heston SDE (cf. (16)) with parameters given in (18). To model the microstructure effects $(\epsilon_{i,n})_{i=1,\dots,n}$, we consider Gaussian and uniform noise with standard deviations $x/5{,}000$ and $x \in \{1, 3, 10\}$. Here, a standard deviations of $1/5{,}000$ refers to a SNR of approximately 15 and represents FGBL data best. We perform a simulation study with 10,000 repetitions. Besides the mean integrated squared error (MISE, cf. (15)), we investigated the behavior of the relative mean integrated squared error (rMISE), given by

$$rMISE = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \frac{\int_0^1 (\hat{\sigma}_i^2(s) - \sigma_i^2(s))^2 ds}{\int_0^1 \sigma_i^4(s) ds},$$

Table 2 Stability under different distributions and levels of noise: MISE (upper row), rMISE (lower row), and respective 95 %-quantiles of the squared errors (in brackets) based on 10,000 repetitions

Std. of noise		1/5,000	3/5,000	10/5,000
MISE·10^{11}	Gaussian	1.41 (3.28)	2.39 (6.04)	5.05 (14.34)
	Uniform	1.40 (3.21)	2.40 (6.10)	5.08 (14.47)
rMISE	Gaussian	0.11 (0.20)	0.19 (0.38)	0.39 (0.94)
	Uniform	0.12 (0.20)	0.19 (0.38)	0.40 (0.97)

where $\hat{\sigma}_i^2$ and σ_i^2 refer to the estimated and the true volatility in run i. Throughout our simulations, we use Haar wavelets and λ_4 as a pre-average function. Following Sect. 3.1, we set $c = 0.3 \cdot \widehat{\text{SNR}}$. The results and empirical 95 %-quantiles are displayed in Table 2. We observe that the outcome is essentially not affected by the distribution. In contrast, the SNR has a large impact on the performance (recall that $\sigma^2 \approx 10^{-5}$). The bad performance of the estimator for the largest standard deviation can be explained by the choice of m, which is inversely proportional to the noise level. In fact the optimal oracle would be $m_{\text{oracle}} = 0.3 \cdot SNR\sqrt{n} \approx 55$. Thus, regarding the problem as a χ_1^2-regression problem (cf. Sect. 2.4), we have to estimate σ^2 based on 55 observations, which is a quite difficult task.

4.2 Robustness

As discussed in Sect. 3.2, there are two major model violations one has to take into account for real data, namely rounding effects and jumps. In a simulation study, we investigate the robustness of ASVE with and without jump detection given data with rounding errors and jumps. The process X is generated from the Heston model (16) with parameters as in (18). This ensures that the SNR lays most of the time between 15 and 20. Mimicking real FGBL prices, the sample size or the number of trades per day is $n = 15{,}000$. Here, rounding means rounding the corresponding price $(110 \exp(Y_{i/n}))$ up to the two decimal places, and afterwards transforming back via $\log(\frac{\cdot}{110})$, that is rounding to full basis points of the price and is not to be confused with rounding of the log price. Notice that FGBL prices are most of the time in the range between 100 and 120. Therefore, 110 is a reasonable starting value (cf. also the upper panel in Fig. 1). The jump process is simulated as a compound Poisson process with constant intensity 3 and jump size distribution $\mathcal{N}(0, 10^{-6})$, resulting in average in three jumps/events per day.

The resulting empirical mean integrated squared errors (MISE) computed on the basis of 10,000 repetitions are displayed in Table 3. Obviously, jumps have a huge influence on ASVE, while rounding effects are negligible (at least regarding the FGBL data sets in Sect. 6). We observe that the bad impact of the jumps is reduced almost completely by the pre-processing of the data.

Table 3 Robustness. Simulation results for the MISE for data generated from the Heston model with additional rounding and jumps for ASVE with and without jump detection

	Pure	Rounded	With jumps	With jumps, rounded
Without jump detection	$1.41 \cdot 10^{-11}$	$1.41 \cdot 10^{-11}$	$12.64 \cdot 10^{-11}$	$12.86 \cdot 10^{-11}$
With jump detection	$1.68 \cdot 10^{-11}$	$1.69 \cdot 10^{-11}$	$1.69 \cdot 10^{-11}$	$1.70 \cdot 10^{-11}$

5 Time Schemes

It has been noticed in the econometrics literature that an increase in volatility might be due to different reasons. One explanation would be that there are larger price changes. Alternatively, the volatility will of course also increase if price changes are of the same size and only the number of trades per time interval goes up (cf. for example [29], Section IV.B). Disentangling the different explanations is quite difficult without an underlying mathematical concept. Nevertheless, determining the source of an increase in volatility is clearly of importance.

A more rigorous treatment of this problem leads to the definition of different notions of time (for instance in [21]). Here, we investigate the most prominent examples: real time and tick time (sometimes also referred to as clock time and transaction time).

Volatility in real time is appealing as it seems very intuitive. In tick time successive ticks are treated as one time unit. By definition, this time scheme does not depend on the speed at which successive trades occur. Consequently, volatility in tick time is independent of the trading intensity and hence measures the volatility of the price changes only. As the trading speed can be estimated directly from the ticks, we argue in this section that tick time volatility is the more natural object. A drawback of tick times is that there is no straightforward extension of the concept to multivariate processes.

Let us clarify the connection between both time schemes in more detail. Denote by t_i, $i = 1, \dots, n$ the ordered $(t_0 < t_1 < t_2 < \dots < t_n)$ sample of trading times. Then, for $i < j$ the time between t_i and t_j equals $\frac{j-i}{n}$ time units in tick time and $t_j - t_i$ time units in real time. With this notation, the tick time model is given by

$$Y^T_{i,n} = X_{t_i} + \epsilon_{i,n}, \quad i = 1, \dots, n. \tag{20}$$

Inspired by the classical high-frequency framework, we think about the trading times as an array, that is $t_i = t_{i,n}$, where the sampling rate gets finer for increasing n. Define the trading intensity ν at time t as

$$\nu(t) = \lim_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[t-\delta_n, t+\delta_n]}(t_i)}{2\delta_n}(t_n - t_0), \tag{21}$$

provided this limit exists and is unique for any sequence $\delta_n \to 0$ and $\delta_n n \to \infty$.

As an example consider the following toy model: Assume that σ is deterministic and there exists a deterministic, differentiable function $h : [0,1] \to [0,1]$ with $h(i/n) = t_{i,n}$ (in particular this implies that h is strictly monotone). Note that in this setting, ν is deterministic as well and given by the derivative of h^{-1}.

Let σ^2_{RT} denote the original (real time) spot volatility. Recall that under tick time, we consider successive trading times as equidistant. Therefore, the tick time spot volatility σ^2_{TT} satisfies for all $i = 1, \ldots, n$

$$\int_0^{i/n} \sigma_{TT}(h(s))dW_s = \int_0^{h(i/n)} \sigma_{RT}(s)dW_s =_{\mathscr{L}} \int_0^{i/n} \sqrt{h'(s)}\sigma_{RT}(h(s))dW_s$$

in law. Thus, the first and the latter integrand are (roughly) equal, that is $\sigma^2_{TT}(h(s)) = h'(s)\sigma^2_{RT}(h(s))$. Rewriting this, we obtain

$$\nu\sigma^2_{TT} = \sigma^2_{RT}, \tag{22}$$

cf. also Dahlhaus and Neddermeyer [21], Section 4. This formula clarifies the connection between tick time and real time volatility. Here, the argument is given for a deterministic trading scheme but extensions to random trading times are possible (cf. the literature on time changed Lévy processes, e.g. in Carr and Wu [16] and Cont and Tankov [20], or the survey article by Veraart and Winkel [50]).

Estimating the real time volatility directly from tick data, we have to construct artificial observations by recording the price each 10th second, for example. This method leads to a loss of information if there are many ticks in one time interval.

Notice that nonparametric estimation of the trading intensity ν is standard using for example (21) together with a proper choice of the bandwidth δ_n. In view of formula (22), it seems therefore more natural to estimate the real time spot volatility as product of $\hat{\sigma}^2_{TT}$ and an estimator of ν. In a simulation study, we estimated the real time volatility via its product representation for Euro-BUND Futures on all days in 2007 (for a description of the data, cf. also Sect. 6). We use Haar wavelets and hence obtain piecewise constant reconstructions. As a measure for the oscillation behavior of the volatility, we take the sum of squared jump sizes of the reconstructions for every of these days. In average, for tick time spot volatility this gives $9.68 \cdot 10^{-11}$ per day, while for real time volatility the corresponding value is $1.98 \cdot 10^{-10}$. This gives some evidence that the tick time volatility is much smoother than its real time counterpart.

As a surprising fact, formula (22) shows that even rates of convergence for estimation of σ^2_{RT} can be much faster than the minimax rates provided σ^2_{TT} is sufficiently smooth. To give an example, assume that σ_{TT} is constant and ν has Hölder continuity $\beta > 1/2$. In this case ν can be estimated with the classical nonparametric rate $n^{-\beta/(2\beta+1)} \ll n^{-1/4}$. Consequently, σ^2_{RT} has also Hölder index β. The rate for estimation of σ^2_{RT} is $n^{-1/4}$ which converges faster to zero than the minimax rate $n^{-\beta/(4\beta+2)}$ (for a derivation of minimax rates see Munk and Schmidt-Hieber [45] and Hoffmann et al. [37]).

To summarize, the tick time volatility is the quantity of interest measuring the volatility of the price changes. Furthermore, the real time volatility can easily be estimated via (22). For these reasons, we restrict ourselves throughout the following to estimation of spot volatility in tick time.

6 Spot Volatility of Euro-BUND Futures

We analyze the spot volatility of Euro-BUND Futures (FGBL) using tick data from Eurex database. The underlying is a 100,000 Euro debt security of the German Federal Government with coupon rate 6 % and maturity 8.5–10.5 years. The price is given in percentage of the par value. The tick times are recorded with precision of 10 ms. The minimum price change is 0.01 % (one basis point), corresponding to 10 Euro, which is comparably large. The number of trades per day varies among 10,000 and 30,000. Observations which are not due to trading are removed from the sample. If there are different FGBL contracts at a time referring to different expiration days, we only consider these belonging to the next possible date. Trading takes places from 8:00 a.m. until 7:00 p.m. Central European Time (CET). For the reconstructions, we restrict ourselves to observations within the time span 9 a.m. to 6 p.m. CET. Outside this period, trading is normally too slow to make use of a high-frequency setting.

During business hours, FGBL prices fit well as an example for high-frequency data. On the one hand, trading is very liquid due to low transaction costs and high trading volume. In average, the holding period is less than 2 days (cf. [27], Figure 4). On the other hand, microstructure effects are present and simple quadratic variation techniques fail as indicated in Fig. 6. In this plot (often referred to as signature

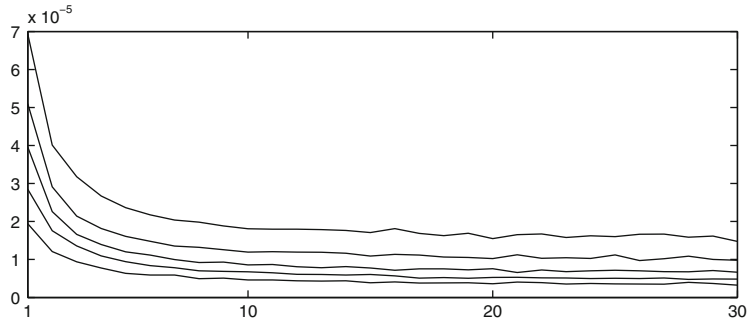


Fig. 6 Realized volatilities of FGBL data from June 4th to June 8th, 2007 for different subsampling frequencies

plot), we investigate how the (integrated) realized volatilities behaves if we consider subsamples of the data with different subsampling frequencies. We observe a rapid increase on small frequencies, that is if more and more data are included. This indicates that microstructure effects have to be taken into account.

In the following, we illustrate the effect of macroeconomic events with unanticipated outcome on spot volatility. As they cause uncertainty, one expects an increase in volatility once they appear. There has been a large body of literature in economics devoted to this subject. Nevertheless, up to now, there seems to be no general consensus quantifying how much the volatility is affected by public announcements. Ederington and Lee [28, 29] claim that volatility is substantially higher for a few minutes after the announcement and is still visible in the data for several hours. They also find evidence that volatility is slightly elevated for some minutes before an announcement. They conclude that macroeconomic announcements are the driving force for volatility. In contrast, in the seminal paper Andersen and Bollerslev [6] daily volatility patterns are found to explain most of the spot volatility behavior, while public announcements have only a secondary effect on overall volatility. In a recent study, Lunde and Zebedee [42] focus on the effects of US monetary policy events on volatility of US equity prices. In accordance with previous work, they conclude that there are spikes in the volatility around macroeconomic announcements, lasting for approximately 15 min. In Jansen and de Haan [39] effects of certain European Central Bank (ECB) announcements on price changes and volatility are studied. Although these papers deal with volatility on relatively short time intervals, none of them accounts for microstructure effects.

To illustrate our method, the 12 days in 2007 (one per month) with an official ECB press conference related to possible changes in key interest rates are studied. During these meetings hold jointly by the president and the vice-president of the European Central Bank, announcements about ECB-policy are made. In Jansen and de Haan [39], press conferences are excluded from the study, but they are very appealing because on the one hand, key interest rates are of major economic importance especially for government bonds like Euro-BUND futures, and on the other hand, the announcement procedure is highly standardized. In fact, on every of the studied dates the decision of the ECB Governing Council on the key interest rates was released on 1.45 p.m. followed by the official press conference starting at 2.30 p.m. and lasting for exactly an hour. The press conference consists of two parts starting with an introductory statement by the ECB president. In a second part, the president and vice-president answer questions of journalists. On every of these events, between 20 and 62 financial analysts are asked in advance to predict possible changes in the key interest rate. Based on these estimates a sample standard deviation is computed which is available at Bloomberg. In the following, we refer to this quantity as *market uncertainty*.

In Fig. 7, ASVE for May 10th, 2007 is displayed. The dashed line represents the time of the announcement, the hatched region refers to the time period of the press conference. On this day, the reconstruction displays an increase in volatility

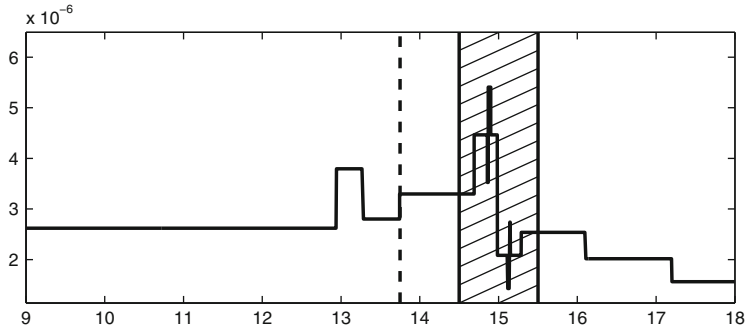


Fig. 7 ASVE for May 10th, 2007. Period of the ECB press conferences is hatched and announcement of not changing the key interest rate is represented by the *dashed line*

Table 4 Features (average, maximum, and total variation) of ASVE for days with ECB press conferences on key interest rates. The second column is an estimate of market uncertainty. Integrated volatility and total variation are normalized by the length of the time interval. All entries related to volatility are multiplied by 10^5

	Market	13.40 pm–13.50 pm			13.45 pm–15.30 pm		
Day	uncertainty	$\int \hat{\sigma}^2$	$\max \hat{\sigma}^2$	TV $\hat{\sigma}^2$	$\int \hat{\sigma}^2$	$\max \hat{\sigma}^2$	TV $\hat{\sigma}^2$
Jan-11	0	0.459	0.459	0	0.435	0.518	0.168
Feb-08	0	0.541	0.541	0	0.509	0.979	1.485
Mar-08	0	0.490	0.490	0	0.497	0.643	0.685
Apr-12	0	0.274	0.331	1.222	0.374	0.698	0.472
May-10	0	0.318	0.330	0.298	0.323	0.541	0.594
Jun-06	0	0.191	0.191	0	0.495	0.677	0.455
Jul-05	0	0.490	0.587	0.772	0.683	1.315	1.045
Aug-02	0.05	0.745	1.286	8.673	1.176	5.749	7.075
Sep-06	0.1	0.906	0.906	0	0.969	2.862	5.626
Oct-04	0.03	0.621	0.621	0	0.701	1.181	0.936
Nov-08	0	0.869	0.869	0	1.020	1.337	0.480
Dec-06	0	1.119	1.119	0	0.958	2.545	3.150
Average of days above		0.585	0.644	0.914	0.678	1.587	1.848
Average of all days		0.515	0.551	0.621	0.552	1.225	1.328
90 %-quantile all days		0.906	0.960	0.661	0.984	2.051	2.609

around the time of the announcement. Furthermore, we observe a higher fluctuation during the press conference. A more thorough analysis is done in Table 4: We observe a slight increase of the spot volatility on most of the considered days in view of average, maximum and total variation (which reflects the volatility of the volatility). On days, where the market uncertainty was nonzero, this effect is even

enhanced. Notice that the integral and TV figures are normalized by the length of the time interval to make them comparable. The results confirm the influence of macroeconomic events on volatility.

7 Generalization to Spot Covolatility Estimation

So far, we considered one-dimensional processes only. As for example in portfolio management, one might more generally be interested in the spot covariance matrix of multi-dimensional (and even very high-dimensional) price processes. There has been a lot of recent interest in this direction. The main additional difficulty is to deal with non-synchronous observations. Synchronization schemes in the context of estimation of the integrated covolatility (the multi-dimensional extension of the integrated volatility) were proposed in Hayashi and Yoshida [36], Ait-Sahalia et al. [4], Christensen et al. [17], Barndorff-Nielsen et al. [10], Zhang [53], and Bibinger [11], among others.

As an outlook, we shortly point out how to construct an estimator of the spot covolatility function κ given synchronous data, that is the covariance function of two log price processes observed at the same time points. For simplicity, we restrict ourselves to the bivariate case. In principle, this estimator can be combined in a second step with any of the synchronization schemes mentioned above.

Assume that we observe two processes

$$Y_{i,n}^{(1)} = X_{i/n}^{(1)} + \epsilon_{i,n}^{(1)}, \quad Y_{i,n}^{(2)} = X_{i/n}^{(2)} + \epsilon_{i,n}^{(2)}, \quad i = 1, \dots, n, \tag{23}$$

where $dX_t^{(1)} = \sigma_t^{(1)} dW_t^{(1)}$ and $dX_t^{(2)} = \sigma_t^{(2)} dW_t^{(2)}$ are two Itô martingales with driving Brownian motions $W^{(1)}, W^{(2)}$, and $\epsilon^{(1)}, \epsilon^{(2)}$ are two independent noise processes each defined analogously to (5). We assume that the spot covolatility function of $X^{(1)}$ and $X^{(2)}$ is given by $\kappa_t \, dt = Cov(dX_t^{(1)}, dX_t^{(2)})$.

For $i = 2, \dots, m$ and $q = 1, 2$, let $\overline{Y}_{i,m}^{(q)}$ be as defined in (7). Then, the wavelet coefficients of the spot covolatility are estimated via

$$\widehat{\langle g, \kappa \rangle} := \sum_{i=2}^{m} g\big(\tfrac{i-1}{m}\big) \overline{Y}_{i,m}^{(1)} \overline{Y}_{i,m}^{(2)}.$$

where again $g \in \{\varphi_{j_0,k}, \psi_{j,k}\}$. Since the noise processes $\epsilon^{(1)}, \epsilon^{(2)}$ are independent, no bias correction is necessary.

For illustration, Fig. 8 shows the reconstruction of the covolatility function of a realization in model (23) using the same thresholding procedure and parameter choices as for ASVE.

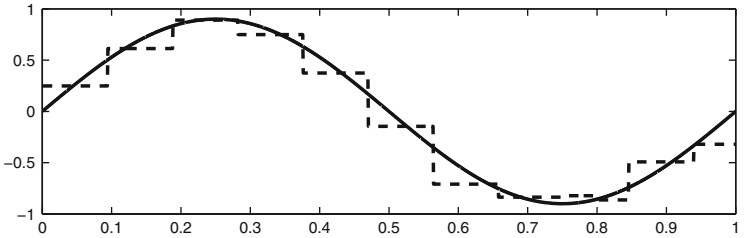


Fig. 8 Reconstruction (*dashed*) and true covolatility function (*solid*) for data following model (23) ($n = 15{,}000$), constant volatilities σ_1, σ_2, and i.i.d. centered Gaussian noise with SNR $= \frac{\sqrt{\langle |\kappa|, 1\rangle}}{\tau} = 20$

Acknowledgements Support of DFG/SNF-Grant FOR 916, DFG postdoctoral fellowship SCHM 2807/1-1, and Volkswagen Foundation is gratefully acknowledged. We appreciate the help of CRC 649 "Economic Risk" for providing us with access to Eurex database. Parts of this work are taken from the PhD thesis Schmidt-Hieber [48]. We thank Marc Hoffmann, Markus Reiß, Markus Bibinger, and an anonymous referee for many helpful remarks and suggestions.

Appendix A: Proof of Lemma 1

The proof is the same as in the PhD thesis Schmidt-Hieber [48]. We include it for sake of completeness.

To keep notation simple, we use the following quantities in the spirit of the definitions of Sect. 2.3: For any process $(A_{i,n}) \in \{(Y_{i,n}), (\epsilon_{i,n}), (X)_{i,n}\}$, define

$$\overline{A}_{i,m} = \overline{A}_{i,m}(\lambda) := \frac{m}{n} \sum_{\frac{j}{n} \in [\frac{i-2}{m}, \frac{i}{m}]} \lambda\big(m\tfrac{j}{n} - (i-2)\big) A_{j,n}.$$

$$\mathfrak{b}(A)_{i,m} = \mathfrak{b}(\lambda, A.)_{i,m} := \frac{m^2}{2n^2} \sum_{\frac{j}{n} \in [\frac{i-2}{m}, \frac{i}{m}]} \lambda^2\big(m\tfrac{j}{n} - (i-2)\big)\big(A_{j,n} - A_{j-1,n}\big)^2.$$

Further, recall that our estimator for the integrated volatility is given by $\widehat{\langle 1, \sigma^2 \rangle} = \sum_{i=2}^{m} \overline{Y}_{i,m}^2 - \mathfrak{b}(Y)_{i,m}$.

To prove the lemma, let us first show that the bias is of smaller order than $n^{-1/4}$. In fact, note that $\mathbb{E}\big[\overline{Y}_{i,m}^2\big] = \mathbb{E}\big[\overline{X}_{i,m}^2\big] + \mathbb{E}\big[\overline{\epsilon}_{i,m}^2\big]$. Clearly, one can bound

$$\Big|\mathbb{E}\big[\overline{\epsilon}_{i,m}^2\big] - \mathbb{E}\big[\mathfrak{b}(\lambda, Y)_{i,m}\big]\Big| = O(\tfrac{1}{n}).$$

Further, Lipschitz continuity of λ together with a Riemann approximation argument gives us

$$\left|\mathbb{E}\big[\,\overline{X}_{i,m}^2\big] - \tfrac{\sigma^2}{m}\right| = \left|\tfrac{\sigma^2}{m}\int_0^2\int_0^2 \lambda(s)\lambda(t)(s\wedge t)dtds - \tfrac{\sigma^2}{m}\right| + O(\tfrac{1}{n}) = O(\tfrac{1}{n}).$$

Here, the last equation is due to partial integration and the definition of a pre-average function (cf. Definition 2.1). Since both approximations are uniformly in i, this shows that the bias is of order $O(n^{-1/2})$.

For the asymptotic variance, first observe that $Var(\sum_{i=2}^m \mathfrak{b}(\lambda, Y)_{i,m}) = o(n^{-1/2})$. Hence,

$$\widehat{Var(\langle 1,\sigma^2\rangle)} = Var(\sum_{i=2}^m \overline{Y}_{i,m}^2) + o\Big(n^{-1/4}\big(Var(\sum_{i=2}^m \overline{Y}_{i,m}^2)\big)^{1/2} + n^{-1/2}\Big),$$

by Cauchy-Schwarz inequality. Recall that for centered Gaussian random variables U and V, $Cov(U^2,V^2) = 2(Cov(U,V))^2$. Therefore, it suffices to compute $Cov(\overline{Y}_{i,m},\overline{Y}_{k,m}) = \mathbb{E}[\overline{Y}_{i,m}\overline{Y}_{k,m}]$.

By the same arguments as above, that is Riemann summation and partial integration, we find

$$\mathbb{E}\Big[\Big|\overline{X}_{i,m}\overline{X}_{k,m} - \int_0^1 \Lambda(ms-(i-2))dX_s\int_0^1 \Lambda(ms-(k-2))dX_s\Big|\Big] \lesssim n^{-1}.$$

Therefore,

$$\mathbb{E}\big[\overline{X}_{i,m}\overline{X}_{k,m}\big] = \sigma^2\int_0^1 \Lambda(ms-(i-2))\Lambda(ms-(k-2))ds + O(n^{-1}),$$

where the last two arguments hold uniformly in i,k.

In order to calculate $\mathbb{E}[\overline{Y}_{i,m}\overline{Y}_{k,m}]$, we must treat three different cases, $|i-k| \geq 2$, $|i-k|=1$ and $i=k$, denoted by I, II and III.

ad I. In this case $(\frac{i-2}{m},\frac{i}{m}]$ and $(\frac{k-2}{m},\frac{k}{m}]$ do not overlap. By the equalities above, it follows $Cov(\overline{Y}_{i,m},\overline{Y}_{k,m}) = O(n^{-1})$.

ad II. Without loss of generality, we set $k = i+1$. Then, we obtain

$$\begin{aligned}
&Cov(\overline{Y}_{i,m},\overline{Y}_{i+1,m}) = \mathbb{E}\big[\overline{X}_{i,m}\overline{X}_{i+1,m}\big] + \mathbb{E}\big[\overline{\epsilon}_{i,m}\overline{\epsilon}_{i+1,m}\big]\\
=&\sigma^2\int_0^1 \Lambda(ms-(i-2))\Lambda(ms-(i-1))ds + O(n^{-1})\\
&+\tau^2\frac{m^2}{n^2}\sum_{\frac{j}{n}\in\left(\frac{i-2}{m},\frac{i}{m}\right]} \lambda(m\tfrac{j}{n}-(i-2))\lambda(m\tfrac{j}{n}-(i-1))\\
=&\frac{\sigma^2}{m}\int_0^1 \Lambda(u)\Lambda(1+u)du + \tau^2\frac{m}{n}\int_0^1 \lambda(u)\lambda(1+u)du + O(n^{-1}),
\end{aligned}$$

where the last inequality can be verified by Riemann summation. Noting that λ is a pre-average function, we obtain $\lambda(1+u) = -\lambda(1-u)$ and

$$Cov(\overline{Y}_{i,m}, \overline{Y}_{i+1,m}) = \frac{\sigma^2}{m}\int_0^1 \Lambda(u)\Lambda(1-u)du - \frac{\tau^2 m}{n}\int_0^1 \lambda(u)\lambda(1-u)du + O(n^{-1}).$$

ad III. It can be shown by redoing the arguments in *II* that

$$Var(\overline{Y}_{i,m}) = Var(\overline{X}_{i,m}) + Var(\overline{\epsilon}_{i,m}) = \frac{\sigma^2}{m}\int_0^2 \Lambda^2(u)du + \tau^2\frac{m}{n}\int_0^2 \lambda^2(u)du + O(n^{-1}).$$

Note that $\|\Lambda\|_{L^2[0,2]} = 1$. Since the above results hold uniformly in i, k, it follows directly that

$$\begin{aligned}
&Var\Big(\sum_{i=2}^{m}\overline{Y}_{i,m}^2\Big)\\
=&\sum_{i,k=2,\ |i-k|\geq 2}^{m} 2\big(Cov(\overline{Y}_{i,m}, \overline{Y}_{k,m})\big)^2\\
&+2\sum_{i=2}^{m-1}2\big(Cov(\overline{Y}_{i,m}, \overline{Y}_{i+1,m})\big)^2 + \sum_{i=2}^{m}2\big(Var(\overline{Y}_{i,m})\big)^2\\
=&O(n^{-1}) + 4\Big(\frac{\sigma^2}{\sqrt{c}}\int_0^1 \Lambda(u)\Lambda(1-u)du - \tau^2 c^{3/2}\int_0^1 \lambda(u)\lambda(1-u)du\Big)^2 n^{-1/2}\\
&+2\Big(\frac{\sigma^2}{\sqrt{c}} + 2\tau^2 c^{3/2}\|\lambda\|^2_{L^2[0,1]}\Big)^2 n^{-1/2}. \qquad \square
\end{aligned}$$

References

1. Ait-Sahalia, Y., & Jacod, J. (2009). Testing for jumps in a discretely observed process. *The Annals of Statistics, 37*, 184–222.
2. Ait-Sahalia, Y., & Yu, J. (2009). High frequency market microstructure noise estimates and liquidity measures. *The Annals of Applied Statistics, 3*, 422–457.
3. Ait-Sahalia, Y., Mykland, P. A., & Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *The Review of Financial Studies, 18*, 351–416.
4. Ait-Sahalia, Y., Fan, J., & Xiu, D. (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association, 105*, 1504–1517.
5. Ait-Sahalia, Y., Jacod, J., & Li, J. (2012). Testing for jumps in noisy high frequency data. *Journal of Econometrics, 168*, 207–222.
6. Andersen, T. G., & Bollerslev, T. (1998). Deutsche Mark-Dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies. *The Journal of Finance, 53*, 219–265.
7. Autin, F., Freyermuth, J. M., & von Sachs, R. (2011). Ideal denoising within a family of tree-structured wavelet estimators. *The Electronic Journal of Statistics, 5*, 829–855.
8. Bandi, F., & Russell, J. (2008). Microstructure noise, realized variance, and optimal sampling. *The Review of Economic Studies, 75*, 339–369.

9. Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Stephard, N. (2008). Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica, 76*(6), 1481–1536.
10. Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2011). Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics, 162*, 149–169.
11. Bibinger, M. (2011). Efficient covariance estimation for asynchronous noisy high-frequency data. *Scandinavian Journal of Statistics, 38*, 23–45.
12. Bollerslev, T., & Todorov, V. (2011). Estimation of jump tails. *Econometrica, 79*, 1727–1783.
13. Cai, T., & Wang, L. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. *The Annals of Statistics, 36*, 2025–2054.
14. Cai, T., & Zhou, H. (2009). A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics, 37*, 569–595.
15. Cai, T., Munk, A., & Schmidt-Hieber, J. (2010). Sharp minimax estimation of the variance of Brownian motion corrupted with Gaussian noise. *Statistica Sinica, 20*, 1011–1024.
16. Carr, P., & Wu, L. (2004). Time-changed Lévy processes and option pricing. *Journal of Financial Economics, 71*(1), 113–141.
17. Christensen, K., Kinnebrock, S., & Podolskij, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics, 159*, 116–133.
18. Cohen, A. (2003). *Numerical analysis of wavelet methods*. Amsterdam/Boston: Elsevier.
19. Cohen, A., Daubechies, I., & Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis, 1*, 54–81.
20. Cont, R., & Tankov, P. (2004). *Financial modelling with jump processes*. Boca Raton: CRC Press.
21. Dahlhaus, R., & Neddermeyer, J. C. (2013). On-line spot volatility-estimation and decomposition with nonlinear market microstructure noise models. ArXiv e-prints. arXiv:1006.1860v4.
22. Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia: SIAM.
23. Delbaen, F., & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische Annalen, 300*, 463–520.
24. Delbaen, F., & Schachermayer, W. (1998). The fundamental theorem of asset pricing for unbounded stochastic processes. *Mathematische Annalen, 312*, 215–250.
25. Donoho, D., & Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika, 81*, 425–455.
26. Donoho, D., Johnstone, I. M., Kerkyacharian, G., & Picard, D. (1995). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society: Series B Statistical Methodology, 57*, 301–369.
27. Dorfleitner, G. (2004). How short-termed is the trading behaviour in Eurex futures markets? *Applied Financial Economics, 14*, 1269–1279.
28. Ederington, L. H., & Lee, J. H. (1993). How markets process information: New releases and volatility. *The Journal of Finance, 48*, 1161–1191.
29. Ederington, L. H., & Lee, J. H. (1995). The short-run dynamics of the price adjustment to new information. *Journal of Financial and Quantitative Analysis, 30*, 117–134.
30. Fan, J., & Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association, 102*, 1349–1362.
31. Fan, J., & Wang, Y. (2008). Spot volatility estimation for high-frequency data. *Statistics and Its Interface, 1*, 279–288.
32. Gloter, A., & Jacod, J. (2001). Diffusions with measurement errors. I. Local asymptotic normality. *ESAIM Probability and Statistics, 5*, 225–242.
33. Gloter, A., & Jacod, J. (2001). Diffusions with measurement errors. II. Optimal estimators. *ESAIM Probability and Statistics, 5*, 243–260.
34. Hasbrouck, J. (1993). Assessing the quality of a security market: A new approach to transaction-cost measurement. *The Review of Financial Studies, 6*, 191–212.

35. Hautsch, N., & Podolskij, M. (2013). Pre-averaging based estimation of quadratic variation in the presence of noise and jumps: Theory, implementation, and empirical evidence. *Journal of Business and Economic Statistics, 31*(2), 165–183.
36. Hayashi, T., & Yoshida, N. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli, 11*, 359–379.
37. Hoffmann, M., Munk, A., & Schmidt-Hieber, J. (2012). Adaptive wavelet estimation of the diffusion coefficient under additive error measurements. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques, 48*(4), 1186–1216.
38. Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., & Vetter, M. (2009). Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic Processes and Their Applications, 119*(7), 2249–2276.
39. Jansen, D., & de Haan, J. (2006). Look who's talking: ECB communication during the first years of EMU. *International Journal of Finance and Economics, 11*, 219–228.
40. Li, W. V., & Shao, Q. M. (2002). A normal comparison inequality and its applications. *Probability Theory and Related Fields, 122*, 494–508.
41. Li, Y., Zhang, Z., & Zheng, X. (2013). Volatility inference in the presence of both endogenous time and microstructure noise. *Stochastic Processes and Their Applications, 123*, 2696–2727.
42. Lunde, A., & Zebedee, A. A. (2009). Intraday volatility responses to monetary policy events. *Financial Markets and Portfolio Management, 23*(4), 383–299.
43. Madahavan, A. (2000). Market microstructure: A survey. *Journal of Financial Markets, 3*, 205–258.
44. Munk, A., & Schmidt-Hieber, J. (2010). Nonparametric estimation of the volatility function in a high-frequency model corrupted by noise. *The Electronic Journal of Statistics, 4*, 781–821.
45. Munk, A., & Schmidt-Hieber, J. (2010). Lower bounds for volatility estimation in microstructure noise models. In J.O. Berger, T.T. Cai, & I.M. Johnstone (Eds.) *Borrowing strength: Theory powering applications - a festschrift for Lawrence D. Brown* (Vol. 6, pp. 43–55). Beachwood: Institute of Mathematical Statistics
46. Podolskij, M., & Vetter, M. (2009). Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. *Bernoulli, 15*, 634–658.
47. Reiß, M. (2011). Asymptotic equivalence for inference on the volatility from noisy observations. *The Annals of Statistics, 39*(2), 772–802.
48. Schmidt-Hieber, J. (2010). Nonparametric methods in spot volatility estimation. PhD thesis, Georg-August-Universität Göttingen.
49. van der Ploeg, A. (2005). *Stochastic volatility and the pricing of financial derivatives* (n°366 of the Tinbergen Institute research series, No. 366).
50. Veraart, A. E., & Winkel, M. (2010). Time change. In R. Cont (Ed.), *Encyclopedia of quantitative finance* (pp. 1812–1816). Chichester: Wiley.
51. Wassermann, L. (2010). *All of nonparametric statistics* (Springer texts in statistics). New York/London: Springer.
52. Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli, 12*, 1019–1043.
53. Zhang, L. (2011). Estimating covariation: Epps effect and microstructure noise. *Journal of Econometrics, 160*, 33–47.
54. Zhang, L., Mykland, P., & Ait-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association, 472*, 1394–1411.
55. Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business and Economic Statistics, 14*, 45–52.

Time Series Prediction via Aggregation: An Oracle Bound Including Numerical Cost

Andres Sanchez-Perez

Abstract We address the problem of forecasting a time series meeting the Causal Bernoulli Shift model, using a parametric set of predictors. The aggregation technique provides a predictor with well established and quite satisfying theoretical properties expressed by an oracle inequality for the prediction risk. The numerical computation of the aggregated predictor usually relies on a Markov chain Monte Carlo method whose convergence should be evaluated. In particular, it is crucial to bound the number of simulations needed to achieve a numerical precision of the same order as the prediction risk. In this direction we present a fairly general result which can be seen as an oracle inequality including the numerical cost of the predictor computation. The numerical cost appears by letting the oracle inequality depend on the number of simulations required in the Monte Carlo approximation. Some numerical experiments are then carried out to support our findings.

1 Introduction

The objective of our work is to forecast a stationary time series $Y = (Y_t)_{t\in\mathbb{Z}}$ taking values in $\mathcal{X} \subseteq \mathbb{R}^r$ with $r \geq 1$. For this purpose we propose and study an aggregation scheme using exponential weights.

Consider a set of individual predictors giving their predictions at each moment t. An aggregation method consists of building a new prediction from this set, which is nearly as good as the best among the individual ones, provided a risk criterion (see [17]). This kind of result is established by oracle inequalities. The power and the beauty of the technique lie in its simplicity and versatility. The more basic and general context of application is individual sequences, where no assumption on the observations is made (see [9] for a comprehensive overview). Nevertheless, results need to be adapted if we set a stochastic model on the observations.

A. Sanchez-Perez (✉)
Institut Mines-Télécom; Télécom ParisTech; CNRS LTCI Télécom ParisTech, 37 rue Dareau, 75014 Paris, France
e-mail: andres.sanchez-perez@telecom-paristech.fr

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_13

The use of exponential weighting in aggregation and its links with the PAC-Bayesian approach has been investigated for example in [5, 8] and [11]. Dependent processes have not received much attention from this viewpoint, except in [1] and [2]. In the present paper we study the properties of the Gibbs predictor, applied to Causal Bernoulli Shifts (CBS). CBS are an example of dependent processes (see [12] and [13]).

Our predictor is expressed as an integral since the set from which we do the aggregation is in general not finite. Large dimension is a trending setup and the computation of this integral is a major issue. We use classical Markov chain Monte Carlo (MCMC) methods to approximate it. Results from Łatuszyński [15, 16] control the number of MCMC iterations to obtain precise bounds for the approximation of the integral. These bounds are in expectation and probability with respect to the distribution of the underlying Markov chain.

In this contribution we first slightly revisit certain lemmas presented in [2, 8] and [20] to derive an oracle bound for the prediction risk of the Gibbs predictor. We stress that the inequality controls the convergence rate of the exact predictor. Our second goal is to investigate the impact of the approximation of the predictor on the convergence guarantees described for its exact version. Combining the PAC-Bayesian bounds with the MCMC control, we then provide an oracle inequality that applies to the MCMC approximation of the predictor, which is actually used in practice.

The paper is organised as follows: we introduce a motivating example and several definitions and assumptions in Sect. 2. In Sect. 3 we describe the methodology of aggregation and provide the oracle inequality for the exact Gibbs predictor. The stochastic approximation is studied in Sect. 4. We state a general proposition independent of the model for the Gibbs predictor. Next, we apply it to the more particular framework delineated in our paper. A concrete case study is analysed in Sect. 5, including some numerical work. A brief discussion follows in Sect. 6. The proofs of most of the results are deferred to Sect. 7.

Throughout the paper, for $\boldsymbol{a} \in \mathbb{R}^q$ with $q \in \mathbb{N}^*$, $\|\boldsymbol{a}\|$ denotes its Euclidean norm, $\|\boldsymbol{a}\| = (\sum_{i=1}^q a_i^2)^{1/2}$ and $\|\boldsymbol{a}\|_1$ its 1-norm $\|\boldsymbol{a}\|_1 = \sum_{i=1}^q |a_i|$. We denote, for $\boldsymbol{a} \in \mathbb{R}^q$ and $\Delta > 0$, $B(\boldsymbol{a}, \Delta) = \{\boldsymbol{a}_1 \in \mathbb{R}^q : \|\boldsymbol{a} - \boldsymbol{a}_1\| \leq \Delta\}$ and $B_1(\boldsymbol{a}, \Delta) = \{\boldsymbol{a}_1 \in \mathbb{R}^q : \|\boldsymbol{a} - \boldsymbol{a}_1\|_1 \leq \Delta\}$ the corresponding balls centered at $\boldsymbol{a}$ of radius $\Delta > 0$. In general bold characters represent column vectors and normal characters their components; for example $\boldsymbol{y} = (y_i)_{i \in \mathbb{Z}}$. The use of subscripts with ':' refers to certain vector components $\boldsymbol{y}_{1:k} = (y_i)_{1 \leq i \leq k}$, or elements of a sequence $X_{1:k} = (X_t)_{1 \leq t \leq k}$. For a random variable U distributed as ν and a measurable function h, $\nu[h(U)]$ or simply $\nu[h]$ stands for the expectation of $h(U)$: $\nu[h] = \int h(u)\nu(\mathrm{d}u)$.

2 Problem Statement and Main Assumptions

Real stable autoregressive processes of a fixed order, referred to as AR(d) processes, are one of the simplest examples of CBS. They are defined as the stationary solution

of

$$X_t = \sum_{j=1}^{d} \theta_j X_{t-j} + \sigma \xi_t \ , \tag{1}$$

where the $(\xi_t)_{t\in\mathbb{Z}}$ are i.i.d. real random variables with $\mathbb{E}[\xi_t] = 0$ and $\mathbb{E}[\xi_t^2] = 1$.

We dispose of several efficient estimates for the parameter $\boldsymbol{\theta} = \left[\theta_1 \ \dots \ \theta_d\right]'$ which can be calculated via simple algorithms as Levinson-Durbin or Burg algorithm for example. From them we derive also efficient predictors. However, as the model is simple to handle, we use it to progressively introduce our general setup.

Denote

$$A(\boldsymbol{\theta}) = \begin{bmatrix} \theta_1 & \theta_2 & \dots & \dots & \theta_d \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \ddots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix} ,$$

$\boldsymbol{X}_{t-1} = \left[X_{t-1} \ \dots \ X_{t-d}\right]'$ and $\boldsymbol{e}_1 = \left[1\ 0 \ \dots \ 0\right]'$ the first canonical vector of $\mathbb{R}^d$. M' represents the transpose of matrix M (including vectors). The recurrence (1) gives

$$X_t = \boldsymbol{\theta}' \boldsymbol{X}_{t-1} + \sigma \xi_t = \sigma \sum_{j=0}^{\infty} \boldsymbol{e}_1' A^j(\boldsymbol{\theta}) \boldsymbol{e}_1 \xi_{t-j} \ . \tag{2}$$

The eigenvalues of $A(\boldsymbol{\theta})$ are the inverses of the roots of the autoregressive polynomial $\boldsymbol{\theta}(z) = 1 - \sum_{k=1}^{d} \theta_k z^k$, then at most δ for some $\delta \in (0,1)$ due to the stability of X (see [7]). In other words $\boldsymbol{\theta} \in s_d(\delta) = \{\boldsymbol{\theta} \ : \ \boldsymbol{\theta}(z) \neq 0 \text{ for } |z| < \delta^{-1}\} \subseteq s_d(1)$. In this context (or even in a more general one, see [14]) for all $\delta_1 \in (\delta, 1)$ there is a constant $\bar{K}$ depending only on $\boldsymbol{\theta}$ and δ_1 such that for all $j \geq 0$

$$\left|\boldsymbol{e}_1' A^j(\boldsymbol{\theta}) \boldsymbol{e}_1\right| \leq \bar{K} \delta_1^j \ , \tag{3}$$

and then, the variance of X_t, denoted γ_0, satisfies $\gamma_0 = \sigma^2 \sum_{j=0}^{\infty} |\boldsymbol{e}_1' A^j(\boldsymbol{\theta}) \boldsymbol{e}_1|^2 \leq \bar{K}^2 \sigma^2 / (1 - \delta_1^2)$.

The following definition allows to introduce the process which interests us.

Definition 1 Let $\mathcal{X}' \subseteq \mathbb{R}^{r'}$ for some $r' \geq 1$ and let $A = (A_j)_{j\geq 0}$ be a sequence of non-negative numbers. A function $H : (\mathcal{X}')^{\mathbb{N}} \to \mathcal{X}$ is said to be A-Lipschitz if

$$\|H(\boldsymbol{u}) - H(\boldsymbol{v})\| \leq \sum_{j=0}^{\infty} A_j \|u_j - v_j\| \ ,$$

for any $\boldsymbol{u} = (u_j)_{j\in\mathbb{N}}, \boldsymbol{v} = (v_j)_{j\in\mathbb{N}} \in (\mathcal{X}')^{\mathbb{N}}$.

Provided $A = (A_j)_{j\geq 0}$ with $A_j \geq 0$ for all $j \geq 0$, the i.i.d. sequence of $\mathcal{X}'$-valued random variables $(\xi_t)_{t\in\mathbb{Z}}$ and $H : (\mathcal{X}')^{\mathbb{N}} \to \mathcal{X}$, we consider that a time series $X = (X_t)_{t\in\mathbb{Z}}$ admitting the following property is a Causal Bernoulli Shift (CBS) with Lipschitz coefficients A and innovations $(\xi_t)_{t\in\mathbb{Z}}$.

(M) The process $X = (X_t)_{t\in\mathbb{Z}}$ meets the representation

$$X_t = H\left(\xi_t, \xi_{t-1}, \xi_{t-2}, \ldots\right), \forall t \in \mathbb{Z},$$

where H is an A-Lipschitz function with the sequence A satisfying

$$\tilde{A}_* = \sum_{j=0}^{\infty} jA_j < \infty . \tag{4}$$

We additionally define

$$A_* = \sum_{j=0}^{\infty} A_j . \tag{5}$$

CBS regroup several types of nonmixing stationary Markov chains, real-valued functional autoregressive models and Volterra processes, among other interesting models (see [10]). Thanks to the representation (2) and the inequality (3) we assert that AR(d) processes are CBS with $A_j = \sigma \bar{K} \delta_1^j$ for $j \geq 0$.

We let ξ denote a random variable distributed as the ξ_ts. Results from [1] and [2] need a control on the exponential moment of ξ in $\zeta = A_*$, which is provided via the following hypothesis.

(I) The innovations $(\xi_t)_{t\in\mathbb{Z}}$ satisfy $\phi(\zeta) = \mathbb{E}\left[\mathrm{e}^{\zeta\|\xi\|}\right] < \infty$.

Bounded or Gaussian innovations trivially satisfy this hypothesis for any $\zeta \in \mathbb{R}$.

Let π_0 denote the probability distribution of the time series Y that we aim to forecast. Observe that for a CBS, π_0 depends only on H and the distribution of ξ. For any $f : \mathcal{X}^{\mathbb{N}^*} \to \mathcal{X}$ measurable and $t \in \mathbb{Z}$ we consider $\hat{Y}_t = f\left((Y_{t-i})_{i\geq 1}\right)$, a possible predictor of Y_t from its past. For a given loss function $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the prediction risk is evaluated by the expectation of $\ell(\hat{Y}_t, Y_t)$

$$R(f) = \mathbb{E}\left[\ell\left(\hat{Y}_t, Y_t\right)\right] = \pi_0\left[\ell\left(\hat{Y}_t, Y_t\right)\right] = \int_{\mathcal{X}^{\mathbb{Z}}} \ell\left(f\left((y_{t-i})_{i\geq 1}\right), y_t\right) \pi_0(\mathrm{d}\mathbf{y}) .$$

We assume in the following that the loss function ℓ fulfills the condition:

(L) For all $y, z \in \mathcal{X}$, $\ell(y, z) = g(y - z)$, for some convex function g which is non-negative, $g(0) = 0$ and K- Lipschitz: $|g(y) - g(z)| \leq K\|y - z\|$.

If $\mathcal{X}$ is a subset of $\mathbb{R}$, $\ell(y, z) = |y - z|$ satisfies 1 with $K = 1$.

From estimators of dimension d for $\boldsymbol{\theta}$ we can build the corresponding linear predictors $f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\theta}' \boldsymbol{y}_{1:d}$. Speaking more broadly, consider a set Θ and associated with it a set of predictors $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$. For each $\boldsymbol{\theta} \in \Theta$ there is a unique $d = d(\boldsymbol{\theta}) \in \mathbb{N}^*$ such that $f_{\boldsymbol{\theta}} : \mathcal{X}^d \to \mathcal{X}$ is a measurable function from which we define

$$\hat{Y}_t^{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(Y_{t-1}, \ldots, Y_{t-d}) \ ,$$

as a predictor of Y_t given its past. We can extend all functions $f_{\boldsymbol{\theta}}$ in a trivial way (using dummy variables) to start from $\mathcal{X}^{\mathbb{N}^*}$. A natural way to evaluate the predictor associated with $\boldsymbol{\theta}$ is to compute the risk $R(\boldsymbol{\theta}) = R(f_{\boldsymbol{\theta}})$. We use the same letter R by an abuse of notation.

We observe $X_{1:T}$ from $X = (X_t)_{t \in \mathbb{Z}}$, an independent copy of Y. A crucial goal of this work is to build a predictor function $\hat{f}_T$ for Y, inferred from the sample $X_{1:T}$ and Θ such that $R(\hat{f}_T)$ is close to $\inf_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta})$ with π_0- probability close to 1.

The set Θ also depends on T, we write $\Theta \equiv \Theta_T$. Let us define

$$d_T = \sup_{\boldsymbol{\theta} \in \Theta_T} d(\boldsymbol{\theta}) \ . \tag{6}$$

The main assumptions on the set of predictors are the following ones.

(P-1) The set $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ is such that for any $\boldsymbol{\theta} \in \Theta_T$ there are $b_1(\boldsymbol{\theta}), \ldots, b_{d(\boldsymbol{\theta})}(\boldsymbol{\theta}) \in \mathbb{R}_+$ satisfying for all $\boldsymbol{y} = (y_i)_{i \in \mathbb{N}^*}, \boldsymbol{z} = (z_i)_{i \in \mathbb{N}^*} \in \mathcal{X}^{\mathbb{N}^*}$,

$$||f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\boldsymbol{\theta}}(\boldsymbol{z})|| \leq \sum_{j=1}^{d(\boldsymbol{\theta})} b_j(\boldsymbol{\theta}) \left|\left| y_j - z_j \right|\right| \ .$$

We assume moreover that $L_T = \sup_{\boldsymbol{\theta} \in \Theta_T} \sum_{j=1}^{d(\boldsymbol{\theta})} b_j(\boldsymbol{\theta}) < \infty$.

(P-2) The inequality $L_T + 1 \leq \log T$ holds for all $T \geq 4$.

In the case where $\mathcal{X} \subseteq \mathbb{R}$ and $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ is such that $\boldsymbol{\theta} \in \mathbb{R}^{d(\boldsymbol{\theta})}$ and $f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\theta}' \boldsymbol{y}_{1:d(\boldsymbol{\theta})}$ for all $\boldsymbol{y} \in \mathbb{R}^{\mathbb{N}}$, we have

$$|f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\boldsymbol{\theta}}(\boldsymbol{z})| \leq \sum_{j=1}^{d(\boldsymbol{\theta})} \left|\theta_j\right| \left|y_j - z_j\right| \ .$$

The last conditions are satisfied by the linear predictors when Θ_T is a subset of the ℓ_1-ball of radius $\log T - 1$ in $\mathbb{R}^{d_T}$.

3 Prediction via Aggregation

The predictor that we propose is defined as an average of predictors $f_{\boldsymbol{\theta}}$ based on the empirical version of the risk,

$$r_T\left(\boldsymbol{\theta}\,|X\right) = \frac{1}{T - d\left(\boldsymbol{\theta}\right)} \sum_{t=d(\boldsymbol{\theta})+1}^{T} \ell\left(\hat{X}_t^{\boldsymbol{\theta}}, X_t\right) \ .$$

where $\hat{X}_t^{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}\left((X_{t-i})_{i\geq 1}\right)$. The function $r_T\left(\boldsymbol{\theta}\,|X\right)$ relies on $X_{1:T}$ and can be computed at stage T; this is in fact a statistic.

We consider a prior probability measure π_T on Θ_T. The prior serves to control the complexity of predictors associated with Θ_T. Using π_T we can construct one predictor in particular, as detailed in the following.

3.1 Gibbs Predictor

For a measure ν and a measurable function h (called energy function) such that $\nu\left[\exp\left(h\right)\right] = \int \exp\left(h\right)\ \mathrm{d}\nu < \infty$, we denote by $\nu\left\{h\right\}$ the measure defined as

$$\nu\left\{h\right\}\left(\mathrm{d}\boldsymbol{\theta}\right) = \frac{\exp\left(h\left(\boldsymbol{\theta}\right)\right)}{\nu\left[\exp\left(h\right)\right]}\nu\left(\mathrm{d}\boldsymbol{\theta}\right) \ .$$

It is known as the Gibbs measure.

Definition 2 (Gibbs predictor) Given $\eta > 0$, called the temperature or the learning rate parameter, we define the Gibbs predictor as the expectation of $f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ is drawn under $\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}$, that is

$$\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|X\right) = \pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left[f_{\cdot}\left(\boldsymbol{y}\right)\right] = \int_{\Theta_T} f_{\boldsymbol{\theta}}\left(\boldsymbol{y}\right) \frac{\exp\left(-\eta r_T\left(\boldsymbol{\theta}\,|X\right)\right)}{\pi_T\left[\exp\left(-\eta r_T\left(\cdot\,|X\right)\right)\right]}\pi_T\left(\mathrm{d}\boldsymbol{\theta}\right) \ . \tag{7}$$

3.2 PAC-Bayesian Inequality

At this point more care must be taken to describe Θ_T. Here and in the following we suppose that

$$\Theta_T \subseteq \mathbb{R}^{n_T} \ \text{ for some } n_T \in \mathbb{N}^* \ . \tag{8}$$

Suppose moreover that Θ_T is equipped with the Borel σ-algebra $\mathcal{B}(\Theta_T)$.

A Lipschitz type hypothesis on $\boldsymbol{\theta}$ guarantees the robustness of the set $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ with respect to the risk R.

(**P**-3) There is $\mathcal{D} < \infty$ such that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_T$,

$$\pi_0 \left[\left| f_{\boldsymbol{\theta}_1} \left((X_{t-i})_{i \geq 1} \right) - f_{\boldsymbol{\theta}_2} \left((X_{t-i})_{i \geq 1} \right) \right| \right] \leq \mathcal{D} d_T^{1/2} \left| \left| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right| \right| ,$$

where d_T is defined in (6).

Linear predictors satisfy this last condition with $\mathcal{D} = \pi_0 [|X_1|]$.

Suppose that the $\boldsymbol{\theta}$ reaching the $\inf_{\boldsymbol{\theta} \in \Theta_T} R(\boldsymbol{\theta})$ has some zero components, i.e. $\text{supp}(\boldsymbol{\theta}) < n_T$. Any prior with a lower bounded density (with respect to the Lebesgue measure) allocates zero mass on lower dimensional subsets of Θ_T. Furthermore, if the density is upper bounded we have $\pi_T[B(\boldsymbol{\theta}, \Delta) \cap \Theta_T] = O(\Delta^{n_T})$ for Δ small enough. As we will notice in the proof of Theorem 1, a bound like the previous one would impose a tighter constraint to n_T. Instead we set the following condition.

(**P**-4) There is a sequence $(\boldsymbol{\theta}_T)_{T \geq 4}$ and constants $\mathcal{C}_1 > 0, \mathcal{C}_2, \mathcal{C}_3 \in (0, 1]$ and $\gamma \geq 1$ such that $\boldsymbol{\theta}_T \in \Theta_T$,

$$R(\boldsymbol{\theta}_T) \leq \inf_{\boldsymbol{\theta} \in \Theta_T} R(\boldsymbol{\theta}) + \mathcal{C}_1 \frac{\log^3(T)}{T^{1/2}} ,$$

$$\text{and} \quad \pi_T \left[B(\boldsymbol{\theta}_T, \Delta) \cap \Theta_T \right] \geq \mathcal{C}_2 \Delta^{n_T^{1/\gamma}}, \forall 0 \leq \Delta \leq \Delta_T = \frac{\mathcal{C}_3}{T} .$$

A concrete example is provided in Sect. 5.

We can now present the main result of this section, our PAC-Bayesian inequality concerning the predictor $\hat{f}_{\eta_T, T}(\cdot | X)$ built following (7) with the learning rate $\eta = \eta_T = T^{1/2}/(4 \log T)$, provided an arbitrary probability measure π_T on Θ_T.

Theorem 1 *Let ℓ be a loss function such that Assumption (**L**) holds. Consider a process $X = (X_t)_{t \in \mathbb{Z}}$ satisfying Assumption (**M**) and let π_0 denote its probability distribution. Assume that the innovations fulfill Assumption (**I**) with $\zeta = A_*$; A_* is defined in (5). For each $T \geq 4$ let $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ be a set of predictors meeting Assumptions (**P-3**), (**P-4**) and (**P-3**) such that d_T, defined in (6), is at most $T/2$. Suppose that the set Θ_T is as in (8) with $n_T \leq \log^\gamma T$ for some $\gamma \geq 1$ and we let π_T be a probability measure on it such that Assumption (**P-4**) holds for the same γ. Then for any $\varepsilon > 0$, with π_0-probability at least $1 - \varepsilon$,*

$$R\left(\hat{f}_{\eta_T, T}(\cdot | X) \right) \leq \inf_{\boldsymbol{\theta} \in \Theta_T} R(f_{\boldsymbol{\theta}}) + \mathcal{E} \frac{\log^3 T}{T^{1/2}} + \frac{8 \log T}{T^{1/2}} \log \left(\frac{1}{\varepsilon} \right) ,$$

where

$$\mathcal{E} = \mathcal{C}_1 + 8 + \frac{2}{\log 2} - \frac{2\log \mathcal{C}_2}{\log^2 2} - \frac{4\log \mathcal{C}_3}{\log 2} + \frac{8K^2\left(A_* + \tilde{A}_*\right)^2}{\tilde{A}_*^2} + \frac{K\mathcal{D}\mathcal{C}_3}{8\log^3 2} + \frac{4K\phi(A_*)}{\log 2} + \frac{2K^2\phi(A_*)}{\log^2 2}, \tag{9}$$

*with $\tilde{A}_*$ defined in (4), K, ϕ and $\mathcal{D}$ in Assumptions (**L**), (**I**) and (**P-3**), respectively, and $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ in Assumption (**P-4**).*

The proof is postponed to Sect. 7.1.

Here however we insist on the fact that this inequality applies to an exact aggregated predictor $\hat{f}_{\eta_T,T}(\cdot|X)$. We need to investigate how these predictors are computed and how practical numerical approximations behave compared to the properties of the exact version.

4 Stochastic Approximation

Once we have the observations $X_{1:T}$, we use the Metropolis – Hastings algorithm to compute $\hat{f}_{\eta,T}(\cdot|X) = \int f_{\boldsymbol{\theta}}(\cdot|X)\,\pi_T\{-\eta r_T(\boldsymbol{\theta}|X)\}(\mathrm{d}\boldsymbol{\theta})$. The Gibbs measure $\pi_T\{-\eta r_T(\cdot|X)\}$ is a distribution on Θ_T whose density $\pi_{\eta,T}(\cdot|X)$ with respect to π_T is proportional to $\exp(-\eta r_T(\cdot|X))$.

4.1 Metropolis: Hastings Algorithm

Given $X \in \mathcal{X}^{\mathbb{Z}}$, the Metropolis-Hastings algorithm generates a Markov chain $\Phi_{\eta,T}(X) = (\boldsymbol{\theta}_{\eta,T,n}(X))_{n\geq 0}$ with kernel $P_{\eta,T}$ (only depending on $X_{1:T}$) having the target distribution $\pi_T\{-\eta r_T(\cdot|X)\}$ as the unique invariant measure, based on the transitions of another Markov chain which serves as a proposal (see [21]). We consider a proposal transition of the form $Q_{\eta,T}(\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}) = q_{\eta,T}(\boldsymbol{\theta}_1, \boldsymbol{\theta})\pi_T(\mathrm{d}\boldsymbol{\theta})$ where the conditional density kernel $q_{\eta,T}$ (possibly also depending on $X_{1:T}$) on $\Theta_T \times \Theta_T$ is such that

$$\beta_{\eta,T}(X) = \inf_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\Theta_T\times\Theta_T} \frac{q_{\eta,T}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)}{\pi_{\eta,T}(\boldsymbol{\theta}_2|X)} \in (0,1)\ . \tag{10}$$

This is the case of the independent Hastings algorithm, where the proposal is i.i.d. with density $q_{\eta,T}$. The condition gets into

$$\beta_{\eta,T}(X) = \inf_{\boldsymbol{\theta}\in\Theta_T} \frac{q_{\eta,T}(\boldsymbol{\theta})}{\pi_{\eta,T}(\boldsymbol{\theta}|X)} \in (0,1)\ . \tag{11}$$

In Sect. 5 we provide an example.

The relation (10) implies that the algorithm is uniformly ergodic, i.e. we have a control in total variation norm ($\|\cdot\|_{TV}$). Thus, the following condition holds (see [18]).

(A) Given $\eta, T > 0$, there is $\beta_{\eta,T} : \mathcal{X}^{\mathbb{Z}} \to (0,1)$ such for any $\boldsymbol{\theta}_0 \in \Theta_T$, $\boldsymbol{x} \in \mathcal{X}^{\mathbb{Z}}$ and $n \in \mathbb{N}$, the chain $\Phi_{\eta,T}(\boldsymbol{x})$ with transition law $P_{\eta,T}$ and invariant distribution $\pi_T\{-\eta r_T(\cdot\,|\boldsymbol{x})\}$ satisfies

$$\left\| P^n_{\eta,T}(\boldsymbol{\theta}_0, \cdot) - \pi_T\{-\eta r_T(\cdot\,|\boldsymbol{x})\} \right\|_{TV} \leq 2\left(1 - \beta_{\eta,T}(\boldsymbol{x})\right)^n .$$

4.2 Theoretical Bounds for the Computation

In [16, Theorem 3.1] we find a bound on the mean square error of approximating one integral by the empirical estimate obtained from the successive samples of certain ergodic Markov chains, including those generated by the MCMC method that we use.

A MCMC method adds a second source of randomness to the forecasting process and our aim is to measure it. Let $\boldsymbol{\theta}_0 \in \cap_{T\geq 1}\Theta_T$, we set $\boldsymbol{\theta}_{\eta,T,0}(\boldsymbol{x}) = \boldsymbol{\theta}_0$ for all $T, \eta > 0$, $\boldsymbol{x} \in \mathcal{X}^{\mathbb{Z}}$. We denote by $\mu_{\eta,T}(\cdot\,|X)$ the probability distribution of the Markov chain $\Phi_{\eta,T}(X)$ with initial point $\boldsymbol{\theta}_0$ and kernel $P_{\eta,T}$.

Let $\nu_{\eta,T}$ denote the probability distribution of $(X, \Phi_{\eta,T}(X))$; it is defined by setting for all sets $A \in (\mathcal{B}(\mathcal{X}))^{\otimes\mathbb{Z}}$ and $B \in (\mathcal{B}(\Theta_T))^{\otimes\mathbb{N}}$

$$\nu_{\eta,T}(A \times B) = \int \mathbb{1}_A(\boldsymbol{x})\, \mathbb{1}_B(\boldsymbol{\phi})\, \mu_{\eta,T}(\mathrm{d}\boldsymbol{\phi}\,|\boldsymbol{x})\, \pi_0(\mathrm{d}\boldsymbol{x}) \tag{12}$$

Given $\Phi_{\eta,T} = (\boldsymbol{\theta}_{\eta,T,n})_{n\geq 0}$, we then define for $n \in \mathbb{N}^*$

$$\bar{f}_{\eta,T,n} = \frac{1}{n}\sum_{i=0}^{n-1} f_{\boldsymbol{\theta}_{\eta,T,i}} . \tag{13}$$

Since our chain depends on X, we make it explicit by using the notation $\bar{f}_{\eta,T,n}(\cdot\,|X)$. The cited [16, Theorem 3.1] leads to a proposition that applies to the numerical approximation of the Gibbs predictor (the proof is in Sect. 7.2). We stress that this is independent of the model (CBS or any), of the set of predictors and of the theoretical guarantees of Theorem 1.

Proposition 1 *Let ℓ be a loss function meeting Assumption (**L**). Consider any process $X = (X_t)_{t\in\mathbb{Z}}$ with an arbitrary probability distribution π_0. Given $T \geq 2$, $\eta > 0$, a set of predictors $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ and $\pi_T \in \mathcal{M}^1_+(\Theta_T)$, let $\hat{f}_{\eta,T}(\cdot\,|X)$ be defined by (7) and let $\bar{f}_{\eta,T,n}(\cdot\,|X)$ be defined by (13). Suppose that $\Phi_{\eta,T}$ meets Assumption (**A**) for η and T with a function $\beta_{\eta,T} : \mathcal{X}^{\mathbb{Z}} \to (0,1)$. Let $\nu_{\eta,T}$ denote*

the probability distribution of $(X, \Phi_{\eta,T}(X))$ *as defined in (14). Then, for all* $n \geq 1$ *and* $D > 0$*, with* $\nu_{\eta,T}$*- probability at least* $\max\{0, 1 - A_{\eta,T}/(Dn^{1/2})\}$ *we have* $|R(\bar{f}_{\eta,T,n}(\cdot\,|X)) - R(\hat{f}_{\eta,T}(\cdot\,|X))| \leq D$*, where*

$$A_{\eta,T} = 3K \int_{\mathcal{X}^{\mathbb{Z}}} \frac{1}{\beta_{\eta,T}(\boldsymbol{x})} \int_{\mathcal{X}^{\mathbb{Z}}} \sup_{\boldsymbol{\theta} \in \Theta_T} \left| f_{\boldsymbol{\theta}}(\boldsymbol{y}) - \hat{f}_{\eta,T}(\boldsymbol{y}\,|\boldsymbol{x}) \right| \pi_0(\mathrm{d}\boldsymbol{y})\, \pi_0(\mathrm{d}\boldsymbol{x}) \,. \tag{14}$$

We denote by $\nu_T = \nu_{\eta_T,T}$ the probability distribution of $(X, \Phi_{\eta,T}(X))$ setting $\eta = \eta_T = T^{1/2}/(4 \log T)$. As Theorem 1 does not involve any simulation, it also holds in ν_T- probability. From this and Proposition 1 a union bound gives us the following.

Theorem 2 *Under the hypothesis of Theorem 1, consider moreover that Assumption* (**A**) *is fulfilled by* $\Phi_{\eta,T}$ *for all* $\eta = \eta_T$ *and* T *with* $T \geq 4$*. Thus, for all* $\varepsilon > 0$ *and* $n \geq M(T, \varepsilon)$*, with* ν_T*- probability at least* $1 - \varepsilon$ *we have*

$$R\left(\bar{f}_{\eta_T,T,n}(\cdot\,|X)\right) \leq \inf_{\boldsymbol{\theta} \in \Theta_T} R(f_{\boldsymbol{\theta}}) + \left(\mathcal{E} + \frac{2}{\log 2} + 2\right) \frac{\log^3 T}{T^{1/2}} + \frac{8 \log T}{T^{1/2}} \log\left(\frac{1}{\varepsilon}\right) ,$$

where $\mathcal{E}$ *is defined in (9) and* $M(T, \varepsilon) = A^2_{\eta_T,T} T/(\varepsilon^2 \log^6 T)$ *with* $A_{\eta,T}$ *as in (14).*

5 Applications to the Autoregressive Process

We carefully recapitulate all the assumptions of Theorem 2 in the context of an autoregressive process. After that, we illustrate numerically the behaviour of the proposed method.

5.1 Theoretical Considerations

Consider a real valued stable autoregressive process of finite order d as defined by (1) with parameter $\boldsymbol{\theta}$ lying in the interior of $s_d(\delta)$ and unit normally distributed innovations (Assumptions (**M**) and (**I**) hold). With the loss function $\ell(y, z) = |y - z|$ Assumption (**L**) holds as well. The linear predictors is the set that we test; they meet Assumption (**P-3**). Without loss of generality assume that $d_T = n_T$. In the described framework we have $\hat{f}_{\eta,T}(\cdot\,|X) = f_{\widehat{\boldsymbol{\theta}}_{\eta,T}(X)}$, where

$$\hat{\boldsymbol{\theta}}_{\eta,T}(X) = \int_{\Theta_T} \boldsymbol{\theta} \frac{\exp(-\eta r_T(\boldsymbol{\theta}\,|X))}{\pi_T[\exp(-\eta r_T(\boldsymbol{\theta}\,|X))]} \pi_T(\mathrm{d}\boldsymbol{\theta}) \,.$$

This $\hat{\boldsymbol{\theta}}_{\eta,T}(X) \in \mathbb{R}^{d_T}$ is known as the Gibbs estimator.

Remark that, by (2) and the normality of the innovations, the risk of any $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{d_T}$ is computed as the absolute moment of a centered Gaussian, namely

$$R\left(f_{\hat{\boldsymbol{\theta}}}\right) = R\left(\hat{\boldsymbol{\theta}}\right) = \frac{\left(2\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \Gamma_T \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) + 2\sigma^2\right)^{1/2}}{\pi^{1/2}}, \tag{15}$$

where $\Gamma_T = (\gamma_{i,j})_{0 \le i,j \le d_T - 1}$ is the covariance matrix of the process. In (15) the vector $\boldsymbol{\theta}$ originally in $\mathbb{R}^d$ is completed by $d_T - d$ zeros.

In this context $\arg\inf_{\boldsymbol{\theta} \in \mathbb{R}^{\mathbb{N}^*}} R(\boldsymbol{\theta}) \in s_d(1)$ gives the true parameter $\boldsymbol{\theta}$ generating the process. Let us verify Assumption (**P-4**) by setting conveniently Θ_T and π_T. Let $\Delta_{d*} > 0$ be such that $B(\boldsymbol{\theta}, \Delta_{d*}) \subseteq s_d(1)$.

We express $\Theta_T = \bigcup_{k=1}^{d_T} \Theta_{k,T}$ where $\boldsymbol{\theta} \in \Theta_{k,T}$ if and only if $d(\boldsymbol{\theta}) = k$. It is interesting to set $\Theta_{k,T}$ as the part of the stability domain of an AR(k) process satisfying Assumptions (**P-3**) and (**P-4**). Consider $\Theta_{1,T} = s_1(1) \times \{0\}^{d_T - 1} \cap B_1(\mathbf{0}, \log T - 1)$ and $\Theta_{k,T} = s_k(1) \times \{0\}^{d_T - k} \cap B_1(\mathbf{0}, \log T - 1) \setminus \Theta_{k-1,T}$ for $k \ge 2$. Assume moreover that $d_T = \lfloor \log^\gamma T \rfloor$.

We write $\pi_T = \sum_{k=1}^{d_T} c_{k,T} \pi_{k,T}$ where for all k, $c_{k,T}\pi_{k,T}$ is the restriction of π_T to $\Theta_{k,T}$ with $c_{k,T}$ a real non negative number and $\pi_{k,T}$ a probability measure on $\Theta_{k,T}$. In this setup $c_{k,T} = \pi_T[\Theta_{k,T}]$ and $\pi_{k,T}[A \cap \Theta_{k,T}] = \pi_T[A \cap \Theta_{k,T}] / c_{k,T}$ if $c_{k,T} > 0$ and $\pi_{k,T}[A \cap \Theta_{k,T}] = 0$ otherwise. The vector $\left[c_{1,T} \ldots c_{d_T,T}\right]$ could be interpreted as a prior on the model order. Set $c_{k,T} = c_k / (\sum_{i=1}^{d_T} c_i)$ where $c_k > 0$ is the k-th term of a convergent series ($\sum_{k=1}^{\infty} c_k = c^* < \infty$).

The distribution $\pi_{k,T}$ is inferred from some transformations explained below. Observe first that if $a \le b$ we have $s_k(a) \subseteq s_k(b)$. If $\boldsymbol{\theta} \in s_k(1)$ then $\left[\lambda\theta_1 \ldots \lambda^k \theta_k\right]' \in s_k(1)$ for any $\lambda \in (-1, 1)$. Let us set

$$\lambda_T(\boldsymbol{\theta}) = \min\left\{1, \frac{\log T - 1}{\|\boldsymbol{\theta}\|_1}\right\}.$$

We define $F_{k,T}(\boldsymbol{\theta}) = \left[\lambda_T(\boldsymbol{\theta})\theta_1 \ldots \lambda_T^k(\boldsymbol{\theta})\theta_k\ 0 \ldots 0\right]' \in \mathbb{R}^{d_T}$. Remark that for any $\boldsymbol{\theta} \in s_k(1)$, $\|F_{k,T}(\boldsymbol{\theta})\|_1 \le \lambda_T(\boldsymbol{\theta})\|\boldsymbol{\theta}\|_1 \le \log T - 1$. This gives us an idea to generate vectors in $\Theta_{k,T}$. Our distribution $\pi_{k,T}$ is deduced from:

Algorithm 1: $\pi_{k,T}$ generation

input an effective dimension k, the number of observations T and $F_{k,T}$;
generate a random $\boldsymbol{\theta}$ uniformly on $s_k(1)$;
return $F_{k,T}(\boldsymbol{\theta})$

The distribution $\pi_{k,T}$ is lower bounded by the uniform distribution on $s_k(1)$.

Provided any $\gamma \ge 1$, let $T_* = \min\{T : d_T \ge d^\gamma, \log T \ge d^{1/2} 2^d\}$. Since $s_k(1) \subseteq B(\mathbf{0}, 2^k - 1)$ (see [19, Lemma 1]) and $k^{1/2}\|\boldsymbol{\theta}\| \ge \|\boldsymbol{\theta}\|_1$ for any $\boldsymbol{\theta} \in \mathbb{R}^k$, the constraint $\|\boldsymbol{\theta}\|_1 \le \log T - 1$ becomes redundant in $\Theta_{k,T}$ for $1 \le k \le d$ and $T \ge T_*$, i.e.

$\Theta_{1,T} = s_1(1) \times \{0\}^{d_T - 1}$ and $\Theta_{k,T} = s_k(1) \times \{0\}^{d_T - k} \backslash \Theta_{k-1,T}$ for $2 \leq k \leq d$. We define the sequence of Assumption (**P-4**) as $\boldsymbol{\theta}_T = \mathbf{0}$ for $T < T_*$ and $\boldsymbol{\theta}_T = \arg\inf_{\boldsymbol{\theta} \in \Theta_T} R(\boldsymbol{\theta})$ for $T \geq T_*$. Remark that the first d components of $\boldsymbol{\theta}_T$ are constant for $T \geq T_*$ (they correspond to the $\boldsymbol{\theta} \in \mathbb{R}^d$ generating the AR(d) process), and the last $d_T - d$ are zero. Let $\Delta_{1*} = 2 \log 2 - 1$. Then, we have for $T < T_*$ and all $\Delta \in [0, \Delta_{1*}]$

$$\pi_T \left[B(\boldsymbol{\theta}_T, \Delta) \cap \Theta_T \right] \geq c_{1,T} \pi_{1,T} \left[B(\mathbf{0}, \Delta) \cap s_1(1) \times \{0\}^{d_T - 1} \right] \geq \frac{c_1}{c^*} \Delta \ .$$

Furthermore, for $T \geq T_*$ and $\Delta \in [0, \Delta_{d*}]$

$$\pi_T \left[B(\boldsymbol{\theta}_T, \Delta) \cap \Theta_T \right] \geq c_{d,T} \pi_{d,T} \left[B(\boldsymbol{\theta}_T, \Delta) \cap s_d(1) \times \{0\}^{d_T - d} \right] \geq \frac{c_d}{2^{d^2} c^*} \Delta^d \ .$$

Assumption (**P-4**) is then fulfilled for any $\gamma \geq 1$ with

$$\begin{aligned}
\mathcal{C}_1 &= \max \left\{ 0, (R(0) - \inf_{\boldsymbol{\theta} \in \Theta_T} R(\boldsymbol{\theta})) T^{1/2} \log^{-3} T, 4 \leq T < T_* \right\} \\
\mathcal{C}_2 &= \min \left\{ 1, \frac{c_1}{c^*}, \frac{c_d}{2^{d^2} c^*} \right\} \\
\mathcal{C}_3 &= \min \{ 1, 4\Delta_{1*}, T_* \Delta_{d*} \} \ .
\end{aligned}$$

Let $q_{\eta,T}$ be the constant function 1, this means that the proposal has the same distribution π_T. Let us bound the ratio (11).

$$\begin{aligned}
\beta_{\eta,T}(X) = \inf_{\boldsymbol{\theta} \in \Theta_T} \frac{q_{\eta,T}(\boldsymbol{\theta})}{\pi_{\eta,T}(\boldsymbol{\theta} \,|X)} &= \inf_{\boldsymbol{\theta} \in \Theta_T} \frac{\sum_{k=1}^{d_T} c_{k,T} \int_{\Theta_{k,T}} \exp(-\eta r_T(z|X)) \, \pi_{k,T}(\mathrm{d}z)}{\exp(-\eta r_T(\boldsymbol{\theta} \,|X))} \\
&\geq \sum_{k=1}^{d_T} c_{k,T} \int_{\Theta_{k,T}} \exp(-\eta r_T(z|X)) \, \pi_{k,T}(\mathrm{d}z) > 0 \ .
\end{aligned} \tag{16}$$

Now note that

$$\left| x_t - f_{\boldsymbol{\theta}} \left((x_{t-i})_{i \geq 1} \right) \right| \leq |x_t| + \sum_{j=1}^{d(\boldsymbol{\theta})} \left| \theta_j \right| \left| x_{t-j} \right| \leq \log T \max_{j=0,\ldots,d(\boldsymbol{\theta})} \left| x_{t-j} \right| \ . \tag{17}$$

Plugging the bound (17) on (16) with $\eta = \eta_T$

$$\beta_{\eta_T,T}(\boldsymbol{x}) \geq \sum_{k=1}^{d_T} c_k \int_{\Theta_k} \exp(-\eta_T r_T(z|\boldsymbol{x})) \, \pi_k(\mathrm{d}z) \geq \exp\left(-\frac{T^{1/2}}{4} \max_{j=0,\ldots,d_T} \left| x_{t-j} \right| \right) \ ,$$

we deduce that

$$\frac{1}{\beta_{\eta_T,T}(\boldsymbol{x})} \leq \sum_{k=0}^{d_T} \exp\left(\frac{T^{1/2}\left|x_{t-j}\right|}{4}\right). \tag{18}$$

Taking (18) into account, setting $\gamma = 1$ (thus $d_T = \lfloor \log T \rfloor$), using Assumption (**P-3**), that $K = 1$ and applying the Cauchy-Schwarz inequality we get

$$\begin{aligned} A_{\eta_T,T} &= 3K \int_{\mathcal{X}^{\mathbb{Z}}} \frac{1}{\beta_{\eta_T,T}(\boldsymbol{x})} \int_{\mathcal{X}^{\mathbb{Z}}} \sup_{\boldsymbol{\theta}\in\Theta_T} \left| f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\hat{\boldsymbol{\theta}}_{\eta_T,T}(\boldsymbol{x})}(\boldsymbol{y}) \right| \pi_0(\mathrm{d}\boldsymbol{y})\, \pi_0(\mathrm{d}\boldsymbol{x}) \\ &\leq 3\left(d_T+1\right) d_T^{1/2} \pi_0\left[\exp\left(\frac{T^{1/2}\left|X_1\right|}{4}\right)\right] \pi_0\left[\left|X_1\right|\right] \sup_{\boldsymbol{\theta}\in\Theta_T} ||\boldsymbol{\theta}|| \\ &\leq 6\log^{3/2} T \pi_0\left[\exp\left(\frac{T^{1/2}\left|X_1\right|}{4}\right)\right] \pi_0\left[\left|X_1\right|\right]. \end{aligned}$$

As X_1 is centered and normally distributed of variance γ_0, $\pi_0\left[\left|X_1\right|\right] = \left(2\gamma_0/\pi\right)^{1/2}$ and $\pi_0[\exp(T^{1/2}\left|X_1\right|/4)] = \gamma_0 T^{1/2}\exp(\gamma_0 T/32)/4$.

From $n \geq M^*(T,\varepsilon) = 9\gamma_0^3 T^2 \exp\left(\gamma_0 T/16\right)/\left(2\pi\varepsilon^2\log^3 T\right)$ the result of Theorem 2 is reached. This bound of $M(T,\varepsilon)$ is prohibitive from a computational viewpoint. That is why we limit the number of iterations to a fixed n^*.

What we obtain from MCMC is $\bar{f}_{\eta_T,T,n}(\boldsymbol{y}\,|X) = \bar{\boldsymbol{\theta}}'_{\eta_T,T,n}(X)\boldsymbol{y}_{1:d_T}$ with $\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X) = \sum_{i=0}^{n-1}\boldsymbol{\theta}_{\eta_T,T,i}(X)/n$. Remark that $\bar{f}_{\eta_T,T,n}(\cdot\,|X) = f_{\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X)}$. The risk is expressed as

$$R\left(\bar{f}_{\eta_T,T,n}(\cdot\,|X)\right) = \frac{\left(2\left(\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X)-\boldsymbol{\theta}\right)'\Gamma(Y)\left(\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X)-\boldsymbol{\theta}\right)+2\sigma^2\right)^{1/2}}{\pi^{1/2}}.$$

5.2 Numerical Work

Consider 100 realisations of an autoregressive processes X simulated with the same $\boldsymbol{\theta} \in s_d(\delta)$ for $d = 8$ and $\delta = 3/4$ and with $\sigma = 1$. Let $\boldsymbol{c}^{(i)}$, $i = 1, 2$ the sequences defining two different priors in the model order:

1. $c_k^{(1)} = k^{-2}$, the sparsity is favoured,
2. $c_k^{(2)} = \mathrm{e}^{-k}$, the sparsity is strongly favoured.

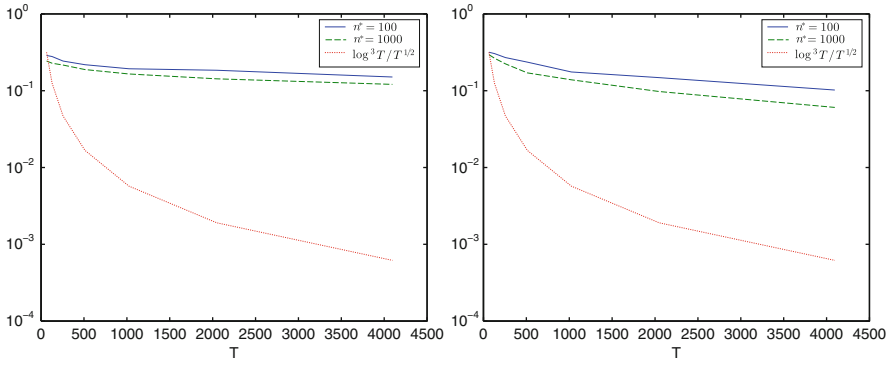


Fig. 1 The plots represent the 0.9-quantiles in data $R(\bar{\boldsymbol{\theta}}_{\eta_T,T,n^*}(X)) - (2/\pi)^{1/2}\sigma^2$ for $T = 32, 64, \ldots, 4{,}096$. The graph on the *left corresponds* to the order prior $c_k^{(1)} = k^{-2}$ while that on the *right corresponds* to $c_k^{(2)} = \mathrm{e}^{-k}$. The *solid curves* were plotted with $n^* = 100$, the *dashed ones* with $n^* = 1{,}000$ and as a reference, the *dotted curve* is proportional to $\log^3 T/T^{1/2}$

For each sequence $\boldsymbol{c}$ and for each value of $T \in \{2^j, j = 6, \ldots, 12\}$ we compute $\bar{\boldsymbol{\theta}}_{\eta_T,T,n^*}$, the MCMC approximation of the Gibbs estimator using Algorithm 2 with $\eta = \eta_T$.

Algorithm 2: Independent Hastings Sampler

input the sample $X_{1:T}$ of X, the prior $\boldsymbol{c}$, the learning rate η, the generators $\pi_{k,T}$ for $k = 1, \ldots, d_T$ and a maximum iterations number n^*;
initialization $\boldsymbol{\theta}_{\eta,T,0} = \mathbf{0}$;
for i=*1* **to** $n^* - 1$ **do**
  generate $k \in \{1, \ldots, d_T\}$ using the prior $\boldsymbol{c}$;
  generate $\boldsymbol{\theta}_{candidate} \sim \pi_{k,T}$;
  generate $U \sim \mathcal{U}(0,1)$;
  if $U \leq \alpha_{\eta,T,X}(\boldsymbol{\theta}_{\eta,T,i-1}, \boldsymbol{\theta}_{candidate})$ **then**
    $\boldsymbol{\theta}_{\eta,T,i} = \boldsymbol{\theta}_{candidate}$ **else**
      $\boldsymbol{\theta}_{\eta,T,i} = \boldsymbol{\theta}_{\eta,T,i-1}$;
return $\bar{\boldsymbol{\theta}}_{\eta,T,n^*}(X) = \sum_{i=0}^{n^*-1} \boldsymbol{\theta}_{\eta,T,k}(X)/n^*$.

The acceptance rate is computed as $\alpha_{\eta,T,X}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \exp(\eta r_T(\boldsymbol{\theta}_1|X) - \eta r_T(\boldsymbol{\theta}_2|X))$.

Algorithm 1 used by the distributions $\pi_{k,T}$ generates uniform random vectors on $s_k(1)$ by the method described in [6]. It relies in the Levinson-Durbin recursion algorithm. We also implemented the numerical improvements of [3].

Set $\varepsilon = 0.1$. Figure 1 displays the $(1-\varepsilon)$-quantiles in data $R(\bar{\boldsymbol{\theta}}_{\eta_T,T,n^*}(X)) - (2/\pi)^{1/2}\sigma^2$ for $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(2)}$ using different values of n^*.

Note that, for the proposed algorithm the prediction risk decreases very slowly when the number T of observations grows and the number of MCMC iterations remains constant. If $n^* = 1{,}000$ the decaying rate is faster than if $n^* = 100$ for smaller values of T. For $T \geq 2{,}000$ we observe that both rates are roughly the same in the logarithmic scale. This behaviour is similar in both cases presented in Fig. 1. As expected, the risk of the approximated predictor does not converge as $\log^3 T/T^{1/2}$.

6 Discussion

There are two sources of error in our method: prediction (of the exact Gibbs predictor) and approximation (using the MCMC). The first one decays when T grows and the obtained guarantees for the second one explode. We found a possibly pessimistic upper bound for $M(T, \epsilon)$. The exponential growing of this bound is the main weakness of our procedure. The use of a better adapted proposal in the MCMC algorithm needs to be investigated. The Metropolis Langevin Algorithm (see [4]) gives us an insight in this direction. However it is encouraging to see that, in the analysed practical case, the risk of $\bar{f}_{\eta_T,T,n^*}(\cdot\,|X)$ does not increase with T.

7 Technical Proofs

7.1 *Proof of Theorem 1*

The proof of Theorem 1 is based on the same tools used by [2] up to Lemma 3. For the sake of completeness we quote the essential ones.

We denote by $\mathcal{M}_+^1(F)$ the set of probability measures on the measurable space $(F, \mathcal{F})$. Let $\rho, \nu \in \mathcal{M}_+^1(F)$, $\mathcal{K}(\rho, \nu)$ stands for the Kullback-Leibler divergence of ν from ρ.

$$\mathcal{K}(\rho, \nu) = \begin{cases} \int \log \frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\boldsymbol{\theta})\, \rho(\mathrm{d}\boldsymbol{\theta}) & \text{, if } \rho \ll \nu \\ +\infty & \text{, otherwise .} \end{cases}$$

The first lemma can be found in [8, Equation 5.2.1].

Lemma 1 (Legendre transform of the Kullback divergence function) *Let $(F, \mathcal{F})$ be any measurable space. For any $\nu \in \mathcal{M}_+^1(F)$ and any measurable function $h : F \to \mathbb{R}$ such that $\nu[\exp(h)] < \infty$ we have,*

$$\nu[\exp(h)] = \exp\left(\sup_{\rho \in \mathcal{M}_+^1(F)} (\rho[h] - \mathcal{K}(\rho, \nu))\right) ,$$

with the convention $\infty - \infty = -\infty$. *Moreover, as soon as* h *is upper-bounded on the support of* ν, *the supremum with respect to* ρ *in the right-hand side is reached by the Gibbs measure* $\nu\{h\}$.

For a fixed $C > 0$, let $\tilde{\xi}_t^{(C)} = \max\{\min\{\xi_t, C\}, -C\}$. Consider $\tilde{X}_t = H(\tilde{\xi}_t^{(C)}, \tilde{\xi}_{t-1}^{(C)}, \ldots)$.

Denote $\tilde{X} = (\tilde{X}_t)_{t\in\mathbb{Z}}$ and by $\tilde{R}(\boldsymbol{\theta})$ and $\tilde{r}_T\left(\boldsymbol{\theta} \left| \tilde{X}\right.\right)$ the respective exact and empirical risks associated with $\tilde{X}$ in $\boldsymbol{\theta}$.

$$\tilde{R}(\boldsymbol{\theta}) = \mathbb{E}\left[\ell\left(\widehat{\tilde{X}}_t^{\boldsymbol{\theta}}, \tilde{X}_t\right)\right],$$

$$\tilde{r}_T\left(\boldsymbol{\theta} \left| \tilde{X}\right.\right) = \frac{1}{T - d(\boldsymbol{\theta})} \sum_{t=d(\boldsymbol{\theta})+1}^{T} \ell\left(\widehat{\tilde{X}}_t^{\boldsymbol{\theta}}, \tilde{X}_t\right),$$

where $\widehat{\tilde{X}}_t^{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}((\tilde{X}_{t-i})_{i\geq 1})$.

This thresholding is interesting because truncated CBS are weakly dependent processes (see [2, Section 4.2]).

A Hoeffding type inequality introduced in [20, Theorem 1] provides useful controls on the difference between empirical and exact risks of a truncated process.

Lemma 2 (Laplace transform of the risk) *Let* ℓ *be a loss function meeting Assumption* (***L***) *and* $X = (X_t)_{t\in\mathbb{Z}}$ *a process satisfying Assumption* (***M***). *For all* $T \geq 2$, *any* $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ *satisfying Assumption* (***P-1***), Θ_T *such that* d_T, *defined in (6), is at most* $T/2$, *any truncation level* $C > 0$, $\eta \geq 0$ *and* $\boldsymbol{\theta} \in \Theta_T$ *we have,*

$$\mathbb{E}\left[\exp\left(\eta\left(\tilde{R}(\boldsymbol{\theta}) - \tilde{r}_T\left(\boldsymbol{\theta} \left| \tilde{X}\right.\right)\right)\right)\right] \leq \exp\left(\frac{4\eta^2 k^2(T, C)}{T}\right), \tag{19}$$

and

$$\mathbb{E}\left[\exp\left(\eta\left(\tilde{r}_T\left(\boldsymbol{\theta} \left| \tilde{X}\right.\right) - \tilde{R}(\boldsymbol{\theta})\right)\right)\right] \leq \exp\left(\frac{4\eta^2 k^2(T, C)}{T}\right), \tag{20}$$

where $k(T, C) = 2^{1/2} CK(1 + L_T)\left(A_* + \tilde{A}_*\right)$. *The constants* $\tilde{A}_*$ *and* A_* *are defined in (4) and (5) respectively,* K *and* L_T *in Assumptions* (***L***) *and* (***P-1***) *respectively.*

The following lemma is a slight modification of [2, Lemma 6.5]. It links the two versions of the empirical risk: original and truncated.

Lemma 3 *Suppose that Assumption* (***L***) *holds for the loss function* ℓ, *Assumption* (***P-1***) *holds for* $X = (X_t)_{t\in\mathbb{Z}}$ *and Assumption* (***I***) *holds for the innovations with* $\zeta = A_*$; A_* *is defined in (5). For all* $T \geq 2$, *any* $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ *meeting Assumption* (***P-1***) *with* Θ_T *such that* d_T, *defined in (6), is at most* $T/2$, *any truncation*

level $C > 0$ and any $0 \leq \eta \leq T/4\,(1 + L_T)$ we have,

$$\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T\left(\boldsymbol{\theta}\,|X\right) - \tilde{r}_T\left(\boldsymbol{\theta}\,\middle|\tilde{X}\right)\right|\right)\right] \leq \exp\left(\eta\varphi\left(T, C, \eta\right)\right)\ ,$$

where

$$\varphi(T, C, \eta) = 2K(1 + L_T)\phi(A_*)\left(\frac{A_* C}{\exp\left(A_* C\right) - 1} + \eta\frac{4K(1 + L_T)}{T}\right)\ ,$$

*with K and L_T defined in Assumptions (**L**) and (**P-1**) respectively.*

Finally we present a result on the aggregated predictor defined in (7). The proof is partially inspired by that of [2, Theorem 3.2].

Lemma 4 *Let ℓ be a loss function such that Assumption (**L**) holds and let $X = (X_t)_{t\in\mathbb{Z}}$ a process satisfying Assumption (**M**) with probability distribution π_0. For each $T \geq 2$ let $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ be a set of predictors and $\pi_T \in \mathcal{M}_+^1\left(\Theta_T\right)$ any prior probability distribution on Θ_T. We build the predictor $\hat{f}_{\eta,T}\left(\cdot\,|X\right)$ following (7) with any $\eta > 0$. For any $\varepsilon > 0$ and any truncation level $C > 0$, with π_0-probability at least $1 - \varepsilon$ we have,*

$$\begin{aligned} R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\right) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left\{\rho\left[R\right] + \frac{2\mathcal{K}\left(\rho, \pi_T\right)}{\eta}\right\} + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\eta} \\ + \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right]\right) + \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\tilde{r}_T - \tilde{R}\right)\right)\right]\right) \\ + \frac{2}{\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T\left(\boldsymbol{\theta}\,|X\right) - \tilde{r}_T\left(\boldsymbol{\theta}\,\middle|\tilde{X}\right)\right|\right)\right]\right)\ . \end{aligned}$$

Proof We use Tonelli's theorem and Jensen's inequality with the convex function g to obtain an upper bound for $R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\right)$

$$\begin{aligned} R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\right) &= \int_{\mathcal{X}^{\mathbb{Z}}} g\left(\int_{\Theta_T}\left(f_{\boldsymbol{\theta}}\left((y_{t-i})_{i\geq 1}\right) - y_t\right)\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left(\mathrm{d}\boldsymbol{\theta}\right)\right)\pi_0\left(\mathrm{d}\mathbf{y}\right) \\ &\leq \int_{\mathcal{X}^{\mathbb{Z}}}\left[\int_{\Theta_T} g\left(f_{\boldsymbol{\theta}}\left((y_{t-i})_{i\geq 1}\right) - y_t\right)\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left(\mathrm{d}\boldsymbol{\theta}\right)\right]\pi_0\left(\mathrm{d}\mathbf{y}\right) \\ &= \int_{\Theta_T}\left[\int_{\mathcal{X}^{\mathbb{Z}}} g\left(f_{\boldsymbol{\theta}}\left((y_{t-i})_{i\geq 1}\right) - y_t\right)\pi_0\left(\mathbf{y}\right)\right]\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left(\mathrm{d}\boldsymbol{\theta}\right) = \pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left[R\right]\ . \end{aligned}$$

In the remainder of this proof we search for upper bounding $\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left[R\right]$.

First, we use the relationship:

$$R - r_T(\cdot|X) = \left(\tilde{R} - \tilde{r}_T\left(\cdot\left|\tilde{X}\right.\right)\right) + \left(R - \tilde{R}\right) - \left(r_T(\cdot|X) - \tilde{r}_T\left(\cdot\left|\tilde{X}\right.\right)\right) . \quad (21)$$

For the sake of simplicity and while it does not disrupt the clarity, we lighten the notation of r_T and $\tilde{r}_T$. We now suppose that in the place of $\boldsymbol{\theta}$ we have a random variable distributed as $\pi_T \in \mathcal{M}_+^1(\Theta_T)$. This is taken into account in the following expectations. The identity (21) and the Cauchy-Schwarz inequality lead to

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{\eta}{2}(R - r_T)\right)\right] &= \mathbb{E}\left[\exp\left(\frac{\eta}{2}\left(\tilde{R} - \tilde{r}_T\right)\right)\exp\left(\frac{\eta}{2}\left(\left(R - \tilde{R}\right) - \left(r_T - \tilde{r}_T\right)\right)\right)\right] \\
&\leq \left(\mathbb{E}\left[\exp\left(\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right]\mathbb{E}\left[\exp\left(\eta\left(\left(R - \tilde{R}\right) - \left(r_T - \tilde{r}_T\right)\right)\right)\right]\right)^{1/2} \\
\leq \left(\mathbb{E}\left[\exp\left(\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right]\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|\left(R - \tilde{R}\right)(\boldsymbol{\theta}) - \left(r_T - \tilde{r}_T\right)(\boldsymbol{\theta})\right|\right)\right]\right)^{1/2} .
\end{aligned} \quad (22)$$

Observe now that $R(\boldsymbol{\theta}) = \mathbb{E}[r_T(\boldsymbol{\theta}|X)]$ and $\tilde{R}(\boldsymbol{\theta}) = \mathbb{E}[\tilde{r}_T(\boldsymbol{\theta}|\tilde{X})]$. Jensen's inequality for the exponential function gives that

$$\begin{aligned}
\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|R(\boldsymbol{\theta}) - \tilde{R}(\boldsymbol{\theta})\right|\right) &\leq \exp\left(\eta\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}|X) - \tilde{r}_T\left(\boldsymbol{\theta}\left|\tilde{X}\right.\right)\right|\right]\right) \\
&\leq \mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}|X) - \tilde{r}_T\left(\boldsymbol{\theta}\left|\tilde{X}\right.\right)\right|\right)\right] .
\end{aligned} \quad (23)$$

From (23) we see that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|\left(R - \tilde{R}\right)(\boldsymbol{\theta}) - \left(r_T - \tilde{r}_T\right)(\boldsymbol{\theta})\right|\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|R(\boldsymbol{\theta}) - \tilde{R}(\boldsymbol{\theta})\right|\right)\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}|X) - \tilde{r}_T\left(\boldsymbol{\theta}\left|\tilde{X}\right.\right)\right|\right)\right] \\
&\leq \left(\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}|X) - \tilde{r}_T\left(\boldsymbol{\theta}\left|\tilde{X}\right.\right)\right|\right)\right]\right)^2 .
\end{aligned} \quad (24)$$

Combining (22) and (24) we obtain

$$\mathbb{E}\left[\exp\left(\frac{\eta}{2}\left(R - r_T\left(\cdot\,|X\right)\right)\right)\right] \leq \left(\mathbb{E}\left[\exp\left(\eta\left(\tilde{R} - \tilde{r}_T\right)\right)\right]\right)^{1/2} \mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T\left(\boldsymbol{\theta}\,|X\right) - \tilde{r}_T\left(\boldsymbol{\theta}\,|\tilde{X}\right)\right|\right)\right]. \tag{25}$$

Let $L_{\eta,T,C} = \log((\mathbb{E}[\exp(\eta(\tilde{R}-\tilde{r}_T))])^{1/2}\mathbb{E}[\exp(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}|r_T(\boldsymbol{\theta}\,|X) - \tilde{r}_T(\boldsymbol{\theta}\,|\tilde{X})|)])$. Remark that the left term of (25) is equal to the integral of the expression enclosed in brackets with respect to the measure $\pi_0 \times \pi_T$. Changing η by 2η and thanks to Lemma 1 we get

$$\pi_0\left[\exp\left(\sup_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left(\eta\rho[R - r_T\left(\cdot\,|X\right)] - \mathcal{K}\left(\rho,\pi_T\right)\right)\right)\right] \leq \exp\left(L_{2\eta,T,C}\right).$$

Markov's inequality implies that for all $\varepsilon > 0$, with π_0- probability at least $1-\varepsilon$

$$\sup_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left(\eta\rho\left[R - r_T\left(\cdot\,|X\right)\right] - \mathcal{K}\left(\rho,\pi_T\right)\right) - \log\left(\frac{1}{\varepsilon}\right) - L_{2\eta,T,C} \leq 0\,.$$

Hence, for any $\pi_T \in \mathcal{M}_+^1\left(\Theta_T\right)$ and $\eta > 0$, with π_0- probability at least $1-\varepsilon$, for all $\rho \in \mathcal{M}_+^1\left(\Theta_T\right)$

$$\rho\left[R - r_T\left(\cdot\,|X\right)\right] - \frac{1}{\eta}\mathcal{K}\left(\rho,\pi_T\right) - \frac{1}{\eta}\log\left(\frac{1}{\varepsilon}\right) - \frac{L_{2\eta,T,C}}{\eta} \leq 0\,. \tag{26}$$

By setting $\rho = \pi_T\{-\eta r_T\left(\cdot\,|X\right)\}$ and relying on Lemma 1, we have

$$\begin{aligned}\mathcal{K}\left(\pi_T\left\{-\eta r_T\right\},\pi_T\right) &= \pi_T\left\{-\eta r_T\right\}\left[\log\frac{\mathrm{d}\pi_T\left\{-\eta r_T\right\}}{\mathrm{d}\pi_T}\right] = \pi_T\left\{-\eta r_T\right\}\left[\log\frac{\exp\left(-\eta r_T\right)}{\pi_T\left[\exp\left(-\eta r_T\right)\right]}\right]\\ &= \pi_T\left\{-\eta r_T\right\}\left[-\eta r_T\right] - \log\left(\pi_T\left[\exp\left(-\eta r_T\right)\right]\right)\\ &= \pi_T\left\{-\eta r_T\right\}\left[-\eta r_T\right] + \inf_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left\{\rho\left[\eta r_T\right] + \mathcal{K}\left(\rho,\pi_T\right)\right\}\end{aligned}$$

Using (26) with $\rho = \pi_T\{-\eta r_T\left(\cdot\,|X\right)\}$ it follows that, with π_0- probability at least $1-\varepsilon$,

$$\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left[R\right] \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left\{\rho\left[r_T\left(\cdot\,|X\right)\right] + \frac{\mathcal{K}\left(\rho,\pi_T\right)}{\eta}\right\} + \frac{\log\left(\frac{1}{\varepsilon}\right)}{\eta} + \frac{L_{2\eta,T,C}}{\eta}\,.$$

To upper bound $\rho[r_T(\cdot|X)]$ we use an upper bond on $\rho\left[r_T(\cdot|X) - R\right]$. We obtain an inequality similar to (26) with $\rho\left[R - r_T(\cdot|X)\right]$ replaced by $\rho\left[r_T(\cdot|X) - R\right]$ and $L_{\eta,T,C}$ replaced by $L'_{\eta,T,C} = \log((\mathbb{E}[\exp(\eta(\tilde{r}_T - \tilde{R}))])^{1/2}\mathbb{E}[\exp(\eta \sup_{\boldsymbol{\theta}\in\Theta_T} |r_T(\boldsymbol{\theta}|X) - \tilde{r}_T(\boldsymbol{\theta}|\tilde{X})|)])$. This provides us another inequality satisfied with π_0- probability at least $1-\varepsilon$. To obtain a π_0- probability of the intersection larger than $1-\varepsilon$ we apply previous computations with $\varepsilon/2$ instead of ε and hence,

$$\pi_T\left\{-\eta r_T\left(\cdot|X\right)\right\}[R] \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left\{\rho\left[R\right] + \frac{2\mathcal{K}\left(\rho,\pi_T\right)}{\eta}\right\} + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\eta}$$
$$+ \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\tilde{R}-\tilde{r}_T\right)\right)\right]\right) + \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\tilde{r}_T-\tilde{R}\right)\right)\right]\right)$$
$$+ \frac{2}{\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T\left(\boldsymbol{\theta}\,|X\right) - \tilde{r}_T\left(\boldsymbol{\theta}\,|\tilde{X}\right)\right|\right)\right]\right).$$

We can now proof Theorem 1.

Proof Let $\pi_{0,C}$ denote the distribution on $\mathcal{X}^{\mathbb{Z}} \times \mathcal{X}^{\mathbb{Z}}$ of the couple $(X, \tilde{X})$. Fubini's theorem and (19) of Lemma 2 imply that

$$\mathbb{E}\left[\exp\left(2\eta\left(\tilde{R}-\tilde{r}_T\right)\right)\right] = \pi_{0,C}\times\pi_T\left[\exp\left(2\eta\left(\tilde{R}-\tilde{r}_T\right)\right)\right] = \pi_T\times\pi_{0,C}\left[\exp\left(2\eta\left(\tilde{R}-\tilde{r}_T\right)\right)\right]$$
$$\leq \exp\left(\frac{16\eta^2k^2(T,C)}{T}\right). \tag{27}$$

Using (20), we analogously get

$$\mathbb{E}\left[\exp\left(2\eta\left(\tilde{r}_T-\tilde{R}\right)\right)\right] \leq \exp\left(\frac{16\eta^2k^2(T,C)}{T}\right). \tag{28}$$

Consider the set of probability measures $\{\rho_{\boldsymbol{\theta}_T,\Delta}, T \geq 2, 0 \leq \Delta \leq \Delta_T\} \subset \mathcal{M}_+^1\left(\Theta_T\right)$, where $\boldsymbol{\theta}_T$ is the parameter defined by Assumption (**P-4**) and $\rho_{\boldsymbol{\theta}_T,\Delta}\left(\boldsymbol{\theta}\right) \propto \pi_T\left(\boldsymbol{\theta}\right)\mathbb{1}_{B(\boldsymbol{\theta}_T,\Delta)\cap\Theta_T}\left(\boldsymbol{\theta}\right)$. Lemma 4, together with Lemma 3, (27) and (28) guarantee that for all $0 < \eta \leq T/8\left(1+L_T\right)$

$$R\left(\hat{f}_{\eta,T}\left(\cdot|X\right)\right) \leq \inf_{0\leq\Delta\leq\Delta_T}\left\{\rho_{\boldsymbol{\theta}_T,\Delta}\left[R\right] + \frac{2\mathcal{K}\left(\rho_{\boldsymbol{\theta}_T,\Delta},\pi_T\right)}{\eta}\right\} + \frac{16\eta k^2(T,C)}{T} + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\eta} +$$
$$4\varphi(T,C,2\eta). \tag{29}$$

Thanks to Assumptions (**L**) and (**P-3**), for any $T \geq 2$ and $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_T, \Delta)$

$$R(\boldsymbol{\theta}) - R(\boldsymbol{\theta}_T) \leq K\pi_0 \left[\left| \left| f_{\boldsymbol{\theta}} \left((Y_{t-i})_{i\geq 1} \right) - f_{\boldsymbol{\theta}_T} \left((Y_{t-i})_{i\geq 1} \right) \right| \right| \right] \leq K\mathcal{D} d_T^{1/2} \Delta \ . \tag{30}$$

For $T \geq 4$ Assumption (**P-4**) gives

$$\mathcal{K}(\rho_{\boldsymbol{\theta}_T,\Delta}, \pi_T) = \log \left(\frac{1}{\pi_T \left[B(\boldsymbol{\theta}_T, \Delta) \cap \Theta_T \right]} \right) \leq -n_T^{1/\gamma} \log(\Delta) - \log(\mathcal{C}_2) \ . \tag{31}$$

Plugging (30) and (31) into (29) and using again Assumption (**P-4**)

$$\begin{aligned} R\left(\hat{f}_{\eta,T}(\cdot|X) \right) \leq R(\boldsymbol{\theta}_T) + \inf_{0\leq\Delta\leq\Delta_T} \left\{ \mathcal{E}_1 d_T^{1/2} \Delta - \frac{2n_T^{1/\gamma} \log(\Delta)}{\eta} \right\} + \frac{\mathcal{E}_2 \eta (1+L_T)^2 C^2}{T} \\ + \frac{\mathcal{E}_3 (1+L_T) C}{\exp(A_* C) - 1} + \frac{2\log\left(\frac{2}{\varepsilon}\right) - 2\log(\mathcal{C}_2)}{\eta} + \frac{\mathcal{E}_4 (1+L_T)^2 \eta}{T} \end{aligned} \tag{32}$$

where $\mathcal{E}_1 = K\mathcal{D}$, $\mathcal{E}_2 = 32K^2 \left(A_* + \tilde{A}_* \right)^2$, $\mathcal{E}_3 = 8K\phi(A_*)A_*$ and $\mathcal{E}_4 = 32K^2\phi(A_*)$.

We upper bound d_T by $T/2$, n_T by $\log^\gamma T$ and substitute $\Delta_T = \mathcal{C}_3/T$. Since it is difficult to minimize the right term of (32) with respect to η and C at the same time, we evaluate them in certain values to obtain a convenient upper bound.

At a fixed ε, the convergence rate of $[2\log(2/\varepsilon) - 2\log(\mathcal{C}_2)]/\eta + \mathcal{E}_4 (1+L_T)^2 \eta/T$ is at best $\log T/T^{1/2}$, and we get it doing $\eta \propto T^{1/2}/\log T$. As $\eta \leq T/8(1+L_T)$ we set $\eta = \eta_T = T^{1/2}/(4\log T)$.

The order of the already chosen terms is $\log^3 T/T^{1/2}$, doing $C = \log T/A_*$ we preserve it. Taking into account that $R(\boldsymbol{\theta}_T) \leq \inf_{\boldsymbol{\theta}\in\Theta_T} R(\boldsymbol{\theta}) + \mathcal{C}_1 \log^3 T/T^{1/2}$ the result follows.

7.2 *Proof of Proposition 1*

Considering that Assumption (**L**) holds we get

$$\left| R\left(\bar{f}_{\eta,T,n}(\cdot|X) \right) - R\left(\hat{f}_{\eta,T}(\cdot|X) \right) \right| \leq K \int_{\mathcal{X}^{\mathbb{Z}}} \left| \bar{f}_{\eta,T,n}(\mathbf{y}|X) - \hat{f}_{\eta,T}(\mathbf{y}|X) \right| \pi_0(\mathrm{d}\mathbf{y})$$

Observe that the last expression depends on $X_{1:T}$ and $\Phi_{\eta,T}(X)$. We bound the expectation to infer a bound in probability.

Tonelli's theorem and Jensen's inequality lead to

$$\nu_{\eta,T}\left[\left|R\left(\bar{f}_{\eta,T,n}\left(\cdot\,|X\right)\right)-R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\right)\right|\right]\leq$$
$$K\int\limits_{\mathcal{X}^{\mathbb{Z}}}\int\limits_{\mathcal{X}^{\mathbb{Z}}}\left(\int\limits_{\Theta_T^{\mathbb{N}}}\left|\bar{f}_{\eta,T,n}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right|^2\mu_{\eta,T}\left(\mathrm{d}\boldsymbol{\phi}\,|\boldsymbol{x}\right)\right)^{1/2}\pi_0\left(\mathrm{d}\boldsymbol{y}\right)\pi_0\left(\mathrm{d}\boldsymbol{x}\right)\,. \tag{33}$$

We are then interested in upper bounding the expression under the square root. To that end, we use [16, Theorem 3.1] which implies that for any $\boldsymbol{x}$

$$\int\limits_{\Theta_T^{\mathbb{N}}}\left|\bar{f}_{\eta,T,n}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right|^2\mu_{\eta,T}\left(\mathrm{d}\boldsymbol{\phi}\,|\boldsymbol{x}\right)\leq$$
$$\sup_{\boldsymbol{\theta}\in\Theta_T}\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{y}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right)^2\left(\frac{4}{\beta_{\eta,T}\left(\boldsymbol{x}\right)}-3\right)\left(\frac{1}{n}+\frac{2}{n^2\beta_{\eta,T}\left(\boldsymbol{x}\right)}\right)\,.$$

Plugging this on (33), using that $n\geq 1$ and that

$$\left(\left(4-3\beta_{\eta,T}\left(\boldsymbol{x}\right)\right)\left(2+\beta_{\eta,T}\left(\boldsymbol{x}\right)\right)\right)^{1/2}\leq 3\,,$$

we obtain the following

$$\nu_{\eta,T}\left[\left|R\left(\bar{f}_{\eta,T,n}\left(\cdot\,|X\right)\right)-R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\right)\right|\right]\leq$$
$$\frac{3K}{n^{1/2}}\int\limits_{\mathcal{X}^{\mathbb{Z}}}\frac{1}{\beta_{\eta,T}\left(\boldsymbol{x}\right)}\int\limits_{\mathcal{X}^{\mathbb{Z}}}\sup_{\boldsymbol{\theta}\in\Theta_T}\left|f_{\boldsymbol{\theta}}\left(\boldsymbol{y}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right|\pi_0\left(\mathrm{d}\boldsymbol{y}\right)\pi_0\left(\mathrm{d}\boldsymbol{x}\right)\,.$$

The result follows from Markov's inequality.

Acknowledgements The author is specially thankful to François Roueff, Christophe Giraud, Peter Weyer-Brown and the two referees for their extremely careful readings and highly pertinent remarks which substantially improved the paper. This work has been partially supported by the Conseil régional d'Île-de-France under a doctoral allowance of its program Réseau de Recherche Doctoral en Mathématiques de l'Île de France (RDM-IdF) for the period 2012–2015 and by the Labex LMH (ANR-11-IDEX-003-02).

References

1. Alquier, P., & Li, X. (2012). Prediction of quantiles by statistical learning and application to GDP forecasting. In J.-G. Ganascia, P. Lenca, & J.-M. Petit (Eds.), *Discovery science* (Volume 7569 of Lecture notes in computer science, pp. 22–36). Berlin/Heidelberg: Springer.
2. Alquier, P., & Wintenberger, O. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli, 18*(3), 883–913.
3. Andrieu, C., & Doucet, A. (1999). An improved method for uniform simulation of stable minimum phase real ARMA (p,q) processes. *IEEE Signal Processing Letters, 6*(6), 142–144.
4. Atchadé, Y. F. (2006). An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in Applied Probability, 8*(2), 235–254.
5. Audibert, J.-Y. (2004). PAC-bayesian statistical learning theory. PhD thesis, Université Pierre et Marie Curie-Paris VI.
6. Beadle, E. R., & Djurić, P. M. (1999). Uniform random parameter generation of stable minimum-phase real ARMA (p,q) processes. *IEEE Signal Processing Letters, 4*(9), 259–261.
7. Brockwell, P. J., & Davis, R. A. (2006). *Time series: Theory and methods* (Springer series in statistics). New York: Springer. Reprint of the second (1991) edition.
8. Catoni, O. (2004). *Statistical learning theory and stochastic optimization* (Volume 1851 of Lecture notes in mathematics). Berlin: Springer. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, 8–25 July 2001.
9. Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge: Cambridge University Press.
10. Coulon-Prieur, C., & Doukhan, P. (2000). A triangular central limit theorem under a new weak dependence condition. *Statistics and Probability Letters, 47*(1), 61–68.
11. Dalalyan, A. S., & Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp PAC-bayesian bounds and sparsity. *Machine Learning, 72*(1–2), 39–61.
12. Dedecker, J., Doukhan, P., Lang, G., León R, J. R., Louhichi, S., & Prieur, C. (2007). *Weak dependence: With examples and applications* (Volume 190 of Lecture notes in statistics). New York: Springer.
13. Dedecker, J., & Prieur, C. (2005). New dependence coefficients. Examples and applications to statistics. *Probability Theory and Related Fields, 132*(2), 203–236.
14. Künsch, H. R. (1995). A note on causal solutions for locally stationary AR-processes. Note from ETH Zürich, available on line here: ftp://ftp.stat.math.ethz.ch/U/hkuensch/localstat-ar.pdf.
15. Łatuszyński, K., Miasojedow, B., & Niemiro, W. (2013). Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli, 19*, 2033–2066.
16. Łatuszyński, K., & Niemiro, W. (2011). Rigorous confidence bounds for MCMC under a geometric drift condition. *Journal of Complexity, 27*(1), 23–38.
17. Leung, G., & Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory, 52*(8), 3396–3410.
18. Mengersen, K. L., & Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics, 24*(1), 101–121.
19. Moulines, E., Priouret, P., & Roueff, F. (2005). On recursive estimation for time varying autoregressive processes. *The Annals of Statistics, 33*(6), 2610–2654.
20. Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Academie des Sciences Paris Series I Mathematics, 330*(10), 905–908.
21. Roberts, G. O., & Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys, 1*, 20–71.

Space-Time Trajectories of Wind Power Generation: Parametrized Precision Matrices Under a Gaussian Copula Approach

Julija Tastu, Pierre Pinson, and Henrik Madsen

Abstract Emphasis is placed on generating space-time trajectories of wind power generation, consisting of paths sampled from high-dimensional joint predictive densities, describing wind power generation at a number of contiguous locations and successive lead times. A modelling approach taking advantage of the sparsity of precision matrices is introduced for the description of the underlying space-time dependence structure. The proposed parametrization of the dependence structure accounts for important process characteristics such as lead-time-dependent conditional precisions and direction-dependent cross-correlations. Estimation is performed in a maximum likelihood framework. Based on a test case application in Denmark, with spatial dependencies over 15 areas and temporal ones for 43 hourly lead times (hence, for a dimension of $n = 645$), it is shown that accounting for space-time effects is crucial for generating skilful trajectories.

1 Introduction

The large-scale integration of wind energy into power systems and electricity markets induces operational and management challenges owing to the stochastic nature of the wind itself, with its variability and limited predictability [1]. Forecasting of wind power generation, at various spatial and temporal scales, is generally seen as a crucial input to the decision-making problems involved [23]. An overview of the history of wind power forecasting, though mainly focused on deterministic approaches, as well as an extensive review of the state of the art in that field,

J. Tastu (✉)
Siemens Wind Power, Borupvang 9, 2750 Ballerup, Danmark
e-mail: julija.tastu@siemens.com

P. Pinson
Technical University of Denmark, Elektrovej 325-058, 2800 Kgs. Lyngby, Danmark
e-mail: ppin@dtu.dk

H. Madsen
Technical University of Denmark, Matematiktorvet 322-218, 2800 Kgs. Lyngby, Danmark
e-mail: hmad@dtu.dk

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_14

are given in [10] and [14], respectively. In parallel, an overview of the current forecasting challenges can be found in [33].

Owing to the complexity of the various decision-making problems at hand, it is preferable that the forecasts provide decision-makers not only with an expected value for future power generation, but also with associated estimates of prediction uncertainty. This calls for probabilistic, rather than point (in the sense of single-valued) forecasting [16]. Example applications of probabilistic forecasting include the design of market offering strategies [6], economic load dispatch and stochastic unit commitment [7, 32], optimal operation of wind-storage systems [9], contingency reserve quantification [4] and assessment of power systems operating costs induced by wind energy [30].

Probabilistic forecasts of wind power generation are commonly generated for each site and lead time of interest, individually. They do not inform about the interdependence structure between potential forecast errors, both in space and in time. Actually, the idea of addressing each site separately can be motivated by the fact that the power curves for the conversion of meteorological variables to power are given by complex non-linear functions of meteorological conditions, number and type of wind turbines, their layout, topographical characteristics, see [22] for instance. Wind power dynamics are generally so site-specific that it is hard to issue high-quality probabilistic forecasts for a large number of sites simultaneously. Similarly, a common practice is to issue power forecasts in the form of marginal predictive densities for each lead time individually, rather than addressing the joint multivariate distribution. The resulting set of marginal predictive densities at N sites and K lead times is a suboptimal input for a substantial share of decision-making problems related to power systems operations and electricity markets, e.g., due to power flows on the network or to inter-temporal constraints for conventional power units to be scheduled. A full picture of the space-time characteristics of the stochastic process is there necessary.

Having a set of marginal distributions for a number of random variables, their joint density can be fully characterized using a copula approach. One important feature of copulas is that they can be used to model dependency between random variables independently of their marginal distribution functions. This is important since, as mentioned previously, modelling wind power generation at individual sites while targeting a specific lead time is already a difficult task. It is thus an advantage to decouple the problem of estimating marginal predictive densities from that related to the space-time dependence structure. Copulas have been widely used in many fields for modelling dependencies among sets of random variables, including a number of problems related to wind power. As an example in [5], predictive densities for wind power generation were built by modelling the relation between wind speed and related power output using copulas. In [31], copulas were employed to estimate system net load distribution when accounting for the dependence structure between wind at different locations, at the level of a country, and its relation to the overall electric consumption. In [20], a copula-based approach was similarly proposed for modelling the spatial dependence in wind speed over the whole European area.

The proposal of issuing trajectories of wind power generation based on predictive marginal densities and a model of their dependence structure was originally described in [36], where the authors focused on a single wind farm, hence considering temporal dependencies only. Based on a Gaussian copula assumption, this temporal dependence was fully specified by an empirical covariance structure to be tracked adaptively and recursively, as in an exponential smoothing framework. Time trajectories of wind power production were then issued by sampling from these multivariate predictive densities. More general parametric and nonparametric approaches were subsequently described in [34], with the main aim of discussing verification of time trajectories. Furthermore, Ref. [26] concentrated on wind power generation at a pair of sites and considered different functional forms of copulas for their dependence, given the lead time. The present study goes along similar lines, even though looking here at the full space-time picture: joint predictive densities of wind power output (and eventually, trajectories) are to be issued based on a set of marginal predictive densities already available for all sites and lead times. The problem then boils down to specifying a model for the dependence structure and to estimating its parameters. Under a Gaussian copula assumption, the modelling approach we propose takes advantage of the sparsity of precision matrices permitting to describe the underlying space-time process. A suitable parametrization of the precision matrix is proposed, hence yielding a more tractable approach even in high dimensions. This proposal goes beyond the conventional assumptions of homogeneous stationary Gaussian Random fields, since the proposed parametrization accounts for the boundary points and considers non-constant conditional variances and direction-dependent conditional correlations.

The paper is structured as following. Section 2 introduces the data set used in the study. The methodology is described in Sect. 3. It consists of some preliminaries and definitions, a short introduction to copula modelling and explanation on how precision matrices relate to the Gaussian copula approach. Further, Sect. 4 presents the proposed parametrization of the dependence structure. The estimation method is discussed in Sect. 5, while the empirical results are given in Sect. 6. The paper ends in Sect. 7 with a set of concluding remarks and perspectives regarding future work.

2 Data

The case study is for western Denmark, covering the Jutland peninsula and the island of Funen, with a nominal capacity close to 2.5 GW. This corresponds to approximately 70 % of the entire wind power capacity installed in Denmark at the time. Even though this nominal capacity regularly evolves due to commissioning and decommissioning of turbines, as well as maintenance operations, it stayed very close to this level over the period considered. Besides the significant share of wind generation, one of the reasons for concentrating on Denmark relates to its climate and terrain characteristics. The territory is small enough for the incoming weather fronts to affect all of its parts. In addition, the terrain is smooth, therefore passing

Fig. 1 Geographical locations of the $N = 15$ control areas of Energinet.dk, the system operator in Denmark, for a nominal capacity close to 2.5 GW

weather fronts do not meet such big obstacles as mountains when propagating over the country. These aspects make the test case an ideal candidate for understanding space-time effects before moving to more complex cases.

The western Denmark area is divided by the system operator, Energinet.dk, into $N = 15$ control zones, as depicted in Fig. 1. For all of these areas, power measurements were made available at an hourly resolution. All measurements and related forecasts were normalized by the nominal capacity of the control area they relate to. Consequently, they are expressed in percentage of nominal capacity, generally denoted by P_n. Point forecasts of wind power generation with an hourly resolution and lead times up to $K = 43$ h were produced with the Wind Power Prediction Tool (WPPT) [29]. This corresponds to the most important lead times, today, for power systems operations in an electricity market environment. The

update frequency of the forecasts is hourly. Marginal predictive densities were generated in a nonparametric framework, for all control zones and hourly lead times up to K hours ahead, based on the adaptive resampling approach described in [35]. It comprises one of the state-of-the-art approaches to generating nonparametric predictive densities of wind power generation, in a fashion similar to adaptive quantile regression [37]. These predictive densities are fully characterized by a set of quantile forecasts with varying nominal levels α, $\alpha \in \{0.05, 0.1, \cdots, 0.95\}$. Related predictive cumulative distribution functions are obtained by linearly interpolating through these sets of quantile forecasts.

The available data covers a period from 1 January 2006 to 24 October 2007. For the purpose of the modelling and forecasting study, this dataset was divided into two subsets. The first one covers a period of 11 months, i.e., from 1 January 2006 to 30 November 2006 (8,016 forecast series), to be used for the data analysis, model building and estimation. The second subset, covering the period between 1 December 2006 and 24 October 2007 (7,872 forecast series), is considered as an evaluation set for the out-of-sample evaluation of the space-time trajectories to be generated, and for comparison with the alternative approaches considered.

3 Methodology

The objective of the methodology introduced here is to generate joint predictive densities describing wind power generation at a number of contiguous locations and for a number of successive lead times, independently from the approach used to originally generate the individual predictive densities. These marginal densities are linked together through a (Gaussian) copula function, for which a parametrization of the precision matrix permits to capture the underlying space-time covariance structure. Our proposal methodology can hence be seen as a two-stage approach to the modelling of joint predictive densities, by first obtaining relevant marginal predictive densities (here, part of the available data described previously), and then estimating the relevant parameters of the chosen copula. Similar approaches are first instance considered in econometrics applications [19].

3.1 Preliminaries and Definitions

Let us first describe the general setup for this forecasting problem. At every time step t, one aims at predicting wind power generation for future times $t+1,\ t+2, \cdots, t+K$ at N contiguous locations. Seeing wind power generation as a stochastic process, there are in total $n = NK$ random variables of interest, denoted in the following by $Y_{t,1}, Y_{t,2}, \cdots, Y_{t,n}$, which we aim at jointly describing given the information available up to time t. For instance here, with 15 zones and 43 lead times, one has $n = 645$. The enumeration is such that $Y_{t,1}, \cdots, Y_{t,K}$ represent wind power

generation at the first location for the lead times $1, \cdots, K$, then $Y_{t,T+1}, \cdots, Y_{t,2T}$ represent wind power generation at the second location for lead times $1, \cdots, K$, and so on. Uppercase letters are used for random variables, while lowercase letters denote the corresponding observations. Bold font is used to emphasize vectors and matrices. For example, $\mathbf{y}_t = [y_{t,1}, y_{t,2}, \cdots, y_{t,n}]^\top$ stands for the realization of $\mathbf{Y}_t$. This translates to generally seeing wind power generation as a vector-valued stochastic process, instead of a spatio-temporal one. This is more for the sake of employing simpler notations, even though in the following sections the space-time structure will be accounted for, by identifying the spatial and temporal neighbourhood of each random variable composing this vector-valued stochastic process.

The aim of the forecaster is to issue a multivariate predictive distribution F_t, describing the random vector $\mathbf{Y}_t = [Y_{t,1}, Y_{t,2}, \cdots, Y_{t,n}]^\top$, conditional to the information available up to time t,

$$F_t(y_1, y_2, \cdots, y_n) = P(Y_{t,1} \leq y_1, Y_{t,2} \leq y_2, \cdots, Y_{t,n} \leq y_n) \,. \tag{1}$$

Proposing a functional form for F_t directly implies a simultaneous description of both the marginal densities as well as of the space-time interdependence structure. They should account for the non-Gaussian and bounded nature of wind power generation, as well as for the complex dynamics of wind power variability. Unfortunately, there is no obvious distribution function which could address these required aspects altogether. Employing a copula-based approach appears to be an appealing solution since allowing decomposing the problem of estimating F_t into two parts.

First, marginal predictive densities, $F_{t,i} = P(Y_{t,i} \leq y_i)$, $i = 1, 2, \cdots, n$, describing wind power generation at each location and for each lead time individually, can be obtained. In contrast to joint multivariate predictive densities, for which very limited literature exist, the case of issuing marginal ones only is increasingly considered, in both parametric and nonparametric framework. Thus, at this point, the forecaster clearly should take advantage of the state-of-the-art methods available for probabilistic wind power forecasting, while concentrating on an appropriate description of the dependence structure.

Subsequently, the marginal predictive densities can be linked together in order to obtain F_t using a copula function. Mathematically the foundation of copulas is given by Sklar's theorem [42], which states that, for any multivariate cumulative distribution function F_t with marginals $F_{t,1}$, $F_{t,2}$,...,$F_{t,n}$ there exists a copula function C such that

$$F_t(y_1, y_2, \cdots, y_n) = C\left(F_{t,1}(y_1), F_{t,2}(y_2), \cdots, F_{t,n}(y_n)\right) \,. \tag{2}$$

In the case where the joint distribution to be modelled involves continuous random variables only, as for wind power generation, the copula C is unique.

3.2 Copulas for Wind Power Data

Several functional forms of copulas have been considered for wind power data. Namely, in Ref. [36] the authors advocate that a Gaussian copula is an adequate choice when generating joint predictive densities, for a single location and a set of successive lead times. In parallel in Ref. [26], different copula functions were compared for the modelling of the dependence between wind power generation at two sites, for a given lead time. The results showed that a Gumbel copula performed best, even though Gaussian and Frank copulas could also fit the data fairly adequately.

When moving to higher dimensions, the construction of Archimedean copulas (e.g., Gumbel) becomes complex. For instance, a traditional approach for constructing the n-variate Gumbel copula requires the nth order derivative of the inverse of the process generating function. Even considering explicit formulas for those derivatives given in Ref. [21], the complexity remains high compared to a Gaussian copula approach. Moreover, Ref. [11] showed that in higher dimensions Gaussian copulas outperformed their Gumbel's counterparts, certainly also owing to the much larger potential number of parameters of Gaussian copulas compared to the single parameter the Gumbel ones. Note that these works and results should be interpreted with care as they might depend upon site characteristics, as well as upon the type of marginal predictive densities employed as input.

These works hint at the fact that a Gaussian copula could be deemed appropriate for describing spatial and temporal dependencies present in wind power data. However, these works have not considered spatio-temporal dependencies. Consequently, in a first step, a preliminary data examination is carried out in order to verify whether or not employing a Gaussian copula could be consistent with the space-time dependence structure observed.

As an example here, consider $Y_{t,(5\times43)+5}$ and $Y_{t,(4\times43)+4}$ representing wind power generation at zone 6 and lead time $t + 5$, and wind power generation at zone 5 at lead time $t + 4$, respectively. The dependence between the random variables $Y_{t,(5\times43)+5}$ and $Y_{t,(4\times43)+4}$ can be graphically illustrated by focusing on the ranks of the uniform variables $F_{t,(5\times43)+5}(Y_{t,(5\times43)+5})$ and $F_{t,(4\times43)+4}(Y_{t,(4\times43)+4})$, for all predictive densities and corresponding realizations available over the first data subset from 1 January 2006 to 30 November 2006.

The scatterplot of the corresponding ranks characterizes the dependence structure between $Y_{t,(5\times43)+5}$ and $Y_{t,(4\times43)+4}$, while the overlaying contour plot represents the so-called empirical copula [13]. This empirical copula is then compared to what would be the corresponding Gaussian copula, as illustrated in Fig. 2. Both patterns are very similar, thus indicating that the Gaussian copula could be seen suitable for describing the spatio-temporal dependence structure. The results obtained while considering different pairs of variables were all deemed qualitatively similar. Obviously, such a visual comparison does not guarantee that employing a Gaussian copula is the best choice for modelling the dependence structure, while different fit evaluation criteria could be used if really aiming to find an optimal copula (as

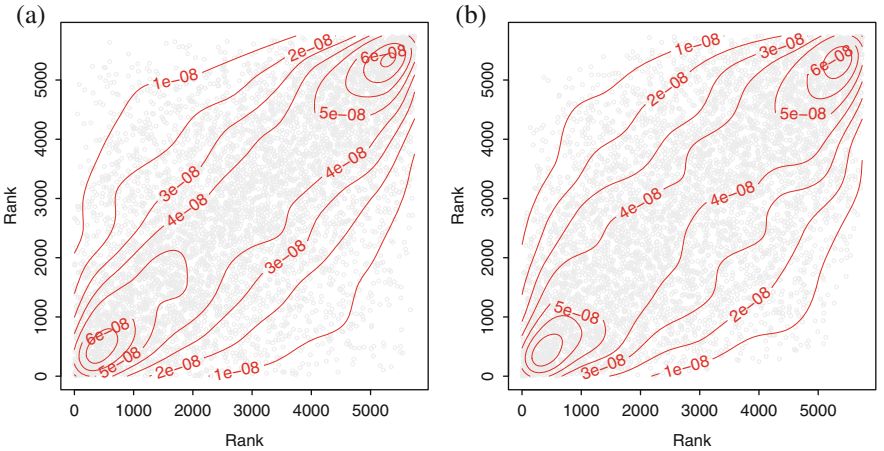


Fig. 2 Comparison of empirical and Gaussian copula for the dependence between $Y_{t,(5\times43)+5}$ and $Y_{t,(4\times43)+4}$. (**a**) A scatterplot with contour overlay for the ranks of $F_{t,(5\times43)+5}(y_{t,(5\times43)+5})$ and $F_{t,(4\times43)+4}(y_{t,(4\times43)+4})$ (empirical copula). In parallel, (**b**) depicts a scatterplot with contour overlay for the simulated bivariate Gaussian process having the same rank correlation

in [11, 26]). Here it is our choice to focus on Gaussian copulas only, owing to the resulting opportunities given in terms of dependence structure modelling with precision matrices only.

3.3 Gaussian Copula

A Gaussian copula is given by

$$C\left(F_{t,1}(y_1), \cdots, F_{t,n}(y_n)\right) = \Phi_{\boldsymbol{\Sigma}}\left(\Phi^{-1}(F_{t,1}(y_1)), \cdots, \Phi^{-1}(F_{t,n}(y_n))\right), \tag{3}$$

where Φ^{-1} denotes the inverse of the standard Gaussian cumulative distribution function and $\Phi_{\boldsymbol{\Sigma}}(.)$ is the n-variate Gaussian distribution function with zero mean, unit marginal variances and a correlation matrix $\boldsymbol{\Sigma}$. Based on the assumption such that the marginal predictive densities are probabilistically calibrated, the random variables defined by $F_{t,i}(Y_{t,i})$, $i = 1, \cdots, n$, are distributed $U[0, 1]$. Following an argument similar to that of [28], it is consequently assumed that a joint multivariate predictive density for $\mathbf{Y}_t$ can be represented by a latent multivariate Gaussian process $\mathbf{X} = [\Phi^{-1}(F_{t,1}(Y_{t,1})), \cdots, \Phi^{-1}(F_{t,n}(Y_{t,n}))]^\top$,

$$\mathbf{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \tag{4}$$

with zero mean, unit marginal variances and a correlation matrix $\boldsymbol{\Sigma}$. The realizations $\mathbf{x}_t = [x_{t,1}, \cdots, x_{t,n}]^\top$ of that process are given by transforming the observations of wind power generation $y_{t,i}$ through the corresponding predictive cumulative distribution functions and through Φ^{-1},

$$x_{t,i} = \Phi^{-1}(F_{t,i}(y_{t,i})), \quad i = 1, \cdots, n\,. \tag{5}$$

The base assumption about calibration of marginal predictive densities is core to the methodology subsequently used for modelling the dependence structure. In practice, it might be very difficult to verify whether these densities are calibrated or not, especially for small samples. Lack of calibration would necessarily translate to $\mathbf{X}$ not being multivariate Gaussian.

Note that, in this setup, even though the marginal distributions $F_{t,i}$ as well as the joint distributions F_t are time-dependent, the underlying dependence structure is fully characterized by the time-invariant correlation matrix $\boldsymbol{\Sigma}$, hence not requiring a time index for either $\mathbf{X}$ or $\boldsymbol{\Sigma}$. Such an assumption may not always hold in practice, since the dependence between the sites and different lead times might change under the influence of meteorological conditions, seasonal effects, changes in the terrain roughness, etc. It is out of scope in this study to address those issues, even though we discuss in Sect. 7 several possible extensions permitting to better capture such variations in the spatio-temporal dependence. The most straightforward one would be to use a sliding-window estimation approach, even though it would clearly increase computational costs.

3.4 Modelling as a Conditional Autoregression

Consider a set of available wind power observations for the vector valued process $\mathbf{Y}_t$, $t = 1, \cdots, T$. This process is transformed so as to obtain the latent multivariate Gaussian one $\mathbf{X}$, and related observations. Emphasis is placed on the correlation structure of $\mathbf{X}$. As can be seen from Fig. 3, the sample correlation matrix $\hat{\boldsymbol{\Sigma}}$ is dense. This directly implies that inference with such a matrix has a computational complexity of $\mathcal{O}(n^3)$. In order to make the proposed methodology applicable for problems of high dimension, instead of modelling the covariance matrix directly, we focus on its inverse, the precision matrix, denoted by $\mathbf{Q}$ [41].

In contrast, the sample precision matrix (see Fig. 4) is very sparse. This suggests considering Gaussian Markov Random Fields (GMRF), allowing us to benefit from computationally efficient algorithms derived for inference with sparse matrices. More specifically, by switching from a dense correlation matrix to its sparse inverse, we reduce the computational complexity from $\mathcal{O}(n^3)$ to a range from $\mathcal{O}(n)$ to $\mathcal{O}(n^{3/2})$, depending on the process characteristics [41].

While a correlation structure tells of global dependencies between marginal dimensions of the vector-valued process, the precision matrix represents conditional

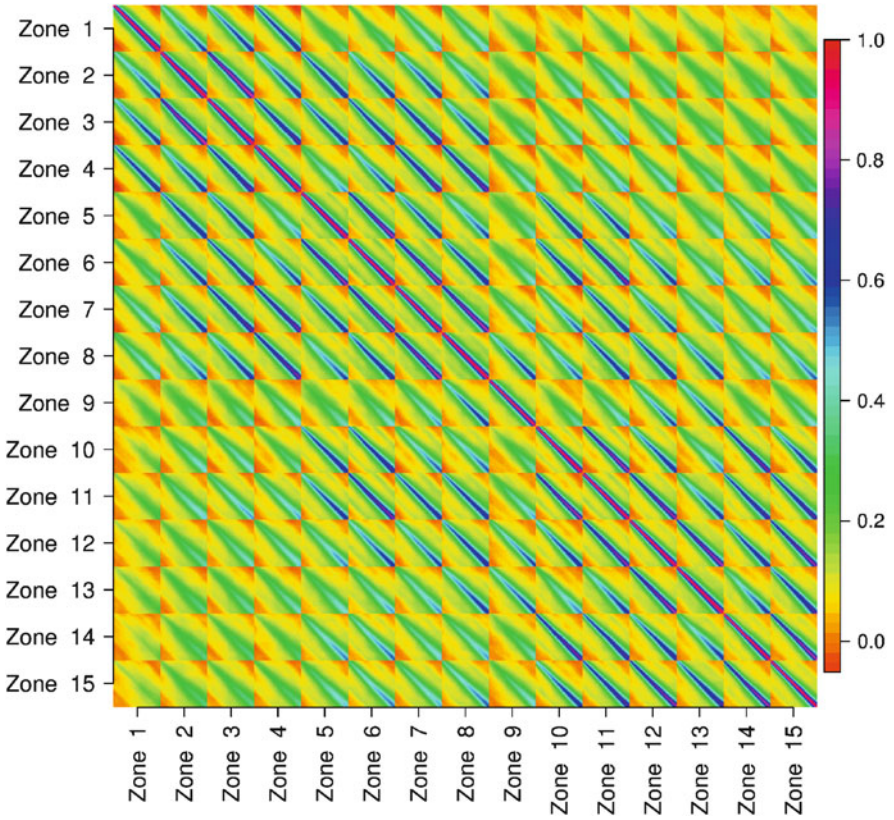


Fig. 3 Sample correlation matrix $\hat{\boldsymbol{\Sigma}}$ over the first subset of data

interdependencies. The elements of the precision matrix have the following interpretation. The diagonal elements of $\mathbf{Q}$ are the conditional precisions of X_i given $\mathbf{X}_{-i} = [X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n]^\top$, while the off-diagonal elements, with a proper scaling, provide information about the conditional correlations between the variables. For a zero mean process such as the one we are dealing with, one has

$$\mathbb{E}(X_i|\mathbf{X}_{-i}) = -\frac{1}{Q_{ii}} \sum_{j \neq i} Q_{ij} X_j \,, \tag{6}$$

$$\mathrm{Var}(X_i|\mathbf{X}_{-i}) = 1/Q_{ii} \,. \tag{7}$$

A very important relation is that $Q_{ij} = 0$ if and only if the elements X_i and X_j are independent, given $\mathbf{X}_{-\{i,j\}}$. Hence, the non-zero pattern of $\mathbf{Q}$ determines the neighbourhood of conditional dependence between variables. This relationship can be used to propose a parametrization of the precision matrix. Of course, one still has to keep in mind that $\mathbf{Q}$ is required to be symmetric positive-definite (SPD).

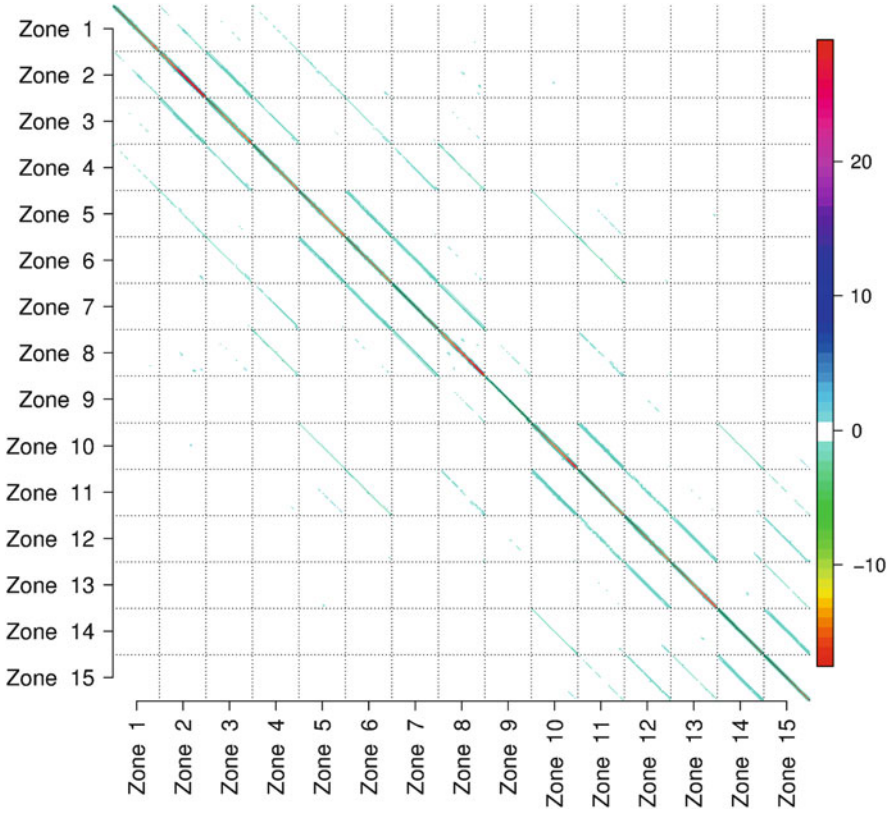


Fig. 4 Sample precision matrix $\hat{\mathbf{Q}}$ over the first subset of data

The relationship given by (6) and (7) is sometimes used for an alternative specification of Gaussian Markov Random Field through the full conditionals. This approach was pioneered by Besag in [3]: the resulting models are also known as conditional autoregressions, abbreviated as CAR. When specifying GMRF through CAR, instead of considering the entries of the precision matrix $\mathbf{Q}$, Q_{ij}, directly, focus is on modelling terms $\kappa_i = Q_{ii}$ and $\beta_{ij} = Q_{ij}/Q_{ii}$, $i, j = 1, \cdots, n$.

From (6) it is seen that the elements β_{ij} are given by the coefficients of the corresponding conditional autoregression models, while κ_i inform on the related variances. This translates to

$$\mathbf{Q} = \boldsymbol{\kappa}\mathbf{B}\,, \tag{8}$$

where $\boldsymbol{\kappa}$ denotes a diagonal matrix of dimension $n \times n$, the diagonal elements of which are given by κ_i, $i = 1, \cdots, n$. $\mathbf{B}$ is a coefficient matrix gathering a set of coefficients β_{ij}, $i, j = 1, \cdots, n$, to be seen as a standardized precision matrix. Such

a CAR specification is generally easier to interpret and we will use it to propose a parametrization for **Q** in this work.

4 Parametrization of the Precision Matrix Q

As the CAR specification of (6) and (7) allows us to decouple the problem of describing **Q** into the matrix of conditional precisions $\boldsymbol{\kappa}$ and the coefficient matrix **B** is presented, their parametrization are presented one after the other below.

4.1 Parametrization of κ

Conventionally, CAR models are given by stationary GMRF. Stationarity implies rather strong assumptions on both the neighbourhood structure and the elements of **Q**. Firstly, the structure of the neighbourhood does not allow for special arrangements for the boundary points. Secondly, the full conditionals have constant parameters not depending on i, i.e., the conditional precisions κ_i, $i = 1, \cdots, n$, are assumed constant. However, our data analysis showed that this assumption was too restrictive in the present case. Indeed, having a closer look at the diagonal of the sample precision matrix $\hat{\mathbf{Q}}$, depicted in Fig. 5, it is clear that its elements are not constant. Instead, except for the boundary points, they exhibit a trend with conditional precision increasing with the lead time.

In addition, the variation patterns for conditional precisions appear similar for all zones. Only the variation pattern for zone 9 looks different. This result is in line with a previous analysis in [43] of the spatial and temporal dynamics of

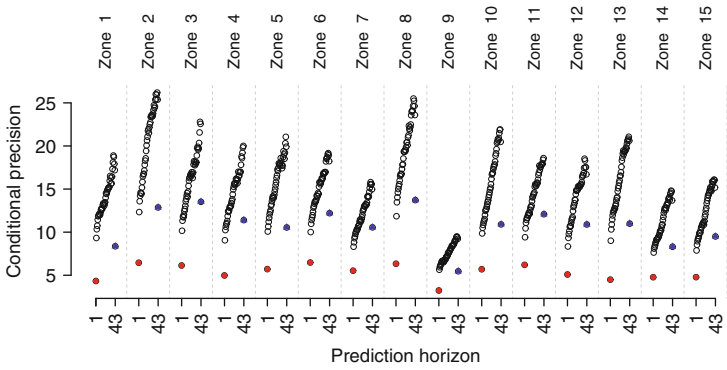


Fig. 5 Diagonal elements of the sample precision matrix, $\hat{\mathbf{Q}}$. Boundary points given by the conditional precisions for lead times 1 and 43 h ahead are marked with *red and blue circles*, respectively

wind power generation for this dataset. On the one hand, that difference might be explained by the smaller area covered by zone 9 in comparison will all the others (and lower nominal capacity), hence leading to more significant local variations, and then lower conditional precisions. On the other hand, that zone also exhibits different characteristics since located off the mainland of Jutland, where offshore wind dynamics can substantially differ from those observed onshore.

If looking at the other zones, the variation patterns observed for the conditional precisions are rather similar. It was not possible to link the differences between zones to the explanatory variables available, either measurements or forecasts. In parallel, even though one might think that the precision pattern could depend on whether a zone is located in the center of the considered territory or on the boundary, such an assumption was not supported by the data. Furthermore, our analysis did not support the alternative assumption such that conditional precisions could depend on the overall level of power variability at that zone. Consequently, our proposed parametrization for conditional precisions is the same for all zones. Further investigation might allow to refine this proposal.

As a result, $\boldsymbol{\kappa}$ is a block diagonal matrix,

$$\boldsymbol{\kappa} = \begin{pmatrix} \boldsymbol{\kappa}_B & & 0 \\ & \boldsymbol{\kappa}_B & \\ & & \ddots \\ 0 & & \boldsymbol{\kappa}_B \end{pmatrix}, \tag{9}$$

where its block element,

$$\boldsymbol{\kappa}_B = \begin{pmatrix} \kappa_1 & & 0 \\ & \kappa_2 & \\ & & \ddots \\ 0 & & \kappa_K \end{pmatrix}, \tag{10}$$

is a diagonal matrix of dimension $K \times K$, repeating N times.

Focusing on a single block element $\boldsymbol{\kappa}_B$, and in view of our observations related to Fig. 5, a parametrization for $\boldsymbol{\kappa}_B$ ought to consider separately the central lead times and the temporal boundaries. These temporal boundaries κ_1 and κ_K cannot be avoided since there cannot be lead times of less than 1 h ahead and more than K hours ahead. As a result for these lead times, the conditional models in (6) and (7) rely on a smaller set of explanatory variables. This in turn leads to lower precision values.

For lead times between 2 to $K-1$ hours ahead, an analysis of Fig. 5 suggests that conditional precisions increase with the lead time, and could be expressed as

$$\kappa_i = \rho^{i-2}, \quad i = 2, \cdots, K-1, \tag{11}$$

with ρ a ratio parameter. This model was found to be more suitable than a simpler linear one, which, visually based on Fig. 5 could also be seen as a relevant candidate.

As a result, this block element $\boldsymbol{\kappa}_B$ is parametrized as

$$\boldsymbol{\kappa}_B = \frac{1}{\sigma^2} \begin{pmatrix} \kappa_1 & & & & \\ & 1 & & 0 & \\ & & \rho & & \\ & & & \ddots & \\ & 0 & & & \rho^{K-2} \\ & & & & & \kappa_K \end{pmatrix} \tag{12}$$

meaning that, overall, the diagonal elements of $\mathbf{Q}$ can be described with four parameters only: the temporal boundaries κ_1 and κ_K, the ratio parameter ρ, and an overall scaling σ^2.

4.2 Parametrization of the Standardized Precision Matrix B

Our next step is to describe a parametrization for $\mathbf{B}$. This requires an analysis of the neighbourhood structure of $\mathbf{Q}$ by identifying which elements are non-zero. We first look at the spatial neighbourhood and then at the temporal one, The final parametrization for $\mathbf{B}$ is finally presented.

4.2.1 Spatial Neighbourhood

Consider a single zone denoted by A. A careful look at Fig. 4 (while also remembering the spatial distributions of these zones shown in Fig. 1) reveals that information at zone A only depends on the local information at A and on the four closest neighbouring zones: northern (N), eastern (E), southern (S) and western (W) neighbours of A. A simplified representation of their spatial neighbourhood effects is given in Fig. 6.

In practice, the spatial neighbours are assigned to each of the zones based on the visual inspection of the spatial distributions of these zones (see Fig. 1), as well

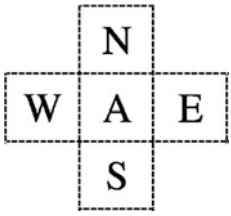


Fig. 6 Neighbourhood specification of a single zone. The focus zone is marked A, while W, N, E and S denote its western, northern, eastern and southern neighbours, respectively

as expert knowledge. For instance, for zone 11, the obvious spatial neighbours are W = 10, E = 12 and S = 15. The N neighbour is more difficult to define, since a unique zone is to be picked, while here both zones 6 and 7 could be seen as appropriate. An analysis of the sample precision matrix allowed to decide on N = 6 in view of a more pronounced dependence.

4.2.2 Temporal Neighbourhood

Figure 4 shows that information observed at zone A at time t is only dependent on a very small amount of elements at zones A, N, E, S, and W. Since precision matrices ought to be symmetric, it is sufficient to focus on the dependency between A and its western and southern neighbours, without direct consideration of the eastern and northern neighbours. Let us zoom into some relevant blocks of the sample coefficient matrix $\hat{\mathbf{B}}$, obtained based on our first subset of data, when focusing on zone 6.

Based on Fig. 7, it appears that the corresponding conditional correlations of zone A with its northern and the western neighbours differ. Information at zone A observed at time t is conditionally dependent only on the simultaneous information at zone N. Meanwhile, the conditional correlation with zone W is significant at times $t - j$, $j = -2, \cdots, 2$. This difference in the dependency patterns can be partly explained by the fact that in Denmark the prevailing winds are westerly. Thus, forecast errors most often propagate from West to East, as discussed in, e.g., [15]. This means that usually zones A and N are influenced by the upcoming weather front simultaneously, while zone W is exposed to it earlier. Of course, one should also keep in mind that in our test case distances between zones A and N are in general larger than those between A and W. That can be another factor influencing the differences.

In general, the results depicted in Fig. 7 show that information corresponding to lead time k for zone A is dependent on the variables at the neighbouring zones corresponding to lead times $k - j$, where $j = -2, \cdots, 2$. Thus, visually the data suggests a second order (temporal) process. In this work both the second ($j = -2, \cdots, 2$) and the first order ($j = -1, 0, 1$) models have been considered. Since the corresponding difference in the performance of the resulting predictive densities was rather minor, in this study the focus is on the first order model ($j = 1$). Extension to higher order models is rather straight-forward and all the discussed parametrization and estimation procedures apply.

In this work a directional non-stationary CAR model, abbreviated as DCAR, is considered. That is, the conditional correlations are made direction-dependent. In this respect the work is inspired by [24] where the authors consider a directional (in space) CAR model. We refer the reader to that work for a clear description of the modelling approach. The current proposal can be viewed as a generalization of the work presented in [24] since space-time neighbourhoods are considered along with the non-constant precisions.

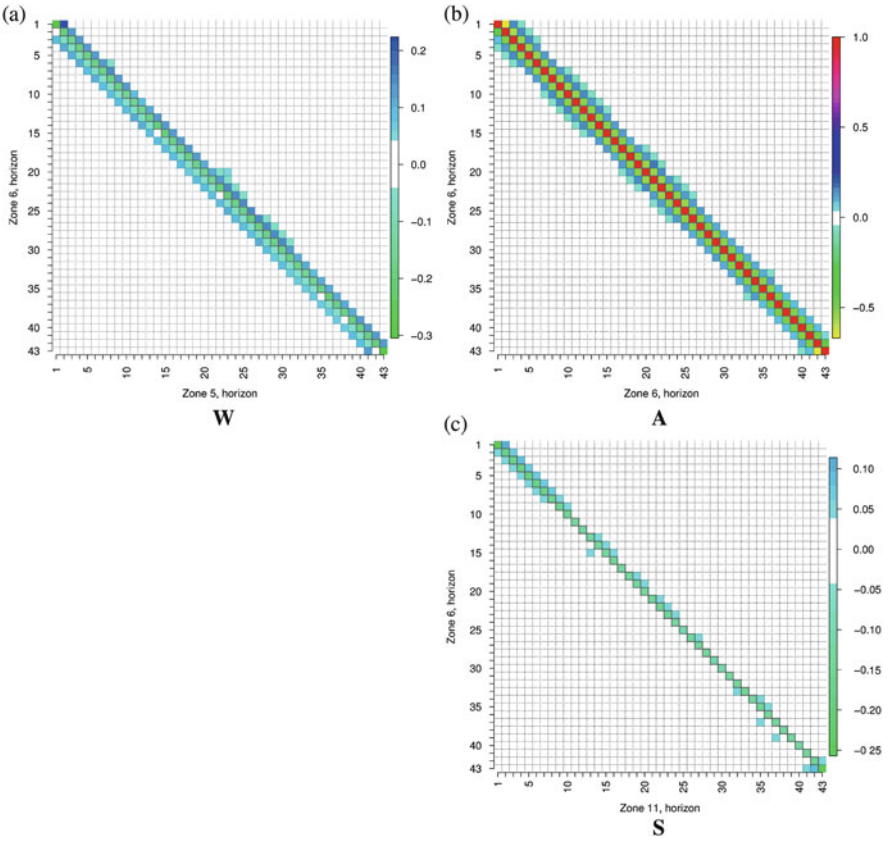


Fig. 7 Zooming on blocks of the standardized sample precision matrix $\hat{\mathbf{B}}$. (**a**) **W**. (**b**) **A**. (**c**) **S**

When considering DCAR models, directional neighbourhoods should be chosen carefully so that each of them forms a (directional) clique. Let us consider two elements from the full random vector $\mathbf{X}$, X_i and X_j. Then, given that X_i is a "west-side" and "one-hour-ago" neighbour of X_j, X_j should be assigned as the "east-side" and "one-hour-ahead" neighbour of X_i. This is essential for ensuring the symmetry of the precision matrix.

4.2.3 Resulting Space-Time Parametrization for B

Our analysis over the first data subset suggested that information for zone A at lead time k conditionally depends on information from zones N, E, S, and W with lead

times $k-1$, k, $k+1$, as well as on the local situation at zone A for lead times $k-1$ and $k+1$. In terms of the CAR specification given in (6) this translates to

$$\begin{aligned}\mathbb{E}\left(X_{t,(A-1)K+k}|\mathbf{X}_{t,-\{(A-1)K+k\}}\right) = &-\sum_{j=\{-1,1\}} a_j X_{t,(A-1)K+(k+j)} \\ &-\sum_{j=\{-1,0,1\}} \left(b_j X_{t,(W-1)K+(k+j)} + b_j^* X_{t,(E-1)K+(k+j)}\right. \\ &\left. + c_j X_{t,(N-1)K+(k+j)} + c_j^* X_{t,(S-1)K+(k+j)}\right) .\end{aligned} \tag{13}$$

In the above, a_j, b_j, b_j^*, c_j and c_j^* denote the coefficients representing conditional dependencies on the local (a) information as well as on the information coming from the west (b), east (b^*), north (c) and south (c^*), respectively. These coefficients are the building blocks for $\mathbf{B}$.

It was found that a_j, b_j, b_j^*, c_j and c_j^* do not depend on the considered lead time k. It can be also seen from Fig. 7 that there is no indication of any increase/decrease of the coefficient values with the lead time. The only values which differ from the general picture are those for the temporal boundaries. This effect is already dealt with when scaling by the corresponding conditional precisions.

In this work it is assumed that the corresponding coefficients are constant for all zones. Further work could be done in order to explain spatial variations in the coefficient values. Some restrictions have to be imposed on the parameters in $\mathbf{B}$ to ensure that the resulting precision matrix $\mathbf{Q}$ is SPD. Imposing symmetry reduces the parameter space significantly, since coefficients a_1 can be derived from a_{-1}, b_j^* from b_{-j} and c_j^* from $c_{-j}, j = -1, 0, 1$. This will be discussed in the following.

For the specific setup of our Danish test case, the resulting standardized precision matrix is defined as

$$\mathbf{B} = \begin{pmatrix}
\mathbf{A} & & \mathbf{S} & & & & & & & & & & \\
 & \mathbf{A} & \mathbf{E} & & \mathbf{S} & & & & & & & & \\
 & \mathbf{W} & \mathbf{A} & \mathbf{E} & & \mathbf{S} & & & & & & 0 & \\
\mathbf{N} & & \mathbf{W} & \mathbf{A} & & & & \mathbf{S} & & & & & \\
 & \mathbf{N} & & & \mathbf{A} & \mathbf{E} & & & & \mathbf{S} & & & \\
 & & \mathbf{N} & & \mathbf{W} & \mathbf{A} & \mathbf{E} & & & & \mathbf{S} & & \\
 & & & & & \mathbf{W} & \mathbf{A} & \mathbf{E} & & & & & \\
 & & & \mathbf{N} & & & \mathbf{W} & \mathbf{A} & \mathbf{E} & & & \mathbf{S} & \\
 & & & & & & & \mathbf{W} & \mathbf{A} & & & & \mathbf{S} \\
 & & & & \mathbf{N} & & & & & \mathbf{A} & \mathbf{E} & & \\
 & & & & & \mathbf{N} & & & & \mathbf{W} & \mathbf{A} & \mathbf{E} & \\
 & & & & & & \mathbf{N} & & & & \mathbf{W} & \mathbf{A} & \mathbf{E} \\
 & 0 & & & & & & \mathbf{N} & & & & \mathbf{W} & \mathbf{A}
\end{pmatrix}, \tag{14}$$

where **W**, **N** represent the blocks describing the conditional dependencies between a given zone its Western and Northern neighbours, respectively, while **A** represent the local dependencies within zone A itself. Note that indices for the zones were added in order to better appraise the structure of this matrix.

The **A**, **W** and **N** blocks are parametrized as

$$\mathbf{A} = \begin{pmatrix} 1 & \frac{a_{-1}}{\kappa_1} & & & 0 \\ a_{-1} & 1 & \rho a_{-1} & & \\ & \ddots & \ddots & \ddots & \\ & & a_{-1} & 1 & \rho a_{-1} \\ 0 & & & \frac{\rho^{K-1} a_{-1}}{\kappa_K} & 1 \end{pmatrix}. \tag{15}$$

One can note that in the upper diagonal of **A** instead of writing a_1 as suggested by (13), we directly express a_1 in terms of a_{-1} in order to ensure symmetry of the resulting **Q**. That is, the upper diagonal of **A** when multiplied by $[\kappa_1, 1, \rho, \cdots \rho^{K-2}]^\top$ has to be equal to the lower diagonal of **A** multiplied by $[1, \rho, \cdots, \rho^{K-2}, \kappa_K]^\top$. From this one can directly obtain the dependency between the upper and the lower diagonals of **A**. Since the elements of $\boldsymbol{\kappa}_B$ increase proportionally by ρ for lead times from 2 to $K-1$, then $a_1 = \rho a_{-1}$ in the corresponding part. In the similar manner, the scaling for the boundary points is a function of κ_1 and κ_K.

$$\mathbf{W} = \begin{pmatrix} b_0 & \frac{b_1}{\kappa_1} & & & 0 \\ b_{-1} & b_0 & b_1 & & \\ & \ddots & \ddots & \ddots & \\ & & b_{-1} & b_0 & b_1 \\ 0 & & & \frac{\rho^{K-1} b_{-1}}{\kappa_K} & b_0 \end{pmatrix}, \tag{16}$$

and

$$\mathbf{N} = \begin{pmatrix} c_0 & \frac{c_1}{\kappa_1} & & & 0 \\ c_{-1} & c_0 & c_1 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{-1} & c_0 & c_1 \\ 0 & & & \frac{\rho^{K-1} c_{-1}}{\kappa_K} & c_0 \end{pmatrix}. \tag{17}$$

For the last two blocks, they are readily obtained with

$$\begin{aligned} \mathbf{E} &= \mathbf{K}^{-1}\mathbf{W}^{\top}\mathbf{K}\,, \\ \mathbf{S} &= \mathbf{K}^{-1}\mathbf{N}^{\top}\mathbf{K}\,, \end{aligned} \tag{18}$$

to ensure symmetry of $\mathbf{Q}$.

Finally, the precision matrix $\mathbf{Q}$ can be fully modelled by a parameter vector $\boldsymbol{\theta}$, as

$$\boldsymbol{\theta} = [\kappa_1,\, \rho,\, \kappa_K,\, \sigma^2,\, a_{-1},\, b_0,\, b_{-1},\, b_1,\, c_0,\, c_{-1},\, c_1]^{\top}\,. \tag{19}$$

It will hence be referred to as $\mathbf{Q}(\boldsymbol{\theta})$ in the following.

5 Estimation

We explain here how to fit the GMRF defined by $\mathbf{Q}(\boldsymbol{\theta})$ to the observations. This task can be divided into two parts. Firstly, the discrepancy measure between observed data and the suggested GMRF is to be chosen. Secondly, one has to insure that the parameter estimates belong to the valid parameter space $\boldsymbol{\Theta}^{+}$, such that the resulting precision matrix is SPD.

5.1 The Valid Parameter Space

The precision matrix $\mathbf{Q}$ was previously described as a function of the parameter vector $\boldsymbol{\theta}$. Symmetry of $\mathbf{Q}(\boldsymbol{\theta})$ is imposed by construction. Hence, we are left with the issue of the positive definiteness of $\mathbf{Q}(\boldsymbol{\theta})$.

Unfortunately, in general, it is hard to determine the valid parameter space $\boldsymbol{\Theta}^{+}$. Analytical results are available for precision matrices that are Toeplitz [39]. These results can be used when working with homogeneous stationary GMRFs, but this is not the case here. An alternative to consider here is to work with a subset $\boldsymbol{\Theta}^{+}$ given by the sufficient condition such that $\mathbf{Q}(\boldsymbol{\theta})$ is diagonal dominant.

Diagonal dominance is easier to treat analytically. On the downside, this approach becomes more and more restrictive for an increasing number of parameters. This issue is discussed in more detail in [39]. For instance, for our particular test case we could see that the assumption of diagonal dominance was too restrictive. And, if no such restriction was imposed, the vector of parameters estimated was far from yielding a diagonal dominant precision matrix.

If the full valid parameter space $\boldsymbol{\Theta}^{+}$ is unknown and its diagonal dominant subset is deemed too restrictive, it is always possible to use a "brute force" approach (following the terminology of [39]). This entails checking if $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}^{+}$ by direct verification of whether the resulting $\mathbf{Q}(\hat{\boldsymbol{\theta}})$ is SPD or not. This is most easily done

by trying to compute the Cholesky factorization which will be successful if and only if $\mathbf{Q}$ is positive definite. This is the approach we decided to use here.

However, it is worth mentioning some advantages given by the diagonal dominance approach over the "brute force" method. An important one is that if one estimates parameters while requiring the diagonal dominance, then one can be sure that if a new territory is to be included, there is no strict necessity (other than aiming for optimality) to re-estimate the parameters. In other words, one can be sure that the vector of parameters already available would guarantee a valid covariance structure for the enlarged lattice. This is not the case if using the "brute force" approach. If aiming to add another zone, it cannot be guaranteed that the previously estimated parameters would result in a valid covariance structure, hence requiring to re-estimate them. Our various experiments showed that such a new set of parameters would be very close to the previous one. This latter vector of parameters could therefore be used as initial condition of the optimization routine used for their re-estimation.

5.2 Choosing an Appropriate Optimization Criterion

When estimating $\boldsymbol{\theta}$ from data, a discrepancy measure between the imposed GMRF and the observations needs to be chosen. Here estimation is carried out in a maximum likelihood framework. In [40], the authors argue that maximum likelihood estimators for GMRF are not robust with respect to model errors and might result in coefficient estimates which do not describe well the global properties of the data. The authors propose a new optimization criterion which resolves this difficulty. The criterion is based on a norm distance between the estimated and the observed correlation structures.

By considering both the norm-based discrepancy optimization and the maximum likelihood approach, we observed that resulting estimates were very similar. Using the maximum likelihood approach was then preferred, following another argument in [40], such that, if a GMRF describes the data adequately, then the maximum likelihood-based inference is more efficient.

5.3 Parameter Estimation Using Maximum Likelihood

Let us focus on a single time t and recall some of the notations introduced in Sect. 3. An observation of the latent Gaussian field $\mathbf{x}_t = [x_{t,1}, x_{t,2}, \cdots, x_{t,n}]^\top$ is obtained by transformation of the wind power observations $[y_{t,1}, y_{t,2}, \cdots, y_{t,n}]^\top$ through the corresponding predictive cumulative distribution functions, as in Eq. (5). In the case where marginal predictive densities are probabilistically calibrated, $\mathbf{x}_t$ is a realization from a multivariate Gaussian distribution with zero mean and correlation matrix given by $\mathbf{Q}^{-1}$. Consequently the log-likelihood contribution given by $\mathbf{x}_t$

writes

$$l_t(\boldsymbol{\theta}) = -\frac{1}{2}\left(n\ln(2\pi) - \ln|\mathbf{Q}(\boldsymbol{\theta})| + \mathbf{x}_t{}^{\top}\mathbf{Q}(\boldsymbol{\theta})\mathbf{x}_t\right) . \tag{20}$$

Given T realizations for the process $\mathbf{X}_t$, the overall likelihood is given by

$$l_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} l_t(\boldsymbol{\theta}) = -\frac{1}{2}\left(nT\ln(2\pi) - T\ln|\mathbf{Q}(\boldsymbol{\theta})| + \sum_{t=1}^{T}\mathbf{x}_t{}^{\top}\mathbf{Q}(\boldsymbol{\theta})\mathbf{x}_t\right) . \tag{21}$$

Then, solving $\frac{\partial l_T(\boldsymbol{\theta})}{\partial\sigma^2} = 0$ with respect to σ^2 yields the following profile maximum likelihood estimator for σ^2,

$$\widehat{\sigma^2} = \frac{\sum_{t=1}^{T}\mathbf{x}_t^{\top}\mathscr{P}\mathbf{x}_t}{Tn} , \tag{22}$$

where

$$\mathscr{P} = \begin{pmatrix} \kappa_1 & & & & \\ & 1 & 0 & & \\ & & \rho & & \\ & & & \ddots & \\ & 0 & & & \rho^{K-2} \\ & & & & & \kappa_K \end{pmatrix} \tag{23}$$

Having the profile likelihood estimate for σ^2, $\mathbf{Q}$ can be seen as a function of the reduced parameter vector $\boldsymbol{\theta}^{-}$:

$$\boldsymbol{\theta}^{-} = [\kappa_1, \rho, \kappa_{43}, a_{-1}, b_0, b_{-1}, b_1, c_0, c_{-1}, c_1]^{\top} . \tag{24}$$

And, finally, an estimate for $\boldsymbol{\theta}^{-}$ is obtained by a numerical optimization of the likelihood function given in (21) with respect to the parameter vector $\boldsymbol{\theta}^{-}$.

The requirement for the resulting $\mathbf{Q}(\hat{\boldsymbol{\theta}})$ to be symmetric positive definite is equivalent to requiring all eigenvalues to be positive. Similarly to [40], we approach the constrained optimization problem as an unconstrained one, adding an infinite penalty if some of the eigenvalues are negative. This approach works well in practice. In parallel, $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) = \mathbf{Q}^{-1}(\hat{\boldsymbol{\theta}})$ is required to have a unit diagonal. In practice this is achieved by the corresponding scaling of $\mathbf{Q}(\hat{\boldsymbol{\theta}})$, as also suggested in [39].

6 Results

6.1 Assessing Global Model Fit

Our verification starts with an examination of the global properties of the estimated dependence structure. This is done in the spirit of [40], i.e., by visually comparing the estimated covariance structure with the sample one, over the first subset of data available and used for modelling and model fitting. The estimated correlation matrix is illustrated in Fig. 8, while the sample one was already shown in Fig. 3. The patterns in these two matrices appear to be very similar.

The motivation for checking the global resemblance between the dependence structures in addition to the overall likelihood evaluation is given by the following. When optimizing the likelihood, the optimal fit is given by fitting the covariances within the neighbourhood exactly, while the remaining ones are determined by the inversion of the fitted precision matrix [40]. This may result in estimates which,

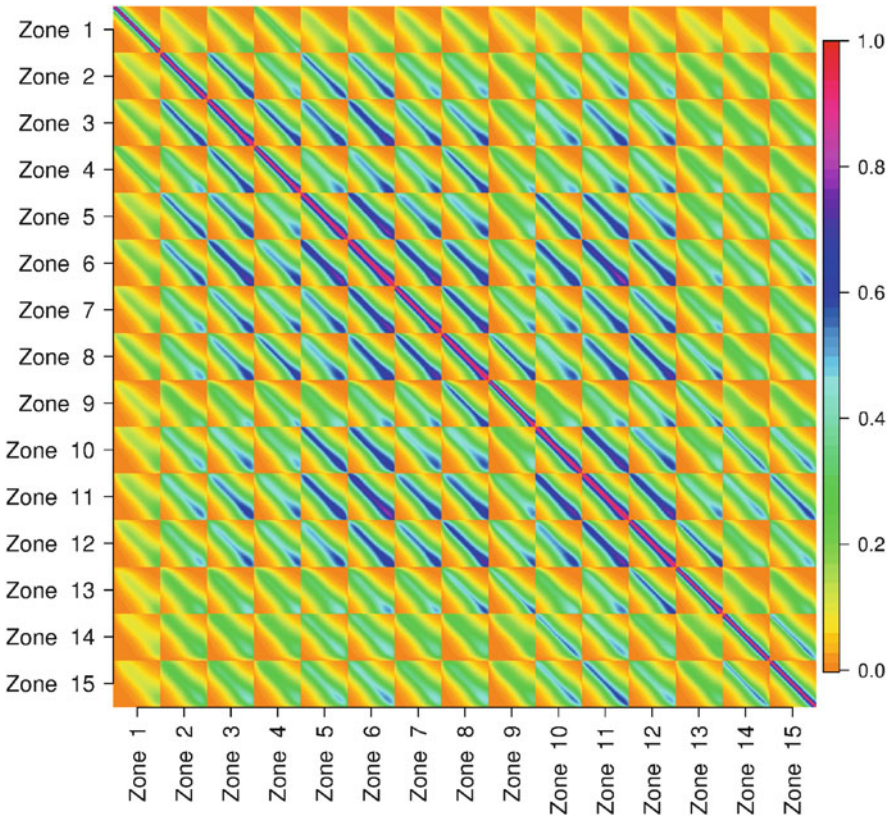


Fig. 8 Estimated correlation matrix

instead of capturing dependencies between all the variable pairs in some reasonable way, capture just some of them with a very high precision, while ignoring the others.

6.2 Assessing Predictive Model Performance

We then turn ourselves to evaluating the predictive performance of our approach. Indeed, so far, all derivations and discussion were for the first subset of data (between 1st January and 30th November 2006). Here, the evaluation uses genuine forecasts generated for the second subset of data (from 30th November 2006 to 24th October 2007), where the available predictive marginal densities, as well as estimated dependence structure, are used to issue multivariate predictive densities with a dimension $n = 645$.

The section starts with a presentation of the benchmark approaches. Further, scores used for the overall quality assessment are discussed. Finally, the empirical results are presented.

6.2.1 Overview of the Models Considered

The following models are considered in this study:

1. *Independent*: The corresponding multivariate predictive densities are based on the assumption that the marginal densities are independent, i.e.,

$$F_t(y_1, y_2, \cdots, y_n) = F_{t,1}(y_1)F_{t,2}(y_2)\cdots F_{t,n}(y_n)\ ; \tag{25}$$

2. *First order time-dependence*: The corresponding multivariate densities are obtained using a Gaussian copula approach. The covariance matrix accounts only for the temporal dependencies while completely ignoring the spatial ones. This is done by constructing the precision matrix $\mathbf{Q}$ as described in the above, but setting $\rho = \kappa_1 = \kappa_K = 1$ and $b_{-1} = b_0 = b_1 = c_{-1} = c_0 = c_1 = 0$. That is, the precision matrix in this case is described by the parameters a_1 and σ^2 only. This model does not allow for any special arrangement for the boundary points. The conditional precisions are assumed to be constant. In other words, this model corresponds to a conventional stationary GMRF defined by the first order autoregressive process in time;
3. *Separable model with first order decays in time and in space* allowing for non-constant conditional precisions: The corresponding multivariate densities are obtained using a Gaussian copula approach. The precision matrix $\mathbf{Q}$ is parametrized as in the above while setting $c_0 = b_0$, $b_1 = b_{-1} = c_1 = c_{-1} = a_1 b_0$. That is, the precision matrix in this case is described the first order time-dependence (given by a_1) and the first order spatial dependence (given by b_0). Additionally, the model gives more flexibility compared with

the conventional separable covariance structures by considering non-constant conditional precisions (modelled by ρ, κ_1 and κ_K). The model does not account for the directional influence, and that is why c_j is set to be equal to b_j with all $j = -1, \cdots, 1$;

4. *Sample correlation*: The corresponding multivariate predictive densities are obtained using a Gaussian copula approach with the correlation structure given by the sample correlation matrix;
5. *Full model*: The first order model which proposed in this study. That is the precision matrix is described by the full parameter vector $\boldsymbol{\theta}$ as given in (19).

6.2.2 Choosing an Appropriate Scoring Rule for the Quality Evaluation

In order to evaluate and compare the overall quality of multivariate probabilistic forecasts proper scoring rules are to be employed [8, 17]. An overview of proper scoring rules used for the multivariate forecast verification is given in [18]. In this work the Logarithmic score is used as a lead score for evaluating the performance of the joint predictive densities. The logarithmic scoring rule, s, is defined as

$$s(p(\mathbf{x}), \mathbf{x}_t) = -\ln(p(\mathbf{x}_t)) \tag{26}$$

Where $p(\mathbf{x})$ stands for the predictive density, which in our case is given by $\mathcal{N}(\mathbf{0}, \mathbf{Q}(\hat{\boldsymbol{\theta}})^{-1})$. $\mathbf{x}_t$ denotes the corresponding observation.

Suppose, the verification set consists of T observations, then the overall empirical score value, S, is given by the average value of the individual $s(p(\mathbf{x}), \mathbf{x}_t)$,

$$S(p(\mathbf{x})) = -\frac{1}{T}\sum_{t=1}^{T} \ln(p(\mathbf{x}_t)) \, . \tag{27}$$

In other words, the Logarithmic score value over the evaluation set is given by the average minus log-likelihood derived from the observations. The score is negatively orientated: the lower its value is, the better.

There are several reasons for choosing the Logarithmic score as the lead evaluation criterion. Firstly, it is consistent with the optimization criterion used when estimating the model parameters. Secondly, allowing for some affine transformations, this is the only local proper score (see Theorem 2 in [2]). Locality means that the score depends on the predictive distribution only through the value which the predictive density attains at the observation [8]. An important advantage of using local scores when dealing with multivariate predictive densities comes with the related computational benefits. When dealing with local scores, there is no need to draw random samples from the predictive density in order to make the evaluation.

For instance, an alternative is to use the Energy score (see detailed information on this in [18]). This score is non-local and is based on the expected Euclidean distance between forecasts and the related observations. Most often, closed-form

expressions for such expectation are not available. One then needs to employ Monte-Carlo methods in order to estimate the score value [18]. In high dimensions, Monte-Carlo techniques result in computational challenges.

A downside of employing local scores is their sensitivity to outliers. For instance, the Logarithmic score is infinite if the forecast assigns a vanishing probability to the event which occurs. In practice, when working with real data, such sensitivity might be a problem.

In this work, we considered both the Energy score and the Logarithmic score for the final density evaluation. In general the results suggested by the two scores were consistent and no contradictions were observed. However, we noticed that the Energy score was not very sensitive to the changes in the correlation structure. That is, the changes in the Energy score when moving from the assumption of independence between the marginal predictive densities to models accounting for the dependence structure were rather small (even though they still proved statistically significant based on Diebold-Mariano test statistics [12]). This is caused by low sensitivity of the Energy score to changes in the dependence structure as argued in [38]. This is another reason to focus on the Logarithmic score further in this study.

6.2.3 Empirical Results

The results from our evaluation of multivariate predictive densities issued based on the approaches mentioned in Sect. 6.2.1 are collated in Table 1, while also describing the complexity of these models in terms of their number of parameters.

One can appreciate the importance of accounting for the dependence structure from the fact that multivariate predictive densities derived from the independence assumption are shown to be of lowest quality. The full model proposed in this study outperforms the other two dependence structures, i.e., both the first-order time-dependence and the separable space-time model. The statistical significance of the improvements was verified using a likelihood-ratio test, similar to that described in [27]. This confirms that letting the related conditional correlations change depending on the direction as well as allowing for non-separable space-time influence results in better quality of the multivariate probabilistic forecasts.

Predictive densities defined by the sample correlation matrix provide the best quality forecasts. This is also expected, since in this study the estimation period

Table 1 Quality assessment of the predictive densities in terms of the Logarithmic score (S)

Model	Nr. of parameters	S
Independent	0	853.14
First order in time	1	409.98
Separable space-time model	6	357.84
Full model	10	318.07
Sample correlation	207,690	267.96

consisted of 1 year of hourly data. Large amount of data made it possible to estimate the covariance structure, even for such a high dimension. However, the main interest in the future is to make the covariance structure dependent on meteorological conditions. In this setup, tracking a sample covariance will become nearly impossible. Thus, the proposed parametrization is crucial for further development of the methodology as it significantly reduces the effective problem dimension.

6.3 Scenario Generation

As an illustration of probabilistic forecasts obtained with the proposed approach, Fig. 9 depicts a set of five scenarios describing wind power generation at zones 6 and 7, for lead times between 1 and 43 h ahead, issued on the 15th of June 2007 at 01:00. The marginal predictive densities originally available are also shown, as well as the power observations obtained a posteriori.

The scenarios generated using our approach respect dependencies both in time and in space. Respecting correlations in time ensures that the corresponding scenarios do not evolve as a random walk whose magnitude would be shaped by predictive marginals. For instance, given that a scenario predicts wind power generation at time $t + k$ to be far from its conditional expectation, then power

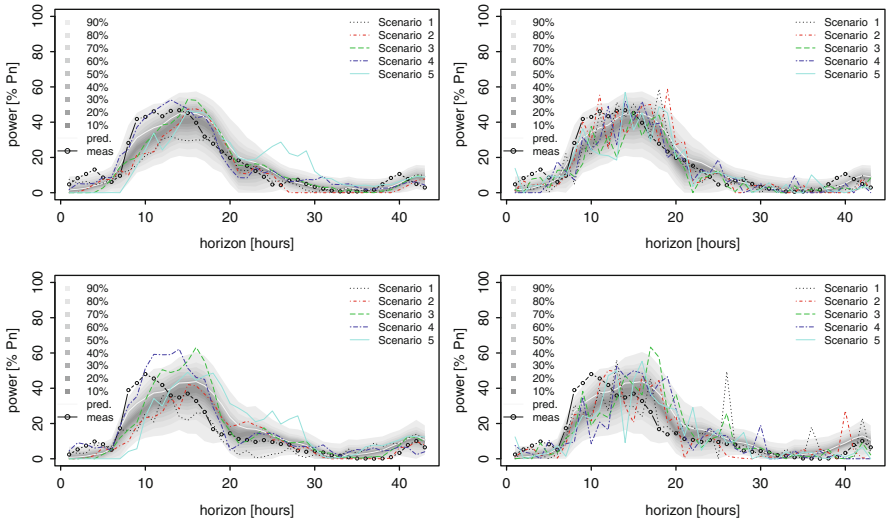


Fig. 9 Scenarios describing wind power generation at zones 6 (*top*) and 7 (*bottom*) from 1 to 43 h ahead issued on the 15th of June, 2007, at 01:00. The scenarios shown in the left are those obtained with our model. Those on the right are obtained under an independence assumption, thus, not respecting neither temporal nor spatial dependencies

generation at time $t + k + 1$ is also expected to deviate substantially from its conditional expectation. As an example see Scenario 5 for zone 6 for lead times between 22 to 30 h ahead. Similarly, respecting spatial dependency between the zones ensures that when large (resp. small) forecast errors are observed at one zone, the errors at the other zone are also expected to be large (resp. small). This is also visible from Fig. 9. For example, in the case of scenario 4, wind power generation deviates a lot from its conditional expectation in both zones 6 and 7, for lead times between 22 to 30 h ahead. In contrast, the corresponding scenarios generated using the independent model do not respect neither temporal, not spatial dependencies in the data.

7 Concluding Remarks and Perspectives

The problem of obtaining multivariate predictive densities of wind power generation was considered, jointly for a number of contiguous spatial areas and lead times, based on already available marginal predictive densities for each individual zone and lead time. A Gaussian copula approach was employed to build such multivariate predictive densities. Our core contribution lies in the proposed parametrization of the dependence structure. More specifically, instead of modelling the covariance matrix directly, focus is given to its inverse (precision matrix). This solution brings substantial practical benefits. For one, the precision matrix is shown to be very sparse. This permits to place ourselves in GMRF framework, hence resulting in computational benefits obtained through faster factorization algorithms available for sparse matrices. Besides, the proposed parametrization allows for more flexibility, since one may readily obtain nonseparable covariance structures.

The data analysis carried out for the Danish dataset revealed that the empirical precision matrix shows non-constant conditional precisions (increasing with the lead time), as well as varying conditional correlations. This hence required to go beyond conventional approaches relying on homogeneous stationary Gaussian fields. We proposed a way to model changes in conditional precisions, also allowing for conditional correlations to change with the direction. Accounting for such directional influence is not only necessary when looking at the data, but it is also quite intuitive, provided that wind power forecast errors propagate in time and in space under the influence of meteorological conditions. Consequently, the application results in terms of predictive performance confirmed that the proposed methodology and precision matrix parametrization could yield benefits in generating high-dimensional multivariate predictive densities of wind power generation, illustrated by lower Logarithmic score values.

Besides the methodological proposal and application results, a number of relevant challenges and perspectives for future work were identified. Firstly, a direct extension of the proposed methodology could consist in conditioning the precision matrix on meteorological conditions. Here, for simplicity, it was considered that the dependence structure was constant through time, with a stationarity assumption for

the underlying process. In practice, however, such a spatio-temporal dependence structure may vary substantially. There are many factors which might influence changes in process dynamics. An obvious one is the influence of the surface and higher-level wind conditions. The influence of wind direction on the spatio-temporal dependence structure could be readily modeled with a regime-switching approach, by allowing the neighbourhood structure to change with wind direction. In other words, instead of distinguishing between "West-East" and "North-South" neighbourhoods, one could instead consider "Upwind"-"Downwind" and "Concurrent"-"Concurrent". Also, it would be interesting to investigate ways to explain the variations in the conditional precisions among the zones. Possibly some clustering techniques could be employed. In parallel, slow variations in the process dynamics could be captured by considering adaptive estimation schemes for the precision matrices.

Further, an interesting challenge will be to move from the lattice setup considered in this study to a fully continuous approach. Following [25], the link between stochastic partial differential equations and some type of precision matrices could be used for such a type of problem. By understanding how the elements of the precision matrix evolve with distance between the zones and prevailing meteorological conditions, one can get a process description via stochastic partial differential equations.

On the forecast verification side, a clear challenge relates to the high dimension of the multivariate predictive densities. Already when working with a dimension $n = 645$, we have faced both methodological and computational issues, in view of the different scoring rules available for multivariate probabilistic forecast verification. Even though both Logarithmic and Energy scores can be used for multivariate probabilistic forecast verification, each of them introduces limitations in the verification exercise. On the one hand, the Energy score, being a non-local score, comes with additional computational costs since its estimation requires Monte Carlo techniques. Furthermore, following [38], this score has low sensitivity to changes in covariance structure. On the other hand, the Logarithmic score is highly sensitive to outliers: this may clearly cause difficulties in practical applications, for which both noisy data and model misspecification may then become problematic. Overall, the field of multivariate probabilistic forecast verification needs increased focus in order to propose theoretically sound and practical ways to thoroughly evaluate high-dimensional forecasts such space-time trajectories of wind power generation.

Acknowledgements The authors were partly supported by the EU Commission through the project SafeWind (ENK7-CT2008-213740), which is hereby acknowledged. Pierre Pinson was additionally supported by the Danish Strategic Research Council under '5s'–Future Electricity Markets (12-132636/DSF). The authors are grateful to Energinet.dk, the transmission system operator in Denmark, for providing the observed power data used in this paper, and to ENFOR A/S for generating the point forecasts of wind power generation used as input. Finally, reviewers and editors are acknowledged for their comments and suggestions on an earlier version of the manuscript.

References

1. Ackermann, T., et al. (2005). *Wind power in power systems*. New York: Wiley.
2. Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics, 7*(3), 686–690.
3. Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B, 36*, 192–236.
4. Bessa, R. J., Matos, M. A., Costa, I. C., Bremermann, L., Franchin, I. G., Pestana, R., Machado, N., Waldl, H. P., & Wichmann, C. (2012). Reserve setting and steady-state security assessment using wind power uncertainty forecast: A case study. *IEEE Transactions on Sustainable Energy, 3*, 827–836.
5. Bessa, R. J., Miranda, V., Botterud, A., Zhou, Z., & Wang, J. (2012). Time-adaptive quantile-copula for wind power probabilistic forecasting. *Renewable Energy, 40*(1), 29–39.
6. Botterud, A., Zhou, Z., Wang, J., Bessa, R. J., Keko, H., Sumaili, J., & Miranda, V. (2012). Wind power trading under uncertainty in LMP markets. *IEEE Transactions on Power Systems, 27*(2), 894–903.
7. Botterud, A., Zhou, Z., Wang, J., Sumaili, J., Keko, H., Mendes, J., Bessa, R. J., & Miranda, V. (2013). Demand dispatch and probabilistic wind power forecasting in unit commitment and economic dispatch: A case study of Illinois. *IEEE Transactions on Sustainable Energy, 4*(1), 250–261.
8. Bröcker, J., & Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting, 22*(2), 382–388.
9. Castronuovo, E. D., Sánchez, I., Usaola, J., Bessa, R., Matos, M., Costa, I. C., Bremermann, L., Lugaro, J., & Kariniotakis, G. (2013). An integrated approach for optimal coordination of wind power and hydro pumping storage. *Wind Energy, 17*(6), 829–852. Available online.
10. Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H., & Feitosa, E. (2008). A review on the young history of the wind power short-term prediction. *Renewable & Sustainable Energy Reviews, 12*(6), 1725–1744.
11. Diaz, G. (2013). A note on the multivariate Archimedian dependence structure in small wind generation sites. *Wind Energy, 17*(8), 1287–1295. Available online (doi:10.1002/we.1633).
12. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263.
13. Genest, C., & Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering, 12*(4), 347–368.
14. Giebel, G., Brownsword, R., Kariniotakis, G., Denhard, M., & Draxl, C. (2011). The state-of-the-art in short-term prediction of wind power–A literature overview (2nd ed.). Technical report, Technical University of Denmark.
15. Girard, R., & Allard, D. (2012). Spatio-temporal propagation of wind power prediction errors. *Wind Energy, 16*(7), 999–1012.
16. Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society: Series A, 171*(2), 319–321.
17. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378.
18. Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., & Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test, 17*(2), 211–235.
19. Hafner, C. M., & Manner, H. (2012). Conditional prediction intervals of wind power generation. *Journal of Applied Econometrics, 27*(2), 269–295.
20. Hagspiel, S., Papaemannouil, A., Schmid, M., & Andersson, G. (2012). Copula-based modeling of stochastic wind power in Europe and implications for the Swiss power grid. *Applied Energy, 96*, 33–44.
21. Hofert, M., Mächler, M., & Mcneil, A. J. (2012). Likelihood inference for archimedean copulas in high dimensions under known margins. *Journal of Multivariate Analysis, 110*, 133–150.

22. Jeon, J., & Taylor, J. W. (2012). Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association, 107*(497), 66–79.
23. Jones, L., & Clark, C. (2011). Wind integration – A survey of global views of grid operators. In *Proceedings of the 10th international workshop on large-scale integration of wind power into power systems*, Aarhus.
24. Kyung, M., & Ghosh, S. K. (2010). Maximum likelihood estimation for directional conditionally autoregressive models. *Journal of Statistical Planning and Inference, 140*(11), 3160–3179.
25. Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B, 73*(4), 423–498.
26. Louie, H. (2014). Evaluation of bivariate archimedean and elliptical copulas to model wind power dependency structures. *Wind Energy, 17*(2), 225–240.
27. Madsen, H., & Thyregod, P. (2011). *Introduction to general and generalized linear models*. Boca Raton: Chapman & Hall/CRC.
28. Möller, A., Lenkoski, A., & Thorarinsdottir, T. L. (2013). Multivariate probabilistic forecasting using bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society, 139*, 982–991.
29. Nielsen, H.Aa., Nielsen, T. S., & Madsen, H. (2011). An overview of wind power forecasts types and their use in large-scale integration of wind power. In *Proceedings of the 10th international workshop on large-scale integration of wind power into power systems*, Aarhus.
30. Ortega-Vazquez, M. A., & Kirschen, D. S. (2010). Assessing the impact of wind power generation on operating costs. *IEEE Transactions on Smart Grid, 1*(3), 295–301.
31. Papaefthymiou, G., & Kurowicka, D. (2009). Using copulas for modeling stochastic dependence in power system uncertainty analysis. *IEEE Transactions on Power Systems, 24*(1), 40–49.
32. Papavasiliou, A., & Oren, S. S. (2013). Multiarea stochastic unit commitment for high wind penetration in a transmission constrained network. *Operations Research, 61*, 578–592.
33. Pinson, P. (2013). Wind energy: Forecasting challenges for its optimal management. *Statistical Science, 28*(4), 564–585.
34. Pinson, P., & Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy, 96*, 12–20.
35. Pinson, P., & Kariniotakis, G. N. (2010). Conditional prediction intervals of wind power generation. *IEEE Transactions on Power Systems, 25*(4), 1845–1856.
36. Pinson, P., Madsen, H., Aa Nielsen, H., Papaefthymiou, G., & Klöckl, B. (2009). From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy, 12*(1), 51–62.
37. Pinson, P., Nielsen, H. A., Møller, J. K., Madsen, H., & Kariniotakis, G. N. (2007). Nonparametric probabilistic forecasts of wind power: Required properties and evaluation. *Wind Energy, 10*(6), 497–516.
38. Pinson, P., & Tastu, J. (2013). Discrimination ability of the Energy score. Technical report, Technical University of Denmark.
39. Rue, H., & Held, L. (2005). *Gaussian Markov random fields: Theory and applications*, vol. 104. Boca Raton: Chapman & Hall.
40. Rue, H., & Tjelmeland, H. (2002). Fitting gaussian markov random fields to gaussian fields. *Scandinavian Journal of Statistics, 29*(1), 31–49.
41. Simpson, D., Lindgren, F., & Rue, H. (2012). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics, 23*(1), 65–74.
42. Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.
43. Tastu, J., Pinson, P., & Madsen, H. (2010). Multivariate conditional parametric models for a spatiotemporal analysis of short-term wind power forecast errors. In *Scientific proceedings of the European wind energy conference*, Warsaw (pp. 77–81).

Game-Theoretically Optimal Reconciliation of Contemporaneous Hierarchical Time Series Forecasts

Tim van Erven and Jairo Cugliari

Abstract In hierarchical time series (HTS) forecasting, the hierarchical relation between multiple time series is exploited to make better forecasts. This hierarchical relation implies one or more aggregate consistency constraints that the series are known to satisfy. Many existing approaches, like for example bottom-up or top-down forecasting, therefore attempt to achieve this goal in a way that guarantees that the forecasts will also be aggregate consistent. We propose to split the problem of HTS into two independent steps: first one comes up with the best possible forecasts for the time series without worrying about aggregate consistency; and then a reconciliation procedure is used to make the forecasts aggregate consistent. We introduce a Game-Theoretically OPtimal (GTOP) reconciliation method, which is guaranteed to only improve any given set of forecasts. This opens up new possibilities for constructing the forecasts. For example, it is not necessary to assume that bottom-level forecasts are unbiased, and aggregate forecasts may be constructed by regressing both on bottom-level forecasts and on other covariates that may only be available at the aggregate level. We illustrate the benefits of our approach both on simulated data and on real electricity consumption data.

1 Introduction

The general setting of *hierarchical time series* (HTS) forecasting has been extensively studied because of its applications to, among others, inventory management for companies [7], euro-area macroeconomic studies [13], forecasting Australian domestic tourism [11], and balancing the national budget of states [4, 16]. As a

T. van Erven (✉)
Université Paris-Sud, 91405 Orsay Cedex, France

INRIA Saclay – Ile de France, Select team, Université Paris-Sud, 91405 Orsay Cedex, France
e-mail: tim@timvanerven.nl

J. Cugliari
Laboratoire ERIC, Université Lumière Lyon2, 69676 Bron Cedex, France
e-mail: Jairo.Cugliari@univ-lyon2.fr

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_15

consequence of the recent deployment of smart grids and autodispatchable sources, HTS have also been introduced in electricity demand forecasting [3], which is essential for electricity companies to reduce electricity production cost and take advantage of market opportunities.

A Motivating Example: Electricity Forecasting The electrical grid induces a hierarchy in which customer demand is viewed at increasing levels of aggregation. One may organize this hierarchy in different ways, but in any case the demand of individual customers is at the bottom, and the top level represents the total demand for the whole system. Depending on the modelling purpose, intermediate levels of aggregation may represent groups of clients that are tied together by geographical proximity, tariff contracts, similar consumption structure or other criteria.

Whereas demand data were previously available only for the whole system, they are now also available at regional (intermediate) levels or even at the individual level, which makes it possible to forecast electricity demand at different levels of aggregation. To this end, it is not only necessary to extend existing prediction models to lower levels of the customer hierarchy, but also to deal with the new possibilities and constraints that are introduced by the hierarchical organization of the predictions. In particular, it may be required that the sum of lower-level forecasts is equal to the prediction for the whole system. This was demanded, for example, in the Global Energy Forecasting Competition 2012 [10], and it also makes intuitive sense that the forecasts should sum in the same way as the real data. Moreover, we show in Theorems 1 and 2 below that this requirement, if enforced using a general method that we will introduce, can only improve the forecasts.

Hierarchical Time Series Electricity demand data that are organized in a customer hierarchy, are a special case of what is known in the literature as *contemporaneous* HTS: each node in the hierarchy corresponds to a time series, and, at any given time, the value of a time series higher up is equal to the sum of its constituent time series. In contrast, there also exist *temporal* HTS, in which time series are aggregated over periods of time, but we will not consider those in this work. For both types of HTS, the question of whether it is better to predict an aggregate time series directly or to derive forecasts from predictions for its constituent series has received a lot of attention, although the consensus appears to be that there is no clear-cut answer. (See [7, 13] for surveys.) A significant theoretical effort has also been made to understand the probability structure of contemporaneous HTS when the constituent series are *auto-regressive moving average* (ARMA) models [8].

HTS Forecasting The most common methods used for hierarchical time series forecasting are *bottom-up*, *top-down* and *middle-out* [3, 7]. The first of these concentrates on the prediction of all the components and uses the sum of these predictions as a forecast of the whole. The second one predicts the top level aggregate and then splits up the prediction into the components according to proportions that may be estimated, for instance, from historical proportions in the time series. The middle out strategy is a combination of the first two: one first obtains predictions at some level of the hierarchy; then one uses the bottom-up strategy to forecast the upper levels and top-down to forecast the lower levels.

As observed by Hyndman et al. [11], all three methods can be viewed as linear mappings from a set of initial forecasts for the time series to *reconciled* estimates that are *aggregate consistent*, which means that the sum of the forecasts of the components of an hierarchical time series is equal to the forecast of the whole. A more sophisticated linear mapping may be obtained by setting up a linear regression problem in which the initial forecasts are viewed as noisy observations of the expected values of the time series [4] (see Sect. 2.3). In this approach, which goes back to Stone et al. [16], it is then inescapable to assume that the initial forecasts are unbiased estimates, so that the noise has mean zero. Assuming furthermore that the covariance matrix Σ of the noise can be accurately estimated for each time step, the outcomes for the time series can be estimated using a *generalized least-squares* (GLS) method, which solves the linear regression problem with aggregate consistency constraints on the solution.

Although the assumption of unbiased initial forecasts rules out using any type of regularized estimator (like, for instance, the LASSO [17] which we consider in Sect. 3.1), it might still be manageable in practice. The difficulty with GLS, however, is estimating Σ, which might be possible on accounting data by laboriously tracing back all the sources of variance in the estimates [6], but does not seem feasible in our motivating example of electricity demand forecasting. (Standard estimators like those of White [18] or MacKinnon and White [14] do not apply, because they estimate an average of Σ over time instead of its value at the current time step.) Alternatively, it has therefore been proposed to make an additional assumption about the covariances of the initial forecasts that allows estimation of Σ to be sidestepped [11], but it is not clear when we can expect this assumption to hold (see Sect. 2.3).

Our Contribution Considering the practical difficulties in applying GLS, and the limited modelling power of bottom-up, top-down, middle-out methods, we try to approach HTS forecasting in a slightly different way. All these previous approaches have been restricted by combining the requirement of aggregate consistency with the goal of sharing information between hierarchical levels. Instead, we propose to separate these steps, which leads to an easier way of thinking about the problem. As our main contribution, we will introduce a Game-Theoretically OPtimal (GTOP) reconciliation method to map any given set of forecasts, which need not be aggregate consistent, to new aggregate consistent forecasts that are guaranteed to be at least as good. As the GTOP method requires no assumptions about the probability structure of the time series or the nature of the given set of forecasts, it leaves a forecaster completely free to use the prediction method of their choice at all levels of the hierarchy without worrying about aggregate consistency or theoretical restrictions like unbiasedness of their forecasts. As illustrated in Sect. 3.2, taking aggregate consistency out of the equation allows one to go beyond simple bottom-up, top-down or middle-out estimators, and consider estimators that use more complicated regression structures, in the same spirit as those considered by Lütkepohl [13, Section 5.3].

Outline In the next section, we present the GTOP method and formally relate it to the GLS approach. Then, in Sect. 3, we demonstrate how GTOP may be applied

with forecasts that do not satisfy the traditional unbiasedness assumption, first on simulated data, and then on real electricity demand data. Finally, Sect. 4 provides an extensive discussion.

2 Game-Theoretically Optimal Reconciliation

We will now introduce the GTOP method, which takes as input a set of forecasts, which need not be aggregate consistent, and produces as output new aggregate consistent forecasts that are guaranteed to be at least as good. In Sect. 2.1, we first present the method for the simplest possible hierarchies, which are composed of two levels only, and then, in Sect. 2.2, we explain how the procedure generalizes in a straightforward way to arbitrary hierarchies. Proofs and computational details are postponed until the end of Sect. 2.2. Finally, in Sect. 2.3, we show how GTOP reconciliation may formally be interpreted as a special case of GLS, although the quantities involved have different interpretations.

2.1 Two-Level Hierarchies

For two-level hierarchies, we will refer to the lower levels as *regions*, in reference to our motivating application of electricity demand forecasting, even though for other applications the lower levels might correspond to something else. Suppose there are K such regions, and we are not only interested in forecasting the values of a time series $(Y_k[t])_{t=1,2,\dots}$ for each individual region $k = 1, \dots, K$, but also in forecasting the sum of the regions $(Y_{\text{tot}}[t])_{t=1,2,\dots}$, where

$$Y_{\text{tot}}[t] = \sum_{k=1}^{K} Y_k[t] \qquad \text{for all } t, \tag{1}$$

as illustrated by Fig. 1.

Having observed the time series for times $1, \dots, t$, together with possible independent variables, we will be concerned with making predictions for their values at time $\tau > t$, but to avoid clutter, we will drop the time index $[\tau]$ from our notation whenever it is sufficiently clear from context. Thus, for any region k, let $\hat{Y}_k \equiv \hat{Y}_k[\tau]$ be the prediction for $Y_k \equiv Y_k[\tau]$, and let $\hat{Y}_{\text{tot}} \equiv \hat{Y}_{\text{tot}}[\tau]$ be the prediction for $Y_{\text{tot}} \equiv Y_{\text{tot}}[\tau]$. Then we evaluate the quality of our prediction for region k by the

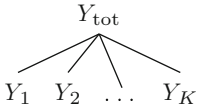


Fig. 1 A two-level hierarchical time series structure

squared loss

$$\ell_k(Y_k, \hat{Y}_k) = a_k(Y_k - \hat{Y}_k)^2,$$

where $a_k > 0$ is a weighting factor that is determined by the operational costs associated with prediction errors in region k. (We give some guidelines for the choice of these weighting factors in Sect. 4.1.) Similarly, our loss in predicting the sum of the regions is

$$\ell_{\text{tot}}(Y_{\text{tot}}, \hat{Y}_{\text{tot}}) = a_{\text{tot}}(Y_{\text{tot}} - \hat{Y}_{\text{tot}})^2,$$

with $a_{\text{tot}} > 0$. Let $\mathbf{Y} = (Y_1, \ldots, Y_K, Y_{\text{tot}})$ and $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_K, \hat{Y}_{\text{tot}})$. Then, all together, our loss at time τ is

$$\ell(\mathbf{Y}, \hat{\mathbf{Y}}) = \sum_{k=1}^{K} \ell_k(Y_k, \hat{Y}_k) + \ell_{\text{tot}}(Y_{\text{tot}}, \hat{Y}_{\text{tot}}).$$

Aggregate Inconsistency In predicting the total Y_{tot}, we might be able to take advantage of covariates that are only available at the aggregate level or there might be noise that cancels out between regions, so that we have to anticipate that $\hat{Y}_{\text{tot}}$ may be a better prediction of Y_{tot} than simply the sum of the regional predictions $\sum_{k=1}^{K} \hat{Y}_k$, and generally we may have $\hat{Y}_{\text{tot}} \neq \sum_{k=1}^{K} \hat{Y}_k$.[1] In light of (1), allowing such an aggregate inconsistency between the regional predictions and the prediction for the total would intuitively seem suboptimal. More importantly, for operational reasons it is sometimes not even allowed. For example, in the Global Energy Forecasting Competition 2012 [10], it was required that the sum of the regional predictions $\hat{Y}_1, \ldots, \hat{Y}_K$ were always equal to the prediction for the total $\hat{Y}_{\text{tot}}$. Or, if the time series represent next year's budgets for different departments, then the budget for the whole organization must typically be equal to the sum of the budgets for the departments.

We are therefore faced with a choice between two options. The first is that we might try to adjust our prediction methods to avoid aggregate inconsistency. But this would introduce complicated dependencies between our prediction methods for the different regions and for the total, and as a consequence it might make our predictions worse. So, alternatively, we might opt to remedy the problem in a post-processing step: first we come up with the best possible predictions $\hat{\mathbf{Y}}$ without worrying about any potential aggregate inconsistency, and then we map these predictions to new predictions $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_K, \tilde{Y}_{\text{tot}})$, which are *aggregate*

[1] It has also been suggested that the *central limit theorem* (CLT) implies that Y_{tot} should be more smooth than the individual regions Y_k [3], and might therefore be easier to predict.

consistent:

$$\tilde{Y}_{\text{tot}} = \sum_{k=1}^{K} \tilde{Y}_k.$$

This is the route we will take in this paper. In fact, it turns out that, for the right mapping, the loss of $\tilde{\mathbf{Y}}$ will always be smaller than the loss of $\hat{\mathbf{Y}}$, no matter what the actual data $\mathbf{Y}$ turn out to be, which provides a formal justification for the intuition that aggregate inconsistent predictions should be avoided.

Mapping to Aggregate Consistent Predictions To map any given predictions $\hat{\mathbf{Y}}$ to aggregate consistent predictions $\tilde{\mathbf{Y}}$, we will use a game-theoretic set-up that is reminiscent of the game-theoretic approach to online learning [5]. In this formulation, we will choose our predictions $\tilde{\mathbf{Y}}$ to achieve the minimum in the following minimax optimization problem:

$$V = \min_{\tilde{\mathbf{Y}} \in \mathscr{A}} \max_{\mathbf{Y} \in \mathscr{A} \cap \mathscr{B}} \Big\{ \ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}}) \Big\}. \tag{2}$$

(The sets $\mathscr{A}$ and $\mathscr{B}$ will be defined below.) This may be interpreted as the *Game-Theoretically OPtimal* (GTOP) move in a zero-sum game in which we first choose $\tilde{\mathbf{Y}}$, then the data $\mathbf{Y}$ are chosen by an adversary, and finally the pay-off is measured by the difference in loss between $\tilde{\mathbf{Y}}$ and the given predictions $\hat{\mathbf{Y}}$. The result is that we will choose $\tilde{\mathbf{Y}}$ to guarantee that $\ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})$ is at most V *no matter what the data* $\mathbf{Y}$ *are*. Satisfyingly, we shall see below that $V \leq 0$, so that the new predictions $\tilde{\mathbf{Y}}$ are always at least as good as the original predictions $\hat{\mathbf{Y}}$.

We have left open the definitions of the sets $\mathscr{A}$ and $\mathscr{B}$, which represent the domains for our predictions and the data. The former of these will represent the set of vectors that are aggregate consistent:

$$\mathscr{A} = \Big\{ (X_1, \ldots, X_K, X_{\text{tot}}) \in \mathbb{R}^{K+1} \mid X_{\text{tot}} = \sum_{k=1}^{K} X_k \Big\}.$$

By definition, both our predictions $\tilde{\mathbf{Y}}$ and the data $\mathbf{Y}$ must be aggregate consistent, so they are restricted to lie in $\mathscr{A}$. In addition, we introduce the set $\mathscr{B}$, which allows us to specify any other information we might have about the data. In the simplest case, we may let $\mathscr{B} = \mathbb{R}^{K+1}$ so that $\mathscr{B}$ imposes no constraints, but if, for example, prediction intervals $[\hat{Y}_k - B_k, \hat{Y}_k + B_k]$ are available for the given predictions, then we may take advantage of that knowledge and define

$$\mathscr{B} = \Big\{ (X_1, \ldots, X_K, X_{\text{tot}}) \in \mathbb{R}^{K+1} \mid X_k \in [\hat{Y}_k - B_k, \hat{Y}_k + B_k] \text{ for } k = 1, \ldots, K \Big\}. \tag{3}$$

We could also add a prediction interval for $\hat{Y}_{\text{tot}}$ as long as we take care that all our prediction intervals together do not contradict aggregate consistency of the data. In

general, we will require that $\mathscr{B} \subseteq \mathbb{R}^{K+1}$ is a *closed* and *convex* set, and $\mathscr{A} \cap \mathscr{B}$ must be non-empty so that $\mathscr{B}$ does not contradict aggregate consistency.

GTOP Predictions as a Projection Let $\|\mathbf{X}\| = (\sum_{i=1}^{d} X_i^2)^{1/2}$ denote the L2-norm of a vector $\mathbf{X} \in \mathbb{R}^d$ for any dimension d. Then the total loss may succinctly be written as

$$\ell(\mathbf{Y}, \hat{\mathbf{Y}}) = \|A\mathbf{Y} - A\hat{\mathbf{Y}}\|^2, \tag{4}$$

where $A = \text{diag}(\sqrt{a_1}, \ldots, \sqrt{a_K}, \sqrt{a_{\text{tot}}})$ is a diagonal $(K+1) \times (K+1)$ matrix that accounts for the weighting factors. In view of the loss, it is quite natural that the GTOP predictions turn out to be equal to the L2-projection

$$\tilde{\mathbf{Y}}_{\text{proj}} = \underset{\tilde{\mathbf{Y}} \in \mathscr{A} \cap \mathscr{B}}{\arg\min} \|A\hat{\mathbf{Y}} - A\tilde{\mathbf{Y}}\|^2 \tag{5}$$

of $\hat{\mathbf{Y}}$ unto $\mathscr{A} \cap \mathscr{B}$ after scaling all dimensions according to A.

Theorem 1 (GTOP: Two-level Hierarchies) *Suppose that $\mathscr{B}$ is a closed, convex set and that $\mathscr{A} \cap \mathscr{B}$ is not empty. Then the projection $\tilde{\mathbf{Y}}_{\text{proj}}$ uniquely exists, the value of (2) is*

$$V = -\|A\tilde{\mathbf{Y}}_{\text{proj}} - A\hat{\mathbf{Y}}\|^2 \leq 0,$$

and the GTOP predictions are $\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}_{\text{proj}}$.

Thus, in a metric that depends on the loss, GTOP makes the minimal possible adjustment of the given predictions $\hat{\mathbf{Y}}$ to make them consistent with what we know about the data. Moreover, the fact that $V \leq 0$ implies that the GTOP predictions are at least as good as the given predictions:

$$\ell(\mathbf{Y}, \tilde{\mathbf{Y}}_{\text{proj}}) \leq \ell(\mathbf{Y}, \hat{\mathbf{Y}}) \qquad \text{for any data } \mathbf{Y} \in \mathscr{A} \cap \mathscr{B}.$$

Theorem 1 will be proved as a special case of Theorem 2 in the next section.

Example 1 If $\mathscr{B} = \mathbb{R}^{K+1}$ does not impose any constraints, then the GTOP predictions are

$$\tilde{Y}_{\text{proj},k} = \hat{Y}_k + \frac{\frac{1}{a_k}}{\sum_{i=1}^{K} \frac{1}{a_i} + \frac{1}{a_{\text{tot}}}} \Delta \quad \text{for } k = 1, \ldots, K,$$

$$\tilde{Y}_{\text{proj,tot}} = \hat{Y}_{\text{tot}} - \frac{\frac{1}{a_{\text{tot}}}}{\sum_{i=1}^{K} \frac{1}{a_i} + \frac{1}{a_{\text{tot}}}} \Delta,$$

where $\Delta = \hat{Y}_{\text{tot}} - \sum_{k=1}^{K} \hat{Y}_k$ measures by how much $\hat{\mathbf{Y}}$ violates aggregate consistency. In particular, if the given predictions $\hat{\mathbf{Y}}$ are already aggregate consistent, i.e. $\hat{Y}_{\text{tot}} =$

$\sum_{k=1}^{K} \hat{Y}_k$, then the GTOP predictions are the same as the given predictions: $\tilde{\mathbf{Y}}_{\text{proj}} = \hat{\mathbf{Y}}$.

Example 2 If $\mathscr{B}$ consists of the prediction intervals specified in (3), then the extreme values $B_1 = \ldots = B_K = 0$ make the GTOP predictions exactly equal to those of the bottom-up forecaster.

Example 3 If $\mathscr{B}$ defines prediction intervals as in (3) and $a_1 = \cdots = a_K = a$ and $B_1 = \cdots = B_K = B$, then the GTOP predictions are

$$\tilde{Y}_{\text{proj},k} = \hat{Y}_k + \left[\frac{\frac{1}{a}}{\frac{K}{a} + \frac{1}{a_{\text{tot}}}} \Delta \right]_B \quad \text{for } k = 1, \ldots, K,$$

$$\tilde{Y}_{\text{proj,tot}} = \sum_{k=1}^{K} \tilde{Y}_{\text{proj},k},$$

where $[x]_B = \max\left\{-B, \min\{B, x\}\right\}$ denotes clipping x to the interval $[-B, B]$ and $\Delta = \hat{Y}_{\text{tot}} - \sum_{k=1}^{K} \hat{Y}_k$.

In general the GTOP predictions $\tilde{\mathbf{Y}}_{\text{proj}}$ do not have a closed-form solution, but, as long as $\mathscr{B}$ can be described by a finite set of inequality constraints, they can be computed using quadratic programming. The details will be discussed at the end of the next section, which generalizes the two-level hierarchies introduced so far to arbitrary summation constraints.

2.2 General Summation Constraints

One might view (1) as forecasting $K + 1$ time series, which are ordered in a hierarchy with two levels, in which the time series $(Y_1[t])$, …, $(Y_K[t])$ for the regions are at the bottom, and their total $(Y_{\text{tot}}[t])$ is at the top (see Fig. 1). More generally, one might imagine having a multi-level hierarchy of any finite number of time series $(Y_1[t]), \ldots, (Y_M[t])$, which are organised in a tree T that represents the hierarchy of aggregation consistency requirements. For example, in Fig. 2 the time series $(Y_1[t])$ might be the expenditure of an entire organisation, the time series $(Y_2[t])$, $(Y_3[t])$, and $(Y_4[t])$ might be the expenditures in different subdivisions within the organization, time series $(Y_5[t])$, $(Y_6[t])$ and $(Y_7[t])$ might represent the expenditures in departments within subdivision $(Y_2[t])$, and similarly $(Y_8[t])$ and $(Y_9[t])$ would be the expenditures in departments within $(Y_3[t])$.

The discussion from the previous section directly extends to multi-level hierarchies as follows. For each time series $m = 1, \ldots, M$, let $c(m) \subset \{1, \ldots, M\}$ denote the set of its children in T. Then aggregate consistency generalizes to the constraint

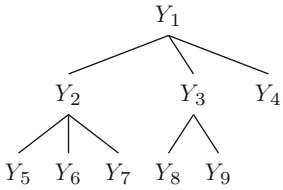


Fig. 2 Example of a multi-level hierarchical time series structure

$$\mathscr{A} = \left\{(X_1, \ldots, X_M) \in \mathbb{R}^M \mid X_m = \sum_{i \in c(m)} X_i \text{ for all } m \text{ such that } c(m) \text{ is non-empty}\right\}.$$

Remark 1 We note that all the constraints $X_m = \sum_{i \in c(m)} X_i$ in $\mathscr{A}$ are *linear equality constraints*. In fact, in all the subsequent developments, including Theorem 2, we can allow $\mathscr{A}$ to be *any set* of linear equality constraints, as long as they are internally consistent, so that $\mathscr{A}$ is not empty. In particular, we could even allow two (or more) predictions for the same time series by regarding the first prediction as a prediction for a time series $(Y_m[t])$ and the second as a prediction for a separate time series $(Y_{m'}[t])$ with the constraint that $Y_m[t] = Y_{m'}[t]$. To keep the exposition focussed, however, we will not explore these possibilities in this paper.

Having defined the structure of the hierarchical time series through $\mathscr{A}$, any additional information we may have about the data can again be represented by choosing a convex, closed set $\mathscr{B} \subseteq \mathbb{R}^M$ which is such that $\mathscr{A} \cap \mathscr{B}$ is non-empty. In particular, $\mathscr{B} = \mathbb{R}^M$ represents having no further information, and prediction intervals can be represented analogously to (3) if they are available.

As in the two-level hierarchy, let $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_M)$ be the original (potentially aggregate inconsistent) predictions for the time series $\mathbf{Y} = (Y_1, \ldots, Y_M)$ at a given time τ. We assign weighting factors $a_m > 0$ to each of the time series $m = 1, \ldots, M$, and we redefine the diagonal matrix $A = \operatorname{diag}(\sqrt{a_1}, \ldots, \sqrt{a_M})$, so that we may write the total loss as in (4). Then the GTOP predictions $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_M)$ are still defined as those achieving the minimum in (2), and the L2-projection $\tilde{\mathbf{Y}}_{\text{proj}}$ is as defined in (5).

Theorem 2 (GTOP: Multi-level Hierarchies) *The exact statement of Theorem 1 still holds for the more general definitions for multi-level hierarchies in this section.*

The proof of Theorems 1 and 2 fundamentally rests on the Pythagorean inequality, which is illustrated by Fig. 3. In fact, this inequality is not restricted to the squared loss we use in this paper, but holds for any loss that is based on a Bregman divergence [5, Section 11.2], so the proof would go through in exactly the same way for such other losses. For example, the Kullback-Leibler divergence, which measures the difference between two probability distributions, is also a Bregman divergence.

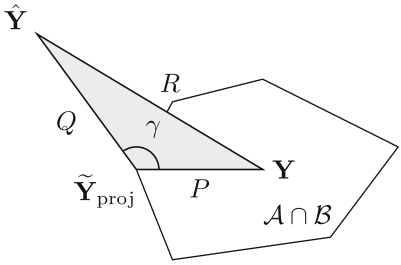


Fig. 3 Illustration of the Pythagorean inequality $P^2 + Q^2 \leq R^2$, where $P = \|A\mathbf{Y} - A\tilde{\mathbf{Y}}_{\text{proj}}\|$, $Q = \|A\tilde{\mathbf{Y}}_{\text{proj}} - A\hat{\mathbf{Y}}\|$ and $R = \|A\mathbf{Y} - A\hat{\mathbf{Y}}\|$. Convexity of $\mathscr{A} \cap \mathscr{B}$ ensures that $\gamma \geq 90°$

Lemma 1 (Pythagorean Inequality) *Suppose that $\mathscr{B}$ is a closed, convex set and that $\mathscr{A} \cap \mathscr{B}$ is non-empty. Then the projection $\tilde{\mathbf{Y}}_{\text{proj}}$ exists and is unique, and*

$$\|A\mathbf{Y} - A\tilde{\mathbf{Y}}_{\text{proj}}\|^2 + \|A\tilde{\mathbf{Y}}_{\text{proj}} - A\hat{\mathbf{Y}}\|^2 \leq \|A\mathbf{Y} - A\hat{\mathbf{Y}}\|^2 \qquad \textit{for all } \mathbf{Y} \in \mathscr{A} \cap \mathscr{B}.$$

Proof The lemma is an instance of the generalized Pythagorean inequality [5, Section 11.2] for the Bregman divergence corresponding to the Legendre function $F(\mathbf{X}) = \|A\mathbf{X}\|^2$, which is strictly convex (as required) because all entries of the matrix A are strictly positive. (The set $\mathscr{A}$ is a hyperplane, so it is closed and convex by construction. The assumptions of the lemma therefore ensure that $\mathscr{A} \cap \mathscr{B}$ is closed, convex and non-empty.) □

Proof (Theorem 2) Let $f(\mathbf{Y}, \tilde{\mathbf{Y}}) = \ell(\mathbf{Y}, \tilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})$. We will show that $(\tilde{\mathbf{Y}}_{\text{proj}}, \tilde{\mathbf{Y}}_{\text{proj}})$ is a saddle-point for f, which implies that playing $\tilde{\mathbf{Y}}_{\text{proj}}$ is the optimal strategy for both players in the zero-sum game and that

$$V = \min_{\tilde{\mathbf{Y}} \in \mathscr{A}} \max_{\mathbf{Y} \in \mathscr{A} \cap \mathscr{B}} f(\mathbf{Y}, \tilde{\mathbf{Y}}) = \max_{\mathbf{Y} \in \mathscr{A} \cap \mathscr{B}} \min_{\tilde{\mathbf{Y}} \in \mathscr{A}} f(\mathbf{Y}, \tilde{\mathbf{Y}}) = f(\tilde{\mathbf{Y}}_{\text{proj}}, \tilde{\mathbf{Y}}_{\text{proj}}) = -\|A\tilde{\mathbf{Y}}_{\text{proj}} - A\hat{\mathbf{Y}}\|^2$$

[15, Lemma 36.2], which is to be shown.

To prove that $(\tilde{\mathbf{Y}}_{\text{proj}}, \tilde{\mathbf{Y}}_{\text{proj}})$ is a saddle-point, we need to show that neither player can improve their pay-off by changing their move. To this end, we first observe that, by the Pythagorean inequality (Lemma 1),

$$f(\mathbf{Y}, \tilde{\mathbf{Y}}_{\text{proj}}) = \|A\mathbf{Y} - A\tilde{\mathbf{Y}}_{\text{proj}}\|^2 - \|A\mathbf{Y} - A\hat{\mathbf{Y}}\|^2 \leq -\|A\tilde{\mathbf{Y}}_{\text{proj}} - A\hat{\mathbf{Y}}\|^2 = f(\tilde{\mathbf{Y}}_{\text{proj}}, \tilde{\mathbf{Y}}_{\text{proj}})$$

for all $\mathbf{Y} \in \mathscr{B} \cap \mathscr{A}$. It follows that the maximum is achieved by $\mathbf{Y} = \tilde{\mathbf{Y}}_{\text{proj}}$. Next, we also have

$$\arg\min_{\tilde{\mathbf{Y}} \in \mathscr{A}} f(\tilde{\mathbf{Y}}_{\text{proj}}, \tilde{\mathbf{Y}}) = \arg\min_{\tilde{\mathbf{Y}} \in \mathscr{A}} \|A\tilde{\mathbf{Y}}_{\text{proj}} - A\tilde{\mathbf{Y}}\|^2 = \tilde{\mathbf{Y}}_{\text{proj}},$$

which completes the proof. □

Efficient Computation For special cases, like the examples in the previous section, the GTOP projection $\tilde{\mathbf{Y}}_{\text{proj}}$ sometimes has a closed form. In general, no closed-form

solution may be available, but $\tilde{\mathbf{Y}}_{\text{proj}}$ can still be computed by finding the solution to the quadratic program

$$\begin{aligned} \min_{\tilde{\mathbf{Y}}} \quad & \|A\hat{\mathbf{Y}} - A\tilde{\mathbf{Y}}\|^2 \\ \text{subject to} \quad & \tilde{\mathbf{Y}} \in \mathscr{A} \cap \mathscr{B}. \end{aligned}$$

Since $\mathscr{A}$ imposes only equality constraints, this quadratic program can be solved efficiently as long as the further constraints imposed by $\mathscr{B}$ are manageable. In particular, if $\mathscr{B}$ imposes only linear inequality constraints, like, for example, in (3), then the solution can be found efficiently using interior point methods [12] or using any of the alternatives suggested by Hazan et al. [9, Section 4]. The experiments in Sect. 3 were all implemented using the `quadprog` package for the `R` programming language, which turned out to be fast enough.

2.3 Formal Relation to Generalized Least-Squares

As discussed in the introduction, HTS has been modelled as a problem of linear regression in the economics literature [4]. It is interesting to compare this approach to GTOP, because the two turn out to be very similar, except that the quantities involved have different interpretations. The linear regression approach models the predictions as functions of the means of the real data

$$\hat{\mathbf{Y}}[\tau] = \mathbb{E}\left\{\mathbf{Y}[\tau]\right\} + \varepsilon[\tau]$$

that are perturbed by a noise vector $\varepsilon[\tau] = (\varepsilon_1[\tau], \ldots, \varepsilon_M[\tau])$, where all distributions and expectations are conditional on all previously observed values of the time series. Then it is assumed that the predictions are *unbiased estimates*, so that the noise variables all have mean zero, and the true means $\mathbb{E}\{\mathbf{Y}[\tau]\}$ can be estimated using the *generalized least-squares* (GLS) estimate

$$\begin{aligned} \min_{\tilde{\mathbf{Y}}} \quad & (\hat{\mathbf{Y}} - \tilde{\mathbf{Y}})^\top \Sigma^{-1} (\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}) \\ \text{subject to} \quad & \tilde{\mathbf{Y}} \in \mathscr{A}, \end{aligned} \tag{6}$$

where $\Sigma \equiv \Sigma[\tau]$ is the $M \times M$ covariance matrix for the noise $\varepsilon[\tau]$ [4]. This reveals an interesting superficial relation between the GTOP forecasts and the GLS estimates: if

$$\Sigma^{-1} = A^\top A \qquad \text{and} \qquad \mathscr{B} = \mathbb{R}^M, \tag{7}$$

then the two coincide! However, the interpretation of A and Σ^{-1} is completely different, and the two procedures serve different purposes: whereas GLS tries to address both reconciliation and the goal of sharing information between hierarchical levels at the same time, the GTOP method is only intended to do reconciliation and requires a separate procedure to share information. The case where the two methods coincide is therefore only a formal coincidence, and one should not assume that the choice $\Sigma^{-1} = A^\top A$ will adequately take care of sharing information between hierarchical levels!

Ordinary Least-squares Given the difficulty of estimating Σ, Hyndman et al. [11] propose an assumption that allows them to sidestep estimation of Σ altogether: they show that, under their assumption, the GLS estimate reduces to the *Ordinary Least-squares* (OLS) estimate obtained from (6) by the choice

$$\Sigma = I,$$

where I is the identity matrix. Via (7) it then follows that the OLS and GTOP forecasts formally coincide when we take all the weighting factors in the definition of the loss to be equal: $a_1 = \ldots = a_M$, and let $\mathscr{B} = \mathbb{R}^M$. Consequently, for two-level hierarchies, OLS can be computed as in Example 1.

The assumption proposed by Hyndman et al. [11] is that, at time τ, the covariance $\mathrm{Cov}(\hat{Y}_m, \hat{Y}_{m'})$ of the predictions for any two time series decomposes as

$$\mathrm{Cov}(\hat{Y}_m, \hat{Y}_{m'}) = \sum_{\substack{i \in S(m) \\ j \in S(m')}} \mathrm{Cov}(\hat{Y}_i, \hat{Y}_j) \quad \text{for all } m, m', \tag{8}$$

where $S(m) \subseteq \{1, \ldots, M\}$ denotes the set of bottom-level time series out of which Y_m is composed. That is, $Y_m = \sum_{i \in S(m)} Y_i$ with Y_i childless (i.e. $c(i) = \emptyset$) for all $i \in S(m)$.

Although the OLS approach appears to work well in practice (see Sect. 3.2), it is not obvious when we can expect (8) to hold. Hyndman et al. [11] motivate it by pointing out that (8) would hold exactly if the forecasts would be exactly aggregate consistent (i.e. $\hat{\mathbf{Y}} \in \mathscr{A}$). Since it is reasonable to assume that the forecasts will be approximately aggregate consistent, it then also seems plausible that (8) will hold approximately. However, this motivation seems insufficient, because reasoning *as if* the forecasts are aggregate consistent leads to conclusions that are too strong: if $\hat{\mathbf{Y}} \in \mathscr{A}$, then any instance of GLS would give the same answer, so it would not matter which Σ we used, and in the experiments in Sect. 3 we see that this clearly does matter.

We therefore prefer to view OLS rather as a special case of GTOP, which will work well when all the weighting factors in the loss are equal and the constraints in $\mathscr{B}$ are vacuous.

3 Experiments

As discussed above, the GTOP method only solves the reconciliation part of HTS forecasting; it does not prescribe how to construct the original predictions $\hat{\mathbf{Y}}$. We will now illustrate how GTOP might be used in practice, taking advantage of the fact that it does not require the original predictions $\hat{\mathbf{Y}}$ to be unbiased. First, in Sect. 3.1, we present a toy example with simulated data, which nevertheless illustrates many of the difficulties one might encounter on real data. Then, in Sect. 3.2, we apply GTOP to real electricity demand data, which motivated its development.

3.1 Simulation Study

We use GTOP with prediction intervals as in (3). We will compare to bottom-up forecasting, and also to the OLS method described in Sect. 2.3, because it appears to work well in practice (see Sect. 3.2) and it is one of the few methods available that does not require estimating any parameters. We do not compare to top-down forecasting, because estimating proportions in top-down forecasting is troublesome in the presence of independent variables (see Sect. 4.2).

Data We consider a two-level hierarchy with two regions, and simulate data according to

$$Y_1[t] = \beta_{1,0} + \beta_{1,1} X[t] + \epsilon_1[t] \qquad Y_2[t] = \beta_{2,0} + \beta_{2,1} X[t] + \epsilon_2[t]$$

where $(X[t])$ is an independent variable, $\beta_1 = (\beta_{1,0}, \beta_{1,1})$ and $\beta_2 = (\beta_{2,0}, \beta_{2,1})$ are coefficients to be estimated, and $(\epsilon_1[t])$ and $(\epsilon_2[t])$ are noise variables. We will take $\beta_1 = \beta_2 = (1, 5)$, and let

$$\epsilon_1[t] = \tau \vartheta_1[t] + \sigma \upsilon[t] \qquad \epsilon_2[t] = \tau \vartheta_2[t] - \sigma \upsilon[t] \qquad \text{for all } t,$$

where $\vartheta_1[t]$, $\vartheta_2[t]$ and $\upsilon[t]$ are uniformly distributed on $[-1, 1]$, independently over t and independently of each other, and τ and σ are scale parameters, for which we will consider different values. Notice that the noise that depends on $\upsilon[t]$ cancels from the total $Y_{\text{tot}}[t] = Y_1[t] + Y_2[t]$, which makes the total easier to predict than the individual regions. We sample a train set of size 100 for the fixed design $(X[t])_{t=1,\ldots,100} = (1/100, 2/100, \ldots, 1)$ and a test set of the same size for $(X[t])_{t=101,\ldots,200} = (1 + 1/100, \ldots, 2)$.

Fitting Models on the Train Set Based on the train set, we find estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of the coefficients β_1 and β_2 by applying the LASSO [17] separately for each of the two regions, using cross-validation to calibrate the amount of penalization. Then

we predict $Y_1[\tau]$ and $Y_2[\tau]$ by

$$\hat{Y}_1[\tau] = \hat{\beta}_{1,0} + \hat{\beta}_{1,1}X[\tau] \qquad \hat{Y}_2[\tau] = \hat{\beta}_{2,0} + \hat{\beta}_{2,1}X[\tau].$$

Remark 2 In general, it is not guaranteed that forecasting the total $Y_{\text{tot}}[\tau]$ directly will give better predictions than the bottom-up forecast [13]. Consequently, if the bottom-up forecast is the best we can come up with, then that is how we should define our prediction for the total, and no further reconciliation is necessary!

If we would use the LASSO directly to predict the total $Y_{\text{tot}}[\tau]$, then, in light of Remark 2, it might not do better than simply using the *bottom-up* forecast $\hat{Y}_1[\tau] + \hat{Y}_2[\tau]$. We can be sure to do better than the bottom-up forecaster, however, by *adding our regional forecasts* $\hat{Y}_1[\tau]$ *and* $\hat{Y}_2[\tau]$ *as covariates*, such that we fit $Y_{\text{tot}}[\tau]$ by

$$\beta_{\text{tot},0} + \beta_{\text{tot},1}X[\tau] + \beta_{\text{tot},2}\hat{Y}_1[\tau] + \beta_{\text{tot},3}\hat{Y}_2[\tau], \tag{9}$$

where $\beta_{\text{tot}} = (\beta_{\text{tot},0}, \beta_{\text{tot},1}, \beta_{\text{tot},2}, \beta_{\text{tot},3})$ are coefficients to be estimated. For $\beta_{\text{tot}} = (0, 0, 1, 1)$ this would exactly give the bottom-up forecast, but now we can also obtain different estimates if the data tell us to use different coefficients. However, to be conservative and take advantage of the prior knowledge that the bottom-up forecast is often quite good, we introduce prior knowledge into the LASSO by regularizing by

$$|\beta_{\text{tot},0}| + |\beta_{\text{tot},1}| + |\beta_{\text{tot},2} - 1| + |\beta_{\text{tot},3} - 1| \tag{10}$$

instead of its standard regularization by $|\beta_{\text{tot},0}| + |\beta_{\text{tot},1}| + |\beta_{\text{tot},2}| + |\beta_{\text{tot},3}|$, which gives it a preference for coefficients that are close to those of the bottom-up forecast. Thus, from the train set, we obtain estimates $\hat{\beta}_{\text{tot}} = (\hat{\beta}_{\text{tot},0}, \hat{\beta}_{\text{tot},1}, \hat{\beta}_{\text{tot},2}, \hat{\beta}_{\text{tot},3})$ for β_{tot}, and we predict $Y_{\text{tot}}[\tau]$ by

$$\hat{Y}_{\text{tot}}[\tau] = \hat{\beta}_{\text{tot},0} + \hat{\beta}_{\text{tot},1}X[\tau] + \hat{\beta}_{\text{tot},2}\hat{Y}_1[\tau] + \hat{\beta}_{\text{tot},3}\hat{Y}_2[\tau].$$

Remark 3 The regularization in (10) can be implemented using standard LASSO software by reparametrizing in terms of $\beta'_{\text{tot}} = (\beta_{\text{tot},0}, \beta_{\text{tot},1}, \beta_{\text{tot},2} - 1, \beta_{\text{tot}_3} - 1)$ and subtracting $\hat{Y}_1[t]$ and $\hat{Y}_2[t]$ from the observation of $Y_{\text{tot}}[t]$ before fitting the model. This gives estimates $\hat{\beta}'_{\text{tot}} = (\hat{\beta}'_{\text{tot},0}, \hat{\beta}'_{\text{tot},1}, \hat{\beta}'_{\text{tot},2}, \hat{\beta}'_{\text{tot},3})$ for β'_{tot}, which we turn back into estimates $\hat{\beta}_{\text{tot}} = (\hat{\beta}'_{\text{tot},0}, \hat{\beta}'_{\text{tot},1}, \hat{\beta}'_{\text{tot},2} + 1, \hat{\beta}'_{\text{tot},3} + 1)$ for β_{tot}.

Reconciliation The procedure outlined above gives us a set of forecasts $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2, \hat{Y}_{\text{tot}})$ for any time τ, but these forecasts need not be aggregate consistent. It therefore remains to reconcile them. We will compare GTOP reconciliation to the bottom-up forecaster and to the OLS method. To apply GTOP, we have to choose the set $\mathscr{B}$, which specifies any prior knowledge we may have about the data. The easiest would be to specify no prior knowledge (by taking $\mathscr{B} = \mathbb{R}^3$), but instead we will opt to define prediction intervals for the two regional predictions as in (3).

We will use the same prediction bounds B_1 and B_2 for the entire test set, which are estimated (somewhat simplistically) by the 95 % quantile of the absolute value of the residuals in the corresponding region in the train set.

Results on the Test Set We evaluate the three reconciliation procedures bottom-up, OLS and GTOP by summing up their losses (4) on the test set, giving the totals L_{BU}, L_{OLS} and L_{GTOP}, which we compare to the sum of the losses $\hat{L}$ for the unreconciled forecasts by computing the *percentage of improvement* $(\hat{L} - L)/\hat{L} \times 100\,\%$ for $L \in \{L_{\mathrm{BU}}, L_{\mathrm{OLS}}, L_{\mathrm{GTOP}}\}$. It remains to define the weighting factors a_1, a_2 and a_{tot} in the loss, and the scales σ and τ for the noise variables. We consider five different sets of weighting factors, where the first three treat the two regions symmetrically (by assigning them both weight 1), which seems the most realistic, and the other two respectively introduce a slight and a very large asymmetry between regions, which is perhaps less realistic, but was necessary to find a case where OLS would beat GTOP. Finally, we always let $\sigma + \tau = 2$, so that the scale of the noise is (somewhat) comparable between experiments. Table 1 shows the median over 100 repetitions of the experiment of the percentages of improvement.

First, we remark that, in all but one of the cases, GTOP reconciliation performs at least as good as or better than OLS and bottom-up, and GTOP is the only of the three methods that always improves on the unreconciled forecasts, as was already guaranteed by Theorems 1 and 2. Moreover, the only instance where OLS performs better than GTOP ($a_1 = 1, a_2 = a_{\mathrm{tot}} = 20$), appears to be the least realistic, because the regions are treated very asymmetrically. For all cases where the weights are equal ($a_1 = a_2 = a_{\mathrm{tot}} = 1$), we see that GTOP and OLS perform exactly the same, which, in light of the equivalence discussed in Sect. 2.3, suggest that the prediction intervals that make up $\mathscr{B}$ do not have a large effect in this case.

Table 1 Percentage of improvement over unreconciled forecasts for simulated data

σ	τ	a_1	a_2	a_{tot}	Bottom-up (%)	OLS (%)	GTOP (%)
0	2	1	1	1	−13.97	**0.40**	**0.40**
0	2	1	1	2	−19.47	−2.35	**0.47**
0	2	1	1	10	−26.62	−7.46	**0.12**
0	2	2	1	5	−22.49	−4.55	**0.23**
0	2	1	20	20	−26.96	−2.69	**0.13**
1	1	1	1	1	−55.51	**5.75**	**5.75**
1	1	1	1	2	−75.09	−6.02	**4.54**
1	1	1	1	10	−141.66	−30.39	**2.41**
1	1	2	1	5	−92.47	−14.09	**3.13**
1	1	1	20	20	−77.18	−2.51	**1.22**
2	0	1	1	1	−94.92	**29.85**	**29.85**
2	0	1	1	2	−184.23	17.57	**34.76**
2	0	1	1	10	−996.22	−79.58	**44.75**
2	0	2	1	5	−319.30	1.32	**35.48**
2	0	1	20	20	−183.95	**23.54**	16.19

Secondly, we note that the unreconciled predictions are much better than the bottom-up forecasts. Because bottom-up and the unreconciled forecasts make the same predictions $\hat{Y}_1$ and $\hat{Y}_2$ for the two regions, this means that the difference must be in the prediction $\hat{Y}_{\text{tot}}$ for the sum of the regions, and so, indeed, the method described in (9) and (10) makes significantly better forecasts than the simple bottom-up forecast $\hat{Y}_1 + \hat{Y}_2$. We also see an overall trend that the scale of the percentages becomes larger as σ increases (or τ decreases), which may be explained by the fact that forecasting Y_{tot} becomes relatively easier, so that the difference between $\hat{Y}_{\text{tot}}$ and $\hat{Y}_1 + \hat{Y}_2$ gets bigger, and the effect of reconciliation gets larger.

3.2 EDF Data

To illustrate how GTOP reconciliation works on real data, we use electricity demand data provided by Électricité de France (EDF). The data are historical demand records ranging from 1 July 2004 to 31 December 2009, and are sampled each 30 min. The total demand is split up into $K = 17$ series, each representing a different electricity tariff. The series are divided into a calibration set (from 1 July 2004 to 31 December 2008) needed by the prediction models, and a validation set (from 1 January 2009 to the end) on which we will measure the performance of GTOP.

Every night at midnight, forecasts are required for the whole next day, i.e. for the next 48 time points. We use a non-parametric function-valued forecasting model by Antoniadis et al. [1], which treats every day as a 48-dimensional vector. The model uses all past data on the calibration and validation sets. For every past day d, it considers day $d + 1$ as a candidate prediction and then it outputs a weighted combination of these candidates in which the weight of day d depends on its similarity to the current day. This forecasting model is used independently on each of the 17 individual series and also on the aggregate series (their total).

We now use bottom-up, OLS and GTOP to reconcile the individual forecasts. Similarly to the simulations in the previous section, the prediction intervals $B_1, \ldots, B_K$ for GTOP are computed as quantiles of the absolute values of the residuals, except that now we only use the past 2 weeks of data from the validation set, and we use the q-th quantile, where q is a parameter. We note that, for the special case $q = 0\,\%$, we would expect B_k to be close to 0, which makes GTOP very similar to the bottom-up forecaster. (See Example 2.)

For each of the three methods, the percentages of improvement on the validation set are computed in the same way as in the simulations in the previous section. Table 2 shows their values for different choices of realistic weighting factors, using $q = 10\,\%$ for GTOP, which was found by optimizing for the weights $a_{\text{tot}} = 17$ and $a_k = 1$ ($k = 1, \ldots, 17$), as will be discussed below.

We see that GTOP consistently outperforms both the bottom-up and the OLS predictor, with gains that increase with a_{tot}. Unlike in the simulations, however, the bottom-up forecaster is comparable to or even better than the unreconciled forecasts in terms of its percentage of improvement. In light of Remark 2, we have therefore

Table 2 Percentage of improvement over unreconciled forecasts for EDF data, using $q = 10\,\%$ for GTOP

a_1	a_2	a_{tot}	Bottom-up (%)	OLS (%)	GTOP (%)
1	1	1	0.98	0.19	**1.62**
1	1	2	1.27	0.27	**1.96**
1	1	10	1.65	0.38	**2.41**
1	1	17	1.70	0.40	**2.47**

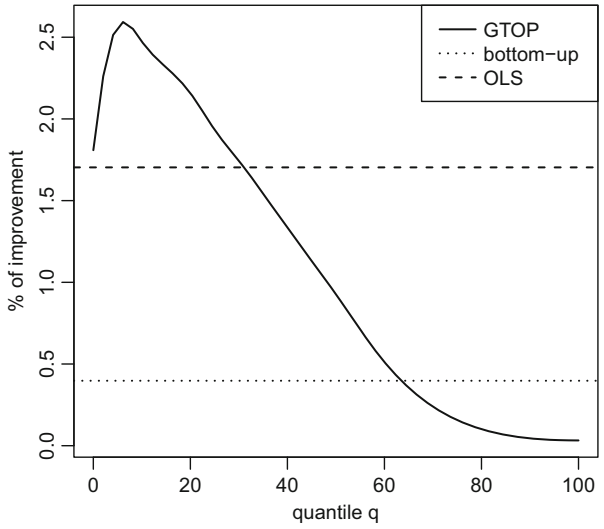


Fig. 4 Percentages of improvement as a function of q for GTOP, OLS and bottom-up, using $a_{\text{tot}} = 17$ and $a_k = 1$ $(k = 1, \ldots, 17)$

considered simply replacing our prediction for the total by the bottom-up predictor, which would make reconciliation unnecessary. However, when, instead of looking at the percentage of improvement, we count the times when the unreconciled forecaster gives a better prediction for the total than the bottom-up forecaster, we see that this is 56 %, so the unreconciled forecaster does predict better than bottom-up slightly more than half of the time, and consequently there is something to gain by using it. As will be discussed next, this does make it necessary to use a small quantile q with GTOP.

Choosing the Quantile To determine which quantile q to choose for GTOP, we plot its percentage of improvement as a function of q for the case $a_{\text{tot}} = 17$ and $a_k = 1$ (see Fig. 4). We see that all values below 60 % improve on the bottom-up forecaster, and that any value below 30 % improves on OLS. The quantile $q \approx 10\,\%$ gives the best results, and, for ease of comparison, we use this same value in all the experiments reported in Table 2. In light of the interpretation of the prediction intervals, it might appear surprising that the optimal value for q would be so small. This can be explained by the fact that the unreconciled forecasts are only better than

bottom-up 56 % of the time, so that a small value of q is beneficial, because it keeps the GTOP forecasts close to the bottom-up ones.

4 Discussion

We now turn to several subjects that we have not been able to treat in full detail in the previous parts of the paper. First, in Sect. 4.1, we discuss appropriate choices for the weighting factors that determine the loss. Then, in Sect. 4.2, we discuss how estimating proportions in top-down forecasting is complicated by the presence of independent variables, and, finally, in Sect. 4.3, we conclude with a summary of the paper and directions for future work.

4.1 How to Choose the Weighting Factors in the Loss

In the General Forecasting Competition 2012 [10], a two-level hierarchy was considered with weights chosen as $a_k = 1$ for $k = 1, \dots, K$ and $a_{\text{tot}} = K$, so that the forecast for the total receives the same weight as all the regional forecasts taken together. At first sight this appears to make sense, because predicting the total is more important than predicting any single region. However, one should also take into account the fact that the errors in the predictions for the total are on a much larger scale than the errors in the predictions for the regions, so that the total is already a dominant factor in the loss without assigning it a larger weight.

To make this argument more precise, let us consider a simplified setting in which we can compute expected losses. To this end, define random variables $\epsilon_k = Y_k - \hat{Y}_k$ for the regional prediction errors at time τ and assume that, conditionally on all prior observations, (1) $\epsilon_1, \dots, \epsilon_K$ are uncorrelated; and (2) the regional predictions are unbiased, so that $\mathbb{E}\{\epsilon_k\} = 0$. Then the expected losses for the regions and the total are

$$\mathbb{E}\,\ell_k(Y_k, \hat{Y}_k) = a_k\,\mathbb{E}\left\{(Y_k - \hat{Y}_k)^2\right\} = a_k\,\mathrm{Var}(\epsilon_k) \qquad (k = 1, \dots, K)$$

$$\mathbb{E}\,\ell_{\text{tot}}(Y_{\text{tot}}, \hat{Y}_{\text{tot}}) = a_{\text{tot}}\,\mathbb{E}\left\{(\sum_k Y_k - \sum_k \hat{Y}_k)^2\right\} = a_{\text{tot}}\,\mathrm{Var}\left(\sum_{k=1}^{K} \epsilon_k\right) = a_{\text{tot}} \sum_{k=1}^{K} \mathrm{Var}(\epsilon_k),$$

where $\mathrm{Var}(Z)$ denotes the variance of a random variable Z.

We see that, even without assigning a larger weight to the total, $\mathbb{E}\,\ell_{\text{tot}}(Y_{\text{tot}}, \hat{Y}_{\text{tot}})$ is already of the same order as the sum of all $\mathbb{E}\,\ell_k(Y_k, \hat{Y}_k)$ together, which suggests that choosing a_{tot} to be 1 or 2 (instead of K) might already be enough to assign sufficient importance to the prediction of the total.

4.2 The Limits of Top-Down Forecasting

As a thought experiment, think of a noiseless situation in which

$$Y_1[t] = X[t], \qquad Y_2[t] = X[t] + 1, \qquad Y_{\text{tot}}[t] = Y_1[t] + Y_2[t] = 2X[t] + 1$$

for some independent variable ($X[t]$). Suppose we use the following top-down approach: first we estimate $Y_{\text{tot}}[\tau]$ by $\hat{Y}_{\text{tot}}[\tau]$ and then we make regional forecasts as $\hat{Y}_1[\tau] = \lambda \hat{Y}_{\text{tot}}[\tau]$ and $\hat{Y}_2[\tau] = (1-\lambda)\hat{Y}_{\text{tot}}[\tau]$ according to a constant λ that we will estimate. Because we are in a noise-free situation, let us assume that estimation is easy, and that we can predict $Y_{\text{tot}}[\tau]$ exactly: $\hat{Y}_{\text{tot}}[\tau] = Y_{\text{tot}}[\tau]$. Moreover, we will assume we can choose λ optimally as well. Then how should λ be chosen? We want to fit:

$$\lambda = \frac{Y_1[t]}{Y_{\text{tot}}[t]} = \frac{1}{2} - \frac{1}{4X[t]+2}, \qquad 1-\lambda = \frac{Y_2[t]}{Y_{\text{tot}}[t]} = \frac{1}{2} + \frac{1}{4X[t]+2}.$$

But now we see that the optimal value for λ depends on $X[t]$, which is not a constant over time! So estimating λ based on historical proportions will not work in the presence of independent variables.

4.3 Summary and Future Work

Unlike previous approaches, like bottom-up, top-down and generalized least-squares forecasting, we propose to split the problem of hierarchical time series forecasting into two parts: first one constructs the best possible forecasts for the time series without worrying about aggregate consistency or theoretical restrictions like unbiasedness, and then one uses the GTOP reconciliation method proposed in Sect. 2 to turn these forecasts into aggregate consistent ones. As shown by Theorems 1 and 2, GTOP reconciliation can only make any given set of forecasts better, and the less consistent the given forecasts are, the larger the improvement guaranteed by GTOP reconciliation.

Our treatment is for the squared loss only, but, as pointed out in Sect. 2, Theorems 1 and 2 readily generalize to any other loss that is based on a Bregman divergence, like for example the Kullback-Leibler divergence. It would be useful to work out this generalization in detail, including the appropriate choice of optimization algorithm to compute the resulting Bregman projection.

In the experiments in Sect. 3, we have proposed some new methods for coming up with the initial forecasts, but although they demonstrate the benefits of GTOP reconciliation, these approaches are still rather simple. In future work, it would therefore be useful to investigate more advanced ways of coming up with initial forecasts, which allow for even more information to be shared between different

time series. For example, it would be natural to use a Bayesian approach to model regions that are geographically close as random instances of the same distribution on regions.

Finally, there seems room to do more with the prediction intervals for the GTOP reconciled predictions as defined in (3). It would be interesting to explore data-driven approaches to constructing these intervals, like for example those proposed by Antoniadis et al.[2].

Acknowledgements The authors would like to thank Mesrob Ohannessian for useful discussions, which led to the closed-form solution for the GTOP predictions in Example 1. We also thank two anonymous referees for useful suggestions to improve the presentation. This work was supported in part by NWO Rubicon grant 680-50-1112.

References

1. Antoniadis, A., Brossat, X., Cugliari, J., & Poggi, J.-M. (2012). Prévision d'un processus à valeurs fonctionnelles en présence de non stationnarités. Application à la consommation d'électricité. *Journal de la Société Française de Statistique, 153*(2), 52–78.
2. Antoniadis, A., Brossat, X., Cugliari, J., & Poggi, J. M. (2013). Une approche fonctionnelle pour la prévision non-paramétrique de la consommation d'électricité. Technical report oai:hal.archives-ouvertes.fr:hal-00814530, Hal, Avril 2013. http://hal.archives-ouvertes.fr/hal-00814530.
3. Borges, C. E., Penya, Y. K., & Fernández, I. (2013). Evaluating combined load forecasting in large power systems and smart grids. *IEEE Transactions on Industrial Informatics, 9*(3), 1570–1577.
4. Byron, R. P. (1978). The estimation of large social account matrices. *Journal of the Royal Statistical Society, Series A, 141*, 359–367.
5. Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge: Cambridge University Press.
6. Chen, B. (2006). A balanced system of industry accounts for the U.S. and structural distribution of statistical discrepancy. Technical report, Bureau of Economic Analysis. http://www.bea.gov/papers/pdf/reconciliation_wp.pdf.
7. Fliedner, G. (1999). An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers & Operations Research, 26*, 1133–1149.
8. Granger, C. W. J. (1988). Aggregation of time series variables—A survey. Discussion paper 1, Federal Reserve Bank of Minneapolis, Institute for Empirical Macroeconomics. http://www.minneapolisfed.org/research/DP/DP1.pdf.
9. Hazan, E., Agarwal, A., & Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning, 69*(2–3), 169–192.
10. Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012. *International Journal of Forecasting, 30*, 357–363.
11. Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis, 55*, 2579–2589.
12. Lobo, M. S., Vandenberghe, L., Boyd, S., & Lebret, H. (1998). Applications of second-order cone programming. *Linear Algebra and its Applications, 284*(1–3), 193–228.
13. Lütkepohl, H. (2009). Forecasting aggregated time series variables: A survey. Working paper EUI ECO: 2009/17, European University Institute. http://hdl.handle.net/1814/11256.

14. MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*, 305–325.
15. Rockafellar, R. T. (1970). *Convex analysis*. Princeton: Princeton University Press.
16. Stone, R., Champernowne, D. G., & Meade, J. E. (1942). The precision of national income estimates. *The Review of Economic Studies, 9*(2), 111–125.
17. Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B, 58*(1), 267–288.
18. White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817–838.

The BAGIDIS Distance: About a Fractal Topology, with Applications to Functional Classification and Prediction

Rainer von Sachs and Catherine Timmermans

Abstract The BAGIDIS (semi-) distance of Timmermans and von Sachs (BAGIDIS: statistically investigating curves with sharp local patterns using a new functional measure of dissimilarity. Under revision. http://www.uclouvain.be/en-369695.html. ISBA Discussion Paper 2013-31, Université catholique de Louvain, 2013) is the central building block of a nonparametric method for comparing curves with sharp local features, with the subsequent goal of classification or prediction. This semi-distance is data-driven and highly adaptive to the curves being studied. Its main originality is its ability to consider simultaneously horizontal and vertical variations of patterns. As such it can handle curves with sharp patterns which are possibly not well-aligned from one curve to another. The distance is based on the signature of the curves in the domain of a generalised wavelet basis, the Unbalanced Haar basis. In this note we give insights on the problem of stability of our proposed algorithm, in the presence of observational noise. For this we use theoretical investigations from Timmermans, Delsol and von Sachs (J Multivar Anal 115:421–444, 2013) on properties of the fractal topology behind our distance-based method. Our results are general enough to be applicable to any method using a distance which relies on a fractal topology.

1 Introduction

In Timmermans and von Sachs [9], a new method for the statistical analysis of differences between curves with sharp local patterns proposes a distance measure between curves which relies on an Unbalanced Haar wavelet decomposition obtained using a modified version of the algorithm by Fryzlewicz [3]. This algorithm allows to describe a curve through a set of points in the so-called breakpoints-details (b,d) plane, where the breakpoints account for the location of level changes in the curve and details account for the amplitude of the latter. The goal has been to

R. von Sachs (✉) • C. Timmermans
Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, 20 Voie du Roman Pays, B-1348 Louvain-la-Neuve, Belgium
e-mail: rvs@uclouvain.be; catherine.timmermans@uclouvain.be

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_16

propose a method capable of statistically investigating datasets of curves with sharp peaks that might be misaligned, thereby overcoming limitations of existing methods. We also recall our paradigm of a "robust" one-step method in order to avoid any preprocessing step – such as Dynamic Time Warping [4] – which would align the curves prior to comparison, for the purpose of, e.g., classification or prediction. This is in particular in order to be able to detect differences between curves due to the presence of features which are actually not aligned.

In this note we address the question of stability of the (b,d) point representation (=the "signature") associated to a given curve if there is some additional noise. In our previous work [9], we have been able to show that this (b,d) point representation is stable in the absence of noise, leading to good "theoretical" performance. This stability has been due to the use of an unambiguous ordered representation of each curve in the (b,d) plane. Hence, it is of importance to examine what happens in the realistic situation of noise.

More particularly, we address the question of robustness of our method towards the following phenomenon of potential feature confusion. Due to noise, curves with a sufficiently similar structure of local events (such as jumps, peaks or troughs) might accidentally be considered as dissimilar because local information might be encoded in a suboptimal, i.e. not unambiguous, way. We investigate the theoretical properties of the BAGIDIS semi-distance in order to handle this situation. With this, we support empirical findings reported in previous work of ours ([7, 9] and [10]) when using the local methods used to process the set of BAGIDIS semi-distances computed on our noisy datasets. Here with "local" we mean essentially nonparametric methods which localise the information in the given data set by using only a fraction of the observations given in a local neighbourhood around the point (or region) of interest. Prominent examples are methods based on Nearest Neighbors (NN), kernels, or Multidimensional Scaling (MDS) which turned out to effectively be sufficiently robust in order to cope with the aforementioned feature confusion. In this article we shed some light on why this happens, leading to the desirable property that makes BAGIDIS better than competitors (e.g. the Euclidean distance) in case of misaligned sharp patterns. We note that the opposite problem of classifying accidentally as similar those curves that would actually be dissimilar in the absence of noise is not the purpose of this examination because this problem, inherent to any distance based classification algorithm, is not caused or amplified by the aforementioned problem of loss of unambiguous ordering due the presence of noise.

Section 2 of this paper reviews what is necessary to recall about the BAGIDIS method. At the end of this section we empirically expose what is behind our problem of robustness towards feature confusion. In Sect. 3 we present our theoretical treatment of the consistency of BAGIDIS in view of this problem, and in fact, any nonparametric method for functional comparisons using a distance which relies on a fractal topology. In particular we give arguments in favour of the BAGIDIS distance compared to the traditional Euclidean distance.

We finish this introduction by noting that extensions of our univariate work to higher dimensions have been provided in Timmermans and Fryzlewicz [8], in the particular context of classification of images.

2 Motivation for and Description of the BAGIDIS Algorithm

We consider series that are made of N regularly spaced measurements of a continuous process (i.e. a curve). Those series are encoded as vectors in $\mathbb{R}^N$. There exists numerous classical methods allowing to measure distances or semi-distances between such series coming from the discretization of a curve. Note that, according to [2], d is a semi-distance on some space $\mathscr{F}$ if

- $\forall \mathbf{x} \in \mathscr{F}, \quad d(\mathbf{x}, \mathbf{x}) = 0$
- $\forall \mathbf{x}^i, \mathbf{x}^j, \mathbf{x}^k \in \mathscr{F}, \quad d(\mathbf{x}^i, \mathbf{x}^j) \leq d(\mathbf{x}^i, \mathbf{x}^k) + d(\mathbf{x}^k, \mathbf{x}^j)$.

Semi-distances are often used [1] when one is interested in comparing the shapes of some groups of curves, but not in comparing their mean level.

2.1 Existing Distance-Based Approaches

We very briefly recall a non-exhaustive collection of some most popular existing distance-based approaches and discuss their properties: (i) Classical l_p distances and their principal components-based extension [6]; (ii) Functional semi-distances [2], taking into account the notion of neighborhood in point-to-point comparisons; (iii) wavelet-based distances: comparing the coefficients of well-suited basis function expansions. Whereas in our work [9] we give a detailed appreciation of these different approaches, here in order to motivate our approach, we contain ourselves to recall some basic visually supported features: in Fig. 1a, we recall that by methods of type (i) the ordering of the series measurements is not taken into account so that the evolutions of two series cannot be compared; (ii) functional approaches happen to fail when dealing with curves with local sharp discontinuities that might not be well aligned from one curve to another one, as illustrated in Fig. 1b; and finally, more particularly, (iii) encoding significant features into a wavelet basis which is not both simultaneously orthogonal and non-dyadic in nature, can lead to shortcomings as illustrated with classical (dyadic) Haar wavelets in Fig. 1c).

2.2 At the Core of the BAGIDIS Method

A major originality of our method to encode closeness of series having a similar discontinuity that is only slightly shifted relies on projections on *orthogonal* basis

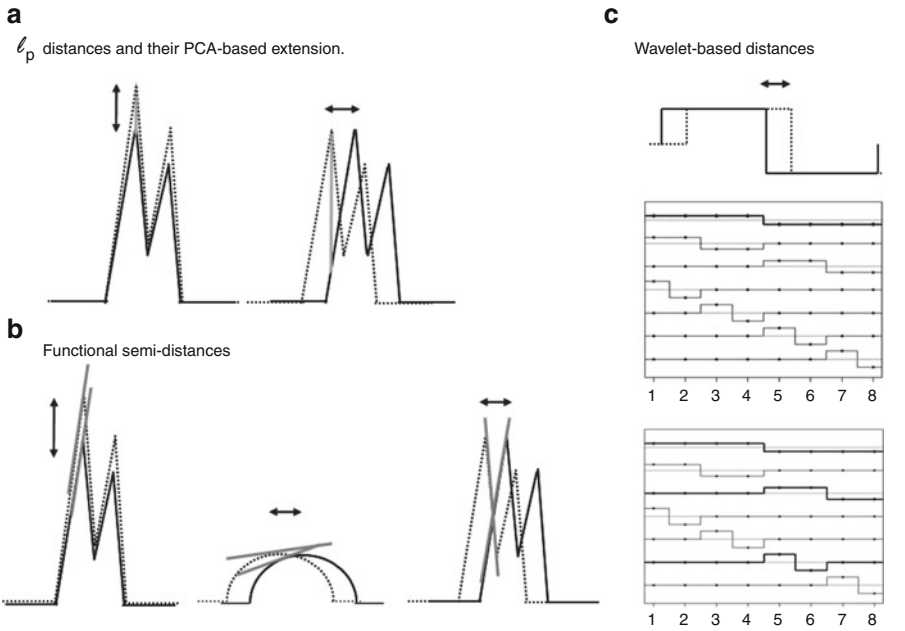


Fig. 1 Schematic illustration of the difficulty for classical methods to take into account horizontal variations of curves. (**a**) l_p distances and PCA-based distances compare curves at each point of measurement, so that patterns that are shifted horizontally are measured distant. An illustrative component of the point-to-point distances is displayed in gray in (**a**). (**b**) Comparing derivatives, as common functional methods do, allows to overcome that difficulty if the patterns are smooth but fails with sharp shifted patterns. Illustrative derivatives are indicated in light gray in (**b**). (**c**) Wavelet-based methods capture well the sharp patterns, but their encoding in the basis expansion differs highly if the location of the discontinuity changes a bit: in (**c**), we illustrate classical Haar-basis expansions of two shifted step function, the only basis vector associated to a non-zero coefficient being highlighted in bold

functions that are *different* from one series to another although providing for a *hierarchy* (essential for the ability of comparing the curve expansions). We describe the main ideas of the method as follows.

(i) We consider a collection of discrete-time series each of which can be pictured as regularly spaced observations of a curve. We note that patterns in a series can be described as a set of level changes.

(ii) We find an *optimal* basis for each curve. As a first step, we want to expand each series in a basis that is best suited to it, in the sense that its first basis vectors should carry the main features of the series, while subsequent basis vectors support less significant patterns. In that respect, we are looking for a basis that is organized in a *hierarchical* way. As a consequence, there will be a particular basis associated to each series. As the series are thought of as described by their level changes, we will consider that the meaningful features for describing

them are both locally important level changes, such as jumps, peaks or troughs, and level changes affecting a large number of data, i.e. discontinuities of the mean level. From this point of view, Unbalanced Haar wavelet bases, to be defined in Sect. 2.3, are the ideal candidates for our expansion. We benefit from their *o*rthogonality property to have no unambiguity in encoding.

(iii) We take advantage of the *hierarchy* of those bases. Given this, we make use of the BAGIDIS semi-distance which is at the core of the BAGIDIS methodology. This semi-distance takes advantage of the hierarchy of the well-adapted unbalanced Haar wavelet bases: basis vectors of similar rank in the hierarchy and their associated coefficients in the expansion of the series are compared to each other, and the resulting differences are weighted according to that rank. This is actually a clue for decrypting the name of the methodology, as the name BAGIDIS stands for *BAsis GIving DIStances*. Section 2.4 recalls the definition of the BAGIDIS semi-distance from [9].

A subsequent interest lies in obtaining some information on the relative importance of horizontal and vertical variations, and on their localization, in order to statistically diagnose whether groups of curves do actually differ and how. Numerous applications to supervised and unsupervised classification and prediction, in the framework of spectroscopy for metabonomic analysis, on analysing solar irradiance time series or on image description and processing, can be found in [7, 9, 10] and [8].

2.3 Finding an Optimal Basis for Each Curve

Given a set of M series $\mathbf{x}^{(i)}$ in $\mathbb{R}^N$, $i = 1 \ldots M$, each of which consists in discrete regularly spaced measurements of a (different) curve, the goal is now to expand each of the series into the Unbalanced Haar wavelet basis that is best suited to it.

2.3.1 Definition of the Unbalanced Haar Wavelet Bases

Unbalanced Haar wavelet bases [5] are orthonormal bases that are made up of one constant vector and a set of Haar-like, i.e. *up-and-down*-shaped, orthonormal wavelets whose discontinuity point between positive and negative parts is not necessarily located at the middle of its support. Using the notation of [3], the general mathematical expression of those Haar-like wavelets is given by

$$\begin{aligned} \boldsymbol{\phi}_{e,b,s}(t) = &\left(\frac{1}{b-s+1} - \frac{1}{e-s+1}\right)^{1/2} \cdot \mathbf{1}_{s \le t \le b} \\ &-\left(\frac{1}{e-b} - \frac{1}{e-s+1}\right)^{1/2} \cdot \mathbf{1}_{b+1 \le t \le e}, \end{aligned} \tag{1}$$

where $t = 1 \ldots N$ is a discrete index along the abscissa axis, and where s, b and e stands for *start*, *breakpoint* and *end* respectively, for some well chosen values of s, b and e along the abscissa axis (see also Figure 1 of [9]). Each wavelet $\boldsymbol{\phi}_{e,b,s}(t)$ is thus associated with a level change from one observation (or group of observations) to the consecutive one, and the projection of the series $\mathbf{x}(t)$ on the wavelet $\boldsymbol{\phi}_{e,b,s}(t)$ encodes the importance of the related level change in the series.

2.3.2 The Basis Pursuit Algorithm and the Property of Hierarchy

In 2007, P. Fryzlewicz [3] proposed an algorithm for building the unbalanced Haar wavelet basis $\{\boldsymbol{\phi}_k\}_{k=0\ldots N-1}$ that is best suited to a given series, according to the principle of hierarchy – namely, the vectors of this basis and their associated coefficients are ordered using information that builds on the importance of the level change they encode for describing the global shape of the series. He called it the *bottom-up unbalanced Haar wavelet transform*, here-after BUUHWT. The resulting expansion is organized in a hierarchical way and avoids the dyadic restriction that is typical for classical wavelets. The family of unbalanced Haar wavelets is thus really adaptive to the shape of the series. A chart-flow diagram of the actual BUUHWT algorithm can also be found in Section 2.1 of [9].

2.3.3 An Example of Bottom-Up Unbalanced Haar Wavelet expansion

Figure 2, *left*, shows the BUUHWT expansion obtained for one particular series. As hoped for and observed by looking at the location of the discontinuity points b between positive and negative parts of the wavelets, the first non-constant vectors support the largest discontinuities of the series and encode therefore the highest peak of the series. Subsequent vectors point to smaller level changes while the few last vectors correspond to zones where there is no level change – as indicated by the associated zero coefficient.

2.3.4 Representing the Series in the *b-d* Plane

Let us denote the optimal Unbalanced Haar wavelet expansion of a series $\mathbf{x}^{(i)}$ as follows:

$$\mathbf{x}^{(i)} = \sum_{k=0}^{N-1} d_k^{(i)} \boldsymbol{\psi}_k^{(i)},$$

where the coefficients $d_k^{(i)}$ are the projections of $\mathbf{x}^{(i)}$ on the corresponding basis vectors $\boldsymbol{\psi}_k^{(i)}$ (i.e. the *detail* coefficients) and where the set of vectors $\{\boldsymbol{\psi}_k^{(i)}\}_{k=0\ldots N-1}$ is the Unbalanced Haar wavelet basis that is best suited to the series $\mathbf{x}^{(i)}$, as obtained

using the BUUHWT algorithm. Let us also denote $b_k^{(i)}$, the breakpoint of the wavelet $\boldsymbol{\psi}_k^{(i)}$, $k = 1 \ldots N-1$, i.e. the value of the highest abscissa where the wavelet $\boldsymbol{\psi}_k^{(i)}$ is strictly positive. An interesting property of the basis $\{\boldsymbol{\psi}_k^{(i)}\}_{k=0\ldots N-1}$, that has been proved by [3], is the following:

Property: The ordered set of breakpoints $\{b_k^{(i)}\}_{k=0\ldots N-1}$ determines the basis $\{\boldsymbol{\psi}_k^{(i)}\}_{k=0\ldots N-1}$ uniquely.

Consequently, the set of pairs $(b_k^{(i)}, d_k^{(i)})_{k=1\ldots N-1}$ determines the shape of the series $\mathbf{x}^{(i)}$ uniquely (i.e., it determines the series, except for a change of the mean level of the series, that is encoded by the additional coefficient $d_0^{(i)}$). This allows us to represent any series $\mathbf{x}$ in the *b-d* plane, i.e. the plane formed by the breakpoints and the details coefficients. An example of such a representation is presented in Fig. 2, *right*.

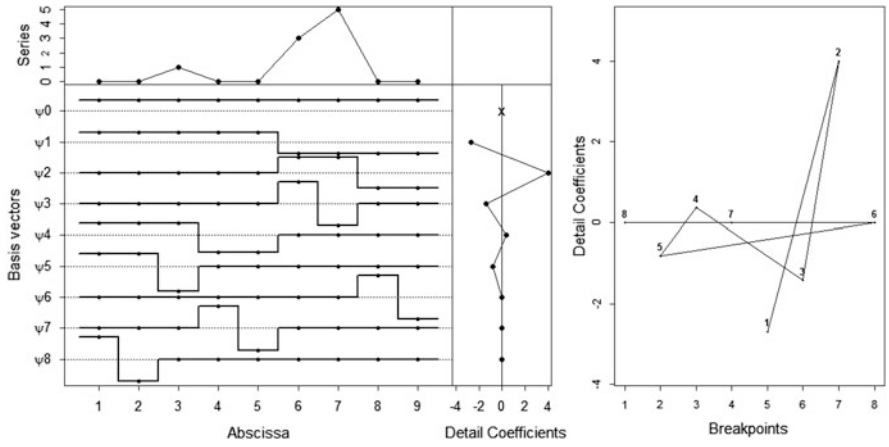


Fig. 2 *Left:* **Illustration of a BUUHWT expansion.** In the upper part of the figure we plot the series. The corresponding abscissa axis at the very bottom is common for that graph and for the graph of basis vectors. The main part of the figure shows the basis vectors of the Unbalanced Haar wavelet basis that is best suited to the series (BUUHWT basis). These vectors are represented rank by rank, as a function of an index along the abscissa axis. *Dotted horizontal lines* indicate the level zero for each rank. Vertically, from top to bottom on the right hand side, we find the detail coefficients associated with the wavelet expansion. Each coefficient is located next to the corresponding basis vector. For graphical convenience, the value of the coefficient d_0 associated with the constant vector $\boldsymbol{\psi}_0$ is not indicated. *Right:* **Representation of a series in the $b-d$ plane.** The same series is plotted in the plane that is defined by the values of its breakpoints and its detail coefficients. Points are numbered according to their rank in the hierarchy

2.4 A Semi-distance Taking Advantage of the Hierarchy of the BUUHWT Expansions

We now have at our disposal all elements to measure the dissimilarity between two curves $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ that are both made of N consecutive observations and whose BUUHWT expansions have been computed. We proceed by calculating the weighted sum of partial distances in the *b-d* plane, i.e. the weighted sum of partial dissimilarities evaluated rank by rank:

$$d_p(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sum_{k=1}^{N-1} w_k \left\| \mathbf{y}_k^{(1)} - \mathbf{y}_k^{(2)} \right\|_p, \quad \text{with} \quad p = 1, 2, \ldots, \infty \tag{2}$$

where $\mathbf{y}_k^{(i)}$ stands for $(b_k^{(i)}, d_k^{(i)})$, $i = 1, 2$, so that $\left\| \mathbf{y}_k^{(1)} - \mathbf{y}_k^{(2)} \right\|_p$ is the distance between the pairs representing the curves at rank k in the *b-d* plane, as measured in any norm $p = 1, 2, \ldots \infty$, and where w_k is a suitably chosen weight function with K non-zero weights.

As becomes clear in the sequel, and in particular in Sect. 3, the number K is quite crucial as it encodes the number of features found to be significant for discrimination. Its choice, and more generally, the choice of the weight function, can be actually be done by cross validation, cf. [7]. In an unsupervised context, one can either use a priori information about how many ranks K should be necessary to encode important local information (such as prominent peaks in the observed signal), or in absence of this, use a uniformly not too badly working weight function which shows a smooth decay towards zero for higher ranks.

Note that we do not consider the rank $k = 0$ in our dissimilarity measure (2), as we are mainly interested in comparing the structures of the series rather than their mean level.

A more general version of definition (2) allows being flexible with respect to scaling effects. In the following, in order to take into account that sensitivity, an additional parameter $\lambda \in [0, 1]$ is introduced that balances between the differences of the details and the differences of the breakpoints:

$$d_p(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sum_{k=1}^{N-1} w_k \left(\lambda \left| b_k^{(1)} - b_k^{(2)} \right|^p + (1 - \lambda). \left| d_k^{(1)} - d_k^{(2)} \right|^p \right)^{1/p}. \tag{3}$$

We refer to [9] for more details on how to choose the parameter λ.

Property: As shown in [9], this measure of dissimilarity is a *semi-distance*.

2.5 Behavior of BAGIDIS in the Presence of Noise

Unambiguity in the absence of noise For the BAGIDIS semi-distance between two curves to be useful, it is necessary that no ambiguity occurs in the hierarchic encoding of the patterns in the algorithm. This constraint essentially translates into the idea that the description of the series should not require more than one level change of any given amplitude. What happens, in an ideal situation of no noise, if a series consists in an exactly symmetric pattern, such as a peak or trough, centered on the middle point of the series (i.e. an even pattern with respect to $\frac{N}{2}$)? We then would observe two level changes of exactly the same amplitude, but opposite signs, and encode the left and right part of the pattern. This ambiguity is solved by defining a left orientation for the BUUHWT algorithm, meaning that a level change of given amplitude that is located to the left of another one with the same amplitude is always encoded first.

The presence of noise The key to applicability of our method to noisy series is the use of a suitable weight function that efficiently filters the noise. In several examples, to be found in [9], we observed a good robustness of the method with respect to the presence of additive noise. Nevertheless, an artefact of the method might occur, for example, in case of a symmetric or quasi-symmetric peak, when there are two detail coefficients being close in absolute value but of opposite sign (compare the extreme case of equality as discussed above). In this case a permutation of the basis vectors might occur in the algorithm constructing the best suited basis, from one series to the other, due to a possible reordering of the amplitudes (in absolute values) of the mentioned noisy coefficients. Consequently a clustering of noisy series might lead to a spurious distinction into two groups.

The robustness property for classification/clustering In case there is a split into two groups A and B, in general, this will not invalidate the analysis of whether or not this split has been spurious: If there are no differences between the groups, a clustering will give two groups (due to the permutations occurring in both series A and series B) but each group will contain a mix of series A and B, so that we will conclude that there are no differences in the distributions of groups A and B. On the opposite, if there are significant differences between groups A and B, a clustering will give four groups, amongst which two are made of series A (the distinction between the groups being an artefact) and two are made of series B. We will thus conclude to the presence of an effect of the A-B factor.

This point is illustrated by the following test-scenarios. A more theoretical treatment what is behind this empirical property is given in Sect. 3.1 and higher.

- Scenario 1: we consider 2 groups of noisy series (noise $N(0, \sigma = 0.5)$) derived from model A and model B, with A different from B (in this example, we take A=(0,0,1,0,0,3,5,0,0) and B=rev(A)). The BUUHWT transforms of those series do suffer from permutations. We compute the dissimilarity matrix between all pairs of series and try to cluster them blindly and provide a multidimensional

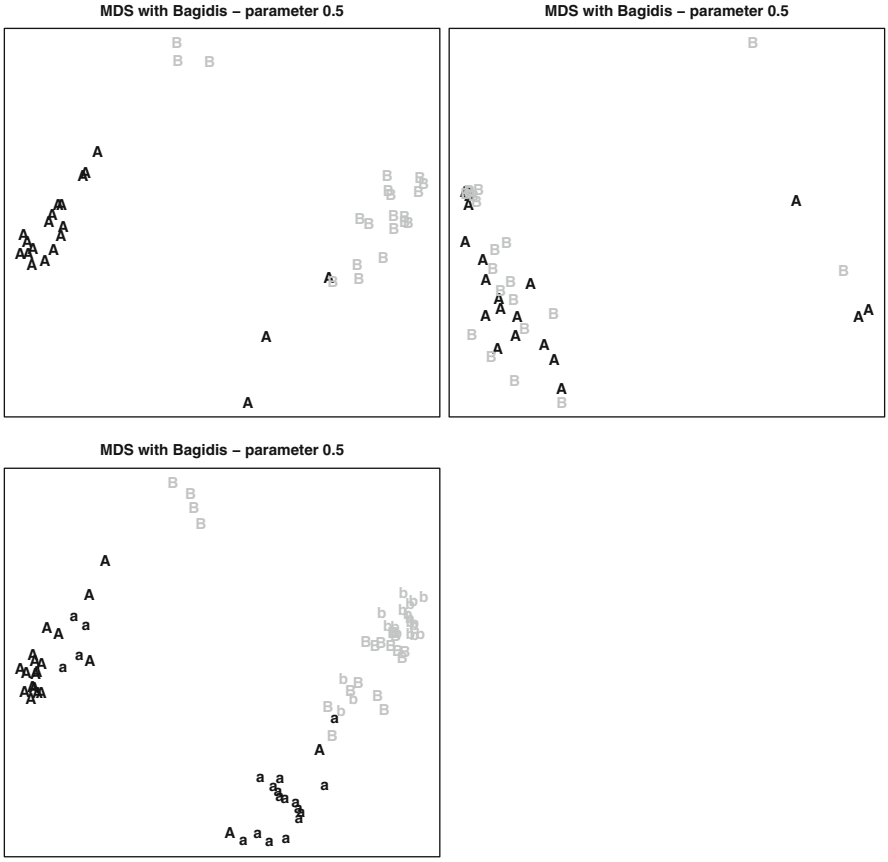


Fig. 3 Study of the applicability of BAGIDIS for noisy series in case of a possible reordering in the hierarchy of the patterns. *From left to right:* MDS representation of the test dataset resulting from Scenario 1 to Scenario 3

scaling (MDS) representation of the dataset (Fig. 3, *top left*). The groups are clearly linearly discriminable, despite a spurious difference which occurs.

- Scenario 2: we perform exactly the same test with A = B, so that we should not detect any effect of the model. Results are shown in Fig. 3, *top right*. Although two groups are distinguished, they both contain series from group A and from group B so that we cannot conclude a difference between the two models.
- Scenario 3: we perform the first test again, with A and B different, with half of the A-curves being shifted by 1 to the left and denoted by *a*, and with half of the B-curves being shifted by 1 to the left and denoted by *b*. We add a Gaussian noise with $\sigma = 0.5$ to all the curves. Results are shown in Fig. 3, *bottom left*. Again, a distinction between the groups (A+a) and (B+b) is visible, despite the spurious difference due to permutation.

In case of a regression or a classification problem, the key for a successful prediction model is that a given permutation should occur with a sufficiently high probability so that each spurious group contains enough representatives. Using a local approach for prediction (such as nonparametric functional regression or k-nearest neighbors) is then possible. In the next section we support this claim theoretically by use of the properties of our semi-distance which are derived from the fact that this distance actually induces a fractal topology [2, 7]. In [7] we derived results on the rate of convergence of the nonparametric functional regression estimate based on BAGIDIS, obtained under mild conditions. In accordance with the above mentioned intuition, the key for the estimator to converge relatively fast is that the probability to find curves in a small ball around the point of prediction is high enough.

3 On a Fractal Topology Induced by the BAGIDIS Semi-distance

We now discuss some properties of our semi-distance which are derived from the fact that this distance actually induces a fractal topology [2, 7]. This allows to address the question evoked at the end of Sect. 2 on the stability of our algorithm with respect to some potential feature confusion, in the sense of comparing closeness of two curves with closeness of their signatures in the (b,d)-plane.

Somewhat naively one could likely be asking the following question:

Property P1: Given a curve x sampled on a grid $\mathbb{N}_{[1;N]}$ and two noisy replications of this series, $x^{(1)} = x + noise$ and $x^{(2)} = x + noise$, we have that the signatures $s^{(1)} = \{(b_k^{(1)}, d_k^{(1)})\}_{k=0}^{N-1}$ and $s^{(2)} = \{(b_k^{(2)}, d_k^{(2)})\}_{k=0}^{N-1}$ are close to each other, at least when the number of sampled points tends to infinity and the noise tends to zero.

However, it is in general not possible to derive a framework for showing such an ideal property that would allow us to include the treatment of the 'breakpoint' components $\{b_k\}$ into an asymptotic result similar to the one of denoising curves by non-linearly selecting (via thresholding, e.g.) the 'best' wavelet coefficients $\{d_k\}$. Luckily, satisfying P1 is not a condition that needs to be fulfilled to address the feature permutation problem of our method. (For a further illustration of this point, we refer to our remark at the end of Sect. 3.1.) Actually, what is necessary for the method to be valid for subsequent classification, clustering or prediction applications, is rather the reverse statement:

Property P2: Given two signatures $s^{(1)} = \{(b_k^{(1)}, d_k^{(1)})\}_{k=0}^{N-1}$ *and* $s^{(2)} = \{(b_k^{(2)}, d_k^{(2)})\}_{k=0}^{N-1}$ *in the (b,d) plane, if* $s^{(1)}$ *and* $s^{(2)}$ *are close enough to each other, then* $x^{(1)}$ *is close to* $x^{(2)}$.

The idea behind this statement lies in the fact that when using local methods for processing the dataset of curves we can base ourselves on the following assumption:

Assumption A1: If the curves $x^{(1)}$ *and* $x^{(2)}$ *are close enough to each other, they will behave similarly with respect to the property we investigate (e.g. associated response in regression, class membership in classification or discrimination, ...).*

As mentioned in the first part of this paper, kernel methods, NN algorithms or radial-basis functions networks are very common examples of local methods, so do distance-based algorithms clearly fall into this category of methods. Property P2 ensures that Assumption A1 can be transposed to curves expressed in the (b,d) plane so that we can use local methods that rely on the BUUHWT expansions of curves.

3.1 Theoretical Results Based on Fractal Topologies

As discussed above, satisfying P2 is a condition that has to be satisfied when considering local methods. For ensuring the efficiency of such methods, the next step is to be sure that there is a sufficient density of observations around each point in the b-d plane at which we want to predict an associated response (class membership, cluster index, scalar response ...). Intuitively, we need to have in our dataset a sufficient number of neighbors that enter into the computation of the local algorithm at hand.

This "density of the space" is a topological property that is measured through the *small ball probability* of finding a curve $x^{(2)}$ around $x^{(1)}$, which is defined [2] as

$$\phi_{D,x^{(1)}}(h) = P(x^{(2)} \in B_D(x^{(1)}, h)),$$

where $B_D(x^{(1)}, h)$ is the ball of radius h centered on $x^{(1)}$ and defined according to the semi-distance D. More generally, for a given semimetric d, the small ball probability $\phi_{d,\chi}(h)$ measures the concentration of the functional variable $\boldsymbol{\chi}$, according to the topology defined by the semimetric.

Intuitively, the higher $\phi_{D,x^{(1)}}(h)$ in a small neighborhood of radius h, the more efficient the method will be in practice. In accordance with this observation, investigating the behavior of $\phi_{D,x^{(1)}}(h)$ when h tends to zero has been shown to be a key step for obtaining the rate of convergence of several local methods in functional data analysis (see for instance the list of references of our paper [7]). In those references, it is shown that a large range of methods are able to enjoy good rates of convergence at point $x^{(1)}$ as soon as the small ball probability function is such that Property P3 is valid for K small enough:

Property P3: there exists a positive constant C such that $\phi_{D,x^{(1)}}(h) \sim C\,h^K$, *when h tends to* 0.

When P3 is satisfied, one says that the (semi-) distance D induces a *fractal topology of order K*. As an illustration in the case of functional regression, Ferraty and Vieu [2] have shown that, under quite general conditions and with the assumption that P3 is satisfied for a certain K, one reaches the near-optimal rate of pointwise convergence of the regression operator r: this latter one is given by $\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+K}}$, with β being a Lipschitz parameter quantifying the smoothness of the regression operator r, and n the number of curves in the training set (this rate of convergence is to be compared with the rate of convergence for a nonparametric multivariate regression directly based on a N-dimensional variable: $\left(\frac{\log n}{n}\right)^{\frac{p}{2p+N}}$, with p the order of differentiability of r).

Theorem 1 of [7] states that Property P3 is satisfied for the BAGIDIS semi-distance in its general form of equation (3),

$$d^B_{w_k,\lambda}(s^{(1)}, s^{(2)}) = \sum_{k=0}^{N-1} w_k \left(\lambda \left|b_k^{(1)} - b_k^{(2)}\right|^2 + (1-\lambda)\left|d_k^{(1)} - d_k^{(2)}\right|^2\right)^{1/2},$$

under quite general conditions, with K being the number of non-zero weights w_k. In other words, we have shown that the BAGIDIS semi-distance induces a *fractal topology* of order K, with K the number of features (ranks) that enter into the comparison of the two curves using this distance.

Combined with property P2, Theorem 1 of [7] ensures good performance (in terms of good rates of convergence) when using local methods relying on BAGIDIS, provided that the number K of significant features in the curves is not too large. In particular this is illustrated in [7] in the setting of a functional regression using BAGIDIS, both theoretically (Theorem 2 of that paper) and by simulation studies.

Given this, we are able now to address the particular concern about (finite sample) situations, where significant local structure of the first and second rank is accidentally permuted due to the influence of noise: how would a subsequent "local"

discrimination method face this situation? We recall our motivating simulation examples of Sect. 2.5 for an illustration of this.

1. Following our intuition, a good performance relies on the fact that the information content is located in a sufficiently small number K of features, whether or not permutations do occur in the dataset.
2. Moreover, the above discussions tell that the key for the successful practical application of BAGIDIS in case of permutations, is that a permutation should occur with a sufficiently high probability so that each spurious groups contains enough representatives for the local method to be efficient in the (b,d) plane. Or conversely, the dataset should be large enough for each spurious group of curves to be sufficiently populated in the observational dataset. As soon as the dataset is large enough, Theorem 1 of [7] guarantees that this will be the case.
3. As for all prediction methods that use a model (link) to associate a response to an explanatory variable, some regularity conditions (e.g. Lipschitz parameter) on the link function are needed. This is the mathematical counterpart of Assumption A1.

We finally add an important remark on whether we would also need to examine the opposite scenario: could it happen that due to the nature of BAGIDIS to need to face spurious permutations, the probability of masking the difference between two curves in the presence of noise is higher than with any other distance-based method? We illustrate this again by discussion of a prominent example: Suppose two curves are meant to be different because they have two consecutive features of a large peak followed by a smaller peak for the first curve, and vice versa for the second. In order that the presence of noise masks this and puts the curves in the same class, the amplitudes for the two consecutive peaks would need to be very close, and consequently no more detected to be significantly different. (Otherwise the information in the d_k-coefficients alone would be sufficient to discriminate the two curves). Hence, in this situation, BAGIDIS is no more sensitive to the noise than e.g. the Euclidean distance.

3.2 Formalising the 'Stability Issue' of the BAGIDIS Algorithm: Proof of Property P2

We now give a more formal treatment of what is behind Property P2 and its proof using properties of fractal topologies.

We recall that we consider two series $x^{(1)}$and $x^{(2)}$ valued in $\mathbb{R}^N$, corresponding to curves sampled on the regular grid $\mathbb{N}_{[1;N]}$. Their expansions in the (b,d) plane are

denoted $s^{(1)} = \{(b_k^{(1)}, d_k^{(1)})\}_{k=0}^{N-1}$ and $s^{(2)} = \{(b_k^{(2)}, d_k^{(2)})\}_{k=0}^{N-1}$ with the conventional notation $b_0^{(1)} = b_0^{(2)} = 0$.

Proving P2 means showing that it is possible to have close proximity in the space of curves measured by a distance such as the Euclidean provided that there exists a small neighborhood around $s^{(1)}$ such that if $s^{(2)}$ is in this neighborhood then $d^{Eucl}(x^{(1)}, x^{(2)}) < \varepsilon$. Mathematically, Property P2 hence translates into

Property P2/math: for all $\varepsilon > 0$, there exists $K \in \mathbb{N}_{[1;N-1]}$ and $\delta > 0$ such that for the BAGIDIS distance based on K non-zero weights w_k, $d^B(s^{(1)}, s^{(2)}) < \delta$ implies $d^{Eucl}(x^{(1)}, x^{(2)}) < \varepsilon$,

which we are going to prove now.

Proof of Property P2

We first consider the simplified version

$$d_K^B(s^{(1)}, s^{(2)}) = \sum_{k=0}^{K} \left(\left| b_k^{(1)} - b_k^{(2)} \right|^2 + \left| d_k^{(1)} - d_k^{(2)} \right|^2 \right)^{1/2}$$

as the measure of the proximity in the (b,d) plane. Recall here, that $K \leq N - 1$.

If there exists $k \in \mathbb{N}_{[1;K]}$ such that $|b_k^{(1)} - b_k^{(2)}| > 1$, where 1 is the sampling step of the grid $\mathbb{N}_{[1;N]}$ on which the curve is observed, then we have

$$d_K^B(s^{(1)}, s^{(2)}) = \sum_{k=0}^{K} \left(\left| b_k^{(1)} - b_k^{(2)} \right|^2 + \left| d_k^{(1)} - d_k^{(2)} \right|^2 \right)^{1/2} \geq \sum_{k=0}^{K} |b_k^{(1)} - b_k^{(2)}| \geq 1.$$

Consequently, it is sufficient to choose $\delta < 1$, so that we have that $b_k^{(1)} = b_k^{(2)}$ for all k in $1, \ldots K$, and such that

1. The expression of the distance reduces to $d_K^B(s^{(1)}, s^{(2)}) = \sum_{k=0}^{K} |d_k^{(1)} - d_k^{(2)}|$.
2. The basis vectors $\psi_k^{(1)}$ and $\psi_k^{(2)}$ are, exactly the same up to rank K. This is because the ordered set of breakpoints $\{(b_k^{(i)})\}_{k=0}^{K}$ combined with the requirements of up-and-down shape, orthonormality and multiscale construction, allows to reconstruct the associated basis vectors $\psi_k^{(i)}$ using a top-down procedure, up to rank K.

We denote $\left\{\psi_k^{(1)}\right\}_{k=0}^K = \left\{\psi_k^{(2)}\right\}_{k=0}^K = \{\psi_k\}_{k=0}^K$, and $\hat{x}^{(1)} = \sum_{k=0}^K d_k^{(1)}\psi_k$ and $\hat{x}^{(2)} = \sum_{k=0}^K d_k^{(2)}\psi_k$. We observe:

1. Using those results and the property of energy conservation in wavelet expansions, we have

$$d^{Eucl}(\hat{x}^{(1)}, \hat{x}^{(2)}) = \sqrt{\sum_{k=0}^{K} |d_k^{(1)} - d_k^{(2)}|^2},$$

when $\delta < 1$. Consequently: for all $\varepsilon_1 > 0$, there exists $\delta_1 \in]0; 1[$ such that if $d_K^B(s^{(1)}, s^{(2)}) < \delta$ then $d^{Eucl}(\hat{x}^{(1)}, \hat{x}^{(2)}) < \varepsilon_1$.
2. By construction of the BUUHWT algorithm, the energy of the signal is concentrated in the first ranks of the expansion. Thus, $\hat{x}^{(1)}$ and $\hat{x}^{(2)}$ may thus be interpreted as the wavelet approximation of $x^{(1)}$and $x^{(2)}$ under some hard thresholding rule such that $d_k = 0$ for all $k > K$. Using the results of [3], we know that such a reconstruction is mean square consistent. Consequently: for all $\varepsilon_2 > 0$, there exists K high enough such that $d^{Eucl}(\hat{x}^{(1)}, x^{(1)}) < \varepsilon_2$ and $d^{Eucl}(\hat{x}^{(2)}, x^{(2)}) < \varepsilon_2$.
3. Because of triangular inequalities, we have

$$d^{Eucl}(x^{(1)}, x^{(2)}) \leq d^{Eucl}(x^{(1)}, \hat{x}^{(1)}) + d^{Eucl}(\hat{x}^{(1)}, \hat{x}^{(2)}) + d^{Eucl}(\hat{x}^{(2)}, x^{(2)}).$$

Combining our three observations above, we have shown P2 in the special case of the simplified version of the BAGIDIS distance. However, considering now its general definition as given by Eq. (3), with $w_k > 0$ for $k = 0 \ldots K$, and $w_k = 0$ elsewhere, similar arguments with slightly more complex notation show again that Property P2 is valid.

We recall that in the statement of Property P2, the closeness of $s^{(1)}$ and $s^{(2)}$ is measured using the BAGIDIS semi-distance whereas the closeness of $x^{(1)}$ and $x^{(2)}$ by their Euclidean distance. This ensures that assumption A1 can be transposed in the (b,d) plane, so that local methods can be used that rely on the signatures of the curves.

Moreover, along with our proof, we showed that in order to ensure our criteria of similarity $d^{Eucl}(x^{(1)}, x^{(2)}) < \varepsilon$ to be satisfied, we constrained $s^{(2)}$ to be in a small ball of maximal radius 1 around $s^{(1)}$, as computed with d^B, which we denote by $B_{d^B}(s^{(1)}, 1)$. As is made clear in our proof, this constraint means that we define a neighborhood in which the main features of the curves are well aligned, because the breakpoints of the curves have to be the same up to rank K.

In the now following discussion, we indicate that if we enlarge our criteria for assessing the proximity of curves so that it allows for $x^{(1)}$ and $x^{(2)}$ to be considered similar although being misaligned, then the neighborhood of $s^{(1)}$ in which we will find $s^{(2)}$ for ensuring the desired proximity of $x^{(1)}$ and $x^{(2)}$ might be larger than $B_{d^B}(s^{(1)}, 1)$. Enlarging our proximity criteria in such a way is desirable. In the (b,d)

plane, a neighborhood larger than $B_{d^B}(s^{(1)}, 1)$ is a neighborhood that might include signatures $s^{(2)}$ the breakpoint component of which is not necessarily the same as for $s^{(1)}$. This translates into the fact that we might use information about series that are potentially misaligned.

3.3 The Case of Possible Misalignments

In our proof of P2 above, we indicate the existence of a small neighborhood, i.e. the small ball $B_{d_K^B}(s^{(1)}, \delta)$ around $s^{(1)}$, which has radius δ, and which is such that if $s^{(2)}$ is in $B_{d_K^B}(s^{(1)}, \delta)$ then the related curves $x^{(1)}$ and $x^{(2)}$ are similar. In our proof, we use an upper bound $\delta < 1$. This bound defines a neighborhood such that $s^{(2)}$ must have the same breakpoints as $s^{(1)}$ up to rank K. This bound appears because we measure the similarity of $x^{(1)}$ and $x^{(2)}$ using the Euclidean distance. Indeed, it ensures that the K main features of $x^{(1)}$ and $x^{(2)}$ are well aligned, which is necessary for the Euclidean distance to detect their closeness.

However, our method has been designed with the aim that the closeness of $s^{(1)}$ and $s^{(2)}$ in the (b,d) plane might also reflect a "visual proximity" of $x^{(1)}$ and $x^{(2)}$ even when the series are misaligned. Therefore, we would ideally like our proof to have the following extension:

Property P2/ideally: Let D be a distance measure between $x^{(1)}$ and $x^{(2)}$ that is relevant even in case of possible misalignment between $x^{(1)}$ and $x^{(2)}$. Then, for all $\varepsilon > 0$, there exists some neighborhood V centered on $s^{(1)}$ such that if $s^{(2)}$ is in V, then $D(x^{(1)}, x^{(2)}) < \varepsilon$.

In this case, V would not necessarily be smaller than $B_{d_K^B}(s^{(1)}, 1)$. Nevertheless, as far as we know, there does not exist a distance measure D that is able to measure the similarity of curves of which the sharp local patterns might be misaligned – which is precisely what motivated us to propose the BAGIDIS semi-distance.

We need thus to define another way to assess that the series $x^{(1)}$ and $x^{(2)}$ are close to each other when $s^{(1)}$ and $s^{(2)}$ are close enough. We propose the following:

We will say that $x^{(1)}$ and $x^{(2)}$ are "globally similar" to each other if

C1 Their "global shape" are similar – this notion is related to the succession of level changes in the series ; we define below how to quantify this intuitive notion.

(continued)

C2 *The main level changes of $x^{(1)}$ are located at abscissas that are not too distant from the abscissas of their counterpart in $x^{(2)}$.*

C3 *The amplitude of the main level changes in $x^{(1)}$ are not too different from the amplitude of their counterpart in $x^{(2)}$.*

Given this, we want to show that

Property P2/extended: There exists some neighborhood V of $s^{(1)}$ such that if $s^{(2)}$ is in V, then $x^{(1)}$ and $x^{(2)}$ are "globally similar".

We say that $x^{(1)}$ and $x^{(2)}$ are "similar in their global shape" if their associated Unbalanced Haar wavelet bases are "similar in structure" up to rank K. By "similar in structure" we mean that the hierarchical trees associated to each of the wavelet bases are identical up to rank K. We encode this hierarchical tree in a way that is common-place in wavelet analysis, but adapted to Unbalanced Haar, which we explain through the following example.

In Fig. 4, on the left hand side is a schematically representation of a basis $B^{(1)} = \{\psi_0^{(1)}, \psi_1^{(1)}, \ldots, \psi_4^{(1)}\}$, where the kth element $\psi_k^{(1)}$ is placed on the $(k+1)$th row, top-down. Here "+" is indicating the positive part of the wavelet and "−" is indicating its negative part. The collection of $\{b_k^{(1)}\}$ denotes the associated breakpoints – it uniquely encodes the associated hierarchical tree $T^{(1)}$ displayed on the right hand side, where we use a notation borrowed from wavelet or regression trees to highlight the implicit hierarchy of splits.

In this case, the hierarchical tree $T^{(1)}$ up to rank $K = 4$ is

$$T_4^{(1)} = \{(1,2); (1.1, 1.2); (1.2.1, 1.2.2); (2.1, 2.2)\}.$$

The kth wavelet basis vector $\psi_k^{(1)}$ is thus associated to the kth pair in the hierarchical tree ($k = 1, \ldots, K$). The elements of this pair are sequences of digits, where the last digit on the left is 1 to indicate that it refers to the positive part of the wavelet at rank k, and the last digit on the right is 2 to indicate that it refers to its negative part. The first digits of each element of the pair are the

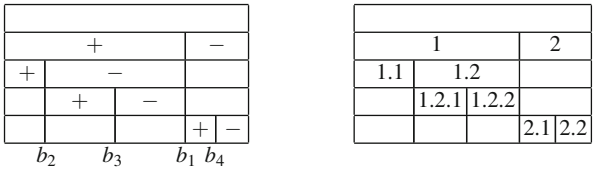


Fig. 4 Basis $B^{(1)} = \{\psi_0^{(1)}, \psi_1^{(1)}, \ldots, \psi_4^{(1)}\}$ (*left*), and its associated hierarchical tree $T^{(1)}$ (*right*)

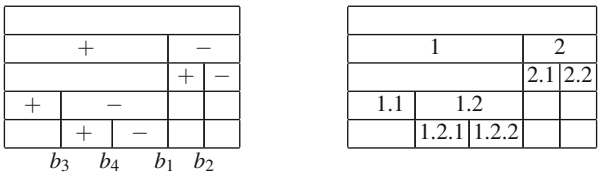


Fig. 5 Basis $B^{(2)} = \{\psi_0^{(2)}, \psi_1^{(2)}, \ldots, \psi_4^{(2)}\}$ (*left*), and its associated hierarchical tree $T^{(2)}$ (*right*)

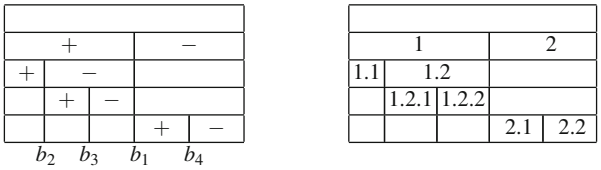


Fig. 6 Basis $B^{(3)} = \{\psi_0^{(3)}, \psi_1^{(3)}, \ldots, \psi_4^{(3)}\}$ (*left*), and its associated hierarchical tree $T^{(3)}$ (*right*)

same and refer to the "block" of the wavelet structure at larger scale (=smaller rank) in which the wavelet $\psi_k^{(1)}$ takes place. This tree structure reveals where the successive wavelets of the basis are located with respect to each other, without indication of the precise extension of their support, nor the precise location of their breakpoint. Thereby it characterizes the "global shape" of the series $x^{(1)}$ with which the basis $B^{(1)}$ is associated. For instance, our tree $T_4^{(1)}$ differs from the tree $T_4^{(2)} = \{(1,2);(2.1,2.2);(1.1,1.2);(1.2.1,1.2.2)\}$, associated to the basis $B^{(2)}$ in Fig. 5 below,

because the latter indicates a higher relative importance of a pattern on the right side of the series $x^{(2)}$ than in $x^{(1)}$ – this is because the element $(2.1, 2.2)$ appears only at rank 4 in $T_4^{(1)}$ while it appears at rank 2 in $T_4^{(2)}$. On the other hand the tree $T_4^{(3)}$ associated to the $B^{(3)}$ in Fig. 6 below

is exactly the same as $T_4^{(1)}$ up to rank 4, even though the patterns this basis supports are not perfectly aligned with the ones that basis $B^{(1)}$ supports.

We now prove *Property P2/extended*.

Proof

C2–C3: It is easy to see that C2 and C3 will be satisfied as soon as we take $s^{(2)}$ in a small enough neighborhood V defined by suitably chosen values δ_k^d and δ_k^b, $k = 1, \ldots, K$, such that $|b_k^{(1)} - b_k^{(2)}| < \delta_k^b$ and $|d_k^{(1)} - d_k^{(2)}| < \delta_k^d$, for $k < K$.

C1: We first observe that the hierarchical tree $T^{(i)}$ (up to rank K) is actually determined from the signature $s^{(i)}$ of a series (up to rank K) by the permutation $P^{(i)}$ we need to apply to the vector $(b_1^{(i)}, b_2^{(i)}, b_3^{(i)}, \ldots, b_K^{(i)})$ in order to sort its elements according to their value (i.e. their position along the breakpoints axis). For instance, for $K = 5$, if $b_2^{(i)} < b_1^{(i)} < b_5^{(i)} < b_4^{(i)} < b_3^{(i)}$, we have

$$(b_2^{(i)}, b_1^{(i)}, b_5^{(i)}, b_4^{(i)}, b_3^{(i)}) = select[(2,1,5,4,3)](b_1^{(i)}, b_2^{(i)}, b_3^{(i)}, b_4^{(i)}, b_5^{(i)}),$$

where $P^{(i)} = (2, 1, 5, 4, 3)$ (and where "select" is used to just denote the mapping of the indices according to this permutation here). The link between $P^{(i)}$ and the hierarchical structure of the wavelet partition arises because we can uniquely reconstruct the basis vectors in a top-down procedure, by using the information of the up-and-down shapes of the wavelets, the location of their breakpoints and their orthonormality.

Then, in order for C1 to be satisfied, it is sufficient to define V as the neighborhood of $s^{(1)}$such that the permutation $P_K^{(2)}$ associated to the K first elements of any $s^{(2)}$ in V is the same as the permutation $P_K^{(1)}$ associated to the K first elements of $s^{(1)}$.

We have thus proved *Property P2/extended*, as it is sufficient to combine the above constraints to define a neighborhood Varound $s^{(1)}$ such that if $s^{(2)}$ is in V, then $x^{(1)}$ and $x^{(2)}$ are "globally similar". As expected, we note that this neighborhood V might possibly be larger than $B_{d_K^B}(s^{(1)}, 1)$, so that it might contain the signatures of curves with main features some of which are misaligned.

This last point is important as it gives support for the large scope of the method and for its performance in investigating datasets of curves of which the main features might be misaligned although similar. It is clearly seen in the practical examples illustrated in our previous work, that the local methods (k-NN, kernels, MDS) that we used to process the set of BAGIDIS semi-distances computed on our datasets have effectively used neighborhoods large enough to include some breakpoint variations. This ability of using the information related to a misaligned feature is precisely the fact that makes BAGIDIS better than competitors in case of misaligned sharp patterns.

Conclusion: By the theoretical property (P2) we have shown that BAGIDIS achieves performances that are consistent with the ones obtained with the Euclidean distance in a local algorithm (while reducing the dimensionality of the problem from N sampled features of a given curve to $K < N$ significant or essential features). On the other hand, our subsequent discussion has shown that BAGIDIS may achieve performances that are superior to the Euclidean distance: local methods operating on the signatures of our curves in the (b,d) plane might be based upon neighborhoods which contain curves that are similar although misaligned.

Hence by this note we have provided the theoretical argument behind what we observed in previous work of ours on a competitive performance of our method when dealing with curves that have possibly misaligned sharp features.

Acknowledgements The first author is particularly grateful to EDF and A. Antoniadis, X. Brossat and J.-M. Poggi for having been given the opportunity to present the BAGIDIS methodology at the generously sponsored WIPFOR 2013 workshop in Paris.

Both authors would also like to acknowledge financial support from the IAP research network grants P06/03 and P07/06 of the Belgian government (Belgian Science Policy).

Finally, useful comments of Melvin Varughese and two anonymous referees have helped to improve the presentation of this note.

References

1. Aneiros-Pérez, G., Cardot, H., Estévez-Pérez, G., & Vieu, P. (2004). Maximum ozone concentration forecasting by functional non-parametric approaches. *Environmetrics, 15*(7), 675–685.
2. Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: Theory and practice* (Springer series in statistics). New York/Secaucus: Springer.
3. Fryzlewicz, P. (2007). Unbalanced Haar technique for non parametric function estimation. *Journal of the American Statistical Association, 102*(480), 1318–1327.
4. Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software, 7*(31), 1–24.
5. Girardi, M., & Sweldens, W. (1997). A new class of unbalanced Haar wavelets that form an unconditional basis for Lp on general measure spaces. *Journal of Fourier Analysis and Applications, 3*(4), 457–474.
6. Jolliffe, I. (2002). *Principal component analysis* (Springer series in statistics, 2nd ed.). New York/Secaucus: Springer.
7. Timmermans, C., Delsol, L., & von Sachs, R. (2013). Using Bagidis in nonparametric functional data analysis: Predicting from curves with sharp local features. *Journal of Multivariate Analysis, 115*, 421–444.
8. Timmermans, C., & Fryzlewicz, P. (2012). SHAH: Shape-adaptive haar wavelet transform for images with application to classification. Under revision http://www.uclouvain.be/en-369695.html. ISBA Discussion Paper 2012-15, Université catholique de Louvain.
9. Timmermans, C., & von Sachs, R. (2013). BAGIDIS: Statistically investigating curves with sharp local patterns using a new functional measure of dissimilarity. Under revision. http://www.uclouvain.be/en-369695.html. ISBA Discussion Paper 2013-31, Université catholique de Louvain.
10. Timmermans, C., de Tullio, P., Lambert, V., Frédérich, M., Rousseau, R., & von Sachs, R. (2012). Advantages of the BAGIDIS methodology for metabonomics analyses: Application to a spectroscopic study of age-related macular degeneration. In *Proceedings of the 12th European symposium on statistical methods for the food industry*, Paris, France (pp. 399–408).