Advances in Computational Management Science 9

Erricos John Kontoghiorghes
Cristian Gatu
Editors

Optimisation, Econometric and Financial Analysis

Springer

Advances in Computational Management Science

9

Editors:
H.M. Amman, Eindhoven, The Netherlands
B. Rustem, London, UK

Erricos John Kontoghiorghes · Cristian Gatu (Eds.)

Optimisation, Econometric and Financial Analysis

Springer

Editors

Prof. Erricos John Kontoghiorghes
University of Cyprus
Department of Public
and Business Administration
75 Kallipoleos St.
CY-1678 Nicosia
Cyprus
erricos@dcs.bbk.ac.uk

School of Computer Science
and Information Systems
Birkbeck College
University of London
Malet Street
London WC1E 7HX
UK

Dr. Cristian Gatu
Université de Neuchatel
Institut d' Informatique
Rue Emile-Argand 11, CP2
CH-2007 Neuchatel
Switzerland
Cristian.Gatu@unine.ch

Library of Congress Control Number: 2006931767

ISSN print edition: 1388-4307
ISBN-10 3-540-36625-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-36625-6 Springer Berlin Heidelberg New York

Typesetting:
Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN: 11801306 VA43/3100/Integra 543210

This book is dedicated to our families

Preface

"Optimisation, Econometric and Financial Analysis" is a volume of the book series on "Advances on Computational Management Science".

Advanced computational methods are often employed for the solution of modelling and decision-making problems. This book addresses issues associated with the interface of computing, optimisation, econometrics and financial modelling. Emphasis is given to computational optimisation methods and techniques.

The first part of the book addresses optimisation problems and decision modelling. Three chapters focus on applications of supply chain and worst-case modelling. The two further chapters consider advances in the methodological aspects of optimisation techniques. The second part of the book is devoted to optimisation heuristics, filtering, signal extraction and various time series models. There are five chapters in this part that cover the application of threshold accepting in econometrics, the investigation of the structure of threshold autoregressive moving average models, the employment of wavelet analysis and signal extraction techniques in time series. The third and final part of the book is about the use of optimisation in portfolio selection and real option modelling. The two chapters in this part consider applications of real investment options in the presence of managerial controls, and random portfolios and their use in measuring investment skills.

London, UK
August 2006

Erricos John Kontoghiorghes
Cristian Gatu

Contents

Part I

Optimisation Models and Methods

Part I

Optimisation Models and Methods

A Supply Chain Network Perspective for Electric Power Generation, Supply, Transmission, and Consumption

Anna Nagurney and Dmytro Matsypura

Department of Finance and Operations Management, Isenberg School of Management, University of Massachusetts, Amherst, MA 01003

Summary. A supply chain network perspective for electric power production, supply, transmission, and consumption is developed. The model is sufficiently general to handle the behavior of the various decision-makers, who operate in a decentralized manner and include power generators, power suppliers, the transmitters, as well as the consumers associated with the demand markets. The optimality conditions are derived, along with the equilibrium state for the electric power supply chain network. The finite-dimensional variational inequality formulation of the equilibrium state is derived, whose solution yields the equilibrium electric power flows transacted between the tiers of the supply chain network as well as the nodal prices. The variational inequality formulation is utilized to provide qualitative properties of the equilibrium electric power flow and price patterns and to propose a computational scheme. The algorithm is then applied to compute the solutions to several numerical examples.

Key words: Electric power, supply chains, networks, variational inequalities, game theory

1 Introduction

The electric power industry in the United States, as well as abroad, is undergoing a transformation from a regulated to a competitive industry. Whereas power generation was once dominated by vertically integrated investor-owned utilities who owned many of the generation capacity, transmission, and distribution facilities, the electric power industry today is characterized by many new companies that produce and market wholesale and retail electric power. In the United States, for example, several factors have made these changes both possible and necessary. First, technological advances have altered the economics of power production. For example, new gas-fired combined cycle power plants are more efficient and less costly than older coal-fired power

plants. In addition, technological advances in electricity transmission equipment have made possible the economic transmission of power over long distances so that customers can now be more selective in choosing an electricity supplier. Secondly, between 1975 and 1985, residential electricity prices and industrial electricity prices in the US rose 13% and 28% in real terms, respectively (US Energy Information Administration, 2002).

Furthermore, the effects of the Public Utilities Regulatory Policies Act of 1978, which encouraged the development of nonutility power producers that used renewable energy to generate power, demonstrated that traditional vertically integrated electric utilities were not the only source of reliable power. Moreover, numerous legislative initiatives have been undertaken by the federal government in order to stimulate the development and strengthening of competitive wholesale power markets. As a consequence, by December 1, 2003, 1310 companies were eligible to sell wholesale power at market-based rates in the US (statistics available at http://www.eia.doe.gov).

The dramatic increase in the number of market participants trading over the past few years, as well as changes to electricity trading patterns have made system reliability more difficult to maintain. The North American Electric Reliability Council (NERC) reported that, "[in recent years] the adequacy of the bulk power transmission system has been challenged to support the movement of power in unprecedented amounts and in unexpected directions" (North American Electric Reliability Council, 1998). Moreover, a US Department of Energy Task Force noted that "there is a critical need to be sure that reliability is not taken for granted as the industry restructures, and thus does not fall through the cracks" (Secretary of Energy Advisory Board's Task Force on Electric System Reliability, 1998).

These concerns have helped to stimulate research activity in the area of electric power supply systems modeling and analysis during the past decade. Several models have been proposed that allow for more decentralization in the markets (see, e.g., Schweppe (1988), Hogan (1992), Chao and Peck (1996), Wu et al. (1996)). Some researchers have suggested different variations of the models depending on the electric power market organizational structure (see, for example, Hobbs (2001)). A wide range of models has been proposed for simulating the interaction of competing generation companies who price strategically (see Kahn (1998) and Hobbs et al. (2000)), as well as those that simulate the exercising of market power on linearized dc networks based on a flexible representation of interactions of competing generating firms (Day et al. (2002)).

Nevertheless, despite all the research and analytical efforts, on August 14, 2003, large portions of the Midwest, the Northeastern United States, and Ontario, Canada, experienced an electric power blackout. The blackout left approximately 50 million people without electricity and affected 61,800 megawatts of electric load (US-Canada Power System Outage Task Force, 2004). In addition, two significant outages during the month of September 2003 occurred abroad: one in England and one, initiated in Switzerland, that

cascaded over much of Italy. The scale of these recent power outages has shown that the reliability of the existing power systems is not adequate and that the latest changes in electric power markets require deep and thorough analysis.

In this chapter, we propose what we believe is a novel approach to the modeling and analysis of electric power markets. In particular, we develop a supply chain network model for electric power generation, supply, transmission, and consumption, which allows for decentralized decision-making, and which differs from recent models (see, e.g., Jing-Yuan and Smeers (1999), Takriti et al. (2000), Boucher and Smeers (2001), and Daxhelet and Smeers (2001)) in that, first and foremost, we consider several different types of decision-makers and model their behavior and interactions explicitly. Moreover, we allow for not only the computation of electric power flows but also the prices associated with the various transactions between the tiers of decision-makers in the electric power supply chain network. Finally, the functional forms that can be handled in our framework are not limited to linear and/or separable functions. For additional background on supply chain network modeling, analysis, and computations, as well as financial engineering, see the annotated bibliography by Geunes and Pardalos (2003). For an overview of electric power systems, see the book by Casazza and Delea (2003). For an edited volume on the deregulation of electric utilities, see Zaccour (1998). For additional background on game theory as it relates to electric power systems, see the edited volume by Singh (1998).

The supply chain network approach permits one to represent the interactions between decision-makers in the market for electric power in terms of network connections, flows, and prices. In addition, we consider noncooperative behavior of decision-makers in the same tier of the supply chain network (such as, for example, the generators, the suppliers, and the demand markets) as well as cooperative behavior between tiers. Furthermore, this approach makes it possible to take advantage of the network topology (which is not limited to a specific number of generators, suppliers, transmitters, and/or demand markets) for computational purposes. Finally, it provides a framework from which a variety of extensions can be constructed to include, among other elements, multicriteria decision-making to incorporate environmental issues, risk and reliability elements, as well as stochastic components, and, in addition, the introduction of explicit dynamics and modeling of disequilibrium behavior.

The chapter is organized as follows. In Sect. 2, we develop the model, describe the various decision-makers and their behavior, and construct the equilibrium conditions, along with the variational inequality formulation. The variables are the equilibrium prices, as well as the equilibrium electricity flows between the tiers of decision-makers. In Sect. 3, we derive qualitative properties of the equilibrium pattern, under appropriate assumptions, notably, the existence and uniqueness of a solution to the governing variational inequality. In Sect. 4, we propose an algorithm, which is then applied to

several illustrative numerical examples in Sect. 5. We conclude the chapter with Sect. 6 in which we summarize our results and suggest directions for future research.

2 The Supply Chain Network Model for Electric Power

In this section, we develop an electric power supply chain network model in which the decision-makers operate in a decentralized manner. In particular, we consider an electric power network economy in which goods and services are limited to electric energy and transmission services. We consider power generators, power suppliers (including power marketers, traders, and brokers), transmission service providers, and consumers (demand markets, or end users). A depiction of the supply chain network for electric power is given in Fig. 1.

Power generators are those decision-makers who own and operate electric generating facilities or power plants. They produce electric power, which, is then sold to the power suppliers. The prices that generators charge for the electricity that they produce is determined by the competitive wholesale market. There is a total of G power generators, depicted as the top tier nodes in Fig. 1, with a typical power generator denoted by g. Power suppliers, in turn, bear a function of an intermediary. They buy electric power from power generators and sell to the consumers at different demand markets. We denote a typical supplier by s and consider a total of S power suppliers. Suppliers are represented by the second tier of nodes in the supply chain network in Fig. 1.

Note that there is a link from each power generator to each supplier in the network in Fig. 1 which represents that a supplier can buy energy from any generator on the wholesale market (equivalently, a generator can sell to any/all the suppliers). Note also that the links between the top tier and the second tier of nodes do not represent the physical connectivity of two particular nodes.

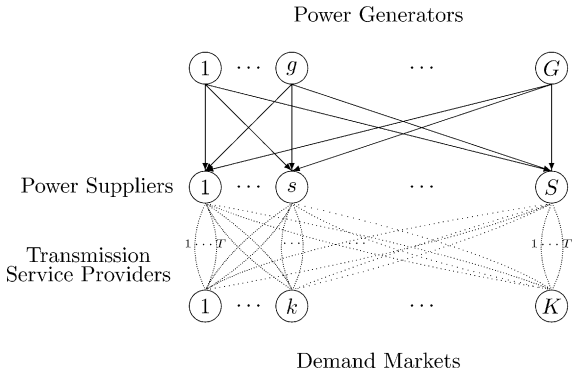

Fig. 1. The electric power supply chain network

Power suppliers do not physically possess electric power at any stage of the supplying process; they only hold the rights for the electric power. Hence, the link connecting a pair of such nodes in the supply chain is a decision-making connectivity link between that pair of nodes.

In order for electricity to be transmitted from a power generator to the point of consumption a transmission service is required. Hence, power suppliers need to buy the transmission services from the transmission service providers. Transmission service providers are those entities that own and operate the electric transmission and distribution systems. These are the companies that distribute electricity from generators via suppliers to demand markets (homes and businesses). Because transmission service providers do not make decisions as to where the electric power will be acquired and to whom it will be delivered, we do not include them in the model explicitly as nodes. Instead, their presence in the market is modeled as different *modes of transaction* (transmission modes) corresponding to distinct links connecting a given supplier node to a given demand market node in Fig. 1. We assume that power suppliers cover the direct cost of the physical transaction of electric power from power generators to the demand markets and, therefore, have to make a decision as to from where to acquire the transmission services (and at what level).

We assume that there are T transmission service providers operating in the supply chain network, with a typical transmission service provider denoted by t. For the sake of generality, we assume that every power supplier can transact with every demand market using any of the transmission service providers or any combination of them. Therefore, there are T links joining every node in the middle tier of the network with every node at the bottom tier (see Fig. 1).

Finally, the last type of decision-maker in the model is the consumers or demand markets. They are depicted as the bottom tier nodes in Fig. 1. These are the points of consumption of electric power. The consumers generate the demand that drives the generation and supply of the electric power in the entire system. There is a total of K demand markets, with a typical demand market denoted by k, and distinguished from the others through the use of appropriate criteria, such as geographic location; the types of consumers; that is, whether they are businesses or households; etc. We assume a competitive electric power market, meaning that the demand markets can choose between different electric power suppliers (power marketers, brokers, etc.).

We also assume that a given power supplier negotiates with the transmission service providers and makes sure that the necessary electric power is delivered. These assumptions fit well into the main idea of the restructuring of the electric power industry that is now being performed in the US, the European Union, and many other countries (see http://www.ferc.gov and http://www.europarl.eu.int).

Clearly, in some situations, some of the links in the supply chain network for electric power in Fig. 1 may not exist (due to, for example, various restrictions, regulations, etc.). This can be handled within our framework by

eliminating the corresponding link for the supply chain network or (see further discussion below) assigning an appropriately high transaction cost associated with that link.

We now turn to the discussion of the behavior of each type of decision-maker and give the optimality conditions.

2.1 The Behavior of Power Generators and their Optimality Conditions

We first start with the description of the behavior of the power generators. Recall that power generators are those decision-makers in the network system, who own and operate electric generating facilities or power plants. They generate electric power and then sell it to the suppliers. Hence, one of the assumptions of our model is that power generators cannot trade directly with the demand markets.

Let q_g denote the nonnegative amount of electricity in *watts* produced by electric power generator g and let q_{gs} denote the nonnegative amount of electricity (also in *watts*), being transacted from power generator g to power supplier s. Note that q_{gs} corresponds to the flow on the link joining node g with node s in Fig. 1. We group the electric power production outputs for all power generators into the vector $q \in R_+^G$. Also, we group all the power flows associated with all the power generators to the suppliers into the column vector $Q^1 \in R_+^{GS}$.

For power generator g, we assume, as given, a power generating cost function denoted by f_g, such that

$$f_g = f_g(q), \quad \forall g. \tag{1}$$

All the power generating functions are assumed to be convex and continuously differentiable. Since generators compete for resources we allow for the general form (1). Of course, a special case is when $f_g = f_g(q_g)$.

Note that we allow each power generating cost function to depend not only on the amount of energy generated by a particular power generator, but also on the amount of energy generated by other power generators. This generalization allows one to model competition.

In addition, while the electric power is being transmitted from node g to node s, there will be some transaction costs associated with the transmission process. Part of these costs will be covered by a power generator. Let c_{gs} denote power generator g's transaction cost function associated with transmitting the electric power to supplier node s. Without loss of generality we let c_{gs} depend on the amount of electric power transmitted from power generator g to power supplier s. Therefore,

$$c_{gs} = c_{gs}(q_{gs}), \quad \forall g, \forall s, \tag{2}$$

and we assume that these functions are convex and continuously differentiable.

Each power generator g faces the conservation of flow constraint given by:

$$\sum_{s=1}^{S} q_{gs} = q_g, \tag{3}$$

that is, a power generator g cannot ship out more electric power than he has produced.

In view of (3) and (1), we may write, without any loss of generality that $f_g = f_g(Q^1)$, for all power generators g; $g = 1, ..., G$. Note that in our framework, as the production output reaches the capacity of a given generator then we expect the production cost to become very large (and, perhaps, even infinite).

2.2 Optimisation Problem of a Power Generator

We assume that a typical power generator g is a profit-maximizer. Let ρ^*_{1gs} denote the price that a power generator g charges a power supplier s per unit of electricity. We later in this section discuss how this price is arrived at. We allow the power generator to set different prices for different power suppliers. Hence, the optimisation problem of the power generator g can be expressed as follows:

$$\text{Maximize} \quad U_g = \sum_{s=1}^{S} \rho^*_{1gs} q_{gs} - f_g(Q^1) - \sum_{s=1}^{S} c_{gs}(q_{gs}) \tag{4}$$

subject to:

$$q_{gs} \geq 0, \qquad \forall s. \tag{5}$$

We assume that the power generators compete in noncooperative manner following the concepts of Nash (1950, 1951) (see also, e.g., Dafermos and Nagurney (1987)). Hence, each power generator seeks to determine his optimal strategy, that is, the generated outputs, given those of the other power generators. The optimality conditions of all power generators g; $g = 1, ..., G$, simultaneously, under the above assumptions (see also Bazaraa et al. (1993), Bertsekas and Tsitsiklis (1997), and Nagurney (1999)), can be compactly expressed as: determine $Q^{1*} \in R^{GS}_+$ satisfying

$$\sum_{g=1}^{G} \sum_{s=1}^{S} \left[\frac{\partial f_g(Q^{1*})}{\partial q_{gs}} + \frac{\partial c_{gs}(q^*_{gs})}{\partial q_{gs}} - \rho^*_{1gs} \right] \times [q_{gs} - q^*_{gs}] \geq 0, \quad \forall Q^1 \in R^{GS}_+. \tag{6}$$

Note that (6) is a variational inequality. Moreover, (6) has a very nice economic interpretation. Indeed, at optimality, if there is a positive flow of electric power between a generator/supplier pair, then the price charged is precisely equal to the sum of the marginal production cost plus the marginal transaction cost; if that sum exceeds the price, then there will be no electric power flow (and, thus, no transaction) between that pair.

2.3 The Behavior of Power Suppliers and their Optimality Conditions

We now turn to the description of the behavior of the power suppliers. The term power supplier refers to power marketers, traders, and brokers, who arrange for the sale and purchase of the output of generators to other suppliers or load-serving entities, or in many cases, serve as load-serving entities themselves. They play a fundamental role in our model since they are responsible for acquiring electricity from power generators and delivering it to the demand markets. Therefore, power suppliers are involved in transactions with both power generators and the demand markets through transmission service providers.

A power supplier s is faced with certain expenses, which may include, for example, the cost of licensing and the costs of maintenance. We refer collectively to such costs as an *operating* cost and denote it by c_s. Let q_{sk}^t denote the amount of electricity being transacted between power supplier s and demand market k via the link corresponding to the transmission service provider t. We group all transactions associated with power supplier s and demand market k into the column vector $q_{sk} \in R_+^T$. We then further group all such vectors associated with all the power suppliers into a column vector $Q^2 \in R_+^{STK}$. For the sake of generality and to enhance the modeling of competition, we assume that

$$c_s = c_s(Q^1, Q^2), \quad \forall s. \tag{7}$$

We also assume that there is another type of cost that a power supplier may face, namely, transaction costs. As mentioned earlier, each power supplier is involved in transacting with both power generators and with the demand markets through transmission service providers. Therefore, there will be costs associated with each such transaction. These costs may include, for example, the expenses associated with maintaining the physical lines, if they belong to the power supplier, or the expenses associated with the transmission service which a power supplier has to purchase. In order to capture all possible scenarios, we will use a transaction cost function of a general form. Let $\hat{c}_{gs}$ denote the transaction cost associated with power supplier s acquiring electric power from power generator g, where we assume that:

$$\hat{c}_{gs} = \hat{c}_{gs}(q_{gs}), \quad \forall g, \forall s. \tag{8}$$

Similarly, let c_{sk}^t denote the transaction cost associated with power supplier s transmitting electric power to demand market k via transmission service provider t, where:

$$c_{sk}^t = c_{sk}^t(q_{sk}^t), \quad \forall s, \forall k, \forall t. \tag{9}$$

We assume that all the above transaction cost functions are convex and continuously differentiable.

Let ρ_{2sk}^t denote the price associated with the transaction from power supplier s to demand market k via transmission service provider t and let ρ_{2sk}^{t*} denote the price actually charged (which we return to later in this section). The total amount of revenue the power supplier obtains from his transactions is equal to the sum over all the modes of transmission and all the demand markets of the price times the amount of electric power transacted with the demand market using the particular transmission mode. Indeed, the total revenue of power supplier s can mathematically be expressed as follows:

$$\sum_{k=1}^{K} \sum_{t=1}^{T} \rho_{2sk}^{t*} q_{sk}^t. \tag{10}$$

Before formulating an optimisation problem of a typical power supplier, let us look closer at the transmission service providers and their role in the electric power supply chain network system.

2.4 Transmission Service Providers

In order for electricity to be transmitted from a given power generator to the point of consumption a transmission service is required. Hence, power suppliers purchase the transmission services from the transmission service providers. Transmission service providers are those entities that own and operate the electric transmission and distribution systems. We assume that the price of transmission service depends on how far the electricity has to be transmitted; in other words, it can be different for different destinations (demand markets or consumers). We also let different transmission service providers have their services priced differently, which can be a result of a different level of quality of service, reliability of the service, etc.

In practice, an electric supply network is operated by an Independent System Operator (ISO) who operates as a disinterested, but efficient entity and does not own network or generation assets. His main objectives are: to provide independent, open and fair access to transmission systems; to facilitate market-based, wholesale electricity rates; and to ensure the effective management and operation of the bulk power system in each region (http://www.isone.org). Therefore, the ISO does not control the electricity rates. Nevertheless, he makes sure that the prices of the transmission services are reasonable and not discriminatory. We model this aspect by having transmission service providers be price-takers meaning that the price of their services is determined and cannot be changed by a transmission service provider himself. Hence, the price of transmission services is fixed. However, it is not constant, since it depends on the amount of electric power transmitted, the distance, etc., and may be calculated for each transmission line separately depending on the criteria listed above. Consequently, as was stated earlier, a transmission service provider does not serve as an explicit decision-maker in the complex network system.

2.5 Optimisation Problem of a Power Supplier

Assuming that a typical power supplier s is a profit-maximizer, we can express the optimisation problem of power supplier s as follows:

$$\text{Maximize} \quad U^s = \sum_{k=1}^{K}\sum_{t=1}^{T}\rho_{2sk}^{t*}q_{sk}^t - c_s(Q^1, Q^2) - \sum_{g=1}^{G}\rho_{1gs}^{*}q_{gs} - \sum_{g=1}^{G}\hat{c}_{gs}(q_{gs}) - \sum_{k=1}^{K}\sum_{t=1}^{T}c_{sk}^t(q_{sk}^t) \tag{11}$$

subject to:

$$\sum_{k=1}^{K}\sum_{t=1}^{T}q_{sk}^t \leq \sum_{g=1}^{G}q_{gs} \tag{12}$$

$$q_{gs} \geq 0, \quad \forall g \tag{13}$$

$$q_{sk}^t \geq 0, \quad \forall k, \forall t. \tag{14}$$

The objective function (11) represents the profit of power supplier s with the first term denoting the revenue and the subsequent terms the various costs and payouts to the generators. Inequality (12) is a conservation of flow inequality which states that a power supplier s cannot provide more electricity than he obtains from the power generators.

We assume that the power suppliers also compete in a noncooperative manner (as we assumed for the power generators). Hence, each power supplier seeks to determine his optimal strategy, that is, the input (accepted) and output flows, given those of the other power suppliers. The optimality conditions of all power suppliers s; $s = 1, ..., S$, simultaneously, under the above assumptions (see also Dafermos and Nagurney (1987) and Nagurney et al. (2002)), can be compactly expressed as: determine $(Q^{1*}, Q^{2*}, \gamma^*) \in R_+^{S(G+KT+1)}$ satisfying

$$\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{\partial c_s(Q^{1*}, Q^{2*})}{\partial q_{sk}^t} + \frac{\partial c_{sk}^t(q_{sk}^{t*})}{\partial q_{sk}^t} - \rho_{2sk}^{t*} + \gamma_s^*\right] \times [q_{sk}^t - q_{sk}^{t*}]$$
$$+ \sum_{s=1}^{S}\sum_{g=1}^{G}\left[\frac{\partial c_s(Q^{1*}, Q^{2*})}{\partial q_{gs}} + \frac{\partial \hat{c}_{gs}(q_{gs}^*)}{\partial q_{gs}} + \rho_{1gs}^* - \gamma_s^*\right] \times [q_{gs} - q_{gs}^*]$$
$$+ \sum_{s=1}^{S}\left[\sum_{g=1}^{G}q_{gs}^* - \sum_{k=1}^{K}\sum_{t=1}^{T}q_{sk}^{t*}\right] \times [\gamma_s - \gamma_s^*] \geq 0,$$
$$\forall (Q^1, Q^2, \gamma) \in R_+^{S(G+KT+1)}, \tag{15}$$

where γ_s^* is the optimal Lagrange multiplier associated with constraint (12), and γ is the corresponding S-dimensional vector of Lagrange multipliers.

Note that γ_s^* serves as a "market-clearing" price in that, if positive, the electric power flow transacted out of supplier s must be equal to that amount accepted by the supplier from all the power generators. Also, note that from (15) we can infer that if there is a positive flow q_{gs}^*, then γ_s^* is precisely equal to the marginal operating cost of supplier s plus the marginal cost associated with this transaction plus the price per unit of electric power paid by supplier s to generator g.

2.6 Equilibrium Conditions for the Demand Markets

We now turn to the description of the equilibrium conditions for the demand markets. Let ρ_{3k} denote the price per unit of electric power associated with the demand market k. Note here that we allow the final price of electric power to be different at different demand markets. We assume that the demand for electric power at each demand market k is elastic and depends not only on the price at the corresponding demand market but may, in general, also depend on the entire vector of the final prices in the supply chain network economy, that is,

$$d_k = d_k(\rho_3), \tag{16}$$

where $\rho_3 = (\rho_{31}, \ldots, \rho_{3k}, \ldots, \rho_{3K})^T$. This level of generality also allows one to facilitate the modeling of competition on the consumption side.

Let $\hat{c}_{sk}^t$ denote the unit transaction cost associated with obtaining the electric power at demand market k from supplier s via transmission mode t, where we assume that this transaction cost is continuous and of the general form:

$$\hat{c}_{sk}^t = \hat{c}_{sk}^t(Q^2), \quad \forall s, \forall k, \forall t. \tag{17}$$

The equilibrium conditions associated with the transactions between power suppliers and demand markets take the following form: We say that a vector $(Q^{2*}, \rho_3^*) \in R_+^{K(ST+1)}$ is an equilibrium vector if for each s, k, t:

$$\rho_{2sk}^{t*} + \hat{c}_{sk}^t(Q^{2*}) \begin{cases} = \rho_{3k}^*, \text{ if } & q_{sk}^{t*} > 0, \\ \geq \rho_{3k}^*, \text{ if } & q_{sk}^{t*} = 0. \end{cases} \tag{18}$$

and

$$d_k(\rho_3^*) \begin{cases} = \sum_{s=1}^S \sum_{t=1}^T q_{sk}^{t*}, \text{ if } & \rho_{3k}^* > 0, \\ \leq \sum_{s=1}^S \sum_{t=1}^T q_{sk}^{t*}, \text{ if } & \rho_{3k}^* = 0. \end{cases} \tag{19}$$

Conditions (18) state that consumers at demand market k will purchase the electric power from power supplier s, if the price charged by the power supplier plus the transaction cost does not exceed the price that the consumers are willing to pay for the electric power. Note that, according to (18), if the transaction costs are identically equal to zero, then the price faced by the consumers for the electric power is the price charged by the power supplier.

Condition (19), on the other hand, states that, if the price the consumers are willing to pay for the electric power at a demand market is positive, then

the amount of the electric power transacted by the power suppliers with the consumers at the demand market is precisely equal to the demand. Conditions (18) and (19) are in concert with the ones in Nagurney et al. (2002), and reflect, spatial price equilibrium (see also, e.g., Nagurney (1999)).

Note that the satisfaction of (18) and (19) is equivalent to the solution of the variational inequality given by: determine $(Q^{2*}, \rho_3^*) \in R_+^{K(ST+1)}$, such that

$$\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\rho_{2sk}^{t*} + \hat{c}_{sk}^{t}(Q^{2*}) - \rho_{3k}^{*}\right] \times \left[q_{sk}^{t} - q_{sk}^{t*}\right] + \sum_{k=1}^{K}\left[\sum_{s=1}^{S}\sum_{t=1}^{T} q_{sk}^{t*} - d_k(\rho_3^*)\right] \times \left[\rho_{3k} - \rho_{3k}^{*}\right] \geq 0, \quad \forall (Q^2, \rho_3) \in R_+^{K(ST+1)}. \tag{20}$$

2.7 The Equilibrium Conditions for the Power Supply Chain Network

In equilibrium, the amounts of electricity transacted between the power generators and the power suppliers must coincide with those that the power suppliers actually accept. In addition, the amounts of the electricity that are obtained by the consumers must be equal to the amounts that the power suppliers actually provide. Hence, although there may be competition between decision-makers at the same tier of nodes of the power supply chain network there must be, in a sense, cooperation between decision-makers associated with pairs of nodes (through positive flows on the links joining them). Thus, in equilibrium, the prices and product flows must satisfy the sum of the optimality conditions (6) and (15), and the equilibrium conditions (20). We make these relationships rigorous through the subsequent definition and variational inequality derivation.

Definition 1 (Equilibrium State). *The equilibrium state of the electric power supply chain network is one where the electric power flows between the tiers of the network coincide and the electric power flows and prices satisfy the sum of conditions (6), (15), and (20).*

We now state and prove:

Theorem 1 (VI Formulation). *The equilibrium conditions governing the power supply chain network according to Definition 1 are equivalent to the solution of the variational inequality given by: determine* $(Q^{1*}, Q^{2*}, \gamma^*, \rho_3^*) \in \mathcal{K}$ *satisfying:*

$$\sum_{g=1}^{G}\sum_{s=1}^{S}\left[\frac{\partial f_g(Q^{1*})}{\partial q_{gs}} + \frac{\partial c_{gs}(q_{gs}^*)}{\partial q_{gs}} + \frac{\partial c_s(Q^{1*}, Q^{2*})}{\partial q_{gs}} + \frac{\partial \hat{c}_{gs}(q_{gs}^*)}{\partial q_{gs}} - \gamma_s^*\right] \times \left[q_{gs} - q_{gs}^*\right]$$

$$+\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{\partial c_s(Q^{1*},Q^{2*})}{\partial q_{sk}^t}+\frac{\partial c_{sk}^t(q_{sk}^{t*})}{\partial q_{sk}^t}+\hat{c}_{sk}^t(Q^{2*})+\gamma_s^*-\rho_{3k}^*\right]\times\left[q_{sk}^t-q_{sk}^{t*}\right]$$
$$+\sum_{s=1}^{S}\left[\sum_{g=1}^{G}q_{gs}^*-\sum_{k=1}^{K}\sum_{t=1}^{T}q_{sk}^{t*}\right]\times[\gamma_s-\gamma_s^*]$$
$$+\sum_{k=1}^{K}\left[\sum_{s=1}^{S}\sum_{t=1}^{T}q_{sk}^{t*}-d_k(\rho_3^*)\right]\times[\rho_{3k}-\rho_{3k}^*]\geq 0,$$
$$\forall(Q^1,Q^2,\gamma,\rho_3)\in\mathcal{K}, \tag{21}$$

where $\mathcal{K}\equiv\{(Q^1,Q^2,\gamma,\rho_3)|(Q^1,Q^2,\gamma,\rho_3)\in R_+^{GS+TSK+S+K}\}$.

Proof. We first establish that the equilibrium conditions imply variational inequality (21). Indeed, summation of inequalities (6), (15), and (20), after algebraic simplifications, yields variational inequality (21).

We now establish the converse, that is, that a solution to variational inequality (21) satisfies the sum of conditions (6), (15), and (20), and is, hence, an equilibrium.

Consider inequality (21). Add term $\rho_{1gs}^*-\rho_{1gs}^*$ to the term in the first set of brackets (preceding the first multiplication sign). Similarly, add term $\rho_{2sk}^{t*}-\rho_{2sk}^{t*}$ to the term in the second set of brackets (preceding the second multiplication sign). The addition of such terms does not change (21) since the value of these terms is zero and yields:

$$\sum_{g=1}^{G}\sum_{s=1}^{S}\left[\frac{\partial f_g(Q^{1*})}{\partial q_{gs}}+\frac{\partial c_{gs}(q_{gs}^*)}{\partial q_{gs}}+\frac{\partial c_s(Q^{1*},Q^{2*})}{\partial q_{gs}}+\frac{\partial \hat{c}_{gs}(q_{gs}^*)}{\partial q_{gs}}-\gamma_s^*+\rho_{1gs}^*-\rho_{1gs}^*\right]\times\left[q_{gs}-q_{gs}^*\right]$$
$$+\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{\partial c_s(Q^{1*},Q^{2*})}{\partial q_{sk}^t}+\frac{\partial c_{sk}^t(q_{sk}^{t*})}{\partial q_{sk}^t}+\hat{c}_{sk}^t(Q^{2*})+\gamma_s^*-\rho_{3k}^*+\rho_{2sk}^{t*}-\rho_{2sk}^{t*}\right]\times\left[q_{sk}^t-q_{sk}^{t*}\right]$$
$$+\sum_{s=1}^{S}\left[\sum_{g=1}^{G}q_{gs}^*-\sum_{k=1}^{K}\sum_{t=1}^{T}q_{sk}^{t*}\right]\times[\gamma_s-\gamma_s^*]$$
$$+\sum_{k=1}^{K}\left[\sum_{s=1}^{S}\sum_{t=1}^{T}q_{sk}^{t*}-d_k(\rho_3^*)\right]\times[\rho_{3k}-\rho_{3k}^*]\geq 0,$$
$$\forall(Q^1,Q^2,\gamma,\rho_3)\in\mathcal{K}, \tag{22}$$

which can be rewritten as:

$$\sum_{g=1}^{G}\sum_{s=1}^{S}\left[\frac{\partial f_g(Q^{1*})}{\partial q_{gs}}+\frac{\partial c_{gs}(q^*_{gs})}{\partial q_{gs}}-\rho^*_{1gs}\right]\times\left[q_{gs}-q^*_{gs}\right]$$
$$+\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{\partial c_s(Q^{1*},Q^{2*})}{\partial q^t_{sk}}+\frac{\partial c^t_{sk}(q^{t*}_{sk})}{\partial q^t_{sk}}-\rho^{t*}_{2sk}+\gamma^*_s\right]\times\left[q^t_{sk}-q^{t*}_{sk}\right]$$
$$+\sum_{s=1}^{S}\sum_{g=1}^{G}\left[\frac{\partial c_s(Q^{1*},Q^{2*})}{\partial q_{gs}}+\frac{\partial \hat{c}_{gs}(q^*_{gs})}{\partial q_{gs}}+\rho^*_{1gs}-\gamma^*_s\right]\times\left[q_{gs}-q^*_{gs}\right]$$
$$+\sum_{s=1}^{S}\left[\sum_{g=1}^{G}q^*_{gs}-\sum_{k=1}^{K}\sum_{t=1}^{T}q^{t*}_{sk}\right]\times\left[\gamma_s-\gamma^*_s\right]$$
$$+\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\rho^{t*}_{2sk}+\hat{c}^t_{sk}(Q^{2*})-\rho^*_{3k}\right]\times\left[q^t_{sk}-q^{t*}_{sk}\right]$$
$$+\sum_{k=1}^{K}\left[\sum_{s=1}^{S}\sum_{t=1}^{T}q^{t*}_{sk}-d_k(\rho^*_3)\right]\times\left[\rho_{3k}-\rho^*_{3k}\right]\geq 0,$$
$$\forall(Q^1,Q^2,\gamma,\rho_3)\in\mathcal{K}. \tag{23}$$

Inequality (23) is a sum of equilibrium conditions (6), (15), and (20). Therefore, the electric power flow and price pattern is an equilibrium according to Definition 1.

The variational inequality problem (21) can be rewritten in standard variational inequality form (cf. Nagurney (1999)) as follows: determine $X^*\in\mathcal{K}$ satisfying

$$\langle F(X^*)^T, X-X^*\rangle\geq 0,\quad \forall X\in\mathcal{K}, \tag{24}$$

where $X\equiv(Q^1,Q^2,\gamma,\rho_3)$, and $F(X)\equiv(F_{gs},F^t_{sk},F_s,F_k)$ where $g=1,...,G$; $s=1,...,S$; $t=1,...,T$; $k=1,...,K$, with the specific components of F given by the functional terms preceding the multiplication signs in (21), respectively. $\langle\cdot,\cdot\rangle$ denotes the inner product in N-dimensional Euclidian space where here $N=GS+SKT+S+K$.

We now describe how to recover the prices associated with the first two tiers of nodes in the power supply chain network. Clearly, the components of the vector ρ^*_3 are obtained directly from the solution to variational inequality (21). In order to recover the second tier prices ρ^*_2 associated with the power suppliers one can (after solving variational inequality (21) for the particular numerical problem) either (cf. (18)) set $\rho^{t*}_{2sk}=\rho^*_{3k}-\hat{c}^t_{sk}(Q^{2*})$ for any s,t,k such that $q^{t*}_{sk}>0$, or (cf. (15)) set $\rho^{t*}_{2sk}=\frac{\partial c_s(Q^{1*},Q^{2*})}{\partial q^t_{sk}}+\frac{\partial c^t_{sk}(q^{t*}_{sk})}{\partial q^t_{sk}}+\gamma^*_s$ for any s,t,k such that $q^{t*}_{sk}>0$.

Similarly, from (6) we can infer that the top tier prices comprising the vector ρ^*_1 can be recovered (once the variational inequality (21) is solved with particular data) in the following way: for any g,s

such that $q^*_{gs} > 0$, set $\rho^*_{1gs} = \frac{\partial f_g(Q^{1*})}{\partial q_{gs}} + \frac{\partial c_{gs}(q^*_{gs})}{\partial q_{gs}}$ or, equivalently, from (15): set $\rho^*_{1gs} = \gamma^*_s - \frac{\partial c_s(Q^{1*}, Q^{2*})}{\partial q_{gs}} - \frac{\partial \hat{c}_{gs}(q^*_{gs})}{\partial q_{gs}}$.

Theorem 2. *The solution to the variational inequality (22) satisfies variational inequalities (6), (15), and (20) (separately) under the condition that vectors ρ^*_1 and ρ^*_2 are derived using the procedure described above.*

Proof. Suppose that $(Q^{1*}, Q^{2*}, \gamma^*, \rho^*_3) \in \mathcal{K}$ is a solution to variational inequality (21). Variational inequality (21) has to hold for all $(Q^1, Q^2, \gamma, \rho_3) \in \mathcal{K}$. Using the procedure for deriving vectors ρ^*_1 and ρ^*_2 one can get (23) from (21). Now, consider expression (23) from the proof of Theorem 1. If one lets $\gamma_s = \gamma^*_s$, $\rho_{3k} = \rho^*_{3k}$, and $q^t_{sk} = q^{t*}_{sk}$ for all s, k, and t in (23), one obtains the following expression:

$$\sum_{g=1}^{G}\sum_{s=1}^{S}\left[\frac{\partial f_g(Q^{1*})}{\partial q_{gs}} + \frac{\partial c_{gs}(q^*_{gs})}{\partial q_{gs}} - \rho^*_{1gs}\right] \times \left[q_{gs} - q^*_{gs}\right] \geq 0, \quad \forall Q^1 \in R^{GS}_+,$$

which is exactly variational inequality (6) and, therefore, a solution to (21) also satisfies (6).

Similarly, if one lets $\rho_{3k} = \rho^*_{3k}$ for all k, $q^t_{sk} = q^{t*}_{sk}$ for all s, k, and t in the fourth functional term (preceding the fourth multiplication sign), and also lets $q_{gs} = q^*_{gs}$ in the first functional term (preceding the first multiplication sign) in (24), one obtains the following expression:

$$\begin{aligned}\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{\partial c_s(Q^{1*}, Q^{2*})}{\partial q^t_{sk}} + \frac{\partial c^t_{sk}(q^{t*}_{sk})}{\partial q^t_{sk}} - \rho^{t*}_{2sk} + \gamma^*_s\right] \times \left[q^t_{sk} - q^{t*}_{sk}\right] \\ + \sum_{s=1}^{S}\sum_{g=1}^{G}\left[\frac{\partial c_s(Q^{1*}, Q^{2*})}{\partial q_{gs}} + \frac{\partial \hat{c}_{gs}(q^*_{gs})}{\partial q_{gs}} + \rho^*_{1gs} - \gamma^*_s\right] \times \left[q_{gs} - q^*_{gs}\right] \\ + \sum_{s=1}^{S}\left[\sum_{g=1}^{G} q^*_{gs} - \sum_{k=1}^{K}\sum_{t=1}^{T} q^{t*}_{sk}\right] \times \left[\gamma_s - \gamma^*_s\right] \geq 0, \\ \forall (Q^1, Q^2, \gamma) \in R^{S(G+KT+1)}_+,\end{aligned}$$

which is exactly variational inequality (15) and, therefore, a solution to (21) also satisfies (15).

Finally, if one lets $\gamma_s = \gamma^*_s$, $q_{gs} = q^*_{gs}$ for all g and s, and also $q^t_{sk} = q^{t*}_{sk}$ for all s, k, and t and substitutes these into the second functional term (preceding the second multiplication sign) in (23), one obtains the following expression:

$$\begin{aligned}\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\rho^{t*}_{2sk} + \hat{c}^t_{sk}(Q^{2*}) - \rho^*_{3k}\right] \times \left[q^t_{sk} - q^{t*}_{sk}\right] \\ + \sum_{k=1}^{K}\left[\sum_{s=1}^{S}\sum_{t=1}^{T} q^{t*}_{sk} - d_k(\rho^*_3)\right] \times \left[\rho_{3k} - \rho^*_{3k}\right] \geq 0, \\ \forall (Q^2, \rho_3) \in R^{K(ST+1)}_+,\end{aligned}$$

which is exactly variational inequality (20) and, hence, a solution to (21) also satisfies (20).

We have, thus, established that a solution to variational inequality (21) also satisfies (6), (15), and (20) separately under the pricing mechanism described above.

3 Qualitative Properties

In this section, we provide some qualitative properties of the solution to variational inequality (24). In particular, we derive existence and uniqueness results.

Since the feasible set is not compact we cannot derive existence simply from the assumption of continuity of the functions. We can, however, impose a rather weak condition to guarantee existence of a solution pattern. Let

$$\mathcal{K}_b = \{(Q^1, Q^2, \gamma, \rho_3) | 0 \leq Q^1 \leq b_1; 0 \leq Q^2 \leq b_2; \\ 0 \leq \gamma \leq b_3; 0 \leq \rho_3 \leq b_4\}, \tag{25}$$

where $b = (b_1, b_2, b_3, b_4) \geq 0$ and $Q^1 \leq b_1$, $Q^2 \leq b_2$, $\gamma \leq b_3$, and $\rho_3 \leq b_4$ means $q_{gs} \leq b_1$, $q^t_{sk} \leq b_2$, $\gamma_s \leq b_3$, and $\rho_{3k} \leq b_4$ for all g, s, k, and t. Then $\mathcal{K}_b$ is a bounded, closed, convex subset of $R_+^{GS+SKT+S+K}$. Therefore, the following variational inequality:

$$\langle F(X^b)^T, X - X^b \rangle \geq 0, \quad \forall X \in \mathcal{K}_b, \tag{26}$$

admits at least one solution $X_b \in \mathcal{K}_b$, from the standard theory of variational inequalities, since $\mathcal{K}_b$ is compact and F is continuous. Following (Kinderlehrer and Stampacchia (1980)) (see also Nagurney (1999)), we then have the following theorems:

Theorem 3 (Existence). *Variational inequality (24) (equivalently (21)) admits a solution if and only if there exists a vector $b > 0$, such that variational inequality (26) admits a solution in $\mathcal{K}_b$ with*

$$Q^{1b} < b_1, \qquad Q^{2b} < b_2, \qquad \gamma^b < b_3, \qquad \rho_3^b < b_4.$$

Theorem 4 (Uniqueness). *Assume that conditions of Theorem 3 hold, that is, variational inequality (26) and, hence, variational inequality (24) admits at least one solution. Suppose that function $F(X)$ that enters variational inequality (24) is strictly monotone on $\mathcal{K}$, that is,*

$$\langle (F(X') - F(X''))^T, X' - X'' \rangle > 0, \quad \forall X', X'' \in \mathcal{K}, \quad X' \neq X''. \tag{27}$$

Then the solution to variational inequality (24) is unique.

4 The Algorithm

In this section, an algorithm is presented that can be applied to solve any variational inequality problem in standard form (see (24)), that is: determine $X^* \in \mathcal{K}$, satisfying:

$$\langle F(X^*)^T, X - X^* \rangle \geq 0, \quad \forall X \in \mathcal{K}. \tag{28}$$

The algorithm is guaranteed to converge provided that the function $F(X)$ that enters the variational inequality is monotone and Lipschitz continuous (and that a solution exists). The algorithm is the modified projection method of Korpelevich (1977) and it has been applied to solve a plethora of network equilibrium problems (see Nagurney and Dong (2002)).

We first provide a definition of a Lipschitz continuous function:

Definition 2 (Lipschitz Continuity). *A function $F(X)$ is Lipschitz continuous, if there exists a constant $L > 0$ such that:*

$$\|F(X') - F(X'')\| \leq L\|X' - X''\|, \quad \forall X', X'' \in \mathcal{K}, \text{ with } L > 0. \tag{29}$$

The statement of the modified projection method is as follows, where $\mathcal{T}$ denotes an iteration counter:

Modified Projection Method

Step 0: Initialization Set $(Q^{10}, Q^{20}, \gamma^0, \rho_3^0) \in \mathcal{K}$. Let $\mathcal{T} = 1$ and let a be a scalar such that $0 < a \leq \frac{1}{L}$, where L is the Lipschitz continuity constant (cf. (29)).

Step 1: Computation Compute $(\bar{Q}^{1\mathcal{T}}, \bar{Q}^{2\mathcal{T}}, \bar{\gamma}^{\mathcal{T}}, \bar{\rho}_3^{\mathcal{T}})$ by solving the variational inequality subproblem:

$$\begin{aligned}
&\sum_{g=1}^{G}\sum_{s=1}^{S}\left[\bar{q}_{gs}^{\mathcal{T}} + a\left(\frac{\partial f_g(Q^{1\mathcal{T}-1})}{\partial q_{gs}} + \frac{\partial c_s(Q^{1\mathcal{T}-1}, Q^{2\mathcal{T}-1})}{\partial q_{gs}}\right.\right.\\
&\left.\left. + \frac{\partial c_{gs}(q_{gs}^{\mathcal{T}-1})}{\partial q_{gs}} + \frac{\partial \hat{c}_{gs}(q_{gs}^{\mathcal{T}-1})}{\partial q_{gs}} - \gamma_s^{\mathcal{T}-1}\right) - q_{gs}^{\mathcal{T}-1}\right] \times \left[q_{gs} - \bar{q}_{gs}^{\mathcal{T}}\right]\\
&+ \sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\bar{q}_{sk}^{t\mathcal{T}} + a\left(\frac{\partial c_s(Q^{1\mathcal{T}-1}, Q^{2\mathcal{T}-1})}{\partial q_{sk}^t} + \frac{\partial c_{sk}^t(q_{sk}^{t\mathcal{T}-1})}{\partial q_{sk}^t}\right.\right.\\
&\left.\left. + \hat{c}_{sk}^t(Q^{2\mathcal{T}-1}) + \gamma_s^{\mathcal{T}-1} - \rho_{3k}^{\mathcal{T}-1}\right) - q_{sk}^{t\mathcal{T}-1}\right] \times \left[q_{sk}^t - \bar{q}_{sk}^{t\mathcal{T}}\right]\\
&+ \sum_{s=1}^{S}\left[\bar{\gamma}_s^{\mathcal{T}} + a\left(\sum_{g=1}^{G} q_{gs}^{\mathcal{T}-1} - \sum_{k=1}^{K}\sum_{t=1}^{T} q_{sk}^{t\mathcal{T}-1}\right) - \gamma_s^{\mathcal{T}-1}\right] \times \left[\gamma_s - \bar{\gamma}_s^{\mathcal{T}}\right]\\
&+ \sum_{k=1}^{K}\left[\bar{\rho}_{3k}^{\mathcal{T}} + a\left(\sum_{s=1}^{S}\sum_{t=1}^{T} q_{sk}^{t\mathcal{T}-1} - d_k(\rho_3^{\mathcal{T}-1})\right) - \rho_{3k}^{\mathcal{T}-1}\right]\\
&\times \left[\rho_{3k} - \bar{\rho}_{3k}^{\mathcal{T}}\right] \geq 0, \qquad \forall (Q^1, Q^2, \gamma, \rho_3) \in \mathcal{K},
\end{aligned} \tag{30}$$

Step 2: Adaptation Compute $(Q^{1\mathcal{T}}, Q^{2\mathcal{T}}, \gamma^{\mathcal{T}}, \rho_3^{\mathcal{T}})$ by solving the variational inequality subproblem:

$$\sum_{g=1}^{G}\sum_{s=1}^{S}\left[q_{gs}^{\mathcal{T}} + a\left(\frac{\partial f_g(\bar{Q}^{1\mathcal{T}})}{\partial q_{gs}} + \frac{\partial c_{gs}(\bar{q}_{gs}^{\mathcal{T}})}{\partial q_{gs}} + \frac{\partial c_s(\bar{Q}^{1\mathcal{T}}, \bar{Q}^{2\mathcal{T}})}{\partial q_{gs}} + \frac{\partial \hat{c}_{gs}(\bar{q}_{gs}^{\mathcal{T}})}{\partial q_{gs}} - \bar{\gamma}_s^{\mathcal{T}}\right) - q_{gs}^{\mathcal{T}-1}\right] \times \left[q_{gs} - q_{gs}^{\mathcal{T}}\right]$$
$$+\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[q_{sk}^{t\mathcal{T}} + a\left(\frac{\partial c_s(\bar{Q}^{1\mathcal{T}}, \bar{Q}^{2\mathcal{T}})}{\partial q_{sk}^t} + \frac{\partial c_{sk}^t(\bar{q}_{sk}^{t\mathcal{T}})}{\partial q_{sk}^t} + \hat{c}_{sk}^t(\bar{Q}^{2\mathcal{T}}) + \bar{\gamma}_s^{\mathcal{T}} - \bar{\rho}_{3k}^{\mathcal{T}}\right) - q_{sk}^{t\mathcal{T}-1}\right] \times \left[q_{sk}^t - q_{sk}^{t\mathcal{T}}\right]$$
$$+\sum_{s=1}^{S}\left[\gamma_s^{\mathcal{T}} + a\left(\sum_{g=1}^{G}\bar{q}_{gs}^{\mathcal{T}} - \sum_{k=1}^{K}\sum_{t=1}^{T}\bar{q}_{sk}^{t\mathcal{T}}\right) - \gamma_s^{\mathcal{T}-1}\right] \times \left[\gamma_s - \gamma_s^{\mathcal{T}}\right]$$
$$+\sum_{k=1}^{K}\left[\rho_{3k}^{\mathcal{T}} + a\left(\sum_{s=1}^{S}\sum_{t=1}^{T}\bar{q}_{sk}^{t\mathcal{T}} - d_k(\bar{\rho}_3^{\mathcal{T}})\right) - \rho_{3k}^{\mathcal{T}-1}\right] \times \left[\rho_{3k} - \rho_{3k}^{\mathcal{T}}\right] \geq 0,$$
$$\forall (Q^1, Q^2, \gamma, \rho_3) \in \mathcal{K}, \tag{31}$$

Step 3: Convergence Verification If $|q_{gs}^{\mathcal{T}} - q_{gs}^{\mathcal{T}-1}| \leq \epsilon$, $|q_{sk}^{t\mathcal{T}} - q_{sk}^{t\mathcal{T}-1}| \leq \epsilon$, $|\gamma_s^{\mathcal{T}} - \gamma_s^{\mathcal{T}-1}| \leq \epsilon$, $|\rho_3^{\mathcal{T}} - \rho_3^{\mathcal{T}-1}| \leq \epsilon$, for all $g = 1, \cdots, G$; $s = 1, ..., S$; $k = 1, \cdots, K$; $t = 1, ..., T$, with $\epsilon > 0$, a prespecified tolerance, then stop; else, set $\mathcal{T} =: \mathcal{T} + 1$, and go to Step 1.

The following theorem states the convergence result for the modified projection method and is due to Korpelevich (1977).

Theorem 5 (Convergence). *Assume that the function that enters the variational inequality (21) (or (24)) has at least one solution and is monotone, that is,*

$$\langle (F(X') - F(X''))^T, X' - X'' \rangle \geq 0, \quad \forall X', X'' \in \mathcal{K}$$

and Lipschitz continuous. Then the modified projection method described above converges to the solution of the variational inequality (21) or (24).

The realization of the modified projection method in the context of the electric power supply chain network model takes on a very elegant form for computational purposes. In particular, the feasible set $\mathcal{K}$ is a Cartesian product, consisting of only nonnegativity constraints on the variables which allows for the network structure to be exploited. Hence, the induced quadratic programming problems in (30) and (31) can be solved explicitly and in closed

form using explicit formulae for the power flows between the tiers of the supply chain network, the demand market prices, and the optimal Lagrange multipliers.

Conditions for F to be monotone and Lipschitz continuous can be obtained from the results in Nagurney et al. (2002).

5 Numerical Examples

In this section, we apply the modified projection method to several numerical examples. The modified projection method was implemented in FORTRAN and the computer system used was a Sun system located at the University of Massachusetts at Amherst.

The convergence criterion utilized was that the absolute value of the flows (Q^1, Q^2) and the prices (γ, ρ_3)between two successive iterations differed by no more than 10^{-4}. For the examples, a was set to .05 in the algorithm, except where noted otherwise. The numerical examples had the network structure depicted in Fig. 2 and consisted of three power generators, two power suppliers, and three demand markets, with a single transmission service provider available to each power supplier.

The modified projection method was initialized by setting all variables equal to zero.

Example 1. The power generating cost functions for the power generators were given by:

$$f_1(q) = 2.5q_1^2 + q_1q_2 + 2q_1, \quad f_2(q) = 2.5q_2^2 + q_1q_2 + 2q_2,$$

$$f_3(q) = .5q_3^2 + .5q_1q_3 + 2q_3.$$

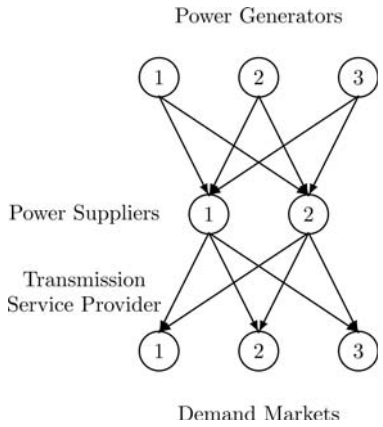

Fig. 2. Electric power supply chain network for the numerical examples

The transaction cost functions faced by the power generators and associated with transacting with the power suppliers were given by:

$$c_{11}(q_{11}) = .5q_{11}^2 + 3.5q_{11}, \quad c_{12}(q_{12}) = .5q_{12}^2 + 3.5q_{12},$$

$$c_{21}(q_{21}) = .5q_{21}^2 + 3.5q_{21}, \quad c_{22}(q_{22}) = .5q_{22}^2 + 3.5q_{22},$$

$$c_{31}(q_{31}) = .5q_{31}^2 + 2q_{31}, \quad c_{32}(q_{32}) = .5q_{32}^2 + 2q_{32}.$$

The operating costs of the power generators, in turn, were given by:

$$c_1(Q^1, Q^2) = .5(\sum_{i=1}^{2} q_{i1})^2, \quad c_2(Q^1, Q^2) = .5(\sum_{i=1}^{2} q_{i2})^2.$$

The demand functions at the demand markets were:

$$d_1(\rho_3) = -2\rho_{31} - 1.5\rho_{32} + 1100, \quad d_2(\rho_3) = -2\rho_{32} - 1.5\rho_{31} + 1100,$$

$$d_3(\rho_3) = -2\rho_{33} - 1.5\rho_{31} + 1200,$$

and the transaction costs between the power suppliers and the consumers at the demand markets were given by:

$$\hat{c}_{11}^1(Q^2) = q_{11}^1 + 5, \quad \hat{c}_{12}^1(Q^2) = q_{12}^1 + 5, \quad \hat{c}_{13}^1(Q^2) = q_{13}^1 + 5,$$

$$\hat{c}_{21}^1(Q^2) = q_{21}^1 + 5, \quad \hat{c}_{22}^1(Q^2) = q_{22}^1 + 5, \quad \hat{c}_{23}^1(Q^2) = q_{23}^1 + 5.$$

All other transaction costs were assumed to be equal to zero.

The modified projection method converged in 232 iterations and yielded the following equilibrium pattern:

$$q_{11}^* = q_{12}^* = q_{21}^* = q_{22}^* = 14.2762; \quad q_{31}^* = q_{32}^* = 57.6051,$$

$$q_{11}^{1*} = q_{12}^{1*} = q_{21}^{1*} = q_{22}^{1*} = 20.3861; \quad q_{31}^{1*} = q_{32}^{1*} = 45.3861.$$

The vector γ^* had components:

$$\gamma_1^* = \gamma_2^* = 277.2487,$$

and the demand prices at the demand markets were:

$$\rho_{31}^* = \rho_{32}^* = 302.6367; \quad \rho_{33}^* = 327.6367.$$

It is easy to verify that the optimality/equilibrium conditions were satisfied with good accuracy.

Example 2. We then constructed the following variant of Example 1. We kept the data identical to that in Example 1 except that we changed the first demand function so that:

$$d_1(\rho_3) = -2\rho_{33} - 1.5\rho_{31} + 1500.$$

The modified projection method converged in 398 iterations, yielding the following new equilibrium pattern:

$$q_{11}^* = q_{12}^* = q_{21}^* = q_{22}^* = 19.5994; \quad q_{31}^* = q_{32}^* = 78.8967,$$

$$q_{11}^{1*} = q_{21}^{1*} = 118.0985,$$

and all other q_{sk}^{1*}s= 0.0000. The vector γ^* had components:

$$\gamma_1^* = \gamma_2^* = 378.3891,$$

and the demand prices at the demand markets were:

$$\rho_{31}^* = 501.4873, \quad \rho_{32}^* = 173.8850, \quad \rho_{33}^* = 223.8850.$$

It is easy to verify that the optimality/equilibrium conditions were satisfied with good accuracy.

Note that with the increased demand at demand market 1 as evidenced through the new demand function, the demand price at that market increased. This was the only demand market that had positive electric power flowing into it; the other two demand markets had zero electric power consumed.

Example 3. We then modified Example 2 as follows: The data were identical to that in Example 2 except that we changed the coefficient preceding the first term in the power generating function associated with the first power generator so that rather than having the term $2.5q_1^2$ in $f_1(q)$ there was now the term $5q_1^2$. We also changed a to .03 since the modified projection method did not converge with $a = .05$. Note that a must lie in a certain range, which is data-dependent, for convergence.

The modified projection method converged in 633 iterations, yielding the following new equilibrium pattern:

$$q_{11}^* = q_{12}^* = 10.3716, \quad q_{21}^* = q_{22}^* = 21.8956, \quad q_{31}^* = q_{32}^* = 84.2407.$$

$$q_{11}^{1*} = q_{21}^{1*} = 116.5115,$$

with all other q_{sk}^{1*}s= 0.0000.

The vector γ^* had components:

$$\gamma_1^* = \gamma_2^* = 383.6027,$$

and the demand prices at the demand markets were:

$$\rho_{31}^* = 505.1135, \quad \rho_{32}^* = 171.1657, \quad \rho_{33}^* = 221.1657.$$

As expected, since the power generating cost function associated with the first power generator increased, the power that he generated decreased; the power generated by the two other power generators, on the other hand, increased. Again, as in Example 2, there was no demand (at the computed equilibrium prices) at the second and third demand markets.

Example 4. The fourth, and final example, was constructed as follows from Example 3. The data were all as in Example 3, but we now assumed that the demand functions were separable; hence, from each of the three demand market functions for electric power in Example 3, we eliminated the term not corresponding to the price at the specific market. In other words, the demand at demand market 1 only depended upon the price at demand market 1; the demand at demand market 2 only depended upon the demand at demand market 2; and the same held for the third demand market.

The modified projection method now converged in 325 iterations and yielded the following equilibrium electric power flow and price pattern:

$$q_{11}^* = q_{12}^* = 14.1801, \quad q_{21}^* = q_{22}^* = 29.9358, \quad q_{31}^* = q_{32}^* = 114.9917,$$

$$q_{11}^{1*} = q_{21}^{1*} = 111.3682, \quad q_{12}^{1*} = q_{22}^{1*} = 11.3683, \quad q_{13}^{1*} = q_{23}^{1*} = 36.3682.$$

The vector γ^* had components:

$$\gamma_1^* = \gamma_2^* = 522.2619,$$

whereas the equilibrium demand prices at the demand markets were now:

$$\rho_{31}^* = 638.6319, \quad \rho_{32}^* = 538.6319, \quad \rho_{33}^* = 563.6319.$$

Observe that since now there were no cross-terms in the demand functions, the electric power flows transacted between the suppliers and the demand markets were all positive. Of course, the incurred demands at both the second and third demand markets also increased. In addition, all the equilibrium flows from the power generators to the suppliers increased since there was increased demands at all the demand markets for electric power.

These numerical examples, although stylized, demonstrate the types of simulations that can be carried out. Indeed, one can easily investigate the effects on the equilibrium power flows and prices of such changes as: changes to the demand functions, to the power generating cost functions, as well as to the other cost functions. In addition, one can easily add or remove various decision-makers by changing the supply chain network structure (with the corresponding addition/removal of appropriate nodes and links) to investigate the effects of such market structure changes.

6 Conclusions and Future Research

In this chapter, we proposed a theoretically rigorous framework for the modeling, qualitative analysis, and computation of solutions to electric power market flows and prices in an equilibrium context based on a supply chain network approach. The theoretical analysis was based on finite-dimensional variational inequality theory.

We modeled the behavior of the decision-makers, derived the optimality conditions as well as the governing equilibrium conditions which reflect competition among decision-makers (in a game-theoretic framework) at the same tier of nodes but cooperation between tiers of nodes. The framework allows for the handling of as many power generators, power suppliers, transmission service providers, and demand markets, as mandated by the specific application. Moreover, the underlying functions associated with electric power generation, transmission, as well as consumption can be nonlinear and non-separable. The formulation of the equilibrium conditions was shown to be equivalent to a finite-dimensional variational inequality problem. The variational inequality problem was then utilized to obtain qualitative properties of the equilibrium flow and price pattern as well as to propose a computational procedure for the numerical determination of the equilibrium electric power prices and flows.

In addition, we illustrated both the model and computational procedure through several numerical examples in which the electric power flows as well as the prices at equilibrium were computed.

As mentioned in the Introduction, there are many ways in which this basic foundational framework can be extended, notably, through the incorporation of multicriteria decision-making associated with the decision-makers (with, for example, such criteria as environmental impacts, reliability, risk, etc.), the introduction of stochastic components, as well as the introduction of dynamics to study the disequilibrium electric power flows and prices.

Acknowledgements

The authors are grateful to the two anonymous referees and to the editor, Erricos Kontoghiorghes, for helpful comments and suggestions. This research was supported, in part, by an AT&T Industrial Ecology Fellowship. This support is gratefully appreciated.

This research was presented at CORS/INFORMS International Meeting, Banff, Alberta, Canada, May 16–19, 2004; and The International Conference on Computing, Communication and Control Technologies: CCCT'04, Austin, Texas, August 14–17, 2004. An earlier version of this work appears in *Proceedings of the International Conference in Computing, Communications and Control Technologies*, Austin, Texas, Vol. VI, 127–134.

References

Bazaraa, M. S., H. D. Sherali, and C. M. Shetty: 1993, *Nonlinear Programming : Theory and Algorithms*. New York, NY: Wiley.

Bertsekas, D. P. and J. N. Tsitsiklis: 1997, *Parallel and Distributed Computation : Numerical Methods*. Belmont, MA: Athena Scientific.

Boucher, J. and Y. Smeers: 2001, 'Alternative Models of Restructured Electricity Systems, Part 1: No Market Power'. *Operations Research* **49**(6), 821–838.

Casazza, J. and F. Delea: 2003, *Understanding Electric Power Systems : An Overview of the Technology and the Marketpalce*. Piscataway, NJ; Hoboken, NJ: IEEE Press; Wiley Interscience.

Chao, H. P. and S. Peck: 1996, 'A Market Mechanism for Electric Power Transmission'. *Journal of Regulatory Economics* **10**(1), 25–59.

Dafermos, S. and A. Nagurney: 1987, 'Oligopolistic and Competitive Behavior of Spatially Separated Markets'. *Regional Science and Urban Economics* **17**(2), 245–254.

Daxhelet, O. and Y. Smeers: 2001, *Variational Inequality Models of Restructured Electric Systems*, Complementarity: Applications, Algorithms and Extensions. Berlin; New York: Springer.

Day, C. J., B. F. Hobbs, and J.-S. Pang: 2002, 'Oligopolistic Competition in Power Networks: A Conjectured Supply Function Approach'. *IEEE Transactions on Power Systems* **17**(3), 597–607.

Geunes, J. and P. M. Pardalos: 2003, 'Network Optimisation in Supply Chain Management and Financial Engineering: An Annotated Bibliography'. *Networks* **42**(2), 66–84.

Hobbs, B. F.: 2001, 'Linear Complementarity Models of Nash-Cournot Competition in Bilateral and POOLCO Power Markets'. *IEEE Transactions on Power Systems* **16**(2), 194–202.

Hobbs, B. F., C. B. Metzler, and J. S. Pang: 2000, 'Strategic Gaming Analysis for Electric Power Systems: an MPEC Approach'. *IEEE Transactions on Power Systems* **15**(2), 638–645.

Hogan, W. W.: 1992, 'Contract Networks for Electric-Power Transmission'. *Journal of Regulatory Economics* **4**(3), 211–242.

Jing-Yuan, W. and Y. Smeers: 1999, 'Spatial Oligopolistic Electricity Models with Cournot Generators and Regulated Transmission Prices'. *Operations Research* **47**(1), 102–112.

Kahn, A. E.: 1998, 'Electric Deregulation: Defining and Ensuring Fair Competition'. *The Electricity Journal* **11**(3), 39–49.

Kinderlehrer, D. and G. Stampacchia: 1980, *An Introduction to Variational Inequalities and Their Applications*, Vol. 88. New York, NY: Academic Press.

Korpelevich, G. M.: 1977, 'Extragradient Method for Finding Saddle Points and Other Problems'. *Matekon* **13**(4), 35–49.

Nagurney, A.: 1999, *Network Economics: A Variational Inequality Approach, Rev. 2nd*, Vol. 10. Boston, MA: Kluwer.

Nagurney, A. and J. Dong: 2002, *Supernetworks: Decision-Making for the Information Age*. Cheltenham, UK; Northampton, MA, USA: Edward Elgar.

Nagurney, A., J. Dong, and D. Zhang: 2002, 'A Supply Chain Network Equilibrium Model'. *Transportation Research: Part E* **38**(5), 281.

Nash, J. F.: 1950, 'Equilibrium Points in n-Person Games'. *Proceedings of the National Academy of Sciences of the United States of America* **36**(1), 48–49.

Nash, J. F.: 1951, 'Non-Cooperative Games'. *Annals of Mathematics* **54**(2), 286–295.

North American Electric Reliability Council: 1998, 'Reliability Assessment 1998–2007'. Technical report.

Schweppe, F. C.: 1988, *Spot Pricing of Electricity*, Vol. SECS 46. Boston, MA: Kluwer.

Secretary of Energy Advisory Board's Task Force on Electric System Reliability: 1998, 'Maintaining Reliability in a Competitive US Electric Industry'. Technical report.

Singh, H.: 1998, *IEEE Tutorial on Game Theory Applications to Power Markets*. IEEE.

Takriti, S., B. Krasenbrink, and L. S. Y. Wu: 2000, 'Incorporating Fuel Constraints and Electricity Spot Prices into the Stochastic Unit Commitment Problem'. *Operations Research* **48**(2), 268–280.

US-Canada Power System Outage Task Force: 2004, 'Final Report on the August 14th, 2003 Blackout in the United States and Canada: Causes and Recommendations'. Technical report.

US Energy Information Administration: 2002, 'The Changing Structure of the Electric Power Industry 2000: An Update'. Technical report.

Wu, F., P. Varaiya, P. Spiller, and S. Oren: 1996, 'Folk Theorems on Transmission Access: Proofs and Counterexamples'. *Journal of Regulatory Economics* **10**(1), 5–23.

Zaccour, G.: 1998, *Deregulation of Electric Utilities*, Vol. 28. Boston, MA: Kluwer.

Worst-Case Modelling for Management Decisions under Incomplete Information, with Application to Electricity Spot Markets

Mercedes Esteban-Bravo[1] and Berc Rustem[2]

[1] Department of Business Administration, Universidad Carlos III de Madrid, Spain
[2] Department of Computing, Imperial College London, UK

Summary. Many economic sectors often collect significantly less data than would be required to analyze related standard decision problems. This is because the demand for some data can be intrusive to the participants of the economy in terms of time and sensitivity. The problem of modelling and solving decision models when relevant empirical information is incomplete is addressed. First, a procedure is presented for adjusting the parameters of a model which is robust against the worst-case values of unobserved data. Second, a scenario tree approach is considered to deal with the randomness of the dynamic economic model and equilibria is computed using an interior-point algorithm. This methodology is implemented in the Australian deregulated electricity market. Although a simplified model of the market and limited information on the production side are considered, the results are very encouraging since the pattern of equilibrium prices is forecasted.

Key words: Economic modelling, equilibrium, worst-case, scenario tree, interior-point methods, electricity spot market

1 Introduction

Decision makers need to build and solve stochastic dynamic decision models to make planning decisions accurately. Three steps are involved. The first is the specification of the structure of the stochastic dynamic decision model reflecting the essential economic considerations. The second step is the calibration of the parameters of the model. The final step is the computation of the model's outcome for forecasting and/or simulating economic problems.

In the first part of this paper, we propose an integrated approach to address this problem. The first task, the specification of the model, involves a trade-off between complexity and realism. A more realistic model is usually

a multistage stochastic problem that will become increasingly impractical as the problem size increases. In general, any multistage stochastic problem is characterized by an underlying exogenous random process whose realizations are data trajectories in a probability space. The decision variables of the model are measurable functions of these realizations. A discrete scenario approximation of the underlying random process is needed for any application of the stochastic problem. This field of research has became very popular due to the large number of finance and engineering applications. For example, (Bounder, 1997), (Kouwenberg, 2001) and (Høyland and Wallace, 2001) developed and employed scenarios trees for a stochastic multistage asset-allocation problem. (Escudero, Fuente, García and Prieto, 1996), among others, considered scenarios trees for planning the production of hydropower systems. We obtain a discrete approximation of the stochastic dynamic problem using the simulation and randomized clustering approach proposed by (Gülpinar, Rustem and Settergren, 2004). In particular, we consider a scenario tree approach to approximate the stochastic random shocks process that affects the market demand.

On the other hand, firms make decisions on production, advertisement, etc. within the constraints of their technological knowledge and financial contracts. In many actual production processes, these constraints contain parameters, often unknown even when they have physical meaning. Decision makers do not usually observe all data required to estimate accurately the parameters of the model. For example, decision makers often lack enough information on the specifications of competitors. In these circumstances, standard econometric techniques cannot help to estimate the parameters of an economic model and still, decision makers require a full specification of the market to design optimal strategies that optimize their returns.

We propose a robust methodology to calibrate the parameters of a model using limited information. The robustness in the calibration of the model is achieved by a worst-case approach. Worst-case modelling essentially consists of designing the model that best fits the available data in view of the worst-case outcome of unobserved decision variables. This is a robust procedure for adjusting parameters with insurance against unknown data.

In the economic context, this approach turns out especially interesting to study situations in which a structural change takes place, for example when there are changes in the technologies of firms or a new firm enters the economy. As a consequence of these exogenous perturbations, the empirical data generating process is modified and classical estimations cannot be made. In this context, a model in which decision makers assess the worst-case effect of the unobserved data is a valuable tool for the decision maker against a risk in future decisions. Worst-case techniques has been applied in n-person games to study decision making in real-world conflict situations (see for example Rosen, 1965). In a worst-case strategy, decision makers seek to minimize the maximum damage their competitor can do. When the competitor can be interpreted as nature, the worst-case strategy seek optimal responses in the

worst-case value of uncertainty. Choosing the worst-case parameters requires the solution of a min-max continuous problem. Pioneering contributions to the study of this problem have been made by (Danskin, 1967) and (Bram, 1966), while computational methods are discussed in (Rustem and Howe, 2002).

The third and final task is the computation of the equilibrium values (decisions and prices) for each scenario. We consider a variant of the interior-point method presented in (Esteban-Bravo, 2004) to compute equilibria of stochastic dynamic models.

In the second part of the paper, we consider the deregulated electricity market in NSW Australia to illustrate the applicability of this methodology. In recent years, the theoretical and empirical study of the electricity market has attracted considerable attention. In particular, the ongoing liberalization process in the electricity markets has created a significant interest in the development of economic models that may represent the behaviour of these markets (a detailed review on this literature can be found in Schweppe, Carmanis, Tabors and Bohn 1988, Kahn 1998, Green 2000, and Boucher and Smeers 2001). One of the key characteristics of these markets is that their databases often collect significantly less variables than necessary for building useful economic models. This is because the demand for some data can be intrusive to the firm in terms of time and sensitivity.

We consider a model that focuses on the effect we hope to study in detail: the process of spot prices. Similar selective approaches are adopted for the decision analysis of dispatchers (Sheblé, 1999), the financial system as a hedge against risk (e.g. Bessembinder and Lemmon, 2002), the externalities given by network effects (e.g. Hobbs 1986 and Jing-Yuan and Smeers 1999).

First, the model developed forecasts daily electricity demand. We assume that the demand is affected by exogenous factors and by an underlying stochastic random process. The discrete outcomes for this random process is generated using the simulation and randomized clustering approach proposed by (Gülpinar, Rustem and Settergren, 2004).

Our model for generators is a simplification of the standard models in the literature. We do not attempt to provide a realistic description of the underlying engineering problems in electricity markets. The literature in this area is extensive (e.g. McCalley and Sheblé 1994). Our aim is to forecast the process of spot prices using limited information on the production side. The knowledge of these prices is the basic descriptive and predictive tool for designing optimal strategies that tackle competition. Some authors have studied spot markets assuming a known probability distribution for spot prices (see, e.g., Neame, Philpott and Pritchard 2003), or considering spot prices as nonstationary stochastic processes (see Valenzuela and Mazumdar 2001, Pritchard and Zakeri 2003, and the references therein). We consider economic equilibrium models to this end. We simplify the effects of the transmission constraints dictated by Kirchoff's laws ((Schweppe, Carmanis, Tabors and Bohn, 1988) and (Hsu, 1997) also consider a simplified model of transmission network). This may be acceptable as we consider managing decisions using

limited information. In any case, the approach presented here can be applied to any other modelling choices which include other phases of the electricity trading and other models of competition (as those presented in Day, Hobbs and Pang 2002).

We apply a worst-case approach to provide indicative values of the parameters in the model using the information available. The worst-case criteria ensures robustness to calibrate these parameters. Robustness is ensured as the best parameter choice is determined simultaneously with the worst-case outcome of unobserved data.

Finally, we compute the expected value of future equilibrium prices and we see that the model captures the essential features of the prices' behaviour. From the analysis of the results, we can conclude that this approach is able to forecast the pattern of equilibrium prices using limited information on the production side.

2 The Methodology

The design of an economic model describing the main features of a certain managerial problem is an essential step for decision makers. The model should allow the practitioner to forecast and design economic policies that reduce, for example, the production cost and market prices. The dynamic stochastic framework has been extensively used in economics to model almost any problem involving sequential decision-making over time and under uncertainty.

Consumers are the agents making consumption plans. Market demand reflects the consumer's decisions as the demand curve shows the quantity of a product demanded in the market over a specified time period and state of nature, at each possible price. Demand could be influenced by income, tastes and the prices of all other goods. The study of demand pattern is one of the key steps in managerial problems.

Firms make decisions on production, advertisement, etc. within the constraints of their technological knowledge and financial contracts. In particular, firms should maximize their expected profits subject to technological and risk constraints. In many actual production processes, these constraints contain parameters, often unknown even when they have physical meaning. Prices could be decision variables as in Cournot models, or could be considered as parameters as in perfect competition models.

Market equilibrium y is a vector of decision variables of agents (consumers and firms) and prices that makes all decisions compatible with one another (i.e. y clears the market in competitive models or y satisfies Nash equilibrium in strategic models). In general, an equilibrium y can be characterized by a system of nonlinear equations $H(\theta, y, x) = 0$, where θ is a vector of parameters, and x is a vector of exogenous variables that affects agents' decisions through technologies and tastes.

To obtain predictive models for decision makers, we face the problem of having to estimate several parameters θ. The optimal determination of these parameters is essential for building economic models that can address a large class of questions. Although some of the parameters can be calibrated easily using the available data, others remain uncertain due to the lack of empirical information. We propose a worst-case strategy to adjust or calibrate these parameters to the model using limited empirical data.

2.1 Worst-Case Modelling

Some of the variables (y, x) can be empirically determined (observed data). Let z be the vector of non-observable variables, r be the vector of observable variables, and let $H(\theta, z, r) = 0$ denote the system of nonlinear equations that characterize an equilibrium of the economy, where θ is a vector of parameters. The aim of the worst-case modelling is essentially to fit the best model (the best choice of parameters θ) to available data in view of the worst-case unobservable decision z. When designing economic models, the worst-case design problem is a continuous minimax problem of the form

$$\min_{\theta\in\Theta, r\in R} \max_{z\in Z} \|r - \widehat{r}\|_2^2 \quad \textit{subject to} \quad H(\theta, z, r) = 0, \tag{1}$$

where $\Theta \subset \mathbb{R}^n$ is the feasible set of parameters, $R \subset \mathbb{R}^m$ is the feasible set of observable variables, $Z \subset \mathbb{R}^l$ is the feasible set of non-observable variables and $\widehat{r}$ is a data sample of r. In other words, our aim is to minimize the maximum deviation for the worst-scenario of realizable decisions. Thus, the optimal solution θ^* to this problem defines a robust optimal specification of the economic model. This criterion for choosing parameters typically can be applied to engineering, economics and finance frameworks.

For solving continuous minimax problems we use the global optimisation algorithm developed by (Žaković and Rustem, 2003). They consider an algorithm for solving semi-infinite programming problem since any continuous minimax problem of the form

$$\min_{\theta\in\Theta} \max_{z\in Z} \{f(\theta, z) : g(\theta, z) = 0, \} \tag{2}$$

can be written as a semi-infinite programming problem. Note that the above problem is equivalent to

$$\min_{\theta\in\Theta,\rho} \quad \left\{\rho : \max_{z\in Z} \{f(\theta, z) \le \rho : g(\theta, z) = 0\}\right\}, \tag{3}$$

and since $\max_{z\in Z} f(\theta, z) \le \rho$ if and only if $f(\theta, z) \le \rho$, for all $z \in Z$, we can solve the alternative semi-infinite problem:

$$\begin{array}{ll} \min_{\theta\in\Theta,\rho} & \rho \\ \textit{subject to} & f(\theta, z) \le \rho, \quad \forall \quad z \in Z, \\ & g(\theta, z) = 0, \quad \forall \quad z \in Z. \end{array} \tag{4}$$

Žaković and Rustem's algorithm involves the use of global optimisation to compute the global worst-case. The global optimisation approach is essential to guarantee the robustness property of the solution of the minimax problems. This is because a crucial step to solve the semi-infinite problem is to find $\theta \in \Theta$, $f(\theta, z) \leq \rho$, $g(\theta, z) = 0$, for all $z \in Z$. To reduce the cost of computing global optima, it is recommended to restrict the domains Θ and Z as much as possible given the information available. The monograph edited by (Pardalos and Resende, 2002) reviews the global optimisation literature (see Chap. 6).

2.2 Modelling the Uncertainty

As discussed in the introduction, the importance of considering uncertainty via scenarios is well known in finance and engineering applications. In this section, we extend the scenario tree methodology to the computation of equilibria in stochastic dynamic economic models. In such models, agents (consumers and firms) face a problem involving sequential decision making over time and under uncertainty. Given the parameters $\theta \in \Theta$ calibrated using the available information, assume that each agent face the decision problem:

$$\max_{x_t} \sum_{t=0}^{T} E\left[U_\theta\left(x_t, a_t, t\right)\right] \quad subject \;\; to \;\; g_\theta\left(x_t, a_t, t\right) \leq 0 \;\; a.e., \tag{5}$$

where $\{x_t\}$ are the decision variables, $\{a_t\}$ are observable Markovian random variables with a continuous distribution function, $U_\theta\left(x_t, a_t, t\right)$ represents the agents' preferences and a.e. denotes "almost everywhere". This decision model will be characterized by the information available at each period of time, among other things. We assume that this information is the same for all agents. Let σ_t be the σ-algebra generated by $\{a_s : 0 \leq s \leq t\}$ and let $\{\sigma_t\}$ be the complete specification of the revelation of information through time, called filtration.

To reduce the cost of computing optima, we approximate the process $\{a_t\}$ and the associated information set $\{\sigma_t\}$ by a discrete process $\{a_{t,s}\}_{s=1}^{S_t}$ of possible outcomes for each t, and a discrete information structure $\{\mathcal{F}_t\}_{t=1}^{T}$. A discrete information structure is formally defined as follows: Given a finite sample space $\Omega = \{\omega_1, \ldots, \omega_M\}$ that represents the states of world, a discrete information structure is a sequence of σ-algebras $\{\mathcal{F}_t\}_{t=1}^{T}$ such that: 1) $\mathcal{F}_1 = \{\Omega, \emptyset\}$, 2) $\mathcal{F}_T = 2^\Omega$, 3) $\mathcal{F}_{t+1}$ is finer than $\mathcal{F}_t$, $\forall t = 1, \ldots, T-1$. The *scenario tree* associated with the discrete information structure $\{\mathcal{F}_t\}_{t=1}^{T}$ is defined as $\Im = \bigcup_{t \in \mathbb{T},\, s \in \mathbb{S}_t} (t, s)$, where $\mathbb{T} = \{0, \ldots, T\}$ and $\mathbb{S}_t = \{1, \ldots, S_t\}$. Each (t, s) is called a *tree node* or *scenario*. For each scenario tree we can define a preorder relation $\succ$ such that $(t, s) \succ (t', s')$ if and only if the node (t', s') comes after (t, s) in the tree, that is, if $t' > t$ and $s' \subset s$.

Two main approaches to generate discrete scenario trees have been considered to date. The first one is known as the optimisation approach. This

method considers the relevant statistical properties of the random variable such as the first four moments of the marginal distributions. Then a nonlinear optimisation problem is formulated where the objective is to minimize the square distance between the statistical properties of the constructed tree and the actual specifications. The second approach is called the simulation approach and only uses the sample from the fitted cumulative distribution function. In this paper, we generate discrete scenario trees using the simulation and randomized clustering approach proposed by (Gülpinar, Rustem and Settergren, 2004). This method is a simulation-based approach that clusters scenarios randomly.

2.3 Computing Stochastic Dynamic Equilibria

Once the uncertainty of the problem is represented by a discrete scenario tree, the stochastic dynamic decision problem of each agent can be written as follows:

$$\max_{x_{t,s}} \sum_{t=0}^{T} \sum_{s=0}^{S_t} \beta_{ts}\, U_\theta\left(x_{t,s}, a_{t,s}, t\right) \quad subject\ \ to\ \ g_\theta\left(x_{t,s}, a_{t,s}, t\right) \leq 0, \forall\left(t, s\right), \tag{6}$$

where $\{\beta_{ts}\}$ are the conditional probabilities associated to state s at each period t, with $\beta_0 = 1$, and $\lambda \geq 0$ denotes the Lagrange's multipliers associated with the inequality constraints. Under appropriate convexity assumptions, equilibria are characterized by the first-order conditions of the agents' problems and the market clearing conditions that define the economic model. These optimality conditions can be seen as a special class of problems known as nonlinear complementarity (complementarity conditions stem from complementarity slackness in the first-order optimality conditions). Mathematically, these problems are stated as follows: find $p^T = (x^T, \lambda^T) \geq 0$ such that $F(p) \geq 0$ and $p^T F(p) = 0$. A nonlinear complementarity problem can be reformulated as a standard system of equations $H(z) = 0$, where

$$H(z) = \begin{pmatrix} p^T F(p) \\ F(p) - s \end{pmatrix}, \tag{7}$$

s are *slack variables* and $z^T = (x^T, \lambda^T, s^T) \geq 0$. Often, the decision variables, the Lagrange's multipliers and the slack variables may take any value within a certain range bounded by positive finite lower and upper bounds, $l \leq z \leq u$. A brief summary of standard approaches for solving these problems can be found in (Esteban-Bravo, 2004). In Chap. 13, (Pardalos and Resende, 2002) provide an excellent introduction to complementarity and related problems.

The final stage of the methodology is the computation of equilibria for the stochastic economic model using the generated scenario tree. In this paper,

we consider a version of the interior point method given in (Esteban-Bravo, 2004). This algorithm can find accurate solutions with little computational cost, what it is a desirable property as the scenario tree can be expanded to arbitrarily large sizes as the temporal horizon increases. The main idea of the algorithm is the application of the Gauss-Newton method to solve the following perturbed system of nonlinear equations,

$$\begin{aligned} J(z_k)^T H(z_k) - w_k^1 + w_k^2 &= 0, \\ (Z_k - L)\, W_k^1 - \mu &= 0, \\ (U - Z_k)\, W_k^2 - \mu &= 0, \\ w_k^1, w_k^2 &> 0, \end{aligned}$$

where $Z_k = diag(z_k)$, $L = diag\,(l)$, $U = diag\,(u)$ and $J(z_k)$ denote the Jacobian matrix of H. Note that when $\mu \to 0$, we compute the original problem. Following the Gauss-Newton approach, the Hessian of the perturbed system is approximated by its first term. As a consequence, this algorithm has the very desirable property that it finds accurate solutions with little computational cost.

3 Modelling the NSW Spot Electricity Market

In this section, we focus on an application of the robust modelling methodology for the deregulated electricity market in NSW, Australia. The deregulated electricity market should be modelled as a sequential trade for goods and assets. A model of sequential markets is a system of reopening spot markets, which is a market for immediate delivery. In other words, a seller and a buyer agree upon a price for a certain amount of electric power (MWs) to be delivered at the current period (in case of electricity markets, in the near future). This agreement is monitored by an independent contract administrator who matches the bids of buyers and sellers.

We consider an economy with three generators that face the NSW Electricity System. In NSW electricity markets, the role of a financial contract is small and, as a consequence, we only focus on spot markets that trade most of the local electricity.

To meet electricity demand and for the spot electricity market to operate efficiently, a reliable forecast of daily electricity demand is required. Typically, the electricity demand is affected by several exogenous variables such as air temperature, and varies seasonally (the total demand will generally be lower over weekend days than weekdays, and higher in summer or winter than in fall or spring). Electricity forecasting process must therefore consider both aspects. The time spans involved in electricity forecasts may range from half an hour to the next few days. The technique described in this paper considers the day-to-day forecast as the aim is to guide decisions on capacity, cost and availability to meet the demand or the necessity to purchase from other

producers. In the very short-term electricity market, the demand varies little in response to price changes so we can say that the demand is not affected by prices, i.e. it is inelastic within the observed range of prices variation.

Generators make decisions about the amount of electricity to produce within the constraints of their technological knowledge. Modelling the technologies of a generation company requires special attention. Generators can produce electricity by means of hydro, thermal and pumped storage plants. A pumped storage hydro plant is designed to save fuel costs by serving the peak load (a high fuel-cost load) with hydro energy and then pumping the water back up into the reservoir at light load periods (a lower cost load). Moreover, generators face uncertainty because of the inflows in the case of hydro generation and the price of fuel in the case of thermal and pumped storage. As the generation system in NSW is overwhelmingly thermally based, we just consider this kind of technology.

3.1 The Demand

The problem of modelling the pattern of the electricity demand has previously been studied in the literature; see e.g. (Rhys, 1984), (Harvey and Koopman, 1993), (Henley and Peirson, 1997), (Valenzuela and Mazumdar, 2000), among others. In this paper, we assume that the daily electricity demand is affected exogenously by air temperature. In addition, we take into account its daily pattern. Note that the total electricity demand is generally lower over weekend days than weekdays, and higher in summer or winter than in fall or spring. Also in the short term we can assume that the aggregate demand for electricity is inelastic, as the quantity of power purchased varies little in response to price changes.

We consider electricity demand data for each day in New South Wales, Australia, between 1999 and 2002 (see http://www.nemmco.com.au/data/). The data sequence starts at January 1, 1999 and ends at April 30, 2002. All values are in MW and according to Eastern Standard Time. The result is a sequence of 1247 values (see Fig. 3 in Appendix). Let x_j denote the electricity demand at the j-th day. The data for temperature are drawn from the file AUCNBERA.txt given in the website http://www.engr.udayton.edu/weather. This dataset contains information on the daily average temperatures for Canberra. Let f_j denote the average temperature (^{o}F) at the j-th day.

With a daily temporal frequency, there are patterns repeated over a stretch of observations. In particular, we observe two seasonal effects: weekly (such as weekdays and weekends) and monthly (such summer and winter). Assuming that these effects follow a deterministic pattern, we consider stational dummy variables. Let $d_{jt} = 1$, if the t-th observation is a j-th day and $d_{jt} = 0$, otherwise, defining the weekly periodic effects, with $j = 1$ for Mondays, $j = 2$ for Tuesdays, and so on; and $\delta_{jt} = 1$, if the t-th observation is a j-th month and $\delta_{jt} = 0$, otherwise, defining the monthly periodic effects, with

$j = 1$ for January, $j = 2$ for February and so on. Thus, given $n = 1247$ pairs of observations (x_j, f_j), with $j = 1, \ldots, n$, we consider the following regression model:

$$\begin{aligned} X_t = \mu + c_1 f_j + c_2 f_j^2 + \sum_{j=1}^{6} \gamma_j \left(d_{jt} - d_{7t}\right) + \sum_{j=1}^{6} \gamma_j' \left(d_{jt} - d_{7t}\right) f_j \\ + \sum_{j=1}^{11} \beta_j \left(\delta_{jt} - \delta_{12t}\right) + \sum_{j=1}^{11} \beta_j' \left(\delta_{jt} - \delta_{12t}\right) f_j \\ + \sum_{j=1}^{6} \gamma_j'' \left(d_{jt} - d_{7t}\right) f_j^2 + \sum_{j=1}^{11} \beta_j'' \left(\delta_{jt} - \delta_{12t}\right) f_j^2 + \varepsilon_t, \end{aligned} \tag{8}$$

where $\{\varepsilon_j\}_j$ is a Gaussian and second order stationary process with zero mean and covariance function, $\gamma(s) = E[\varepsilon_t \varepsilon_{t-s}]$. This specification avoids the multicollinearity problems derived from the fact that $\sum_{j=1}^{7} d_{jt} = 1$ and $\sum_{j=1}^{12} \delta_{jt} = 1$. In particular, we consider a linear regression model using all the dummies variables and assuming that $\sum_{j=1}^{7} \gamma_j = 0$, and $\sum_{j=1}^{12} \beta_j = 0$. For an introduction to the estimation of this type of models see e.g. (Brockwell and Davis, 1987).

The regression parameters were estimated by the ordinary least squares (OLS) method using STATA (see http://www.stata.com/). Least-square regression estimations can be found in Appendix. The degree of explanation of this model is quite significant, as its R-Squared and adjusted R-Squared values are 0.7793 and 0.7695, respectively.

The study of the plots of residual autocorrelation and partial autocorrelation estimates (see below Fig. 4 in Appendix) suggests an autoregressive AR(1) model for the perturbation ε_t. Model (2) considers an autoregressive AR(1) specification for the process $\{\varepsilon_j\}_j$, $\varepsilon_j = \tau \varepsilon_{j-1} + a_j$, where $|\tau| < 1$ and $\{a_j\}_j$ are independent identically distributed disturbances with zero mean and constant variance, σ_a^2. Using STATA to estimate Model (2) by the OLS method, its coefficient estimation is $\widehat{\tau} = 0.60069$, with $\sigma_a^2 = 1.5402e$, and its R-Squared and adjusted R-Squared values are 0.3687 and 0.3682, respectively. Simple and partial autocorrelations of its residuals, shown in Fig. 5 in Appendix, reveal that Model (1) and Model (2) fit data.

The Markovian process $\{a_t\}_t$ is approximated by a discrete scenario tree $\{a_{ts}\}$ as presented in Sect. 2.2. Following the simulation and randomized clustering approach proposed in (Gülpinar, Rustem and Settergren, 2004) and given the covariance matrix σ_a^2, we construct a tree with a planning horizon of $T + 1$ days (today and T future periods of time) and a branching structure of $1 - 2 - 4 - 6$. This means that the tree has an initial node at day 0, 2 at day 1,... The scenario tree provides information about the probabilities β_{ts} associated with the different states s at each period t, with $\beta_0 = 1$, and the

$AR(1)$ stochastic process of error terms a_{ts} at each state s and period t. The values of these elements can be found in Appendix.

3.2 The Problem of the Electricity Generators

As we mentioned before, in the very short-term electricity market, the demand varies little in response to price changes. In this applications where the planning time horizon is assumed to be three days or periods of time, we should consider a pure competitive behaviour of generators rather than oligopolistic strategies. Therefore, in a deregulated environment, the purpose of the short-term generator is to maximize its expected profit on its technological constraints over a time period of length $T+1$, today and the planning time horizon. This means that each generator collects its revenue from selling electricity at spot prices in the spot market. There are network capacity constraints affecting generators and therefore the total amount of electricity that these generators can produce will be bound by the network externalities. The notation used to present the problem of the electricity generators is the following, at each period $t = 0, 1, \ldots, T$:

Decision variables:

p_t, spot price,

y_{jt}, spot electricity production of generator j,

w_{jt}, input of generator j.

Parameters:

T, maximum number of periods,

J, number of generators,

$0 \leq r$, discount rate for generators,

q_{jt}, unit generation cost (input's price) of generator j,

A_j, a_j, parameters associated with the technology of generator j,

N, maximum capacity of the network,

M, rate limit to generation over two periods,

l_j, u_j, minimum and maximum of generation capacity of generator j, respectively.

The generation constraints are:

Cobb-Douglas type technological constraint: $y_{jt} \leq A_j w_{jt}^{a_j}$.

Network capacity constraint: $\sum_{j' \neq j} y_{j't} + y_{jt} \leq N$.

Rate limit to generation over two consecutive periods: $y_{jt} - y_{jt-1} \leq M$.

We consider a market with three generators, $j = 1, \ldots, J$ with $J = 3$, that aim to maximize the expected revenues and minimize the expected costs:

$$\sum_{t=0}^{T} \left(\frac{1}{1+r} \right)^t \left[p_t \cdot y_{jt} - q_{jt} \cdot w_{jt} \right]. \tag{9}$$

Thus, the decision problem of each generator is given by

$$\begin{aligned} \max_{y_j, w_j} \quad & \sum_{t=0}^{T} \left(\tfrac{1}{1+r}\right)^t [p_t \cdot y_{jt} - q_{jt} \cdot w_{jt}] \\ \textit{subject to } & y_{jt} \leq A_j w_{jt}^{a_j}, \forall t, \\ & \sum_{j' \neq j}^{T} y_{j't} + y_{jt} \leq N, \forall t, \\ & y_{jt} - y_{jt-1} \leq M, \forall t, \\ & l_j \leq y_{jt} \leq u_j, \forall t, \\ & 0 \leq w_j. \end{aligned} \tag{10}$$

Most of studies on the generator's problem have been based on a cost function (i.e. the minimum cost of producing a given level of output from a specific set of inputs), even though this formulation is equivalent to the one that uses technology constraints (this is proved by dual arguments, see e.g. Varian, 1992 and Mas-Colell, Whinston and Green, 1995). But, in case of having limited information, the formulation with technological constraints is more recommendable as the specification of cost functions requires detailed information on the labor costs, inputs costs, and buildings and machinery amortization, among others. In particular, we choose a Cobb-Douglas technology as this specification is characterized by a ready capability to adapt to new, different, or changing requirements. For example, they can exhibit increasing, decreasing or constant return to scale depending on the values of their parameters. Furthermore, we can readily derive the analytical form of its associated cost function.

On the other hand, the modelling considered here has strong simplifications on the transmission side although these simplifications could have effects for the analysis. This is because we lack sufficient information to calibrate this externality.

Next, we introduce the concept of equilibrium, the basic descriptive and predictive tool for economists. The equilibrium of this economy is a vector prices p^* and an allocation $\left(y_j^*, w_j^*\right)$ for all $j = 1, \ldots, J$, that satisfies:

- For each $j = 1, \ldots, J$, $\left(y_j^*, w_j^*\right)$ is the solution of Problem (10).
- Generators fulfil market demand, i.e.

$$\sum_j y_{jt}^* = \widehat{X}_t, \forall t, \tag{11}$$

where $\widehat{X}$ is the estimation of the market demand determined by Model (8).

Then, under appropriate convexity assumptions, equilibria can be characterized by the first order conditions of all generators' problems (10) and the market clearing conditions (11). In other words, the vector (p^*, y^*, w^*) is an

equilibrium if, for all $j = 1, \ldots, J$, there exist Lagrange multiplier vectors $\gamma_j^1, \gamma_j^2, \gamma_j^3 \geq 0$, such that:

$$\begin{aligned}
&\left(\tfrac{1}{1+r}\right)^t \; p_t^* - \gamma_{jt}^1 - \gamma_{jt}^2 + \gamma_{jt}^3 = 0, \forall t, \\
&-\left(\tfrac{1}{1+r}\right)^t \; q_{jt} + \gamma_{jt}^1 A_j a_j w_{jt}^{*a_j - 1} = 0, \forall t, \\
&y_{jt}^* - A_j w_{jt}^{*a_j} + h_{jt}^1 = 0, \forall t, \\
&\gamma_{jt}^1 h_{jt}^1 = 0, \forall t, \\
&\textstyle\sum_j y_{jt}^* + h_{jt}^2 - N = 0, \forall t, \\
&\gamma_{jt}^2 h_{jt}^2 = 0, \forall t, \\
&y_{jt}^* - y_{jt-1}^* + h_{jt}^3 - M = 0, \forall t, \\
&\gamma_{jt}^3 h_{jt}^3 = 0, \forall t, \\
&\textstyle\sum_j y_{jt}^* = \widehat{X}_t, \forall t, \\
&l_j \leq y_{jt}^* \leq u_j, \\
&0 \leq w_j^*,
\end{aligned} \tag{12}$$

where h^1, h^2 and $h^3 \geq 0$ are slack variables.

3.3 Worst-Case Calibration

To obtain predictive decision models for generators, we are faced with the problem of having to estimate several parameters. The optimal calibration of these parameters is the aim of this section.

As we mentioned before, some of the parameters can be calibrated easily. Given that the planning horizon of electricity generators considered is short, $T = 3$, the impact of the discount factor parameter is small. In this model, we set the discount factor as $r = 0.05$ for all generators.

Fuel prices are subject to a substantial margin of error. However, in the case of coal, prices are determined in a world market and the data can be found in http://www.world bank.org/prospects/pinksheets0. In this model we assume that fuel prices for each generator are given as $q_{1t} = 25.6$, $q_{2t} = 26$, $q_{3t} = 15$, for all $t = 0, 1, \ldots, T$.

One of the most important parameters in the management of the electricity generation is the maximum capacity of the network N. As $\max |X_t| = 455670$, where X_t is the observed electricity demand, estimates of this parameter can be specified as $N = 456000$. The rate limit to generation over two periods M plays also an important role in the generation of electricity. As $\max |X_t - X_{t-1}| = 84622.3$, we set $M = 85000$. Generation capacity is also constrained by lower and upper bounds: $l_j = 0$ for all $j = 1, 2, 3$ and $u_1 = 350000$, $u_2 = 220000$, $u_3 = 280000$.

To calibrate the parameters that remain uncertain, we will consider the worst-case modelling presented in Sect. 2.1. In this context, the parameters should satisfy the optimality conditions (12), the available information is the daily average price p^o and daily observed demand X^o, and the worst-case unobservable decision corresponds to the input's decision variable w.

As we can determine parameters N and M, we will not consider their associated constraints in the calibration analysis. In addition, it is predictable that the variable $h^1 = 0$ (generators are willing to generate the maximum amount of electricity) which implies $y_{jt} = A_j w_{jt}^{a_j}$ for all t and j. Therefore, the optimal conditions (12) can be simplified as follows:

$$\begin{aligned} p_t A_j a_j w_{jt}^{a_j - 1} - q_{jt} &= 0, \forall t, \\ \sum_j A_j w_{jt}^{a_j} - X_t^o &= 0, \forall t, \end{aligned}$$

where X_t^o is the daily observed demand at each day t. Let $C(w, p, A, a)$ denote this system of nonlinear equations.

Thus, we define the best choice of parameters $\{A_j\}$, $\{a_j\}$ in view of the worst-case unobservable decisions p, $\{w_j\}$ as the solution of the minimax problem:

$$\begin{array}{ll} \min_{A_j, a_j, p \geq 0} \max_w & \|p - p^o\| \\ subject\ \ to & C(w, p, A, a) = 0, \end{array} \tag{13}$$

given the observed demand X_t^o and the average price p_t^o at each day t. In particular, we consider the following observed data:

day	28/4/2002	29/4/2002	30/4/2002
X_t^o	334443.7	382222.6	389739.8
p_t^o	23.53	32.67	25.15

(14)

As recommended before to guarantee little computational cost, we suggest to restrict the interval of the variables $\{A_j\}$, $\{a_j\}$, $\{w_j\}$, p given the information available. In the context of the Australian electricity market, the bounds should be:

	A_1	A_2	A_3	a_1	a_2	a_3	$\{w_{jt}\}$	p
lower bound	16500	16500	18000	0.1	0.1	0.1	10000	15
upper bound	18000	18000	20000	1.0	1.0	1.0	17000	70

(15)

Therefore, the solution to Problem (13) is

	$j = 1$	$j = 2$	$j = 3$
A_j	18000	18000	20000
a_j	0.194774	0.195836	0.152826.

(16)

3.4 Computing Equilibrium

Given the scenario tree computed in Sect. 3.1, the stochastic version of the generators' problem (10) is defined as:

$$\begin{aligned}
&\max_{y_j,w_j} \quad \sum_{t=0}^{T}\sum_{s=0}^{S_t}\left(\tfrac{1}{1+r}\right)^t \beta^{t,s}\ \left[p_{t,s}\cdot y_{jt,s} - q_{jt,s}\cdot w_{jt,s}\right]\\
&\textit{subject to}\\
&\qquad\qquad y_{jt,s} - A_j w_{jt,s}^{a_j} \le 0,\ \forall t,s,\\
&\qquad\qquad \textstyle\sum_{j'\neq j} y_{j't,s} + y_{jt,s} \le N,\ \forall t,s,\\
&\qquad\qquad y_{jt,s} - y_{jt-1,s(t-1)} \le M,\ \forall t,s,\\
&\qquad\qquad l_j \le y_{jt,s} \le u_j,\ \forall t,s,\\
&\qquad\qquad 0 \le w_{jt,s},\ \forall t,s,
\end{aligned} \tag{17}$$

where $s\,(t-1)$ is the predecessor state. Problem (17) can be transformed into an equality constrained problem by introducing slack variables h_j^1, h_j^2, $h_j^3 \ge 0$, and a barrier function that penalizes the infeasibility of the inequality constraints in the slack h_j^1, h_j^2, $h_j^3 \ge 0$ and decision variables $0 \le w_j$ and $l_j \le y_j \le u_j$. Thus, the transformed problem is defined as follows:

$$\begin{aligned}
&\max_{y_j,w_j} \quad \sum_{t=0}^{T}\sum_{s=0}^{S_t}\left(\tfrac{1}{1+r}\right)^t \beta^{t,s}\ \left[p_{t,s}\cdot y_{jt,s} - q_{jt,s}\cdot w_{jt,s}\right]\\
&\qquad\quad -\mu\ \textstyle\sum_{t=0}^{T}\sum_{s=0}^{S_t}[\log\left(u_j - y_{jt,s}\right) + \log\left(y_{jt,s} - l_j\right) + \log\left(w_{jt,s}\right)\\
&\qquad\quad + \textstyle\sum_{m=1}^{3}\log\left(h_{jt,s}^m\right)]\\
&\textit{subject to}\\
&\qquad\qquad y_{jt,s} - A_j w_{jt,s}^{a_j} + h_{jt,s}^1 = 0, \forall t,\ \forall s,\\
&\qquad\qquad \textstyle\sum_{j'\neq j} y_{j't,s} + y_{jt,s} + h_{jt,s}^2 - N = 0, \forall t,\ \forall s,\\
&\qquad\qquad y_{jt,s} - y_{jt-1,s} + h_{jt,s(t-1)}^3 - M = 0, \forall t,\ \forall s.
\end{aligned} \tag{18}$$

Under appropriate convexity assumptions, the vector (y_j, w_j, h_j) is said to satisfy the necessary and sufficient conditions of optimality for Problem (18) if there exist Lagrange multiplier vectors $\gamma_j^1, \gamma_j^2, \gamma_j^3 \ge 0$ such that for all j, all t, all s :

$$\begin{aligned}
&\left(\tfrac{1}{1+r}\right)^t\ \beta^{t,s}p_{t,s} + \mu\left(u_j - y_{jt,s}\right)^{-1} - \mu\left(y_{jt,s} - l_j\right)^{-1} - \gamma_{jt,s}^1 - \gamma_{jt,s}^2 + \gamma_{jt,s}^3 = 0,\\
&-\left(\tfrac{1}{1+r}\right)^t\ \beta^{t,s}q_{jt,s} - \mu w_{jt,s}^{-1} + \gamma_{jt,s}^1\ A_j a_j w_{jt,s}^{a_j-1} = 0,\\
&-\mu\left(h_{jt,s}^m\right)^{-1} - \gamma_{jt,s}^m = 0,\ \forall m = 1,2,3,\\
&y_{jt,s} - A_j w_{jt,s}^{a_j} + h_{jt,s}^1 = 0,\\
&\textstyle\sum_j y_{jt,s} + h_{jt,s}^2 - N = 0,\\
&y_{jt,s} - y_{jt-1,s} + h_{jt,s(t-1)}^3 - M = 0.
\end{aligned} \tag{19}$$

Let $Z^1 = \mu\,(U-Y)^{-1}$, $Z^2 = \mu\,(Y-L)^{-1}$, $Z^3 = \mu W^{-1}$ and $Z^{4m} = \mu\,(H^m)^{-1}$,where $Y = diag(y)$, $L = diag\,(l)$, $U = diag\,(u)$, $W = diag(w)$, and $H^m = diag(h^m)$ for all $m = 1, 2, 3$. Then, the above conditions can be written as:

$$\begin{aligned}
&\left(\tfrac{1}{1+r}\right)^t\ \beta^{t,s}p_{t,s} + z^1_{jt,s} - z^2_{jt,s} - \gamma^1_{jt,s} - \gamma^2_{jt,s} + \gamma^3_{jt,s} = 0,\\
&-\left(\tfrac{1}{1+r}\right)^t\ \beta^{t,s}q_{jt,s} - z^3_{jt,s} + \gamma^1_{jt,s}\ A_j a_j w^{a_j-1}_{jt,s} = 0,\\
&-z^{4m}_{jt,s} - \gamma^m_{jt,s} = 0, \forall m = 1,2,3,\\
&y_{jt,s} - A_j w^{a_j}_{jt,s} + h^1_{jt,s} = 0,\\
&\textstyle\sum_j y_{jt,s} + h^2_{jt,s} - N = 0,\\
&y_{jt,s} - y_{jt-1,s(t-1)} + h^3_{jt,s} - M = 0,\\
&(U_j - Y_j)\,Z^1_j e - \mu e = 0,\\
&(Y_j - L_j)\,Z^2_j e - \mu e = 0,\\
&W_j Z^3_j e - \mu e = 0,\\
&H^m_j Z^{4m}_j e - \mu e = 0, \forall m = 1,2,3,
\end{aligned} \tag{20}$$

for all $j = 1, 2, 3$, all $s = \{1, \ldots S_t\}$, where $S_t = 2$, and all $t = 0, 1, 2, 3$.

Assume that we aim to forecast the prices, inputs and electricity outputs in equilibrium for the days May 1, May 2 and May 3, 2002, given the temperature data $f_1 = 45.5$, $f_2 = 41.4$, $f_3 = 42.0$. Using the initial point $\xi_0 = 1^T$, the interior-point algorithm converges to the equilibrium given in Appendix.

To show the accuracy of the computed equilibrium, we consider the expected value of the computed equilibrium prices. Given the probabilities β_{ts} associated with the different states s at each period t, by Bayes' rule, we calculate the marginal probability $\pi_{ts} = \prod_{(t,s)\succ(t',s')}\beta_{t's'}$ (see Appendix). Then, the expected value of the computed equilibrium prices $E\,[p_t] = \sum_{s=1}^{S_t}\pi_{ts}p_{ts}$ and the actual prices for $t = 1, 2, 3$ (which can be found in http://www. nemmco.com.au/data/) are shown in Fig. 1, what reveals that the model captures the essential features of the price's behaviour.

Let now assume that we aim to forecast the prices, inputs and outputs in equilibrium for the days October 1, October 2 and October 3, 2002, given the temperature data $f_1 = 58.8$, $f_2 = 49$, $f_3 = 48.0$. The actual and forecast equilibrium prices are shown in Fig. 2.

Note that the accuracy of the prediction depends on the data used to calibrate the parameters of the model. A structural change in the market can affect the prices and productions in equilibrium, and in that case an updated calibration of the model should be considered using the new information.

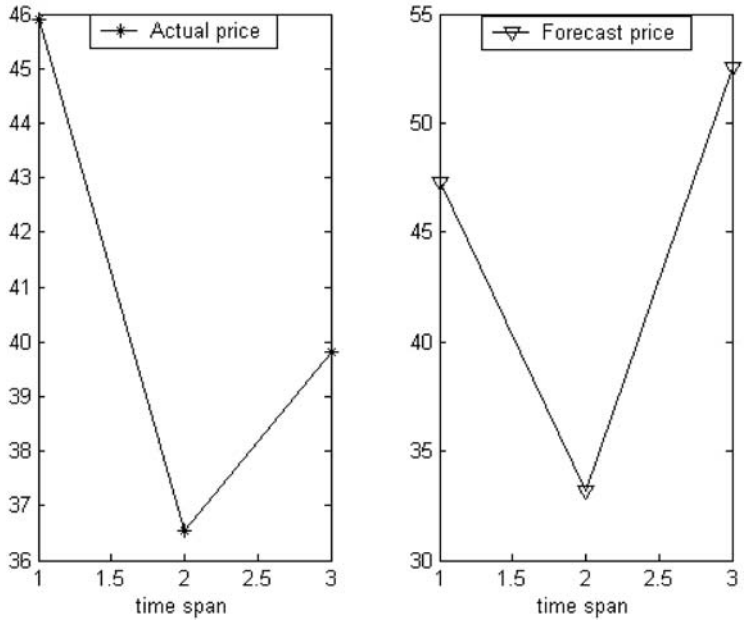

Fig. 1. Actual and computed prices for May 1st, 2nd, 3rd, 2002

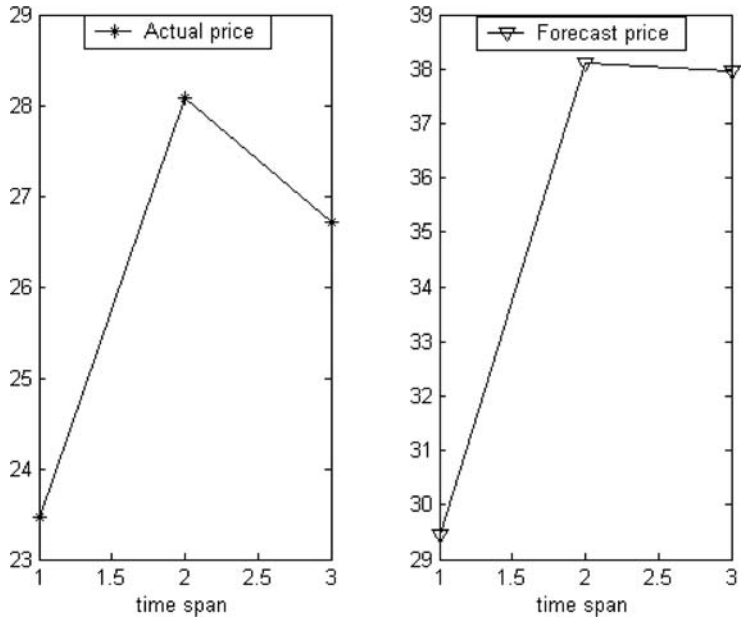

Fig. 2. Actual and computed prices for October 1st, 2nd, 3rd, 2002

4 Summary and Conclusions

This paper presents a methodology to build and solve stochastic dynamic economic models using limited data information. The approach has the potential for application in many economics sectors as practitioners often face the problem of having significantly less data than necessary for analyzing standard decision problems.

Decision-makers require the use of a stochastic dynamic complex model to approximate economic problems in a realistic way, and they often lack sufficient information to estimate the parameters involved in the model accurately. In this paper, we present a robust procedure for calibrating the parameters of model that best fits to the available data. The robust calibration of the model is achieved by a worst-case approach, involving the computation of a minimax problem. Also, we consider a scenario tree approach to model the underlying randomness of the demand. We generate scenarios using the simulation and randomized clustering approach and then, we compute equilibria by means of the interior-point approach. This algorithm can find accurate solutions incurring little computational cost.

We illustrate the performance of the method considering the NSW Australian deregulated electricity market. From the analysis of the results, we can conclude that this approach is able to forecast the pattern of equilibrium prices using limited information on the production side.

Appendix

The least-square regression coefficients of Model (8) are:

$\widehat{\mu} = 420453.7,$
$\widehat{\gamma} = (-96356.6, -16410.4, 86594.03,\ 38223.82, 17360.08, 37401.59)^T,$
$\widehat{\beta} = (-481045.5, 12954.2, -25529.96,\ -27617.52, 265764.7, 136627.8,$
$37069.41, -55922.99, -35578.56, -679.53, 98605.83)^T,$
$\widehat{c}_1 = -1976.111,$
$\widehat{\gamma}' = (2243.55, 642.10, -2502.44,\ -997.01, -129.57, -1008.41)^T,$
$\widehat{\beta}' = (12235.02, -4899.86, -94.86, 63.39, -7934.55,$
$-2853.79, 1208.76, 5784.12, 2677.32, 93.16494, -4471.44)^T,$
$\widehat{c}_2 = 16.45,$
$\widehat{\gamma}'' = (-20.82, -4.14, 21.37,\ 10.16, 1.53, 8.87),$
$\widehat{\beta}'' = (-75.68, 46.27, 8.56, 5.59,\ 59.65, 17.30, -22.62,$
$-82.25, -35.93, -3.36, 46.17)^T.$

The values of the probabilities β_{ts} associated with the different states s at each period t, with $\beta_0 = 1$, and the $AR(1)$ stochastic process of error terms a_{ts} at each state s and period t are:

for $t = 1$,

	1	2
β_s	0.26	0.74
a_s	0.73	0.59

(21)

for $t = 2$,

	1	2	3	4
β_s	0.79	0.20	0.79	0.20
a_s	0.67	0.51	0.60	0.80

(22)

for $t = 3$,

	1	2	3	4	5	6	7	8
β_s	0.47	0.52	0.62	0.37	0.40	0.59	0.6	0.4
a_s	0.54	0.72	0.68	0.54	0.75	0.55	0.57	073

(23)

The computed values of equilibrium are:
for $t = 0$, $p_0^* = 49.39$, and

	1	2	3
y_{j0}^*	1.51e5	1.53e5	1.05e5
w_{j0}^*	57.10e3	57.22e2	53.07e3

(24)

for $t = 1$,
$p_1^* = (47.29, 47.29)$, and

s=1	1	2	3	s=2	1	2	3
y_{js}^*	1.50e5	1.52e5	1.04e5	y_{js}^*	1.50e5	1.52e5	1.046e5
w_{js}^*	54.11e3	54.21e3	50.42e3	w_{js}^*	54.11e3	54.21e3	50.42e3

(25)

for $t = 2$,
$p_2^* = (33.19, 33.19, 33.19, 33.19)$, and

s=1	1	2	3	s=2	1	2	3
y_{js}^*	1.38e5	1.39e5	9.81e4	y_{js}^*	1.38e5	1.39e5	9.81e4
w_{js}^*	34.85e3	34.90e3	33.19e3	w_{js}^*	34.85e3	34.90e3	33.19e3

(26)

s=3	1	2	3	s=4	1	2	3
y_{js}^*	1.38e5	1.39e5	9.81e4	y_{js}^*	1.38e5	1.39e5	9.81e4
w_{js}^*	34.85e3	34.90e3	33.19e3	w_{js}^*	34.85e3	34.90e3	33.19e3

(27)

for $t = 3$,
$p_3^* = (52.53, 52.53, 52.53, 52.53, 52.53, 52.53, 52.53, 52.53)$, and

s=1	1	2	3	s=2	1	2	3
y_{js}^*	1.54e5	1.56e5	1.06e5	y_{js}^*	1.54e5	1.56e5	1.06e5
w_{js}^*	61.64e3	61.77e3	57.07e3	w_{js}^*	61.65e3	61.77e3	57.07e3

(28)

s=3	1	2	3	s=4	1	2	3
y_{js}^*	1.54e5	1.56e5	1.06e5	y_{js}^*	1.54e5	1.56e5	1.06e5
w_{js}^*	61.65e3	61.77e3	57.07e3	w_{js}^*	61.64e3	61.77e3	57.07e3

(29)

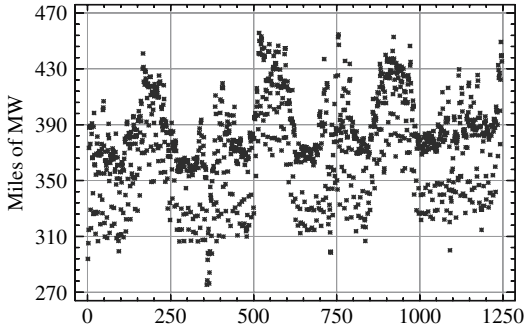

Fig. 3. Daily electricity demand in NSW, Australia

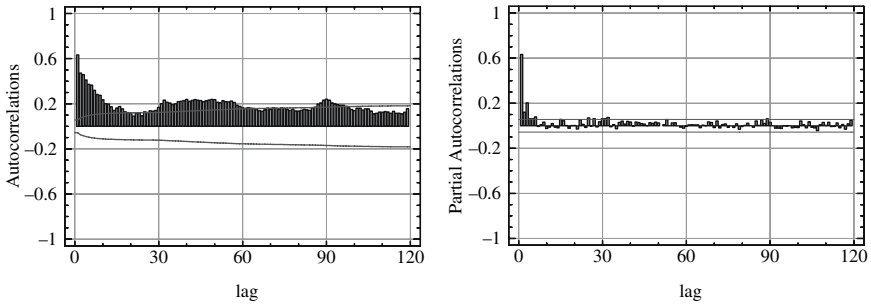

Fig. 4. Estimated autocorrelations and partial autocorrelations for residuals of Model (1)

s=5	1	2	3	s=6	1	2	3
y^*_{js}	1.54e5	1.56e5	1.06e5	y^*_{js}	1.54e5	1.56e5	1.06e5
w^*_{js}	61.65e3	61.77e3	57.07e3	w^*_{js}	61.64e3	61.77e3	57.07e3

(30)

s=7	1	2	3	s=8	1	2	3
y^*_{js}	1.54e5	1.563e5	1.06e5	y^*_{js}	1.54e5	1.56e5	1.06e5
w^*_{js}	61.65e3	61.77e3	57.07e3	w_{*js}	61.65e3	61.77e3	57.07e3

(31)

The marginal probabilities π_{ts} are:
for $t = 1$,

	1	2
π_{ts}	0.22	0.78

(32)

for $t = 2$,

	1	2	3	4
π_{ts}	0.132	0.088	0.632	0.148

(33)

for $t = 3$,

	1	2	3	4
π_{ts}	0.068	0.064	0.072	0.016

(34)

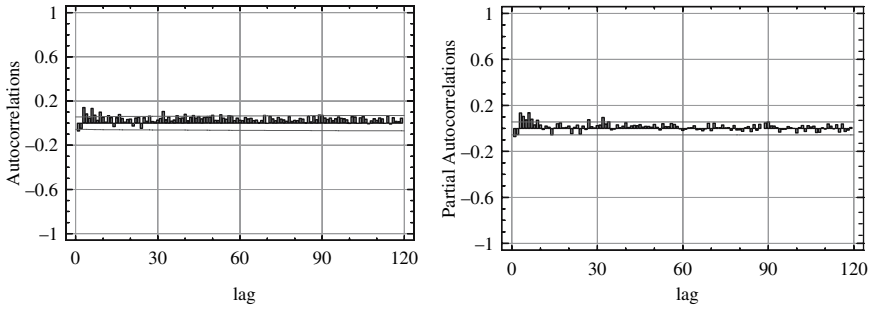

Fig. 5. Estimated autocorrelations and partial autocorrelations for residuals of Model (2)

Acknowledgements

This research has been supported by a Marie Curie Fellowship of the European Community programme IHP under contract number HPMF-CT-2000-00781 and the Ministerio de Educacion y Ciencia of Spain, through project SEJ-2004-00672.

References

Bessembinder, H. and L. Lemmon. Equilibrium Pricing and Optimal Hedging in Electricity Forward Markets. *The Journal of Finance*, LVII:1347–1382, 2002.

Boucher, J. and Y. Smeers. Alternative models of restructured electricity systems, Part 1: No market power. *Operations Research*, 49:821–838, 2001.

Bounder, G. C. E. A hybrid simulation/optimisation scenario model for asset/liability management. *European Journal of Operational Research*, 99:126–135, 1997.

Bram, J. The Lagrange multiplier theorem for max-min with several constraints. *SIAM J. Appl. Math.*, 14:665–667, 1966.

Brockwell, P. J. and R. A. Davis. Time series: theory and methods. Springer Series in Statistics. Springer-Verlag, Berlin, Heidelberg New York Tokyo, 1987.

Danskin, J. M. The Theory of Max-Min. Springer-Verlag, Berlin, Heidelberg New York Tokyo, 1967.

Day, C. J., B. F. Hobbs and J.-S. Pang Oligopolistic Competition in Power Networks: A Conjectured Supply Function Approach. *IEEE Trans. Power Systems*, 17:597–607, 2002.

Escudero, L. F., J. L. de la Fuente, C. García and F.-J. Prieto Hydropower generation management under uncertainty via scenario analysis and parallel computation. *IEEE Trans. Power Systems*, 11:683–689, 1996.

Esteban-Bravo, M. Computing equilibria in general equilibrium models via interior-point method. *Computational Economics*, 23:147–171, 2004.

Green, R. Competition in Generation: The Economic Foundations. *Proceedings of the IEEE*, 88:128–139, 2000.

Gülpinar, N., B. Rustem and R. Settergren. Simulation and Optimisation Approaches to Scenario Tree Generation. *Journal of Economics Dynamics and Control*, 28:1291–1315, 2004.

Harvey, A. and S. J. Koopman. Forecasting Hourly Electricity Demand Using Time-Varying Splines. *Journal of the American Statistical Association*, 88:1228–1236, 1993.

Henley, A. and J. Peirson. Non-linearities in electricity demand and temperature: parametric versus non-parametric methods. *Oxford Bulletin of Economics and Statistics*, 59:149–162, 1997.

Hobbs, B. F. Network models of spatial oligopoly with an application to deregulation of electricity generation. *Operations Research*, 34:395–409, 1986.

Høyland, K., and S. W. Wallace Generating Scenario Tree for Multistage Decision Problems. *Management Science*, 47:295–307, 2001.

Hsu, M. An introduction to the pricing of electric power transmission. *Utilities Policy*, 6:257–270, 1997.

Jing-Yuan, W. and Y. Smeers. Spatial oligopolistic electricity models with Cournot generators and regulated transmission prices. *Operations Research*, 47:102–112, 1999.

Kahn, E. P. Numerical Techniques for Analyzing Market Power in Electricity. *The Electricity Journal*, 11:34–43, 1998.

Kouwenberg, R. Scenario generation and stochastic programming models for asset liability management. *European Journal of Operational Research*, 134:279–292, 2001.

Mas-Colell, A., M. D. Whinston and J. R. Green. Microeconomic Theory. Oxford University Press, New York, 1995.

McCalley, J. D. and G. B. Sheblé. Competitive electric energy systems: engineering issues in the great experiment. Tutorial paper presented at the 4th International Conference of Probabilistic Methods Applied to Power Systems, 1994.

Neame P. J., A. B. Philpott and G. Pritchard. Offer stack optimisation in electricity pool markets. *Operations Research*, 51:397–408, 2003.

Pardalos P. M. and G. C. Resende. Handbook of Applied Optimisation. Oxford University Press, New York, 2002.

Pritchard, G. and Zakeri, G. Market Offering Strategies for Hydroelectric Generators. *Operations Research*, 51:602–612, 2003.

Rhys, J. M. W. Techniques for Forecasting Electricity Demand. *Statistician*, 33:23–33, 1984.

Rosen, J. B. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33:520–534, 1965.

Rustem, B. and M. B. Howe. Algorithms for Worst-Case Design and Applications to Risk Management. Princeton University Press,Princeton and Oxford, 2002.

Schweppe, F. C., M. C.Carmanis, R. B. Tabors and R. E. Bohn Spot Pricing of Electricity. Kluwer Academic Publishers, Boston, Mass., 1988.

Sheblé, G. B. Decision Analysis Tools for GENCO Dispatchers. *IEEE Transactions on Power Systems*, 14:745–750, 1999.

Valenzuela, J. and M. Mazumdar. Statistical analysis of electric power production costs. *IIE Transactions*, 32:1139–1148, 2000.

Valenzuela, J. and M. Mazumdar. Making unit commitment decisions when electricity is traded at spot market prices. Proceedings of the 2001 IEEE Power Engineering Society Winter Meeting, Feb 01, Columbus, Ohio, 2001.

Varian, H. Microeconomics Analysis. Norton, New York, 1992.
Wright, M. H. Interior methods for constrained optimisation. *Acta Numerica*, 341–407, 1991.
Žaković, S. and B. Rustem. Semi-infinite Programming and Applications to Minimax Problems. *Annals of Operations Research*, 124:81–110, 2003.

An Approximate Winner Determination Algorithm for Hybrid Procurement Mechanisms in Logistics

Chetan Yadati[1], Carlos A.S. Oliveira[2] and Panos M. Pardalos[3]

[1] School of Industrial Engineering and Management, Oklahoma State University, 322 Engineering North, Stillwater, OK, 74078, USA, Email: chetan.yadati@okstate.edu
[2] School of Industrial Engineering and Management, Oklahoma State University, 322 Engineering North, Stillwater, OK, 74078, USA, Email: carlos.oliveira@okstate.edu
[3] Department of Industrial and systems Engineering, University of Florida, Gainesville, FL 32611 USA, Email: pardalos@ufl.edu

Summary. Logistics services form the backbone of every supply chain. Given their importance in the operation of corporations, it is interesting to determine efficient methods for optimal service procurement. A typical problem faced by most managers of global firms is studied: given a set of service providers with respective quantity-discount curves, the objective is to compute the set of logistics services that should be procured from each provider, such that the overall supply chain efficiency requirements are met. Although this is a very common problem, it is actually intractable when the number of logistics providers and their services is large enough. An auction based mechanism to model this situation is developed, using a hybrid auction approach. Integer programming formulations for the problem are presented, which try to explore the combinatorial features of the problem. In order to allow for the efficient computation of large instances, a heuristic algorithm to the winner determination problem is presented. The proposed polynomial algorithm is applied to a large number of test instances. Results demonstrate that close to optimal solutions are achieved by the algorithm in reasonable time, even for large instances typically occurring in real applications.

Key words: Logistics, auction, procurement, integer programming, heuristics

1 Introduction

An auction is a common method for setting prices of commodities that have an undetermined or variable value. This is a mechanism that is frequently used when there is a large number of suppliers interested in acquiring a given

product. The most common auctions are *sequential auctions*, in which the products are auctioned in sequence until they are all sold. In this type of auction, determination of the winner is trivial because the highest bid gets the item. More involved auctions exist however, where for example the price paid is the price of the second highest bid. Such types of auctions will not be considered here.

1.1 General Combinatorial Auctions

Standard auctions have disadvantages when the number of items available is large. It is difficult to fulfill the requirements of interested parties, which may have complementarity or substitutability issues. *Complementarity* occurs when two items complement each other, therefore their combined valuation for the supplier is higher than the sum of the separate costs. *Substitutability* is the opposite situation, when two items have features that are substitute, and thus their combined value is less than the sum of individual values.

To avoid such problems, modern auctions have introduced the idea of bidding on sets of items, instead of single items. The term *combinatorial auction* is generally used in this case. Combinatorial auctions can be of several types, depending on the specific mechanism or protocol that is used for its accomplishment. A description of the types of combinatorial auctions is beyond the scope of this paper, and we refer the reader to surveys on the topic, such as Vries and Vohra (2003), Rothkopf et al. (1998), Pekec and Rothkopf (2003), Narahari and Dayama (2005), and Sandholm (2000).

A well known application of combinatorial auctions is the widely publicized Federal Communications Commission auction of wireless communication spectrum (Crampton, 1997, Cramton, 1998), performed between 1994 and 1996. Combinatorial auction mechanisms have also been designed for other problems, such as airport time slot allocation Rassenti et al. (1982).

Closely related to combinatorial auctions is the problem of procuring services for a company Sheffi (2004). In this case, the goal is to buy products or services from a set of providers, with the goal of minimizing the total cost of procured items. This form of inverse auction has been frequently used in the last few years by companies that want to find the most competitive prices for the services they need. Examples of companies that have recently used this method include Home Depot, Wal-Mart, and several others Sheffi (2004).

1.2 The Procurement Model for Transportation

A prime example of this type of inverse auction is in the procurement of transportation services. A set of lanes, connecting distribution points, is given and the auctioneer receives bids for the prices of transporting goods on subsets of the lanes. Service providers consider several characteristics of the lanes in which they bid: complementarity is a frequent issue arising in this type of

application, since transportation services can become less expensive if they are restricted to smaller geographical locations. Another type of complementarity that is frequently considered is related to circuit formation: it is most often desirable to have ways of returning the fleet to its original location without additional costs. The result of such a transportation auction is a partition of the available lanes, where the total cost the auctioneer needs to pay for the services procured over the partition is minimized.

A *hybrid procurement mechanism* is one in which each supplier gives not only a price, but also a supply curve, depicting different prices for different quantities of items. This occurs normally when providers are able to give discounts for additional items serviced. For example, a bidding company may be able to give a 20% reduction for each item serviced above the limit of 200 units. We assume that in this case the auctioneer needs to pay the full price for any quantity up to 200 units, with the discount being applied to units above the limit.

For a more concrete example of such hybrid mechanisms, consider a typical enterprise having distribution centers over a large geographical area. Such a company needs to transport its products to all its distribution centers. The enterprise manufactures products of various types, each with different physical dimensions and hence occupies different volumes when packed for transportation. Notice that it is also common to ship certain products as bundles of items rather than single units. Assume that there are multiple logistics providers offering to carry out logistic operations over different subsets of distributors (geographical area). The goal of the procurement model we consider is to provide a decision framework so that the appropriate set of providers are located. The major issues concerning this auction model are:

Combinations of regions: Enterprises ideally would like their products to be distributed over a large geographical area. However, logistics providers for various reasons exhibit preferences for certain regions. Thus, the decision maker is faced with choosing a set of providers to distribute his products over the entire market geographically in a an economic way.

Combinations of volumes: It is natural that market preferences vary with geographical regions. To suit this requirement, the volumes of goods being sent to each region will also vary. Logistics providers prefer full units of transport, i.e., full truck loads to fractional truckloads. Thus, it is common among logistics providers to offer volume discounts. The decision maker then has to choose between a combination of volumes and providers to be used to transport his goods.

In this paper we consider the problem of winner determination for procurement performed using hybrid combinatorial auctions. This is a difficult problem in most cases, being known to be NP-hard. We start with a formal definition of the problem in Sect. 2. Practical applications of combinatorial auctions and hybrid procurement are introduced in Sect. 3. We then discuss previous work done in this area in Sect. 4, and proceed to provide some new techniques to solve the problem in practice. First, we introduce integer

programming formulations for the problem in Sect. 5. Then, we propose in Sect. 6 a heuristic algorithm that can provide near optimal solutions for many of the instances tested. Experimental results with the methods presented in the previous sections are then presented and analyzed in Sect. 7. Finally, concluding remarks and open questions are provided in Sect. 8.

2 Winner Determination Problem

The main problem arising on the execution of combinatorial auctions is winner determination. Contrary to single item auctions where bids are made for single items, each bid in a combinatorial auction can be made on an arbitrary set of items. The number of such sets is exponential in the number of items, therefore even evaluating the price achievable by each subset is out of question for any relatively large auction. Since items can be part of multiple sets, the issue is to determine a partition of the original set of items such that the total revenue is maximized. Once a partition has been determined, the winner of each set can be easily chosen as the supplier that gives the highest valuation for that set (with ties being broken according to pre-specified rules).

In the context of procurement, the winner determination problem is to determine a partition into subsets of the original items that must be procured, such that the associated cost of the resulting partition is as small as possible. As we are considering hybrid procurement mechanisms, this choice is still further complicated by the fact that multiple items can be serviced at different prices, according to the discount model provided by each supplier.

To formalize the problem, let I be the set of items, and d_i the quantity of the i-th item that must be procured. There are m suppliers, and each supplier makes a bid for a subset B_j of I, for $j \in \{1, \ldots, m\}$. The j-th supplier also provides a discount curve with n_j different prices. These are represented by p_{jk} for volumes between v_{k-1} and v_k, and $k \in \{1, \ldots, n_j\}$.

A solution for this problem is a partition $\mathcal{A}$ of I. This partition satisfies (1) if $A, B \in \mathcal{A}$ and $A \neq B$ then $A \cap B = \emptyset$ and (2) $\bigcup_{A \in \mathcal{A}} A = I$. The solution also need to specify the number r_{ij} of items of type i that will be procured from provider j, such that $r_{ij} \leq d_i$ and $\sum_{j \in \{1,\ldots,m\}} r_{ij} = d_i$ for each item i. From now on, we assume that the total supply (from all bidders) is greater than or equal to the demand, in order to guarantee that there is a feasible solution for the resulting problem.

3 Applications

Logistic services form the backbone of every supply chain network. A recent survey reports that third party logistics users expected logistics expenditures to represent approximately 7% of their organizations' anticipated total sales for 2004. Major driving factors to adopt third party logistics include emphasis

on improved supply chain management, enhancing customer service, reducing cost, consolidations, mergers, and acquisitions. Other common factors affecting the cost of operations include rapidly accelerating new product introductions, implementation of new information technologies, and the rising of new markets due to globalization. With so many driving factors, it is natural for enterprises to try to choose an optimal set of providers for servicing their logistics needs.

As a way of reducing the costs associated with global logistics, companies have been increasingly adopting the procurement of services based on combinatorial auction methods. For example, an early application of combinatorial auctions is presented by Moore et al. (1991) on the Reynolds Metals Company. Elmaghraby and Keskinocak (2003) describe the use of combinatorial procurement as one of the key strategies employed by the Home Depot company.

An important application of combinatorial procurement models is in the transportation industry. In this application, lanes connecting important points are procured among several transportation companies. Ledyard et al. (2002) provided an example of combinatorial procurement for transportation problems. Sheffi (2004) discussed several other procurement problems being solved by companies such as LogiCorp and Logistics.com.

Hybrid (also known as quantity-discount) procurement has also been very important for the logistics operation of several companies. Hybrid procurement has been used with success for example by Mars Inc., as reported by Hohner et al. (2003). Due to this success in modeling complex operations of the procurement process, combinatorial auctions have been the most well known paradigm used in the literature for solving this kind of problem, as discussed in the next section.

4 Previous Work

The winner determination problem for general combinatorial auctions has been studied by several researchers. The most well known approach to winner determination is to use a model based on the set packing problem. Given a set I and a collection $\mathcal{C}$ of subsets $S_i \subseteq I$, for $i \in \{1, \ldots, n\}$, a *set packing* is a set $P \subseteq \mathcal{C}$ such that for $A, B \in P$, with $A \neq B$, we have $A \cap B = \emptyset$. Given a cost for each set in $\mathcal{C}$, the *set packing problem* asks for the set packing with maximum cost.

The resemblance between the set packing problem and the winner determination problem is clear, once we interpret I as the set of items and $S_j \in \mathcal{S}$ as the subsets that the j-th supplier is interested in, for $j \in \{1, \ldots, m\}$. This modeling approach has been used by most algorithms for winner determination.

A popular, although not very efficient, algorithm for the set packing problem is based on integer programming. The integer programming

formulation for set packing can be described as follows. Let x_S be equal to 1 if the set S is selected as occurring in the partition, and $x_S = 0$ otherwise. Let c_S be the cost associated with using set S in the solution. With these definitions, we can write the problem as

$$\max \quad \sum_{S \in \mathcal{C}} x_S c_S$$

subject to

$$\sum_{S: i \in S} x_S \leq 1 \qquad \text{for each item } i \in I$$

$$x_S \in \{0, 1\}, \text{for each } S \subset I.$$

This classical formulation is also called a *packing formulation*, since we are allowed to add each element $i \in I$ to at most one set.

The winner determination problem is known to be NP-hard, by reduction to the set packing problem Garey and Johnson (1979). Approximation is also hard for this problem, with the best possible algorithm achieving only a factor $n^{1-\epsilon}$ approximation (for any $\epsilon > 0$) unless NP = ZPP Arora and Lund (1996).

Despite its general intractability, various researchers have designed algorithms to give heuristic solutions to the winner determination problem. A through discussion of some of the most common algorithms is presented by Sandholm (2002). Recent methods include *combinatorial auctions multi-unit search* Leyton-Brown et al. (2000), *branch on bids* Leyton-Brown et al. (2000), and *combinatorial auction branch on bids* Sandholm et al. (2005), among others. The basic idea behind these algorithms is to prune searching in a way that minimizes the chance of missing an optimal result. The first algorithm guarantees an optimal solution, however its applicability is limited by the inherent complexity of dynamic programming when employed to solve NP-hard optimisation problems. The two latter algorithms rely on tree structures that allow branching on specific bids during the optimisation process. These algorithms derive bounds based on the expected improvement of the existing optimal value on the subtrees of the branch-and-bound data structure, and prune them accordingly.

Other algorithms, such as *limited discrepancy search* Sakurai et al. (2000), limit the search efforts to the region of the decision tree that is most likely to contain the optimal result. This algorithm starts by selecting only the best nodes initially and much later expands the search to include other nodes in case of necessity. Notice that the above mentioned algorithms all rely on developing more efficient branch and bound techniques to solve the winner determination problem. They usually apply depth first search methods to explore the decision tree. Although these algorithms perform well in many cases, their worst case behavior is still exponential.

Additionally, several heuristic algorithms have been proposed for the winner determination problem in combinatorial auctions. A good discussion of heuristics, as well as other approaches such as dynamic programming, can

be found in Sandholm (2002). Vries and Vohra (2003) is a survey of models for combinatorial auctions problems, with the description of several special cases that can be used to speed up the running time of the general algorithms for winner determination.

It is important to remark that the algorithms above are designed for auction problems where a single price is given by service providers. Therefore, they are not directly applicable for the situation we are considering in this paper, where the bids contain not only a fixed price but a curve of discount-prices per volume.

5 Integer Programming Formulation

5.1 First Integer Programming Formulation

We now turn to the discussion of mathematical models for the combinatorial procurement problem. We propose a mathematical formulation based on linear integer programming, as described bellow. We assume there are n items to be procured and m providers. Let d_i be the demand for the i-th item. Each provider j ($j \in \{1, \ldots, m\}$) gives a quote composed of prices p_{jk} for volumes between $v_{j(k-1)}$ and v_{jk}, for all $k \in \{1, \ldots, n_j\}$ (we assume $v_{j0} = 0$ and $v_{k-1} < v_k$). Each unit of service of provider j has q_{ij} items of the i-th type.

Let x_{jk}, for $j \in \{1, \ldots, m\}$, and $k \in \{1, \ldots, n_j\}$, be an integer variable equal to the quantity selected from the k-th part of the discount function quoted by provider j.

The choice of the exact part of the discount function that must be selected from each provider is encoded using a binary variable w_{jk}, for $j \in \{1, \ldots, m\}$, and $k \in \{1, \ldots, n_j\}$. This variable is 1 whenever the quantity available in one of the segments of the curve is not completely selected.

Using the variables described above, the integer programming formulation for the problem is

$$\min \quad \sum_{j=1}^{m} \sum_{k=1}^{n_j} p_{jk}\, x_{jk} \tag{1}$$

subject to

$$w_{jk} \geq [(v_{jk} - v_{j(k-1)}) - x_{jk}]/K \qquad j = 1, \ldots, m,\ k = 1, \ldots, n_j \tag{2}$$

$$w_{jk} \leq (v_{jk} - v_{j(k-1)}) - x_{jk} \qquad j = 1, \ldots, m,\ k = 1, \ldots, n_j \tag{3}$$

$$x_{j(k+1)} \leq (1 - w_{jk})(v_{j(k+1)} - v_{j(k)}) \qquad j = 1, \ldots, m,\ k = 1, \ldots, n_j \tag{4}$$

$$x_{j1} \leq v_{j1} \qquad j = 1, \ldots, m \tag{5}$$

$$\sum_{j=1}^{m} \sum_{k=1}^{n_j} q_{ij} x_{jk} \geq d_i \qquad i = 1, \ldots, n \tag{6}$$

$$x_{jk} \in Z_+ \text{ and } w_{jk} \in \{0, 1\} \quad j = 1, \ldots, m,\ k = 1, \ldots, n_j, \tag{7}$$

where $K \geq \max\{v_{jk} - v_{j(k-1)}\}$, for all $j \in \{1, \ldots, m\}$, $k \in \{1, \ldots, n_j\}$.

The objective function (1) specifies that the total cost of the items obtained from the selected suppliers is minimized. Constraints (2) and (3) determine the value of binary variable w_{jk}, for each supplier and section of the discount curve, as previously explained. If a segment k of the discount domain for supplier j is not completely obtained, then w_{jk} is equal to one. Constraint (4) uses the value of the variable w_{ij} to determine if a product can be acquired at a given price level. Constraint (5) is similar to (4) but is necessary only for the first section of the domain of the discount function. Constraint (6) enforces the demand satisfaction requirements. Finally, Constraint (7) states the feasible domain for each variable in the formulation.

By inspection, the number of constraints in this formulation is of the order $O(n + mN)$, where N is defined as $\max_{1\leq j\leq m}(n_j)$. The formulation has also $2\sum_{j=1}^{m} n_j$ variables, of which $\sum_{j=1}^{m} n_j$ are binary and the remaining are integer variables.

5.2 Second Integer Programming Formulation

A second integer programming formulation for the procurement problem can be defined as follows. Let us introduce binary variables x_{jk} with the value 1 meaning that a quote from the j-th supplier was accepted at the k-th level of its discount curve. Let z_{jk} represent the amount of items procured from the j-th supplier, from the k-th part of its discount curve. Then, the second formulation is

$$\min \quad \sum_{j=1}^{m}\sum_{k=1}^{n_j}((\sum_{\ell=1}^{k-1}(v_{j\ell} - v_{j(\ell-1)})p_{j\ell})x_{jk} + p_{jk}\, z_{jk}) \tag{8}$$

subject to

$$\sum_{j=1}^{m}\sum_{k=1}^{n_j} \quad q_{ij}((\sum_{\ell=1}^{k-1}(v_{j\ell} - v_{j(\ell-1)}))x_{jk} + z_{jk}) \geq d_i \qquad i = 1, \ldots, n \tag{9}$$

$$z_{jk} \leq (v_{jk} - v_{j(k-1)})x_{jk} \qquad j = 1, \ldots, m, \text{ and } k = 1, \ldots, n_j \tag{10}$$

$$\sum_{k=1}^{n_j} x_{jk} \leq 1 \qquad j = 1, \ldots, m \tag{11}$$

$$x_{jk} \in \{0, 1\} \quad \text{and} \quad z_{jk} \in Z_+ \qquad j = 1, \ldots, m,\ k = 1, \ldots, n_j. \tag{12}$$

The main difference between the latter formulation and the former one is that the selection made by variable x_{jk} determines only the exact part of the discount curve where the quantity we want from supplier j is located. The remaining quantities are found implicitly, using a summation over the previous sections of the domain. The objective function (8) uses this idea to compute the total price paid by the selected items. Constraint (9) guarantees that the demand is satisfied by the total items procured. Constraint (10)

determines the feasible bounds for each quantity procured from supplier j at discount level k. Constraint (11) defines the main property of variables x_{jk}, by selecting at most one variable for each supply j. Finally, the domains of variables x_{jk} and z_{jk} are determined by constraint (12).

The number of constraints in this formulation can easily be seen to be of the order $O(n + mN)$, where N is defined as $\max_{1 \leq j \leq m}(n_j)$. Similarly to the previous formulation, there are $\sum_{j=1}^{m} n_j$ binary variables and $\sum_{j=1}^{m} n_j$ integer variables. This formulation is a little more compact than the previous one, and therefore we selected it to perform computational experiments, as shown in Sect. 8.

6 A Heuristic for Winner Determination

Due to the complexity of the winner determination problem, it is unlikely that an exact integer programming formulation for large instances can be solved in practice. However, most problems occurring on the industry are of large scale; to overcome this difficulty we propose a heuristic that has polynomial time complexity, but that provides very good solution for the instances tested in our computational experiments.

6.1 The Costliest Item Heuristic

The heuristic proposed is based on the following idea: instead of finding the optimum solution we can just select, for each item procured, the provider that gives the best price for that item. Although this might be suboptimal for some combination of items, in practice the algorithm can provide a good enough solution for most practical purposes (as will be shown in Sect. 7). The heuristic tries to satisfy the demand of the costliest items first, hence the name used. Given a supplier $j \in \{1, \ldots, m\}$, let k_d be the minimum value k such that $\sum_{\ell=1}^{k}(v_{j\ell} - v_{j(\ell-1)})q_{ij} \geq d$ is satisfied. Then, we define the total price $\mathrm{TP}(j, d)$ necessary to satisfy the demand d as

$$\mathrm{TP}(j, d) = \sum_{\ell=1}^{k_d - 1} (v_{j\ell} - v_{j(\ell-1)})p_{jk} + \frac{(d - v_{jk_d})p_{jk_d}}{q_{ij}}.$$

Let π_j be the quantity that has already been procured from supplier j at some point in the algorithm. Then we define a function $P(\cdot, \cdot)$ representing the average cost at which the demand of item i can be satisfied by supplier j. We let

$$P(i, j) = \frac{TP(j, d_i + \pi_j q_{ij}) - TP(j, \pi_j q_{ij})}{d_i},$$

if there is enough capacity to satisfy the whole demand, i.e., $d_i < (v_{jn_j} - \pi_j)q_{ij}$. Otherwise, we define

$$P(i,j) = \frac{\mathrm{TP}(j, v_{n_j} q_{ij}) - \mathrm{TP}(j, \pi_j q_{ij})}{(v_{n_j} - \pi_j) q_{ij}}.$$

Using the notation defined above, we provide a formal description of the algorithm in Fig. 1. At the beginning, we are given a vector of item demands and the discount curves for each of the suppliers. The idea is to start finding the minimum cost needed to satisfy the demand of one item completely, without bothering about the demand satisfaction of other items. This is a type of greedy procedure, where there is no guarantee that the computed costs are optimal. This first phase of the algorithm is called the *cost computing phase* (Fig. 2). Once the cost ordered items are obtained, the algorithm proceeds by satisfying the demand for the costliest item first. In the next iteration the algorithm satisfies the next costliest item using the minimum available cost, until all items are satisfied in this way. This stage of the algorithm is called the *demand satisfaction phase* (Fig. 3). By combining the two phases described above, and based on the demand and prices, we select locally the supplier that will provide the best price for the next item procured. After such a supplier is found, we update the vector of demands accordingly, removing the items that have been previously selected.

```
1  Input: demands d_i, for i ∈ {1,...,n}.
2  Input: prices p_jk, for j ∈ {1,...,m} and k ∈ {1,...,n_j}.
3  for i ∈ {1,...,n} do
4     cost_i ← ∞
5  end
6  while there is d_i ≥ 0, for i ∈ {1,...,n} do
7     Cost ordering phase
8     Demand satisfaction phase
9  end
```

Fig. 1. Costliest item heuristic

```
1   Input: demands d_i, for i ∈ {1,...,n}.
2   Input: prices p_jk, for j ∈ {1,...,m} and k ∈ {1,...,n_j}.
3   Output: vector cost, with minimum costs for all items.
4   for all i ∈ {1,...,n} such that d_i > 0 do
5      for all quotes j ∈ {1,...,m} do
6         if cost_i > P(i,j) then
7            cost_i ← P(i,j)
8         end
9      end
10  end
```

Fig. 2. Cost computing phase

Input: demands d_i, for $i \in \{1, \ldots, n\}$.
Input: prices p_{jk}, for $j \in \{1, \ldots, m\}$ and $k \in \{1, \ldots, n_j\}$.
Input: vector *cost*, with minimum costs for all items.
Output: vector $\pi \in Z^m$, with quantities procured from each supplier.
$i' \leftarrow \arg\max_{1 \leq i \leq n} cost_i$
while $d_{i'} > 0$ **do**
  $j' \leftarrow \arg\min_{1 \leq j \leq m} P(i', j)$
  $\delta \leftarrow \min(d_{i'}, v_{j'n_{j'}} - \pi_{j'})$
  $\pi_{j'} \leftarrow \pi_{j'} + \delta$
  $d_{i'} \leftarrow d_{i'} - \delta$
end
Return π

Fig. 3. Demand satisfaction phase

6.2 Variations of the Proposed Heuristic

The method used (highest cost) to select the item that must be satisfied next was quite arbitrary, since this can in practice be determined in several ways. For example, one may try instead to satisfy the demand as fast as possible, by selecting the item that has higher demand. One can also try to give precedence to items high higher average cost across providers. Such policies may prove to be more effective on different instances of the problem, and should be implemented according to the requirements of the real instances solved. Thus, in addition to the method used above, we tried to determine the procurement costs using various alternative methods of selecting the next item. Examples of such policies are *lowest cost first*, *highest volume first*, and *lowest volume first*. We performed a set of computational experiments with these alternate methods, which are described in the next section.

7 Computational Experiments

7.1 Test Environment

In this section we describe the computational experiments that have been performed with the proposed algorithms. Our main goal when designing the computational experiments has been to determine in practice the efficiency of the methods previously discussed.

With this objective in mind, the second integer programming model discussed in Sect. 5.2 was solved using Dash Xpress, a commercial solver from Dash optimisation. The model was implemented using Mosel, a modeling language for mathematical programming, available with the solver.

Both the costliest item first heuristic and the generator of random instances was implemented using Java 1.4. Each of the random instances had a set of 20 bidders and the quantities for each of the items ranged from 10 to 60 units.

The machine used had a Pentium 4 processor with clock speed of 1.59 GHz, and 512MB of main memory. The Java programs and the integer programming models were executed on the same machine. (The instances used in this paper are available at the address http://www.okstate.edu/ceat/iem/iepeople/oliveira/papers/procurement).

7.2 Comparison of Heuristic Policies

The results of the first test performed with the heuristic is presented in Fig. 4. In this test, we have used several policies for the same heuristic, and tested them against a set of instances of the winner determination problem. The instances tested ranged in size from 1 to 22 items, for a fixed number of bidders. The graph is shown in a logarithm scale to facilitate the visualization of the smaller values. The results of this first experiment show that the heuristic works best when the costliest item is selected at each iteration. Other policies used included selecting at each step the item with the highest volume, the one with the lowest volume, and the cheapest item. They all proved to be of lower quality when compared to the simple decision of finding the costliest item.

The second best heuristic policy from the results is clearly the *cheapest item* policy. Its results are the best when a small number of items is involved. It them becomes inferior to *costliest item*, but generally follows the same pattern of increasing costs, meaning that there is a correlation between the values achieved by the costliest item and the cheapest items policies.

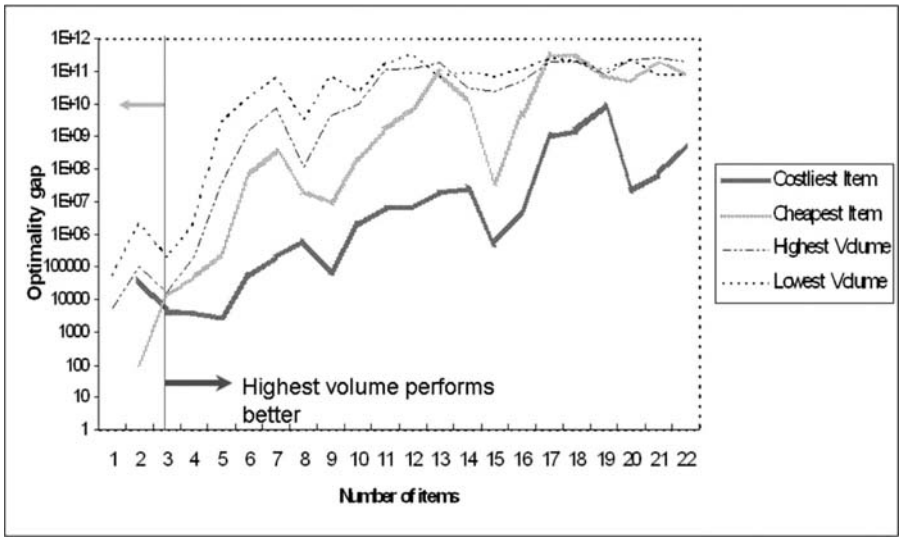

Fig. 4. Comparison of different policies for costliest item first heuristic

7.3 Comparison of MIP and Heuristic

The number of items to be procured in the experiment ranged from 3 to 43. The number of bidders was kept constant, as a way of controlling the variability of the instances. The Tables 1 and 2 summarize the results found in our computational tests. The *MIP* column reports the results found by our second mixed integer programming model, with a maximum time limit set to two hours (the heuristic took less than ten minutes in most cases). Therefore, the results in this column are not optimal in most of the cases and provide only a lower bound for the exact solution. When integer solutions were found, they have been reported instead – this is the reason why there are some rows where the MIPS value is larger than the corresponding heuristic solution value. The *heuristic* column represents the results found by the best policy for the winner determination heuristic. The columns *difference* and *%difference* show the variation between the results of the integer programming model and the heuristic, both in cost units and in percentage. We notice that the gap between the heuristic and the integer programming solutions are in the range of 0 to 60%. However, as the partial solutions for the MIP give only a lower bound for the the optimum solution, it may be possible that the heuristic values are much closer to the optimum solution.

Table 1. Comparison of results for the MIP and the proposed heuristic

n	MIP	heuristic	difference	%difference
3	159.4	153.4	−6.0	−3.8
4	142.8	179.9	37.2	26.0
5	169.0	264.9	95.8	56.7
6	243.2	260.1	16.8	6.9
7	341.1	329.0	−12.1	−3.5
8	149.4	147.3	−2.1	−1.4
9	224.8	311.6	86.8	38.6
10	234.3	226.3	−8.0	−3.4
11	339.9	486.9	147.0	43.2
12	280.1	389.5	109.4	39.1
13	225.5	318.9	93.4	41.4
14	295.8	406.1	110.4	37.3
15	294.9	428.6	133.7	45.3
16	294.6	476.7	182.1	61.8
17	359.4	539.8	180.4	50.2
18	267.5	345.3	77.9	29.1
19	257.2	323.8	66.6	25.9
20	276.5	410.6	134.1	48.5
21	286.2	386.9	100.7	35.2

Table 2. Comparison of results for the MIP and the proposed heuristic (continued)

n	MIP	heuristic	difference	%difference
22	279.3	427.1	147.8	52.9
23	305.7	377.6	71.9	23.5
24	339.1	484.2	145.1	42.8
25	257.1	398.0	140.9	54.8
26	344.8	402.9	58.2	16.9
27	251.1	418.1	167.0	66.5
28	299.7	422.4	122.6	40.9
29	286.7	430.3	143.6	50.1
30	335.8	455.0	119.2	35.5
31	322.6	457.1	134.5	41.7
32	318.2	435.9	117.7	37.0
33	337.5	467.9	130.4	38.6
34	326.9	492.9	166.1	50.8
35	330.6	504.0	173.4	52.4
36	315.5	404.6	89.1	28.2
37	342.5	491.5	149.0	43.5
38	322.9	418.1	95.2	29.5
39	363.8	532.9	169.1	46.5
40	334.2	486.5	152.4	45.6
41	288.4	484.0	195.6	67.8
42	333.9	516.9	183.0	54.8
43	382.2	612.4	230.2	60.2

8 Concluding Remarks

Hybrid procurement is an important problem, especially considering the complexity of modern supply chains. It involves the solution of a complex winner determination problem, which is in general NP-hard. We considered in this paper the winner determination problem for the case in which the suppliers provide a quantity-discount curve of prices, instead of a single bid. The winner determination problem in this case is still NP-hard, since this is a generalization of the normal bidding process.

We provided mathematical programming formulations for winner determination applied to hybrid procurement mechanisms. We also presented a heuristic solution scheme, where solutions are constructed in a greedy manner. We illustrated the solution procedure using several instances of the problem, and showed in practice that some selection policies are more effective in the determination of the winner for such auctions.

It remains an open question if the mathematical programming models provided in this paper can be improved in order to reduce the time necessary to find an optimal solution. It would also be interesting to provide even better heuristics for this problem, probably using a general meta-heuristic scheme such as genetic algorithms or tabu search.

Acknowledgments

The second author was partially supported by a NSF Industry/University Cooperative Research Center grants (MCEC and CELDi). The third author was partially supported by NSF and US Air Force grants.

References

A. Pekec and H. Rothkopf: 2003, 'Combinatorial Auction Design'. *Management Science* **49**, 1485–1503.

Arora, S. and C. Lund: 1996, 'Hardness of Approximations'. In: D. Hochbaum (ed.): *Approximation Algorithms for NP-hard Problems*. PWS Publishing.

Cramton, P.: 1998, 'The Efficiency of the FCC Spectrum Auctions'. *Journal of Law and Economics* **41**, 727–736.

de Vries, S. and R. Vohra: 2003, 'Combinatorial Auctions: A Survey'. *INFORMS Journal on Computing* **15**(3).

Elmaghraby, W. and P. Keskinocak: 2003, 'Combinatorial Auctions in Procurement'. In: C. Billington, T. Harrison, H. Lee, and J. Neale (eds.): *The Practice of Supply Chain Management*. Kluwer Academic Publishers.

Garey, M. and D. Johnson: 1979, *Computers and Intractability: A Guide to the theory of NP-completeness*. San Francisco: W.H. Freeman.

Hohner, G., J. Rich, E. Ng, G. Reid, A. Davenport, J. Kalagnanam, H. Lee, and C. An: 2003, 'Combinatorial and Quantity-Discount Procurement Auctions Benefit Mars, Incorporated and Its Suppliers'. *Interfaces* **33**(1), 23–35.

Ledyard, J., M. Olson, D. Porter, J. Swanson, and D. Torma: 2002, 'The First Use of a Combined-Value Auction for Transportation Services'. *Interfaces* **32**(5), 4–12.

Leyton-Brown, K., Y. Shoham, and M. Tennenholtz: 2000, 'An Algorithm for Multi-Unit Combinatorial Auctions'. In: *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. pp. 56–61.

Moore, E., J. Warmke, and L. Gorban: 1991, 'The indispensable role of management science in centralizing freight operations at Reynolds Metals Company'. *Interfaces* **21**.

P.C. Crampton: 1997, 'The FCC Spectrum Auction: an early assessment'. *Journal of Economics and Management Strategy* **6(3)**, 431–495.

Rassenti, S., V. L. Smith, and R. L.Bulfin: 1982, 'A combinatorial auction mechanism for airport time slot allocation'. *Bell Journal of Economics* **13**, 402–417.

Rothkopf, M. H., A. Pekec, and R. M. Harstad: 1998, 'Computationally Manageable Combinational Auctions'. *Manage. Sci.* **44**(8), 1131–1147.

Sakurai, Y., M. Yokoo, and K. Kamei: 2000, 'An Efficient Approximate Algorithm for Winner Determination in Combinatorial Auctions'. In: *Proceedings of the ACM Conference on Electronic Commerce*. p. 30.

Sandholm, T.: 2000, 'Approaches to winner determination in combinatorial auctions'. *Decis. Support Syst.* **28**(1–2), 165–176.

Sandholm, T.: 2002, 'Algorithm for optimal winner determination in combinatorial auctions'. *Artificial Intelligence* **135**, 1–54.

Sandholm, T., S. Suri, A. Gilpin, and D. Levine: 2005, 'CABOB: A Fast Optimal Algorithm for Winner Determination in Combinatorial Auctions'. *Management Science* **51(3)**, 374–390.

Sheffi, Y.: 2004, 'Combinatorial auctions in the procurement of transportation services'. *Interfaces* **34**(4), 245–252.

Y. Narahari and P. Dayama: 2005, 'Combinatorial Auctions for Electronic Business'. *Sadhana Special issue on Electronic commerce* **30**, 179–212.

Proximal-ACCPM: A Versatile Oracle Based Optimisation Method

Frédéric Babonneau, Cesar Beltran, Alain Haurie, Claude Tadonki
and Jean-Philippe Vial

Logilab, HEC, Université de Genève, 40 Bd du Pont d'Arve, CH-1211 Geneva, Switzerland.

Summary. Oracle Based Optimisation (OBO) conveniently designates an approach to handle a class of convex optimisation problems in which the information pertaining to the function to be minimized and/or to the feasible set takes the form of a linear outer approximation revealed by an oracle. Three representative examples are introduced to show how one can cast difficult problems in this format, and solve them. An efficient method, Proximal-ACCPM, is presented to trigger the OBO approach. Numerical results for these examples are provided to illustrate the behavior of the method. This paper summarizes several contributions with the OBO approach and aims to give, in a single report, enough information on the method and its implementation to facilitate new applications.

Key words: Non-differentiable optimisation, cutting plane methods, interior-point methods, Proximal-ACCPM, multicommodity flow, p-median, integrated assessment models

1 Introduction

Oracle Based optimisation (OBO) conveniently designates an approach to handle a class of convex optimisation problems in which the information pertaining to the function to be minimized and/or to the feasible set takes the form of a linear outer approximation revealed by an oracle. By oracle, we mean a black-box scheme that returns appropriate information on the problem at so-called query points. In convex unconstrained optimisation, this information takes the form of a linear support for the epigraph set of the function to be minimized. This class of problems is known as "Nondifferentiable Convex Optimisation". We use the terminology OBO to emphasize the principle of the method — a dialog between an optimizer and an oracle — and the fact that we can handle more general classes of problems.

The goal of this paper is two-fold. We first intend to present an efficient method, Proximal-ACCPM, that implements an OBO approach. We give a

concise but accurate description of the analytic center cutting plane method (ACCPM), and more precisely of its recent enhancements that include a proximal term (Proximal-ACCPM) and a logarithmic barrier on the epigraph of the smooth component of the objective function. The main issue in a cutting plane method is to decide where to query the oracle in order to improve a current polyhedral approximation of the problem. Proximal-ACCPM selects the analytic center of this polyhedral set, that is, the point that minimizes the logarithmic barrier function on that set, augmented with a proximal term. This choice is efficient since it usually requires relatively few query points to achieve an accurate approximation of an optimal solution. Proximal-ACCPM relies on the interior-point methodology to compute the query points. This methodology is well suited to handle nonlinear information and makes it easy to implement the extensions we discuss in the paper.

Our second goal is to provide a set of application problems that are very different in nature and thus illustrate the versatility of the method. This choice does not cover the full range of applications successfully handled with Proximal-ACCPM. Yet it gives a flavor of what can be done and hopefully it will convince readers to develop applications of their own.

In this paper we do not deal with the convergence issue. The pseudo-polynomial complexity of the method on the feasibility problem has been proved in (Goffin et al., 1996; Nesterov, 1995). It straightforwardly extends to optimality problems by casting the latter in the format of a pure feasibility problem. The proofs are involved but the principles underlying the method are relatively simple. Neither will we review the literature on nondifferentiable convex optimisation. The field is large and we content ourselves with referring to survey papers (Lemaréchal, 1989; Goffin and Vial, 2002). In this presentation we concentrate on the description of the method with some recent extensions and we illustrate its implementation and performance on three large-scale applications recently reported in the literature.

The paper is organized as follows. In Sect. 2 we present the framework of Oracle Base Optimisation. Section 3 provides a succinct description of Proximal-ACCPM. Two enhancements of the method are discussed. None of them is really new, but we believe that they crucially contribute to the overall efficiency of the implementation. We also discuss how to compute a lower bound and thus obtain a reliable stopping criterion. Section 4 deals with three examples. The first one, the well-known multicommodity flow problem, is representative of large-scale continuous optimisation. The method has been applied to the linear (Babonneau et al., 2006) and the nonlinear (Babonneau and Vial, 2005) cases. The nonlinear version of the multicommodity flow problem we present here is particularly interesting, because part of the problem structure need not be revealed by a first-order oracle. As it is presented in Sect. 3, Proximal-ACCPM directly incorporates the nonlinear information and thus achieves a significant gain of efficiency.

The second application is the p-median problem, a combinatorial optimisation problem that is solved by Lagrangian relaxation. This example

illustrates how powerful is Lagrangian relaxation to generate lower bounds for the optimal value of this combinatorial problem. These bounds are further used in an enumerative scheme which computes an optimal integer solution. In the same subsection we present the new concept of semi-Lagrangian relaxation, recently introduced in (Beltran et al., 2004). There, it is shown that using semi-Lagrangian relaxation permits us to solve to optimality the original combinatorial problem without resorting to an enumerative scheme.

Our last application deals with air quality control in urban regions and the coupling of modules in Integrated Assessment Models (IAM). The economic activity creates pollutant emissions that are spatially distributed. Geographic and climate models translate those primary pollutant emissions into ozone concentrations which determine air quality. The objective of the study is to find an optimal adjustment of the economic activity that results in acceptable ozone concentrations. The modeling approach consists in coupling two models, a techno-economic model and a climate model, to properly handle the interaction between the economic activity and the air quality. From a methodological point of view, this approach is original as it allows the coupling of two models that have totally different natures.

2 Oracle Based Optimisation

Oracle based optimisation deals with the convex programming problem

$$\min\{f(u) = f_1(u) + f_2(u) \mid u \in U \subset \mathbb{R}^n\}, \tag{1}$$

where f_1 is a convex function, f_2 is a twice differentiable convex function and U is a convex set. We assume that $f_1(u)$ and U are revealed by a first order oracle while $f_2(u)$ is accessed through a second order oracle in an explicit way. By oracle, we mean a black-box procedure which at any *query point* u returns the information described in Definitions 1 and 2 below.

Definition 1. *A first-order oracle for problem (1) is a black box procedure with the following property. When queried at u, the oracle returns 1 or 2.*

1. *$u \notin U$ and (a, c) is such that $a^T u' - c \leq 0, \forall u' \in U$ (feasibility cut). In that case, we set $f_1(u) = +\infty$.*
2. *$u \in U$ and (a, c) is such that $a^T u' - c \leq f_1(u'), \forall u' \in U$ (optimality cut). In general, $a \in \partial f_1(u)$, $c = a^T u - f_1(u)$, but this is not necessarily so. The cut may have no intersection with the epigraph set (i.e., may be situated strictly below that set).*

Definition 2. *A second-order oracle for problem (1) is a black-box procedure with the following property. When queried at u, the oracle returns the function value and the first and second derivatives of $f_2(u)$.*

In the traditional OBO approach, the function f_2 is handled in the same way as f_1, that is by means of a first-order oracle. This approach looses information. In this paper, we exploit the explicit knowledge of the function f_2 and its derivatives in the form of a barrier on the epigraph set.

Assumption 1. *The function f_2 is such that the logarithmic barrier $-\log(\zeta - f_2(u))$ on the epigraph set of f_2, $\{(u,\zeta) \mid \zeta \geq f_2(u),\ u \in U\}$, is self-concordant.*

Remark 1. The concept of self-concordant function has been introduced in (Nesterov and Nemirovski, 1994) to extend the theory of interior-point methods for linear programming to a more general class of functions. The condition links the second and third derivatives of the function. For a thorough but more readable presentation of the theory of self-concordant functions we refer to (Nesterov, 2004).

In many applications, the objective function f_1 is a strictly positively weighted sum of p nonsmooth convex functions

$$f_1(u) = \sum_{i=1}^{p} \pi_i f_{1i}(u).$$

In that expression, we can consider that $f_1(u)$ is revealed by p independent first-order oracles. The epigraph of the function f is the set defined by $\{(u,z,\zeta) \mid \pi^T z \geq f_1(u),\ \zeta \geq f_2(u)\}$. Using this property, problem (1) can also be written in as

$$\begin{aligned} &\min \pi^T z + \zeta \\ &\text{s.t. } f_{1j}(u) - z_j \leq 0,\ j = 1,\ldots,p, \\ &\quad f_2(u) - \zeta \leq 0, \\ &\quad u \in U. \end{aligned} \tag{2}$$

This formulation is conveniently named the *disaggregate mode*.

The first order oracle is used to build a polyhedral approximation of the epigraph of f_1. Suppose the oracle has been queried at $u^k,\ k = 1,\ldots,\kappa$, and has returned feasibility and/or optimality cuts associated with those points. The corresponding inequalities are collected in

$$A^T u - E^T z \leq c.$$

In that definition, the subgradients a of the function f_1 form the matrix A while E is a binary matrix that is constructed as follows. If the objective f_1 is treated in an aggregate mode ($p = 1$), then E is a binary row vector. An entry one in E indicates that the z variable is present in the cut, implying that the cut is an *optimality cut*. In contrast, a zero indicates that the cut is a *feasibility cut*. If the objective f_1 is disaggregated into p components, row j of E corresponds to a variable z_j and each column corresponds to a cut. An

entry one in row j and column k indicates that the cut k is an optimality cut for $f_{1j}(u)$. If column k is a null vector, then cut k is a feasibility cut.

Let $\bar{\theta}$ be the best recorded value such that $\bar{\theta} = \min_{k \leq \kappa}\{f_1(u^k) + f_2(u^k)\}$. In view of the above definitions, we can define the localization set $\mathcal{L}_\kappa$ as

$$\mathcal{L}_\kappa = \left\{(u, z, \zeta) \mid A^T u - E^T z \leq c,\ f_2(u) \leq \zeta,\ \pi^T z + \zeta \leq \bar{\theta}\right\},$$

which is a subset of an outer approximation of the epigraph of f that contains all optimal pairs $(u^*, f(u^*))$. Thus, the search for a new query point should be confined to the localization set. Among possible solution methods for (1), we briefly sketch cutting plane schemes which work as follows:

1. Select a query point in the localization set.
2. Send the query point to the first order oracle and get back the optimality/ feasible cuts.
3. Send the query point to the second order oracle to compute the objective function f_2.
4. Update the lower and upper bounds and the localization set.
5. Test termination.

The main issue in the design of a cutting plane scheme is step 1. Different choices lead to different results. In that paper, we propose a particular method, named *Proximal-ACCPM*, that selects the analytic center of the localization set as the new query point.

3 Proximal-ACCPM

It is well-known that efficient methods for non differentiable convex optimisation rely on some regularization schemes to select the query point. We discuss here such a scheme; it is based on the concept of proximal analytic center which corresponds to the minimum of the standard logarithmic barrier augmented with a proximal term.

3.1 Proximal Analytic Center

We associate with the localization set a standard (weighted) logarithmic barrier

$$F(s_0, s, \sigma) = -w_0 \log s_0 - \sum_{i=1}^{\kappa} w_i \log s_i - \omega \log \sigma, \tag{3}$$

with $(s_0, s, \sigma) > 0$ defined by

$$\begin{aligned} s_0 &= \bar{\theta} - \pi^T z - \zeta, \\ s_i &= c_i - (A^T u - E^T z)_i,\ i \in K = \{1, \ldots, \kappa\}, \\ \sigma &= \zeta - f_2(u). \end{aligned}$$

The barrier function is augmented with a proximal term to yield the augmented barrier

$$\Psi(u, s_0, s, \sigma) = \frac{\rho}{2}||u - \bar{u}||^2 + F(s_0, s, \sigma), \tag{4}$$

where $\bar{u} \in \mathbb{R}^n$ is the query point that has achieved the best objective value $\bar{\theta}$. We name it the proximal reference point. The proximal analytic center is defined as the solution of

$$\begin{aligned} \min_{u,z,\zeta,s_0,s,\sigma} \quad & \Psi(u, s_0, s, \sigma) \\ \text{s.t. } & s_0 + \pi^T z + \zeta = \bar{\theta}, \\ & s_i + (A^T u - E^T z)_i = c_i, \qquad i \in K = \{1, \ldots, \kappa\}, \\ & \sigma + (f_2(u) - \zeta) = 0, \\ & s_0 > 0,\ s > 0,\ \sigma > 0. \end{aligned} \tag{5}$$

If $(u, z, \zeta, s_0, s, \sigma)$ is feasible to (5), then (5) is equivalent to minimizing $\Phi(u, z, \zeta) = \Psi(u, s_0, s, \sigma)$, in which s_0, s and σ are replaced by their value in u, z and ζ. Note that the localization set is not necessarily compact, but it is easy to show that, thanks to the proximal term, the generalized analytic center exists and is unique.

In the next paragraphs, we shall use the following notation. Given a vector $s > 0$, S is the diagonal matrix whose main diagonal is s. We also use $s^{-1} = S^{-1}e$ to denote the vector whose coordinates are the inverse of the coordinates of s. Similarly, $s^{-2} = S^{-2}e$. Finally, given two vectors x and y of same dimension, xy denotes their component-wise product. With this notation, the first order optimality conditions for (5) are

$$\rho(u - \bar{u}) + Aws^{-1} + \omega f_2'(u)\sigma^{-1} = 0, \tag{6}$$

$$\pi w_0 s_0^{-1} - Ews^{-1} = 0, \tag{7}$$

$$w_0 s_0^{-1} - \omega\sigma^{-1} = 0, \tag{8}$$

$$s_0 + \pi^T z + \zeta - \bar{\theta} = 0, \tag{9}$$

$$s_i + (A^T u - E^T z)_i - c_i = 0, \quad i \in K = \{1, \ldots, \kappa\}, \tag{10}$$

$$\sigma + f_2(u) - \zeta = 0. \tag{11}$$

The algorithm that computes the analytic center is essentially a Newton method applied to (6)–(11). We shall see later how the vector $\xi = ws^{-1}$ is used to derive a lower bound for the optimal solution.

In view of Assumption (1), Φ is self-concordant; Newton's method is thus polynomially convergent (Nesterov, 2004). For the sake of simplicity, let us define $v = (u, z, \zeta)$. In the case when v is feasible to (5) the Newton direction is

$$dv = -[\Phi''(v)]^{-1}\Phi'(v).$$

The damped Newton method for computing the proximal analytic center consists in taking damped steps to preserve feasibility of v. The aim is to

achieve a sufficient decrease of Φ, until the domain of quadratic convergence is reached. Let

$$\lambda(v) = ([\Phi''(v)]^{-1}\Phi'(v))^T\Phi'(v) = -dv^T\Phi'(v). \tag{12}$$

As long as $\lambda(v) > \frac{3-\sqrt{5}}{2}$ a step of length $(1+\lambda(v))^{-1}$ preserves feasibility and induces a decrease of Φ by an absolute constant. When $\lambda(v) \leq \frac{3-\sqrt{5}}{2}$ a full step is feasible and the method converges quadratically. The method has polynomial complexity.

The stopping criterion is triggered by the proximity measure. When $\lambda(v)$ falls below the threshold value $\eta < \frac{3-\sqrt{5}}{2}$, the search for the proximal analytic center stops. In practice, the much looser criterion $\eta = 0.99$ suffices.

3.2 Infeasible Newton's Method

Unfortunately we haven't easy access to feasible solution for problem (5). In cutting plane schemes, new constraints cut off the current iterate from the new localization set and there is no direct way to retrieve feasibility if the cuts are deep. Since we can't anymore eliminate the variables (s_0, s, σ), we can't apply a feasible Newton method to minimize Φ. Thus, we propose an infeasible start Newton method for (5), which aims to achieve feasibility and optimality simultaneously in the extended space $(u, z, \zeta, s_0, s, \sigma)$.

In the course of the optimisation process, the first order conditions (6)–(11) are never satisfied. However, we can assume that $(s_0, s, \sigma) > 0$. We introduce the residuals $r = (r_u, r_z, r_\zeta, r_{s_0}, r_s, r_\sigma)$ and write

$$\rho(u-\bar{u}) + Aws^{-1} + \omega f_2'(u)\sigma^{-1} = -r_u, \tag{13}$$
$$w_0\pi s_0^{-1} - Ews^{-1} = -r_z, \tag{14}$$
$$w_0 s_0^{-1} - \omega\sigma^{-1} = -r_\zeta, \tag{15}$$
$$s_0 + \pi^T z + \zeta - \bar{\theta} = -r_{s_0}, \tag{16}$$
$$s_i + (A^T u - E^T z)_i - c_i = -r_{s_i}, \quad i \in K = \{1, \ldots, \kappa\}, \tag{17}$$
$$\sigma + f_2(u) - \zeta = -r_\sigma. \tag{18}$$

The Newton direction associated to (13)–(18) is given by

$$P \begin{pmatrix} du \\ dz \\ d\zeta \\ ds_0 \\ ds \\ d\sigma \end{pmatrix} = \begin{pmatrix} r_u \\ r_z \\ r_\zeta \\ r_{s_0} \\ r_s \\ r_\sigma \end{pmatrix}, \tag{19}$$

where

$$P = \begin{pmatrix} \rho I + \omega f_2(u)''\sigma^{-1} & 0 & 0 & 0 & -AS^{-2} & \omega f_2(u)'\sigma^{-2} \\ 0 & 0 & 0 & -w_0\pi s_0^{-2} & E^T S^{-2} & 0 \\ 0 & 0 & 0 & -w_0 s_0^{-2} & 0 & \omega\sigma^{-2} \\ 0 & \pi^T & 1 & 1 & 0 & 0 \\ A^T & -E^T & 0 & 0 & I & 0 \\ f_2'(u) & 0 & -1 & 0 & 0 & 1 \end{pmatrix}.$$

Since (9) and (10) are linear, a full Newton step, i.e., a step of length 1, yields a point that is feasible with respect to these equations. However, the same step does not yield a feasible point with respect to the nonlinear equation (11). Thus, the method remains essentially infeasible and we cannot use the proximity measure λ to determine the steplength α_{step}. Instead, we use the following empirical rule. Let

$$\alpha_{max} = \max(\alpha \mid s + \alpha ds > 0,\ s_0 + \alpha ds_0 > 0,\ \sigma + \alpha d\sigma > 0),$$

the selected step is

$$\alpha_{step} = \min(1,\ \gamma\alpha_{max});$$

where the parameter γ is a safeguard to stay away from the boundary of the domain. In practice, we take $\gamma = 0.95$.

When $f_2(u)$ is linear (or constant), it may be the case that (9) and (10) become satisfied. Instead of using the default step length $(1 + \lambda(v))^{-1}$, as prescribed by the theory, we perform the one-dimensional linesearch

$$\alpha_{step} = \arg\min \Psi(v + \alpha dv).$$

As mentioned earlier, the query point is not feasible for the new cuts returned by the first order oracle. Finding a good starting value for $s_{\kappa+1}$ and/or s_0 after a cut has been added is an issue. Though (Goffin and Vial, 1999) proposes a scheme that preserves the polynomial complexity of the method, in our practical implementation we use a simple heuristic that turns out to be very efficient.

To summarize, a basic step of the Newton iteration is

1. Send the current point u to the second order oracle to compute the objectif function $f_2(u)$ and its first and second derivatives.
2. Compute the Newton step $(du, dz, d\zeta, ds_0, ds, d\sigma)$ by (19).
3. Compute a step length α_{step} to update $(u, z, \zeta, s_0, s, \sigma)$.
4. Test termination.

3.3 Lower Bound

A lower bound for (1) permits a measure of progress to optimality. We now explain a way to generate such a bound. The first step in the derivation of

the lower bound consists in introducing the perturbed function $f(u) - r^T u$, where r is a vector to be specified later. The second step is to replace the non-smooth function $f_1(u)$ by its current polyhedral approximation. This is done by replacing $f_1(u)$ by $\pi^T z$ under the constraints $A^T u - E^T z \leq c$. We thus have the bounding inequality

$$f(u) - r^T u \geq \min_{u,z} \{\pi^T z + f_2(u) - r^T u \mid A^T u - E^T z \leq c\}.$$

In view of the convexity of f_2, we may write

$$\begin{aligned} f(u) - r^T u \geq\ & f_2(u^c) - f_2'(u^c)^T u^c + \\ & \min_{u,z} \{\pi^T z + f_2'(u^c) u - r^T u \mid A^T u - E^T z \leq c\}, \end{aligned}$$

where u^c is a point of choice (e.g., approximate analytic center). By duality we obtain

$$\begin{aligned} f(u) - r^T u &\geq f_2(u^c) - f_2'(u^c)^T u^c + \\ &\quad \min_{u,z} \max_{\xi \geq 0} \{(f_2'(u^c) + A\xi)^T u + (\pi - E)^T z - c^T \xi - r^T u\}, \\ &= f_2(u^c) - f_2'(u^c)^T u^c + \max_{\xi \geq 0} \Big\{ -c^T \xi + \\ &\qquad + \min_{u,z} \left[(f_2'(u^c) + A\xi - r)^T u + (\pi - E\xi)^T z \right] \Big\}. \end{aligned} \tag{20}$$

If $\xi \geq 0$ is such that $f_2'(u^c) + A\xi = r$ and $E\xi = \pi$, then

$$f(u) \geq f_2(u^c) - f_2'(u^c)^T u^c + r^T u - c^T \xi.$$

We now show how one can get such a vector ξ at the end of the iterations that compute the proximal analytic center. In view of (14), we let $\xi = \xi^c = w(s^c)^{-1} > 0$ and we scale ξ^c by using the special structure of the matrix E to have $\pi - E\xi^c = 0$ and we define $r = f_2'(u^c) + A\xi^c$. In view of the optimality conditions (6) and (6) one may expect r to be small. We obtain the bound for the optimal objective function value by

$$\begin{aligned} f(u^*) &\geq f_2(u^c) - f_2'(u^c)^T u^c - c^T \xi^c + r^T u^*, \\ &\geq f_2(u^c) - f_2'(u^c)^T u^c - c^T \xi^c + r^T (u^* - u^c) + r^T u^c, \\ &\geq f_2(u^c) - f_2'(u^c)^T u^c + r^T u^c - c^T \xi^c - ||r|| \delta. \end{aligned} \tag{21}$$

The last inequality follows from Cauchy-Schwartz and $\delta \geq ||u^* - u^c||$ is an upper bound on the distance of the current point u^c to the optimal set. Finding a good value for δ cannot be done on theoretical grounds. It is essentially problem dependent. In practice, we obtained good results by taking the "empirical" value $\delta = 5 \times ||u^c - \bar{u}||$.

If the variable u is constrained to be nonnegative in (1), we can further improve the computation of the lower bound by taking

$r = -\min\{0, f_2'(u^c) + A\xi^c\}$, where the min operator is taken component-wise. In that case, the coefficient of u in the inner minimization is always nonnegative and $(f_2'(u^c) + A\xi - r)^T u = 0$ at the solution of (20). This remark is particularly useful when $r = 0$. Then we obtain the exact lower bound $f_2(u^c) - f_2'(u^c)^T u^c - c^T \xi^c$.

3.4 Implementation

Since the oracle is entirely user-defined, we do not include it in the description. The code has two main blocks: the first one computes query points; the second one organizes the dialog between the oracle and the query point generator. The code also includes an important initialization block.

Initialization This module initializes the instance and the various parameters.

Query point generator This modules includes two submodules: the first one creates the localization set based on the information sent by the cut manager; the second one computes approximate proximal analytic centers.

Manager This module keeps track of the cuts generated by the oracle and of the current primal and dual coordinates of the analytic center. It also controls the parameters that are dynamically adjusted and computes criteria values that can be used by the user to stop the algorithm. Finally, it acts as a filter between the oracle and the query point generator.

Two parameters of Proximal-ACCPM are often critical in the applications: the weight w_0 on the epigraph cut in (3) and the coefficient ρ of the proximal term in (4). The general strategy is to assign to w_0 a value equal to the number of generated cuts (Goffin and Vial, 2002). The management of the proximal term is more problem dependent. This point will be briefly commented in the next section. When the problem to be solved has no box constraints on the variables (e.g., when relaxing equality constraints in Lagrangian relaxation) the computation of the Newton direction in Proximal-ACCPM can be made more efficient than in plain ACCPM (du Merle and Vial, 2002).

The code is written in Matlab; it has around 700 lines of code in the query point generator and 400 in the manager. Matlab is particularly efficient in dealing with linear algebra. Not much gain can be expected by translating the code into C++. However, a C version would make it easier to link Proximal-ACCPM with oracles written in C or FORTRAN or to do an embedding of Proximal-ACCPM within a larger optimisation scheme (e.g., a branch and bound scheme). The code is the result of a continuing development efforts by teams at Logilab partly supported by Swiss NSF.

4 Applications

We have seen that oracle based optimisation is relevant when it is possible to approximate the epigraph set of the function to be minimized, and the feasible set, by polyhedral sets. Let us list a few techniques that lead to this

situation: Lagrangian relaxation (Geoffrion, 1974), Lagrangian decomposition (Guignard and Kim, 1987), column generation (Barnhart et al., 1998), Benders' decomposition (Benders, 2005), dual gap function in variational inequalities (Nesterov and Vial, 1999), etc. In this section we present three representative applications, one in large-scale nonlinear continuous optimisation, one in combinatorial optimisation and one dealing with the coupling of economic and environmental models. Those problems have been fully treated in (Babonneau et al., 2006; Babonneau and Vial, 2005; Beltran et al., 2004; Carlson et al., 2004).

In each case, we give a brief presentation of the problem and report a sample of numerical results. This will give the reader an idea of the type of problems that can be solved with Proximal-ACCPM. When the numerical results are displayed in a table, we give the following information: problem identification, denoted 'Problem ID', number of outer iterations (equivalently, the number of oracle calls), denoted 'Outer', number of inner iterations (equivalently, the number of Newton iterations to compute an analytic center), denoted 'Inner', total CPU time in second, denoted 'CPU' and the fraction of the CPU time spent in the oracle, denoted '%Oracle'.

4.1 Multicommodity Flow Problems

Given a network represented by the directed graph $\mathcal{G}(\mathcal{N}, \mathcal{A})$, with node set $\mathcal{N}$ and arc set $\mathcal{A}$, the multicommodity flow problem consists in shipping some commodity flows from sources to sinks such that the demands for each commodities are satisfied, the arc flow constraints are met and the total cost flow is minimum. The arc-flow formulation of the multicommodity flow problem is

$$\min \quad \sum_{a \in \mathcal{A}} f_a(y_a) \tag{22}$$

$$\text{s.t.} \sum_{k \in \mathcal{K}} x_a^k = y_a, \quad \forall a \in \mathcal{A}, \tag{23}$$

$$N x^k = d_k \delta^k, \quad \forall k \in \mathcal{K}, \tag{24}$$

$$x_a^k \geq 0, \qquad \forall a \in \mathcal{A},\ \forall k \in \mathcal{K}. \tag{25}$$

Here, N is the network matrix; $\mathcal{K}$ is the set of commodities; d_k is the demand for commodity k; and δ^k is vector with only two non-zeros components: a 1 at the supply node and a -1 at the demand node. The variable x_a^k is the flow of commodity k on arc a of the network and x^k is the vector of x_a^k. The objective function f is a congestion function on the arcs.

For the sake of simpler notation we write problem (22)–(25) in the more compact formulation

$$\min\{f(y) \mid Bx = y,\ x \in X\}, \tag{26}$$

where X represents the set of feasible flows that meet the demands with respect to the network constraints. Bx defines the load flow.

The standard Lagrangian relaxation of (26) assigns the dual variables u to the coupling constraints $Bx = y$ and relaxes them. The Lagrangian problem is

$$\max_u \mathcal{L}(u), \tag{27}$$

where

$$\begin{aligned}\mathcal{L}(u) &= \min_{x\in X, y} f(y) + u^T(Bx - y),\\ &= \min_y (f(y) - u^T y) + \min_{x\in X} u^T Bx,\\ &= -f_*(u) + \min_{x\in X} u^T Bx.\end{aligned}$$

The function $f_*(u)$ is the Fenchel conjugate of f; it is convex. In the multicommodity case, the second part of the Lagrangian is a sum of $|\mathcal{K}|$ shortest path problems. We denote

$$\mathrm{SP}(\bar{u}) = \min_{x\in X} (B^T \bar{u})^T x. \tag{28}$$

We recall that in Proximal-ACCPM, we treat the negative of the objective function (27). Let $\bar{x}$ be an optimal solution returned by the oracle (28) at a given point $\bar{u}$. Since $\mathrm{SP}(u)$ results from a minimization problem, the inequality $\mathrm{SP}(u) \leq (B\bar{x})^T u$ provides a linear upper estimate of the concave function $\mathrm{SP}(u)$. The solution computed by the oracle $-f_*(\bar{u}) + (B\bar{x})^T \bar{u}$ produces a lower bound for the original problems. Instead of using (21) to compute an upper bound, we use the variable ξ to compute a feasible solution to (22) (It can be shown).

For the nonlinear multicommodity flow problem, we use the most widely used function in telecommunications, the so-called Kleinrock congestion function:

$$f(y) = \frac{y}{c - y},$$

where c is the vector of capacities on the arcs. The conjugate function is

$$f_*(u) = 2\sqrt{c^T u} - c^T u - 1, \quad \forall u \geq \frac{1}{c}.$$

For the linear case, the objective function is

$$f(y) = \begin{cases} t^T y, \ 0 \leq y \leq c,\\ +\infty, \text{ otherwise,}\end{cases}$$

where c is the vector of capacities and t the vector of unit shipping cost on the arcs. The conjugate function is

$$f_*(u) = c^T u, \quad \forall u \geq 0.$$

To get a feel for the numerical performance, we pick few examples that have been solved in (Babonneau et al., 2006; Babonneau and Vial, 2005). We

select 3 types of problems. `Planar` and `Grid` instances are telecommunications networks while `Winnipeg`, `Barcelona` and `Chicago` are transportation problems. Table 1 gives for each problem the number of nodes, the number of arcs, and the number of commodities. The oracle is a shortest path problem solved with Dijkstra algorithm. The code is written in C. The tests were performed on a PC (Pentium IV, 2.8 GHz, 2 Gb of RAM) under Linux operating system.

Table 2 shows the numerical results to solve the linear and the nonlinear case with a relative otpimality gap less than 10^{-5}. We followed different strategies in the management of the proximal term, depending on whether the problem is linear or not. In the linear case, a constant value for the proximal parameter, say $\rho = 10^{-2}$ is suitable. In the nonlinear case, the proximal parameter is dynamically adjusted, according to success or failure in improving the value of the Lagrangian dual objective (lower bound). We start with $\rho = 1$ and multiply the current ρ by 10 in case of a 3 consecutive failures, up to the limit value $\rho = 10^{10}$.

Table 1. Test problems

Problem ID	# nodes	# arcs	# commodities
`planar500`	500	2842	3525
`planar800`	800	4388	12756
`planar1000`	1000	5200	20026
`grid12`	900	3480	6000
`grid13`	900	3480	12000
`grid14`	1225	4760	16000
`grid15`	1225	4760	32000
`Winnipeg`	1067	2975	4345
`Barcelona`	1020	2522	7922
`Chicago`	933	2950	93513

Table 2. Numerical results

	Linear case				Nonlinear case			
Problem ID	Outer	Inner	CPU	%Oracle	Outer	Inner	CPU	%Oracle
`planar500`	229	744	88.7	21	127	324	32.2	37
`planar800`	415	1182	557.2	16	182	429	110.5	40
`planar1000`	1303	2817	7846.7	12	381	869	568.1	26
`grid12`	509	1341	658.5	18	201	409	106.7	41
`grid13`	673	1629	1226.8	12	222	454	128.7	39
`grid14`	462	1363	843.6	22	204	414	173.2	48
`grid15`	520	1450	1055.1	20	203	414	172.8	48
`Winnipeg`	224	592	81.2	18	338	988	215.0	14
`Barcelona`	157	421	35.9	23	253	678	101.1	15
`Chicago`	180	493	79.2	47	145	370	48.6	41

The results in Table 2 have been further improved by means of column elimination and an active set strategy. With these enhancements, the method could solve huge instances with up to 40,000 arcs and 2,000,000 commodities. It has also been compared to other state-of-the-art methods. It appears to be very competitive, especially in the linear case, where it turns out to be from 4 to 30 times faster than the best known results. (For more details, see (Babonneau et al., 2006; Babonneau and Vial, 2005).)

Let us also mention that the impact of the proximal term has been analyzed to some depth in the two papers cited above. The introduction of a proximal term in ACCPM instead of box constraints on the variables has proved to be beneficial in almost all cases. It never appeared to be detrimental. On nonlinear multicommodity flow problems or on linear problems with an advanced treatment (column elimination, active set strategy) the version with the proximal term outperformed the version with box constraints.

4.2 Lagrangian Relaxations of the p-median Problem

In the p-median problem the objective is to open p 'facilities' from a set of m candidate facilities relative to a set of n 'customers', and to assign each customer to a single facility. The cost of an assignment is the sum of the shortest distances c_{ij} from a customer to a facility. The distance is sometimes weighted by an appropriate factor, e.g., the demand at a customer node. The objective is to minimize this sum. Applications of the p-median problem can be found in cluster analysis, facility location, optimal diversity management problem, etc. (Briant and Naddef, 2004). The p-median problem is NP-hard (Kariv and Hakimi, 1979).

The p-median problem can be formulated as follows

$$\min_{x,y} \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} \tag{29}$$

$$\text{s.t.} \sum_{i=1}^{m} x_{ij} = 1, \quad \forall j, \tag{30}$$

$$\sum_{i=1}^{m} y_i = p, \tag{31}$$

$$x_{ij} \le y_i, \quad \forall i, j, \tag{32}$$

$$x_{ij}, y_i \in \{0, 1\}, \tag{33}$$

where $x_{ij} = 1$ if facility i serves the customer j, otherwise $x_{ij} = 0$ and $y_i = 1$ if we open facility i, otherwise $y_i = 0$.

In the following two sections we formulate the (standard) Lagrangian relaxation of the p-median problem, and the semi-Lagrangian relaxation.

Standard Lagrangian Relaxation of the p-median Problem

In this section we focus on the resolution of the (standard) Lagrangian relaxation (LR) of the p-median problem by means of Proximal-ACCPM. To this end, we relax constraints (30) and (31) in problem (29)–(33), to yield the dual problem

$$\max_{u,v} \mathcal{L}_1(u,v), \tag{34}$$

and the oracle

$$\mathcal{L}_1(u,v) = \min_{x,y} \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} + \sum_{j=1}^{n} u_j(1 - \sum_{i=1}^{m} x_{ij}) + v(p - \sum_{i=1}^{m} y_i) \tag{35}$$

$$\text{s.t.} \quad x_{ij} \leq y_i, \quad \forall i,j, \tag{36}$$

$$x_{ij}, y_i \in \{0,1\}, \tag{37}$$

where $u \in \mathbb{R}^n$ is associated to the constraints $\sum_{i=1}^{m} x_{ij} = 1, \quad j = 1,\ldots,n$, and $v \in \mathbb{R}$ to the constraint $\sum_{i=1}^{m} y_i = p$.

We name *Oracle 1* this oracle; it is trivially solvable. Its optimal solution is also optimal for its linear relaxation. Consequently, the optimum of $\mathcal{L}_1$ coincides with the optimum of the linear relaxation of (29).

To show Proximal-ACCPM performance when solving the standard Lagrangian relaxation (35), we take a few examples reported in (du Merle and Vial, 2002). In this technical report, several p-median problems based on data from the *traveling salesman problem* (TSP) library (Reinelt, 2001) are solved. Instances of the grid problem, where the customers are regularly spaced points on square, are also solved. In Table 3 we show the results for ten representative instances (Proximal-ACCPM stopping criterion set equal to 10^{-6}). In this case, the proximal parameter is set to $\rho = 1$ initially and is dynamically adjusted by multiplicative factors 2 and 0.5 depending on the success or failure in improving the objective of the Lagrangian dual objective. The updating is limited by the bounds 10^{-6} and 10^4. Programs have been

Table 3. Numerical results

Problem ID	n	p	Outer	Inner	CPU	%Oracle
`Grid1521`	1521	10	348	902	132	33
`Grid1849`	1849	10	417	1042	241	32
`Grid2025`	2025	10	382	961	229	37
`Grid2304`	2304	10	448	1111	370	34
`Grid2500`	2500	10	440	1095	428	34
`TSP1817`	1817	10	1070	2303	1861	10
`TSP2103`	2103	10	316	701	156	48
`TSP2152`	2152	10	196	430	98	51
`TSP2319`	2319	10	369	775	237	46
`TSP3038`	3038	10	127	292	102	62

written in MATLAB and run in a PC (Pentium-III PC, 800 MHz, with 256 Mb of RAM) under the Linux operating system.

Semi-Lagrangian Relaxation of the p-median Problem

The standard Lagrangian relaxation is commonly used in combinatorial optimisation to generate lower bounds for a minimization problem. An optimal integer solution is obtained by a branch and bound scheme. The semi-Lagrangian relaxation (SLR) is a more powerful scheme, introduced in (Beltran et al., 2004), that generates an optimal integer solution for (linear) combinatorial problems with equality constraints.

To strengthen $\mathcal{L}_1$, the SLR introduces in problem (29)–(33) the redundant constraints $\sum_i x_{ij} \leq 1$, $j = 1, \ldots, n$, and $\sum_i y_i \leq p$. After relaxing (30–31), we obtain the SLR dual problem

$$\max \mathcal{L}_3(u, v), \tag{38}$$

and the new oracle

$$\mathcal{L}_3(u,v) = \min_{x,y} \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} + \sum_{j=1}^{n} u_j(1 - \sum_{i=1}^{m} x_{ij}) + v(p - \sum_{i=1}^{m} y_i) \tag{39}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} x_{ij} \leq 1, \quad \forall j, \tag{40}$$

$$\sum_{i=1}^{m} y_i \leq p, \tag{41}$$

$$x_{ij} \leq y_i, \quad \forall i, j, \tag{42}$$

$$x_{ij}, y_i \in \{0, 1\}. \tag{43}$$

This oracle, which we name *Oracle 3*, is much more difficult than Oracle 1 (in fact, Oracle 3 is NP-hard). To cope with this difficulty one can use an intermediate oracle (*Oracle 2*) defined as the Oracle 3 but without constraint (41). We denote $\mathcal{L}_2$ the associated dual function. In general, Oracle 2 is easier to solve than Oracle 3, especially in cases where the p-median underlying graph associated to Oracle 2 decomposes into independent subgraphs. In such situation, we solve an integer problem per subgraph (see (Beltran et al., 2004) for more details).

It can be seen that solving the SLR dual problem (39) completely solves the p-median problem. Based on this result, we design a branch-and-bound free procedure to completely solve the p-median problem. This procedure successively maximizes the dual functions $\mathcal{L}_i(u, v), i = 1, 2, 3$. In this succession of three dual problems, the optimal solution of one dual problem is used as the starting point for the next dual problem. After solving the last dual problem ($\mathcal{L}_3(u, v)$) we obtain, as a by-product, an optimal integer solution for (29).

These dual problems are solved by means of Proximal-ACCPM. Oracle 2 and 3 are solved by means of CPLEX 8.1. Note that, although our procedure is branch-and-bound free, CPLEX is, of course, based on a sophisticated branch-and-bound procedure.

If we are not able to solve the three dual problems we will only have a lower bound of the p-median optimal value. In this case, we will compute an integer solution for the p-median problem by means of an heuristic as for example the "Variable Neighborhood Decomposition Search" (VNDS) (Hansen et al., 2001). The quality or the integer solution will be determined by the dual lower bound.

In Tables 4 and 5 we show the results (solution quality and performance) for 10 representative examples of the 44 instances tested in (Beltran et al., 2004). These instances can be found in the TSPLIB (Reinelt, 2001) and range from 1304 to 3795 customers, which implies 2 to 14 million binary variables. The proximal parameter is set to the constant value $\rho = 10^{-2}$ for problems with Oracle 2 and Oracle 3. In these tables 'Or.' stands for Oracle, 'VNDS' for variable neighborhood decomposition search, 'SLR' for semi-Lagrangian relaxation and 'ANIS' for averaged number of independent subgraphs. '%Opt.' gives the quality of the solution and is computed as

$$100 \times \left(1 - \frac{\text{'Upper bound'} - \text{'Lower bound'}}{\text{'Lower bound'}}\right).$$

Programs have been written in MATLAB and run on a PC (Pentium-IV Xeon PC, 2.4 GHz, with 6 Gb of RAM) under the Linux operating system. Note that in some cases the Oracle 3 is not called. The reason is either because the problem has been completely solved by the second dual problem or the CPU time limit has been reached when solving the second dual problem.

Table 4. Solution quality

Instance			Lower bound			Upper bound		%Opt.
Problem ID	n	p	Or. 1	Or. 2	Or. 3	Value	Method	
rl1304	1304	10	2131787.5	2133534	–	2134295	VNDS	99.96
rl1304	1304	500	97008.9	97024	–	97024	SLR	100
vm1748	1748	10	2982731.0	2983645	–	2983645	SLR	100
vm1748	1748	500	176976.2	176986	176986	176986	SLR	100
d2103	2103	10	687263.3	687321	–	687321	SLR	100
d2103	2103	500	63938.4	64006	64006	64006	SLR	100
pcb3038	3038	5	1777657.0	1777677	–	1777835	VNDS	99.99
pcb3038	3038	500	134771.8	134798	134798	136179	VNDS	98.98
fl3795	3795	150	65837.6	65868	–	65868	SLR	100
fl3795	3795	500	25972.0	25976	25976	25976	SPR	100

Table 5. Performance

Instance			Outer			ANIS	CPU			
Problem ID	n	p	Or. 1	Or. 2	Or. 3		Or. 1	Or. 2	Or. 3	Total
rl1304	1304	10	390	35	0	1	95	17241	0	17336
rl1304	1304	500	133	15	0	143	8	40	0	48
vm1748	1748	10	500	21	0	1	174	3771	0	3945
vm1748	1748	500	146	15	2	131	14	61	22	97
d2103	2103	10	241	7	0	2	41	504	0	545
d2103	2103	500	500	26	2	39	143	10086	4309	14538
pcb3038	3038	5	341	5	0	1	111	1988	0	2099
pcb3038	3038	500	211	17	2	38	56	3269	3900	7225
fl3795	3795	150	1000	27	0	17	1100	39199	0	40299
fl3795	3795	500	500	38	1	25	259	2531	218	3008

4.3 Coupling Economic and Environmental Models

Integrated assessment of environmental (IAM) policies is becoming an important priority due to the social need for local air pollution control or global climate change mitigation. Typically an IAM will combine an economic model and an environmental model to yield an evaluation of the costs and benefits associated with some environmental goals, given the technological and economic choices that are available. In this section we present a successful implementation using Proximal-ACCPM in this context.

In (Haurie et al., 2004), it has been proposed to use an oracle-based method to couple an Eulerian air quality model and a techno-economic model of energy choices in an urban region. The implementation of the approach has been further developed and tested in (Carlson et al., 2004). Ozone (O_3) pollution is usually modelled in so-called Eulerian models that represent the transport of primary pollutants (typically NO_x and VOCs) and the air photochemistry under various weather conditions and for the specific topography of the region considered. These models take the form of large scale distributed parameter systems that are run over specific "weather episodes" (for example a two-day summer sunny period which may amplify the probability of ozone peaks in green areas). These simulations serve to build air-quality indicators like, e.g. the *ozone concentration peak* or *the average over a threshold* (AOT) during an episode. On the other side techno-economic models are dynamic capacity expansion and production models, also called activity analysis models. A typical example is MARKAL, initially developed to represent energy-technology choices at a country level (see (Fishbone and Abilock, 1981), (Berger et al., 1992)) and also adapted to the description of these choices at a city level in (Fragnière and Haurie, 1996a) and (Fragnière and Haurie, 1996b). In a MARKAL model the planning horizon is in general defined as 9 periods of 5 years. The model finds, for specified demands in

energy services, world prices of imported energy and given a gamut of technology choices, an investment plan and a production program that minimize a system-wide total discounted cost while satisfying some pollutant emissions limits.

From this brief description of the two categories of models, the reader may realize that they belong to very different worlds. The interaction of the models in a coupling procedure can be schematized as follows. The economic model produces a vector of pollutants emissions per sector of activity. These emissions are then distributed over time and space using patterns that depend on the type of activity. For instance, global urban heating emissions are easily dispatched in space using the geographical distribution of buildings. They are also distributed in time to follow a yearly seasonal pattern. The other important cause of emissions is the volume of traffic. The economic activity analysis proposes a list of technologies used in different transport sectors (cars, public transport, taxis, etc), resulting in a global emission level for each of these sectors. To obtain the spatio-temporal distribution of these emissions due to traffic one resorts to a complex congestion model of traffic, that essentially computes traffic equilibria. These different sources of pollutant emissions are then combined into a spatio-temporal distribution map of emissions. The last step in the analysis consists in simulations performed with the Eulerian model to compute air quality indices on a set of critical episodes. The combination of models that eventually produces the air quality indices is complex, but at the end one can effectively compute air quality indices as a function of the global emissions of pollutants by sector of economic activity. Clearly, one cannot expect this function to be linear. Even worse, the computation may be very time consuming.

We have described a one-way interaction of the models, starting from the economic model and ending with air quality indices. Let us now describe the feedback from the air quality assessment. Indeed, one may want to limit peaks of pollution. This can be translated into upper limits on the air quality indices. We now study this reverse mechanism and show how the complete problem can be recast in the format of problem (1). Let us first schematize the economic activity analysis as the linear program

$$\min\{c^T x \mid Ax = a,\ x \geq 0\}. \tag{44}$$

We shall refer to it as the E^3 model. The economic activity x induces a vector y of pollutant emissions. This vector is indexed by sector of activity. In the paradigm of linear activity analysis, the total emission vector is assumed to be a linear function of the economic activity level, say

$$y = Bx.$$

The complex transformation of the vector y of sectorial emissions into air quality indices is represented by a vector function $\Pi(y)$. In (Carlson et al., 2004) it is shown that one can compute the function value and estimate its

gradient at any point y. If $\bar{\Pi}$ is the bound imposed on the air quality indices (higher indices imply lower air quality), we can represent our complex problem as the mathematical programming problem

$$\min\{c^T x \mid Ax = a,\ Bx - y = 0,\ \Pi(y) \leq \bar{\Pi},\ x \geq 0\}. \tag{45}$$

This large-scale highly nonlinear model is intractable by standard optimisation tools. However, it is quite easily amenable to an Oracle Based Optimisation approach. To this end, we introduce the function

$$f(y) = \min\{c^T x \mid Ax = a,\ Bx = y,\ x \geq 0\}, \tag{46}$$

and the set

$$Y = \{y \mid \Pi(y) \leq \bar{\Pi}\}. \tag{47}$$

Our original problem can now be written as

$$\min\{f(y) \mid y \in Y\}.$$

It remains to show that the above problem is of the same type as (1). It is a well-known fact of convex analysis that the function $f(y)$ is convex (this is easily seen by considering the dual of the linear program that defines f) and that one can compute a subgradient at each point of the domain of the function. Unfortunately, one cannot make a similar statement on Y. Being the result of such a complex transformation process, $\Pi(y)$ is likely to be nonconvex. However, one can hope that in the range of values that are of interest the nonconvexity is mild. This is supported by empirical evidence. A gradient is also estimated by a finite difference scheme.

Even in presence of mild nonconvexity, one cannot exclude pathology in running Proximal-ACCPM. A separating hyperplane for the set Y may turn out to cut off part of the set, and exclude a point that was proved to be feasible earlier. To cope with this difficulty, the authors of (Carlson et al., 2004) simply shifted the plane to maintain feasibility. They also made problem (46) easier by assuming monotonicity that made it possible to replace the equality constraint $Bx = y$ by $Bx \leq y$.

As the air chemistry description actually involves nonlinear functions, we have implemented a technique of successive local linearizations of the air pollution dynamic equations. The details of the implementation are given in (Carlson et al., 2004). In a particular simulation based on data describing the Geneva (Switzerland) region, a solution to the reduced order optimisation problem is obtained through Proximal-ACCPM, with 30 calls to the oracles (24 feasibility cuts and 6 optimality cuts were peformed). A feasibility cut (call to the air quality oracle) takes 30 minutes computing time (SUN Ultra-80, Ultrasparc driver) whereas an optimality cut (call to the techno-economic model) takes 10 seconds.

This application demonstrates the possibilities offered by an OBO method to tackle Integrated Assessment Models where part of the modeling is a large-scale simulator of complex physics and chemistry processes. Since Proximal-ACCPM keeps the number of oracle calls to a small or moderate size it permits

the use of these simulators in the design of some oracles and therefore it realizes the coupling that is the essence of IAMs.

Remark 2. A similar implementation has been realized recently for an IAM of climate change policies. It is reported in (Drouet et al., 2005a; Drouet et al., 2005b). In that case the coupling is realized between an economic growth model and an intermediate complexity climate model. This second successful experience that we will not further described here confirms the potential of OBO techniques for the exploitation of complex and large-scale IAMs.

5 Conclusion

In this paper we have presented Proximal-ACCPM, an efficient method for convex nondifferentiable optimisation, and discussed three large-scale applications that are representative of an oracle based optimisation approach. Our presentation of Proximal-ACCPM focuses on the necessary information for an efficient implementation. It also includes recent extensions, in particular an explicit treatment of second-order information when this information is available. The three examples we selected have recently been reported in the literature. They are genuinely very large-scale problems. The first two are solved using a classical transformation known as Lagrangian relaxation. The transformed problem has much smaller dimension, thousands of variables instead of millions, but one can only collect information about it via a first-order oracle. It is shown that Proximal-ACCPM is powerful enough to solve huge instances of these problems. The third application fully exploits the concept of oracle based optimisation to organize a dialog between two large-scale models that have totally different natures, a techno-economic model and a large-scale simulator of complex physics and chemistry processes. The exchanges between the two models are performed through few variables and each model is treated as a first-order oracle vis-à-vis these variables. These oracles, and especially the simulator, are computationally costly. To make the OBO approach successful, one needs a method that keeps the number of calls to the oracles as low as possible. Proximal-ACCPM does the job.

Acknowledgements

The work was partially supported by the Swiss NSF (Fonds National Suisse de la Recherche Scientifique, grant # 12-57093.99).

References

Babonneau, F., O. du Merle and J.-P. Vial. Solving large scale linear multicommodity flow problems with an active set strategy and Proximal-ACCPM. *Operations Research*, 54(1):184–197, 2006.

Babonneau, F. and J.-P. Vial. ACCPM with a nonlinear constraint and an active set strategy to solve nonlinear multicommodity flow problems. Tech. rep., HEC/ Logilab, University of Geneva, 40 Bd du Pont dArve, CH-1211, 2005. To appear in *Mathematical Programming*.

Barnhart, C., E.L. Johnson, G.L. Nemhauser, M.W.P. Savelsbergh and P.H. Vance. Branch-and-price: Column generation for solving huge integer programs. *Operations Research*, 46(3):316–329, 1998.

Beltran, C., L. Drouet, N.R. Edwards, A. Haurie, J.-P. Vial and D.S. Zachary. An oracle method to couple climate and economic dynamics, Chap. 3 in A. Haurie and L. Viguier (eds.), *The Coupling of Climate and Economic Dynamics*, Springer, 2005.

Beltran, C., C. Tadonki and J.-P. Vial. Solving the p-median problem with a semi-lagrangian relaxation. *Computational Optimization and Applications*, 35 (2006).

Benders, J.F. Partitioning procedures for solving mixed-variables programming problems. *Computational Management Science*, 2:3–19, 2005. Initially appeared in *Numerische Mathematik*, 4: 238–252, 1962.

Berger, C., R. Dubois, A. Haurie, E. Lessard, R. Loulou and J.-P. Waaub. Canadian MARKAL: An advanced linear programming system for energy and environmental modelling. *INFOR*, 30(3):222–239, 1992.

Briant, O. and D. Naddef. The optimal diversity management problem. *Operations research*, 52(4), 2004.

Carlson, D., A. Haurie, J.-P. Vial and D.S. Zachary. Large scale convex optimisation methods for air quality policy assessment. *Automatica*, 40:385–395, 2004.

Drouet, L., N.R. Edwards and A. Haurie. Coupling climate and economic models: a convex optimization approach. *Environmental Modeling and Assessment*, Vol.11, pp.101–114, 2006.

du Merle, O. and J.-P. Vial. Proximal ACCPM, a cutting plane method for column generation and Lagrangian relaxation: application to the p-median problem. Technical report, Logilab, University of Geneva, 40 Bd du Pont d'Arve, CH-1211 Geneva, Switzerland, 2002.

Fishbone, L.G. and H. Abilock. MARKAL, a linear programming model for energy systems analysis: Technical description of the BNL version. *International Journal of Energy Research*, 5:353–375, 1981.

Fragnière, E. and A. Haurie. MARKAL-Geneva: A model to assess energy-environment choices for a Swiss Canton. In C. Carraro and A. Haurie, editors, *Operations Research and Environmental Management*, Vol. 5 of *The FEEM/KLUWER International Series on Economics, Energy and Environment*. Kluwer Academic Publishers, 1996.

Fragnière, E. and A. Haurie. A stochastic programming model for energy/ environment choices under uncertainty. *International Journal of Environment and Pollution*, 6(4–6):587–603, 1996.

Geoffrion, A.M. Lagrangean relaxation for integer programming. *Mathematical Programming Study*, 2:82–114, 1974.

Goffin, J.-L., Z. Q. Luo and Y. Ye. Complexity analysis of an interior point cutting plane method for convex feasibility problems. *SIAM Journal on Optimisation*, 69:638–652, 1996.

Goffin, J.-L. and J.-P. Vial. Shallow, deep and very deep cuts in the analytic center cutting plane method. *Mathematical Programming*, 84:89–103, 1999.

Goffin, J.-L. and J.-P. Vial. Convex nondifferentiable optimisation: A survey focussed on the analytic center cutting plane method. *Optimisation Methods and Software*, 174:805–867, 2002.

Guignard, M. and S. Kim. Lagrangean decomposition: a model yielding stronger Lagrangean bounds. *Mathematical Programming*, 39:215–228, 1987.

Hansen, P., N. Mladenovic and D. Perez-Brito. Variable neighborhood decomposition search. *Journal of Heuristics*, 7:335–350, 2001.

Haurie, A., J. Kübler, A. Clappier and H. van den Bergh. A metamodeling approach for integrated assessment of air quality policies. *Environmental Modeling and Assessment*, 9:1–122, 2004.

Kariv, O. and L. Hakimi. An algorithmic approach to network location problems. ii: the p-medians. *SIAM Journal of Applied Mathematics*, 37(3):539–560, 1979.

Lemaréchal, C. Nondifferentiable optimisation. In G.L. Nemhauser, A.H.G Rinnooy Kan, and M.J. Todd, editors, *Handbooks in Operations Research and Management Science*, Vol. 1, pp. 529–572. North-Holland, 1989.

Nesterov, Y. Complexity estimates of some cutting plane methods based on the analytic center. *Mathematical Programming*, 69:149–176, 1995.

Nesterov, Y. *Introductory Lectures on Convex Optimisation, a Basic Course*, Vol. 87 of *Applied Optimisation*. Kluwer Academic Publishers, 2004.

Nesterov, Y. and A. Nemirovski. *Interior Point Polynomial Algorithms in Convex Programming: Theory and Applications*. SIAM, Philadelphia, Penn., 1994.

Nesterov, Y. and J.-P. Vial. Homogeneous analytic center cutting plane methods for convex problems and variational inequalities. *SIAM Journal on Optimisation*, 9:707–728, 1999.

Reinelt G. Tsplib, 2001. http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95.

A Survey of Different Integer Programming Formulations of the Travelling Salesman Problem

A.J. Orman[1] and H.P. Williams[2]

[1] Shell Gas and Power International B.V. The Hague, Netherlands
[2] London School of Economics, Houghton Street, London, WC2A 2AE

Summary. Eight distinct (and in some cases little known) formulations of the Travelling Salesman Problem as an Integer Programme are given. Apart from the standard formulation all the formulations are 'compact' in the sense that the number of constraints and variables is a polynomial function of the number of cities in the problem. Comparisons of the formulations are made by projecting out variables in order to produce polytopes in the same space. It is then possible to compare the strengths of the Linear Programming relaxations. These results are illustrated by computational results on a small problem.

Key words: Travelling salesman problem, polytopes

1 Introduction

In this paper we survey eight different formulations of the Asymmetric Travelling Salesman Problem (ATSP) as an Integer Programme (IP). We choose to treat the Asymmetric case as being more general than the Symmetric case. Some of the work has been published elsewhere by other authors. Our purpose is, however, to provide new results as well as present a unifying framework, by projecting all the formulations into the same space.

In Sect. 2 we present the eight formulations classifying them as 'conventional' (**C**), "sequential" (**S**), "flow based" (**F**) and "time staged" (**T**). The reasons for these terms will become apparent. In order to facilitate comparison between the formulations, in some cases we introduce extra variables which equate to expressions within the models. This enables us, in Sect. 3, to compare the Linear Programming (LP) relaxations of all the formulations by projecting out all, but the, common variables. Such comparisons have already been done for some of the formulations by Padberg and Sung (1991), Wong (1980) and Langevin et al. (1990).

Some of the time staged formulations have also been compared by Gouveia and Voss (1995) and discussed by Picard and Queyranne (1978).

The sequential formulation has also been improved by Gouveia and Pires (2001). The extra variables incorporated in this formulation have been used by Sherali and Driscoll (2002) to further tighten the Linear Programming relaxation.

Comparisons have also been made for some formulations of the Symmetric TSP by Carr (1996) and Arthanari and Usha (2000). We unify all these results in the same framework.

In Sect. 4 we present computational results on a small illustrative example in order to verify the results of Sect. 3.

2 Eight Formulations of the ATSP

In all our formulations we will take the set of cities as $N = \{1, 2, ..., n\}$ and define variables

$$\begin{aligned} x_{ij} &= 1 \text{ iff arc } (i,j) \text{ is a link in the tour} \\ &= 0 \text{ otherwise} (i \neq j) \end{aligned}$$

c_{ij} will be taken as the length of arc(i, j)

The objective function will be:

$$\text{Minimise} \sum_{\substack{i,j \\ i \neq j}} c_{ij} x_{ij} \tag{1}$$

2.1 Conventional Formulation (C) (Dantzig, Fulkerson and Johnson (1954))

$$\sum_{\substack{j \\ j \neq i}} x_{ij} = 1 \qquad \forall\, i \in N \tag{2}$$

$$\sum_{\substack{i \\ i \neq j}} x_{ij} = 1 \qquad \forall\, j \in N \tag{3}$$

$$\sum_{\substack{i,j \in M \\ i \neq j}} x_{ij} \leq |M| - 1 \qquad \forall\, M \subset N \textit{ such that } \{1\} \notin M, |M| \geq 2 \tag{4}$$

(*the symbol '$\subset$' represents proper inclusion)*

This formulation has $2^{n-1} + n - 1$ constraints and $n(n-1)$ 0–1 variables.

The exponential number of constraints makes it impractical to solve directly. Hence, the usual procedure is to apply the Assignment constraints

(2) and (3) and append only those Subtour Elimination constraints (4) when violated. Alternatively, different relaxations such as the LP relaxation or the Spanning-2 Tree relaxation can be applied and solved iteratively. A reference to these methods is Lawler et al. (1995).

A variant of the above formulation (which we will not classify as a different formulation) is to replace constraints (4) by:

$$\sum_{\substack{i\in M\\ j\in \overline{M}}} x_{ij} \geq 1 \qquad \forall\ M \subset N \text{ where } \{1\} \notin M \text{ and } \overline{M} = N - M \tag{5}$$

Constraints (5) can be obtained by adding constraints (2) for $i \in M$ and subtracting from (4).

2.2 Sequential Formulation (S) (Miller, Tucker and Zemlin (1960))

Constraints (2) and (3) are retained but we introduce (continuous) variables

$$u_i = \text{sequence in which city } i \text{ is visited } (i \neq 1)$$

and constraints

$$u_i - u_j + nx_{ij} \leq n - 1 \qquad \forall\ i, j \in N - \{1\}, i \neq j \tag{6}$$

This formulation has $n^2 - n + 2$ constraints, $n(n-1)$ 0–1 variables and $(n-1)$ continuous variables.

2.3 Flow Based Formulations

Single Commodity Flow (**F1**) (Gavish and Graves (1978))
Constraints (2) and (3) are retained but we also introduce (continuous) variables:

$$y_{ij} = \text{'Flow' in an arc } (i,j) i \neq j$$

and constraints:

$$y_{ij} \leq (n-1)x_{ij} \qquad \forall\ i, j \in N, i \neq j \tag{7}$$

$$\sum_{\substack{j\\ j\neq 1}} y_{1j} = n - 1 \tag{8}$$

$$\sum_{\substack{i\\ i\neq j}} y_{ij} - \sum_{\substack{k\\ i\neq k}} y_{jk} = 1 \qquad \forall\ j \in N - \{1\} \tag{9}$$

Constraints (8) and (9) restrict $n-1$ units of a single commodity to flow into city 1 and 1 unit to flow out of each of the other cities. Flow can only take place in an arc if it exists by virtue of constraints (7).

It is possible to improve this formulation (**F1'**) by tightening constraints (7) for $i \neq 1$ to:

$$y_{ij} \leq (n-2)x_{ij} \qquad \forall\ i, j \in N - \{1\}, i \neq j \tag{10}$$

This relies on the observation that at most $n-2$ units can flow along any arc not out of city 1. We are not aware of any other authors having recognised this improvement.

This formulation has $n(n+2)$ constraints, $n(n-1)$ 0–1 variables and $n(n-1)$ continuous variables.

Two Commodity Flow (**F2**) (Finke, Claus and Gunn (1983))
Constraints (2) and (3) are retained but we also introduce (continuous) variables:

$$y_{ij} = \text{'Flow' of commodity 1 in arc } (i,j) i \neq j$$

$$z_{ij} = \text{'Flow' of commodity 2 in arc } (i,j) i \neq j$$

and constraints:

$$\sum_{\substack{j \\ j \neq 1}} (y_{1j} - y_{j1}) = n-1 \tag{11}$$

$$\sum_{j} (y_{ij} - y_{ji}) = 1 \qquad \forall\, i \in N - \{1\}, i \neq j \tag{12}$$

$$\sum_{\substack{j \\ j \neq 1}} (z_{1j} - z_{j1}) = -(n-1) \tag{13}$$

$$\sum_{j} (z_{ij} - z_{ji}) = -1 \qquad \forall\, i \in N - \{1\}, i \neq j \tag{14}$$

$$\sum_{j} (y_{ij} + z_{ij}) = n-1 \qquad \forall\, i \in N \tag{15}$$

$$y_{ij} + z_{ij} = (n-1)x_{ij} \qquad \forall\, i,j \in N \tag{16}$$

Constraints (11) and (12) force $(n-1)$ units of commodity 1 to flow in at city 1 and 1 unit to flow out at every other city. Constraints (13) and (14) force $(n-1)$ units of commodity 2 to flow out at city 1 and 1 unit to flow in at every other city. Constraints (15) force exactly $(n-1)$ units of combined commodity in each arc. Constraints (16) only allow flow in an arc if present.

This formulation has $n(n+4)$ constraints, $n(n-1)$ 0–1 variables and $2n(n-1)$ continuous variables.

Multi-Commodity Flow (**F3**) (Wong (1980) and Claus (1984))
Constraints (2) and (3) are retained but we also introduce (continuous) variables:

$$y_{ij}^{k} = \text{'Flow' of commodity } k \text{ in arc } (i,j)\ \kappa \in N - \{1\}$$

and constraints:

$$y_{ij}^k \leq x_{ij} \qquad \forall\, i,j,k \in N, k \neq 1 \tag{17}$$

$$\sum_i y_{1i}^k = 1 \qquad \forall\, k \in N - \{1\} \tag{18}$$

$$\sum_i y_{i1}^k = 0 \qquad \forall\, k \in N - \{1\} \tag{19}$$

$$\sum_i y_{ik}^k = 1 \qquad \forall\, k \in N - \{1\} \tag{20}$$

$$\sum_j y_{kj}^k = 0 \qquad \forall\, k \in N - \{1\} \tag{21}$$

$$\sum_i y_{ij}^k - \sum_i y_{ji}^k = 0 \qquad \forall\, j,k \in N - \{1\}, j \neq k \tag{22}$$

Constraints (17) only allow flow in an arc which is present. Constraints (18) force exactly one unit of each commodity to flow in at city 1 and constraints (19) prevent any commodity out at city 1. Constraints (20) force exactly one unit of commodity k to flow out at city k and constraints (21) prevent any of commodity k flowing in at city k. Constraints (22) force 'material' balance for all commodities at each city, apart from city 1 and for commodity k at city k.

This formulation has $n^3 + n^2 + 6n - 3$ constraints, $n(n-1)$ $0-1$ variables and $n(n-1)^2$ continuous variables.

2.4 Timed Staged Formulations

1st Stage Dependent **T1** (Fox, Gavish and Graves (1980))
In order to facilitate comparisons with the other formulations it is convenient, but not necessary, to retain the variables x_{ij} (linked to the other variables by constraints (25)) and constraints (2) and (3). We introduce 0–1 integer variables:

$$\begin{aligned} y_{ij}^t &= 1 \qquad \text{if arc } (i,j) \text{ is traversed at stage } t \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

and constraints:

$$\sum_{i,j,t} y_{ij}^t = n \tag{23}$$

$$\sum_{\substack{j,t \\ t \geq 2}} t y_{ij}^t - \sum_{k,t} t y_{ki}^t = 1 \qquad \forall\, i \in N - \{1\} \tag{24}$$

$$x_{ij} - \sum_t y_{ij}^t = 0 \qquad \forall\, i,j \in N, i \neq j \tag{25}$$

In addition we impose the conditions:

$$y_{il}^t = 0 \forall\ t \neq n, \quad y_{ij}^t = 0 \forall\ t \neq 1, \quad y_{ij}^l = 0 \forall\ i \neq 1, \quad i \neq j \tag{26}$$

Constraints (24) guarantee that if a city is entered at stage t it is left at stage $t+1$. Removing certain variables by conditions (26) forces city 1 to be left only at stage 1 and entered only at stage n.

It is not necessary to place upper bounds of 1 on the variables x_{ij}, and this condition may be violated in the LP relaxation.

This model has $n(n+2)$ constraints and $n(n-1)(n+1)$ 0–1 variables. Clearly, but for constraints (25) and variables x_{ij} this model would be even more compact having only n constraints and $n(n-1)$ variables. This is a remarkable formulation for this reason although, as will be shown in the next section it is also remarkably bad in terms of the strength of its Linear Programming relaxation and therefore the slowness of its overall running time.

2nd Stage Dependent **T2** (Fox, Gavish and Graves (1980))
We use the same variables as in **T1** and constraints (2), (3) and (25) together with:

$$\sum_{\substack{i,t \\ i \neq j}} y_{ij}^t = 1 \qquad \forall\ j \in N \tag{27}$$

$$\sum_{\substack{j,t \\ j \neq i}} y_{ij}^t = 1 \qquad \forall\ i \in N \tag{28}$$

$$\sum_{i, j \neq i} y_{ij}^t = 1 \qquad \forall\ t \in N \tag{29}$$

$$\sum_{\substack{j,t \\ t \geq 2}} t y_{ij}^t - \sum_{k,t} t y_{ki}^t = 1 \qquad \forall\ i \in N - \{1\} \tag{30}$$

Clearly this is a disaggregated form of **T1.**

This model has $4n-1$ constraints and $n(n-1)(n+1)$ 0–1 variables. Again but for the constraints (25) and variables x_{ij} this would be smaller. In fact the y_{ij}^t variables can, in this formulation, be regarded as continuous.

3rd Stage Dependent **T3** (Vajda (1961))
We use the same variables as in **T1** and **T2** and constraints (2), (3) and (25) together with:

$$\sum_j y_{1j}^1 = 1 \tag{31}$$

$$\sum_i y_{i1}^n = 1 \tag{32}$$

$$\sum_j y_{ij}^t - \sum_k y_{ki}^{t-1} = 0 \qquad \forall\ i, t \in N - \{1\} \tag{33}$$

Constraint (31) forces city 1 to be left at stage 1 and constraint (32) forces it to be entered at stage n. Constraints (33) have the same effect as (24).

This model has $2n^2 - n + 3$ constraints and $n(n-1)(n+1)$ 0–1 variables which again could be reduced by leaving out constraints (25) and variables x_{ij} . Again the y_{ij}^t variables can be regarded as continuous.

All the formulations, apart from **C**, have a polynomial (in n) number of constraints. This makes them superficially more attractive than **C**. However, the number of constraints may still be large, for practically sized n, and the LP relaxations weaker. These considerations are discussed in the next section.

3 Comparison of LP Formulations

All formulations presented in Sect. 2 can be expressed in the form:

$$\begin{aligned} &\text{Minimise} \quad \underline{c}.\underline{x} \\ &\text{subject to } A\underline{x} + B\underline{y} \sim b \quad \text{where '}\sim\text{' represents '}<=\text{' and '}=\text{' relations.} \\ &\qquad\qquad \underline{x}, \underline{y} \geq 0 \end{aligned} \tag{34}$$

$\underline{x}$ is the vector of variables $x_{i,j}$ and $\underline{y}$ the different vectors used in the formulations **S**, **F** and **T**. In the case of **S** and **F** $\underline{y}$ represents continuous variables but in the case of **T** integer variables.

In order to facilitate comparisons between the formulation **S**, **F** with **C** we can project out the continuous variables $\underline{y}$ to create a model involving only $\underline{x}$. The size of the polytopes of the associated LP relaxations can then be compared. We will denote the polytope of the resultant LP relaxation of a (projected) model **M** as **P**(**M**). In the case of formulation **T1** the variables $\underline{y}$ must be integer. The projection out of such variables is more complex and may not even result in an IP (see Kirby and Williams (1997)). However, we can still project out the variables $\underline{y}$ from the LP relaxation and return an IP. The LP relaxation of this IP will be weaker than that resulting from the true projection. It will still, however, be a valid comparator of computational difficulty when LP based IP methods are used. Therefore we will continue to use the notation **P**(**M**) for the resulting polytope when projecting out the LP relaxations of the variables $\underline{y}$ in **T1.**

In order to project out the variables $\underline{y}$ in all the formulations we can use Fourier-Motzkin elimination (see Williams (1986)) or equivalently full Benders Decomposition (1962). Martin (1999) gives a full general description of the methods of projection. We do not reproduce the derivation of the methods here but simply restate them. The projection out of the variables $\underline{y}$ is effected by finding all real vectors $\underline{w}$, of appropriate dimension, such that,

$$\underline{w}'B \geq \underline{0} \tag{35}$$

Where $\underline{w}$ has non-negative entries corresponding to rows of (34) with '$\leq$' constraints and unconstrained entries in rows with '=' constraints. The set of $\underline{w}$ satisfying (35) form a convex polyhedral cone and can be characterised by its extreme rays. It is therefore sufficient to seek the finite set of $\underline{w}$ representing extreme rays, which are what would be obtained by (restricted) Fourier-Motzkin elimination. We denote these as rows of the matrix Q. Applying Q to (34) gives:

$$QA\underline{x} \leq Q\underline{b} \tag{36}$$

as an alternative formulation to **C**. Of course, as would be expected, (36) will have an exponential number of constraints, unlike (34), but is in the same space as **C**.

We present the effect of the matrix Q for each of the formulations **S**, **F** and **T** of Sect. 2.

Formulation **S**

The effect of Q is to eliminate $u_2, u_3, ..., u_n$ from all the inequalities in (6). This is done by adding those inequalities around each directed cycle $M \subset N$, where $1 \notin M$. This results in inequalities (for each subset $M \subset N$ by virtue of (2) and (3))

$$x_{i1i2} + x_{i2i3} + \cdots + x_{i|M|i1} \leq |M| - \frac{|M|}{n} \tag{37}$$

(together with (2), (3) and non-negativity).

$$Clearly\, |M| - 1 < |M| - \frac{|M|}{n} \text{ since } M \subset N$$

Since cycles are subsets of their associated sets this demonstrates that

$$\mathbf{P}(\mathbf{S}) \supset \mathbf{P}(\mathbf{C}) \tag{38}$$

(strict inclusion can be proved by numerical examples).

Therefore the LP relaxation associated with **S** will be weaker than that associated with **C**. This result has already been obtained by Wong (1980) and Padberg and Sung (1991).

Formulation **F1**

The effect of Q is to, for each (subset) $M \subset N$, where $1 \notin M$, create

$$\sum_{i,j \in M} x_{ij} \leq |M| - \frac{|M|}{n-1} \tag{39}$$

$$\text{Clearly } |M| - 1 < |M| - \frac{|M|}{(n-1)} < |M| - \frac{|M|}{n}$$

demonstrating that

$$\mathbf{P}(\mathbf{S}) \supset \mathbf{P}(\mathbf{F1}) \supset \mathbf{P}(\mathbf{C}) \tag{40}$$

(Strict inclusion can again be proved by numerical examples).

This result is also obtained by Wong (1980).

Applying the same elimination procedure to the modified formulation (**F1'**) we obtain

$$\frac{1}{n-1} \sum_{\substack{i \in \overline{M} - \{1\} \\ j \in \overline{M}}} x_{ij} + \sum_{i,j \in \overline{M}} x_{ij} \leq |M| - \frac{|M|}{n-1} \tag{41}$$

Clearly, by virtue of (2) and (3), $\dfrac{1}{n-1} \displaystyle\sum_{\substack{i \in \overline{M} - \{1\} \\ j \in M}} x_{ij} \leq 1 - \dfrac{|M|}{n-1}$

$$\text{Hence } \mathbf{P}(\mathbf{F1}) \supset \mathbf{P}(\mathbf{F1}') \tag{42}$$

(Strict inclusion can again be proved by numerical examples).

Formulation **F2**

If z_{ij} are interpreted as the 'slack' variables in (16) we can use (16) to substitute them out reducing this formulation to **F1**. This demonstrates that

$$\mathbf{P}(\mathbf{F2}) = \mathbf{P}(\mathbf{F1}) \tag{43}$$

This result is also given by Langevin et al. (1990).

Formulation **F3**

The effect of Q is to, for each $M \subset N$,where $1 \notin M$, create

$$\sum_{i,j \in M} x_{ij} \leq |M| - 1 \tag{44}$$

i.e. constraints (4) of formulation **C**.

$$\text{Hence } \mathbf{P}(\mathbf{F3}) = \mathbf{P}(\mathbf{C}) \tag{45}$$

This remarkable result is also obtained by Wong (1980) and Padberg and Sung (1991)

Formulation **T1**

The effect of Q is to, for each $M \subset N$, where $1 \notin M$, create

$$\sum_{\substack{i \in M \\ j \in \overline{M}}} x_{ij} \geq \frac{|M|}{n} \tag{46}$$

and

$$\sum_{i,j \in N} x_{ij} = n \tag{47}$$

In the absence of assignment constraints, in this formulation, it is not possible to convert (46) to a form similar to (4). We therefore express it in a form similar to (5). Representing constraints (37) in a similar form to (5) demonstrates that

$$\mathbf{P}(\mathbf{S}) \subset \mathbf{P}(\mathbf{T1}) \tag{48}$$

Formulation **T2**
The effect of Q is to, for each subset M of $N - \{1\}$, create

$$\frac{1}{n-1} \sum_{\substack{i \in M \\ j \in \overline{M} - \{1\}}} x_{ij} + \frac{1}{n-1} \sum_{\substack{i \in \overline{M} - \{1\} \\ j \in M}} x_{ij} + \sum_{i,j \in M} x_{ij} \leq |M| - \frac{|M|}{n-1} \tag{49}$$

However, other constraints are also created which, to date, it has not been possible to obtain through the combinatorial explosion resulting from projection. Padberg and Sung give constraints equivalent to (49) as the projection of **T1.** This is clearly wrong.

$$\text{Hence } \mathbf{P}(\mathbf{T2}) \subset \mathbf{P}(\mathbf{F1'}) \tag{50}$$

Again strict inclusion can be proved by numerical examples.

Formulation **T3**
We have again not been able to discover the full effect of Q. However, one of the effects of the projection is to produce constraints (49) but there are others

$$\text{Hence } \mathbf{P}(\mathbf{C}) \subset \mathbf{P}(\mathbf{T3}) \subset \mathbf{P}(\mathbf{T2}) \tag{51}$$

Numerical examples demonstrate that the inclusion is strict.

4 Computational Results

In order to demonstrate the comparative sizes of different formulations and the relative strengths of their LP relaxations we give results below for a 10 city TSP.

These results were obtained using the NEWMAGIC modelling language and EMSOL optimiser.

Model	Size	LP Obj.	Iterations	Time (secs)	IP Obj.	Nodes	Time (secs)
C	502*90						
Conventional		766	37	1	766	0	1
	Ass.	804	40	1	804	0	1
	relaxation	835	43	1	835	0	1
	+ subtours (5)	878	48	1	881	9	1
	+ subtours (3)						
	+ subtours (2)						
S	92*99	773.6	77	3	881	665	16
Sequential							
F1	120*180	794.22	148	1	881	449	13
1 Commodity							
F1'	120*180	794.89	142	1	881	369	11
Modified							
F2	140*270	794.22	229	2	881	373	12
2 Commodity							
F3	857*900	878	1024	7	881	9	13
Multi Commodity							
T1	10*990	364.5	25	1	No solution after 12 hours		
1st Stage Dependent							
T2	120*990	799.46	246	18	881	2011	451
2nd Stage Dependent							
T3	193*990	804.5	307	5	881	145	27
3rd Stage Dependent							

5 Concluding Remarks

Eight formulations of the ATSP as an IP have been compared. Unlike other published work in this area the authors provide a unifying framework, in the form of projection, to conduct the comparison. Verification of the results are obtained through a numerical example.

The authors are now investigating, in the first instance, strategies for the manual introduction of the sub-tour elimination constraints with a view to developing a fully automated procedure. This work is being done using the NEWMAGIC modelling language.

Acknowledgement

The second author would like to acknowledge the help of EPSRC Grant EP/C530578/1in preparing the final version of this paper.

References

Arthanari, T.S. and M.Usha (2000) An alternate formulation of the symmetric travelling salesman problem and its properties, *Discrete Applied Mathematics*, **98**, 173–190.

Benders, J.F. (1962) Partitioning procedure for solving mixed-variable programming problems, *Numer. Math.*, **4**, 238–252.

Carr, R. (1996) Separating over classes of TSP inequalities defined by 0 node-lifting in polynomial time, preprint.

Claus, A. (1984) A new formulation for the travelling salesman problem, *SIAM J. Alg. Disc. Math.*, **5**, 21–25.

Dantzig, G.B., D.R. Fulkerson and S.M. Johnson (1954) Solutions of a large scale travelling salesman problem, *Ops. Res.*, **2**, 393–410.

Finke, G., A. Claus and E. Gunn (1983) A two-commodity network flow approach to the travelling salesman problem, *Combinatorics, Graph Theory and Computing, Proc. 14th South Eastern Conf.*, Atlantic University, Florida.

Fox, K.R., B. Gavish and S.C. Graves (1980) An n-constraint formulation of the (time-dependent) travelling salesman problem, *Ops. Res.*, **28**, 1018–1021.

Gavish, B. and S.C. Graves (1978) The travelling salesman problem and related problems, *Working Paper OR-078-78*, Operations Research Center, MIT, Cambridge, MA.

Gouveia, L. and S. Voss (1995) A classification of formulations for the (time-dependent) travelling salesman problem, *European Journal of OR*, **83**, 69–82.

Gouveia, L. and J.M. Pires (2001) The asymmetric travelling salesman problem: on generalisations of disaggregated Miller-Tucker-Zemlin constraints, *Discrete Applied Mathematics*, **112(1–3)**, 129–145.

Kirby, D and H.P. Williams (1997) Representing integral monoids by inequalities *J. Comb.Math. & Comb. Comp.*, **23**, 87–95.

Langevin, A., F. Soumis and J. Desrosiers (1990) Classification of travelling salesman formulations, *OR Letters*, **9**, 127–132.

Lawler, E.L., J.K. Lenstra, A.H.G. Rinnooy Kan and D.B. Shmoys (1995) (Eds.) *The Travelling Salesman Problem*, Wiley, Chichester.

Martin R.K. (1999) *Large Scale Linear and Integer Optimisation: A Unified Approach*, Kluwer, Boston.

Miller C.E., A.W. Tucker and R.A. Zemlin (1960) Integer programming formulation of travelling salesman problems, *J. ACM*, **3**, 326–329.

Padberg M. and T.-Y. Sung (1991) An analytical comparison of different formulations of the travelling salesman problem, *Math. Prog.*, **52**, 315–358.

Picard, J. and M. Queyranne (1978), The time-dependent travelling salesman problem and its application to the tardiness problem on one-machine scheduling, *Operations Research*, **26**, 86–110.

Sherali, H.D. and P.J. Driscoll (2002), On tightening the relaxations of Miller-Tucker-Zemlin formulations for asymmetric travelling salesman problems, *Operations Research*, **50(4)**, 656–669.

Vajda S., (1961) *Mathematical Programming*, Addison-Wesley, London.
Williams, H.P. (1986) Fourier's method of linear programming and its dual, *American Math. Monthly*, **93**, 681–695.
Wong, R.T. (1980) Integer programming formulations of the travelling salesman problem, *Proc. IEEE Conf. on Circuits and Computers*, 149–152.

Part II

Econometric Modelling and Prediction

Part II

Econometric Modelling and Prediction

The Threshold Accepting Optimisation Algorithm in Economics and Statistics*

Peter Winker[1] and Dietmar Maringer[2]

[1] Faculty of Economics, University of Giessen
[2] CCFEA, University of Essex

Summary. Threshold Accepting (TA) is a powerful optimisation heuristic from the class of evolutionary algorithms. Using several examples from economics, econometrics and statistics, the issues related to implementations of TA are discussed and demonstrated. A problem specific implementation involves the definition of a local structure on the search space, the analysis of the objective function and of constraints, if relevant, and the generation of a sequence of threshold values to be used in the acceptance-rejection-step of the algorithm. A routine approach towards setting these implementation specific details for TA is presented, which will be partially data driven. Furthermore, fine tuning of parameters and the cost and benefit of restart versions of stochastic optimisation heuristics will be discussed.

Key words: Heuristic optimisation, threshold accepting

1 Introduction

Threshold accepting is an optimisation heuristic. Reasonable features of such optimisation heuristics include the following (Barr et al., 1995, p. 12). Firstly, they should aim at good approximations to the global optimum. Secondly, they should be robust to changes in problem characteristics, tuning parameters and changes in the constraints. Thirdly, they should be easy to implement to many problem instances, including new ones. Finally, a necessary requirement is that the solution approach consists of a procedure which does not depend on individual subjective elements. We will try to demonstrate that a suitable implementation of threshold accepting fulfills these requirements.

Threshold accepting is a modification of the more often used simulated annealing (Kirkpatrick et al., 1983) using a deterministic acceptance

* We are indebted to Manfred Gilli, the editor and two anonymous referees for valuable comments on a preliminary draft of this contribution.

criterion instead of the probabilistic one in simulated annealing. It also belongs to the class of local search methods (Aarts and Lenstra, 1997, p. 2). A classification of optimisation heuristics can be found in Winker and Gilli (2004), and a more detailed description of the threshold accepting algorithm is provided by Winker (2001).

Classical or standard optimisation techniques such as Newton's method are mostly based on differential calculus and first order conditions. However, this strategy requires the search space Ω to be continuous and to have just one global optimum. Many of the problems arising in statistics and economics exhibit objective functions with several local optima or discontinuities. A classification of optimisation problems and some references to such cases are provided by Winker and Gilli (2004). Applied on these problems, classical optimisation techniques might report the local optimum next to the starting point – provided it was able to converge in the first place. It therefore seems adequate to extend the portfolio of optimisation techniques applied in these fields by optimisation heuristics. There are a large number of problems in economics and statistics, including Maximum Likelihood Estimations, GMM, numerical models in economics, e.g., for computable general equilibrium models or quantitative game theory (Judd, 1998, pp. 133ff and 187ff, respectively), which are documented, for which standard optimisation approaches may fail to provide solutions at all or would require tremendous amounts of computing resources. E.g., (Brooks et al., 2001) found that commonly used econometric software may fail for a rather simple maximum likelihood estimation for the parameters of a GARCH model whereas threshold accepting is capable of finding significantly better results as Maringer (2005) reports. Therefore, the question as to whether new optimisation paradigms could be useful in economics and statistics has to be answered by a clear–cut "yes".

During the last 15 years, threshold accepting has been successfully applied to many different problems ranging from classical operations research to economics and statistics. In fact, the algorithm has been introduced with an application to the famous traveling salesman problem by Dueck and Scheuer (1990). It appears that simulated annealing is still more widespread in its use, but there exist also a number of implementations of threshold accepting both in traditional operational research applications and for more specific problems from economics and statistics. The second implementation of threshold accepting, described by Dueck and Wirsching (1991), covers multi–constraint 0–1 knapsack problems and has been included in a comparative study by Hanafi et al. (1996). Some further early applications in the area of operational research are cited in the bibliography provided by Osman and Laporte (1996, p. 547). A more recent survey is provided in Winker (2001).

Although we do not aim at providing a complete overview on applications of threshold accepting in statistics and economics, some further fields of application seem noteworthy. Dueck and Winker (1992) have applied threshold accepting to portfolio optimisation for different risk measures, an approach taken up by Gilli and Këllezi (2002a, 2002b), recently. Winker (1995,

2000) introduces an application of threshold accepting to lag structure identification in VAR–models. Finally, threshold accepting has been applied with great success in the construction of low discrepancy experimental designs. First, Winker and Fang (1997a) obtain lower bounds for the star–discrepancy and Winker and Fang (1997b) use the approach to obtain low discrepancy U–type designs for the star–discrepancy. Next, Fang et al. (2000) extend the analysis to several modifications of the L_2–discrepancy, while Fang et al. (2002), Fang et al. (2003), and Fang et al. (2005) consider the centered and wrap–around L_2–discrepancy allowing to obtain lower bounds for the objective function (see subsection 2.3).

We will mention some further applications in the text when they are used as examples to demonstrate specific settings and approaches for a successful implementation of threshold accepting.

1.1 Basic Features of Threshold Accepting

Much akin to its ancestor simulated annealing, threshold accepting is a typical local search heuristic that iteratively suggests slight random modifications to the current solution and by doing so gradually moves through the search space. TA is therefore well suited for problems where the solution space has a local structure and a notion of neighborhood around solutions can be introduced.

The second crucial property TA shares with simulated annealing (and most other heuristic search strategies) is that not only modifications for the better are accepted, but also for the worse in order to escape local optima. However, while simulated annealing uses a probabilistic criterion to decide whether to accept or reject a suggested "uphill move", TA has the deterministic criterion of a threshold value for impairments: Whenever the suggested modification improves the objective function or its degradation does not exceed a given threshold value, this modification is accepted; if the modification would degrade the objective function by more than the threshold, it is rejected. This threshold is not kept fixed in the course of iterations, but forms a "threshold sequence" which usually makes the criterion rather tolerant in early iterations and increasingly restrictive in the later iterations. By this strategy, the algorithm can be shown to converge asymptotically to the global optimum (Althöfer and Koschnik, 1991).

1.2 Pseudo Code

Algorithm 1 provides the pseudo–code for a prototype threshold accepting implementation for a minimization problem.

Thereby, f represents the objective function, which has to be minimized over the search space Ω. Of course, by replacing f with $-f$, the algorithm can also be applied to maximization problems.

Threshold accepting performs a refined local search on the search space Ω. It starts with a (randomly) generated feasible solution x^c (2:) and continues

Algorithm 1 Pseudo–code for Threshold Accepting

```
1:  Initialize n_R, n_{S_r} and the sequence of thresholds τ_r, r = 1, 2, ..., n_R
2:  Choose (randomly) feasible solution x^c ∈ Ω
3:  for r = 1 to n_R do
4:     for i = 1 to n_{S_r} do
5:        Choose (randomly) neighbor x^n of x^c
6:        if Δf = f(x^n) − f(x^c) < τ_r then
7:           x^c = x^n
8:        else
9:           leave x^c unchanged
10:       end if
11:    end for
12: end for
```

by iterating local search steps. For each step, a new candidate solution x^n has to be chosen in the neighborhood of the current solution x^c (5:). Then, the value of the objective function of both candidate solutions is compared (6:). The new candidate solution is accepted if it is better than x^c, but also if it is not much worse. The extent of an accepted worsening is limited by the current value of the threshold sequence (τ_r), which decreases to zero during the course of iterations.

The performance of the threshold accepting implementation depends on a number of settings. In particular, the definition of neighborhoods for the choice of x^n, the sequence of threshold values τ_r and, finally, the total number of iterations are most relevant. We will come back to all of these factors in the following sections of this contribution.

1.3 The Basic Ingredients

Approaching an optimisation problem with TA demands two basic types of ingredients: ones that characterize the problem, and those that are needed for the heuristic search. The former group usually covers a proper problem statement by giving the decision variables, x, and the search space, Ω, the constraints the decision variables must meet, and the objective function, $f(x)$. Section 2 will present several relevant aspects in this respect.

For the TA implementation, the first basic ingredient is a concept of the local structure or the neighborhood of the current solution, $\mathcal{N}(x^c)$, within which new solutions are generated. What makes a suitable neighborhood depends on the optimisation problem. Nonetheless there are some general requirements and approaches; Section 3 addresses these issues. The second crucial ingredient is the design of the acceptance criterion. Section 4 describes how to find an appropriate threshold sequence and related issues. The third

crucial ingredient for a TA is the decision of how to use the available computational time by setting the number of iterations per run and the total number of runs on the one hand and detecting when to halt a run and restart the search process on the other hand; Section 5 has more details on this issue.

2 Objective Function and Constraints

2.1 Objective Function

Obviously, the objective function f and the search space Ω are problem specific. Given that threshold accepting is an iterative local search procedure, it does not require the objective function f to be smooth or even differentiable. However, it has to evaluate f for many different elements $x \in \Omega$. Therefore, the efficiency of the algorithm will depend heavily on the fast calculation of $f(x)$ for any given $x \in \Omega$. Furthermore, if $f(x)$ cannot be calculated exactly, the quality of any approximation has to be taken into account.

This statement appears to be trivial for any optimisation problem. However, in practice it is not. We will discuss two issues related to the objective function. First, although the objective function might be calculable in principle, the cost for doing so in terms of computational load can be quite high. Local updating, considered in more detail in subsection 3.2, often provides a remarkable speed up. The idea of local updating stems from the observation that if $x^n \in \mathcal{N}(x^c)$ is a neighbor of x^c, it is quite similar. Consequently, the objective function value for x^n could be similar to $f(x^c)$ as well. If it is possible to evaluate the difference directly, a tremendous speed up can result. For example, instead of recalculating a complete tour for the traveling salesman problem, if x^n and x^c differ only by the ordering of a few cities, it is possible to calculate directly the difference in tour length resulting from these few differences. We will come back to this idea in subsection 3.2 as it is closely linked to the definition of local neighborhoods.

Second, there exist applications where the objective function itself cannot be easily evaluated. For example, in uniform design a given number of points has to be found in a discrete multi-dimensional space such that these points are as uniformly distributed as possible. A classical measure of the quality of such designs, i.e., the uniformity of these points, is the so called "star–discrepancy" which is described, e.g., in Winker and Fang (1997a). However, in order to evaluate this measure, a complex combinatorial problem has to be solved. Consequently, Winker and Fang (1997a) proposed to use threshold accepting to obtain a lower bound for this objective function. Now, if one would be interested in obtaining a low discrepancy design under the star–discrepancy, each evaluation of the objective function would correspond to a run of the threshold accepting heuristic itself. Fortunately, for this problem other measures of discrepancy have been developed which are much easier to compute. A similar problem comes up in the context of simulation models.

If the value of the objective function is obtained by running a simulation model, again the computational complexity of the algorithm becomes quite substantial. In addition, the value of the objective function provided by the simulation will include some Monte Carlo variance. Gilli and Winker (2003) discuss how this Monte Carlo variance can be taken into account in a threshold accepting implementation.

2.2 Constraints

In most applications, the search space Ω is not a standard space like $\{0,1\}^k$ or $\mathbb{R}^k$, but only a subset of such a space resulting from some explicit constraints. If there is a large number of different constraints, this subspace might be not connected or it might prove difficult to generate elements in Ω. Also, the step of selecting $x^n \in \mathcal{N}(x^c)$ can become quite time consuming. Furthermore, the algorithm risks to get stuck in some part of the search space where no good solution can be found.

In these cases, a superior approach consists in considering the whole space $\{0,1\}^k$ or $\mathbb{R}^k$ as search space and to add a penalty term to the objective function if $x^c \notin \Omega$. If the penalty term is set at a very high level from the very beginning as sometimes suggested, this approach will just mimic the standard case, i.e., will face the same difficulties. Thus, it appears to be reasonable to start with a small penalty in order to enable the algorithm to access different parts of the search space. While the algorithm proceeds, the penalty term has to increase in order to make sure that the final solution obtained by the algorithm will be a feasible one.

2.3 Lower Bounds

Optimisation heuristics like threshold accepting often provide high quality approximations to the global optimum for a given problem instance. However, given that the procedure is stochastic and convergence to the global optimum can only be expected asymptotically (see Sect. 4), missing information about the quality of an actually found solution is often considered to be a major drawback of optimisation heuristics. Of course, optimisation heuristics share this potential drawback with classical optimisation approaches. If, for example, a numerical procedure detects a solution of a maximum likelihood problem, this solution is determined by the first order condition. However, this condition does not guarantee a global optimum unless the function is globally convex which is rather a rare exception than the rule.

In particular for combinatorial optimisation problems, lower bounds might provide a helpful tool in this context. For some problems it is possible to derive minimum values of the objective function for each instance without calculating an optimum solution. Provided that such a lower bound exists, any solution obtained by threshold accepting can be compared with this value. If the lower bound is met, the current solution is a global optimum. In this case,

a further analysis of the problem instance is only required if one expects to have multiple global optima and one is interested in identifying the optimizing set instead of just a single optimum solution. If the lower bound is not met, the difference of the objective function to this lower bound provides an indicator of the maximum improvement which might be obtained by further runs of the algorithm. However, the existence of a lower bound does not imply that this lower bound can actually be reached. For the traveling salesman problem, e.g., a trivial lower bound for the round trip is the sum of the distances to the closest neighboring point for each point of the problem set. Obviously, no tour can be shorter than this sum, but in general, it has to be much longer.

Fang et al. (2003, 2005) provide theoretical lower bounds for some instances of the uniform design problem. Consequently, it is possible to prove that some of the designs obtained by threshold accepting represent global optima, while others differ to a small extent from the lower bounds. Again, it is not guaranteed that the lower bounds can be reached at all. Nevertheless, it has been shown that the designs obtained by the threshold accepting heuristic are not farther from a global optimum than a few percentage points in terms of the objective function.

To sum up this argument, although it is still a rare situation to have access to theoretical lower bounds for optimisation problems arising in economics and statistics, such results are extremely helpful for evaluating the quality of the results obtained by optimisation heuristics. Consequently, some effort should be devoted to the generation of lower bounds.

3 Local Structure and Updating

3.1 Neighborhoods

As the classification as a *local* search heuristic suggests, threshold accepting requires some notion of closeness or neighborhood for all elements of the search space Ω. For this purpose, for each element $x \in \Omega$ a neighborhood $\mathcal{N}(x) \subset \Omega$ has to be defined. Of course, given the typical size of Ω, this assignment of neighborhoods cannot be done element by element, but has to follow some algorithmic approach. Furthermore, in each iteration, an element x^n in the neighborhood of the current solution x^c has to be generated. Thus, the neighborhoods have to be constructed in a way which makes the search or construction of such neighboring elements a simple task in terms of computational complexity.

While for some of the classical combinatoric optimisation problems like the traveling salesman problem there exist well–known standard concepts for constructing solutions which are neighbors to a current solution, this is not the case for most of the new optimisation problems studied in economics and statistics during the last decade. However, most of these problems allow for the application of a general concept given that the search space Ω is either

a subset of some real valued vector space $\mathbb{R}^k$ or of a discrete search space $\{0,1\}^k$. For these instances, ε–spheres provide a well–known concept of neighborhood on the vector space corresponding to a notion of distance provided by the Euclidean and Hamming metric (Hamming, 1950), respectively. Given a current solution x^c, a new element is considered a neighbor if the distance between both elements is smaller than ε for the given distance measure. This concept is easily transferred to the subspace Ω: x^n is a neighbor to $x^c \in \Omega$ if it satisfies the distance condition and is an element of Ω itself (Winker, 2001, pp. 117ff). Thus, we define

$$\mathcal{N}(x^c) = \{x^n | x^n \in \Omega, \parallel x^n - x^c \parallel < \varepsilon\}, \tag{1}$$

where $\parallel \cdot \parallel$ stands for the distance measure.

Although being a quite general approach, the proposed construction of neighborhoods by projection of ε–spheres onto Ω will not always work. In particular, one has to check whether the resulting neighborhoods are non–trivial, i.e., contain more than a single element for reasonable choices of ε. Furthermore, the objective function should exhibit local behavior with regard to the chosen neighborhoods, i.e., for the elements in $\mathcal{N}(x^n)$, the mean value of the objective function should be closer to $f(x^c)$ than for randomly selected elements in Ω. Both requirements result in a trade–off between large neighborhoods, which guarantee non–trivial projections, and small neighborhoods coming together with a real local behavior of the objective function.

A further argument with regard to the choice of (the size of) neighborhoods is closely connected to the features of the algorithm itself. While larger neighborhoods allow for fast movements through the search space, they also increase the peril that a global optimum is simply stepped over. Smaller neighborhoods, on the other hand, increase the number of iterations required to trespass a certain distance, e.g., in order to escape a local optimum: To escape a local optimum, a sequence of (interim) impairments of the objective function has to be accepted; the smaller the neighborhoods are, the longer this sequence is. Consequently, for smaller neighborhoods, the threshold sequence has to be more tolerant in order to be able to escape local optima.

In order to illustrate the idea of generating local neighborhoods, we consider the example of optimal aggregation of time series discussed by Chipman and Winker (2005). The authors analyze the aggregation of time series which is considered to be a central but still mainly unsolved problem in econometrics. In the specific setting considered in their paper, namely the international transmission of prices, aggregation boils down to the forming of groups of commodities and replacing the disaggregate time series by sums or weighted averages of the variables in each group. If one is interested in choosing the modes of aggregation, i.e., the composition of the groups, optimally with regard to a measure of mean–square forecast error, a highly complex integer optimisation problem results. In fact, it has been shown that this problem is NP–complete (Winker, 2001, Chap. 13.9). Thus, it appears adequate to tackle the problem with an optimisation heuristic like threshold accepting.

For this example, the search space is given by the set of all proper grouping matrices, which is a subspace of $\{0,1\}^{6\times 42}$ for the actual application. Thus, although the search space is finite, an exact solution by means of enumeration is not possible. The objective function is a measure of the aggregation bias in forecasting which results from using the model aggregated to only six aggregate groups as suggested by the official statistics as compared to the disaggregate data for 42 commodities. Unfortunately, the evaluation of the objective function requires some time consuming matrix inversion. Consequently, the number of iterations for this application has to be much smaller than for some of the other applications of threshold accepting mentioned.

Given that the search space is defined as a subspace of some $\{0,1\}^k$, the neighborhood concept is based on the projection of ε–spheres with regard to the Hamming distance. In this example, the Hamming distance d_H between two grouping matrices $H = (h_{ij})$ and $\tilde{H} = (\tilde{h}_{ij})$ is given by the number of differing entries:

$$d_H(H,\tilde{H}) = \sum_{i=1}^{m}\sum_{j=1}^{m^*} \mid h_{ij} - \tilde{h}_{ij} \mid \ . \tag{2}$$

Figure 1 shows the histogram of relative local differences of the objective function for three different neighborhood definitions, each based on 50 000 pairs of proper grouping matrices (H_k^1, H_k^2). For the trivial neighborhood represented by the top most panel, H_k^2 is randomly generated. This corresponds to setting $\varepsilon \to \infty$. The large dispersion of these relative deviations indicates that the probability of finding an acceptable new grouping in such a neighborhood is rather small unless the acceptance criterion becomes very loose, since no really local structure is imposed.

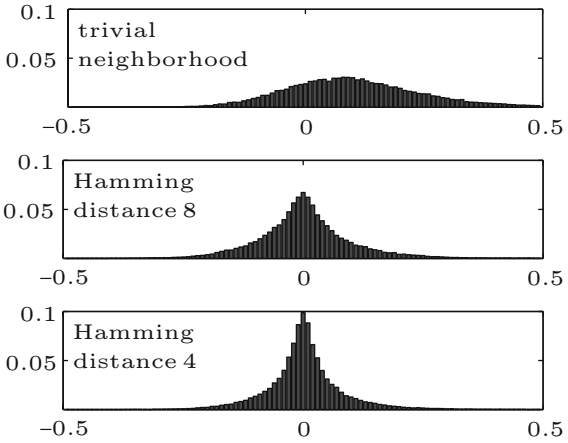

Fig. 1. Local differences for different neighborhood definitions

In contrast, the lower two panels provide histograms for a Hamming distance of 4 and 8, respectively. In these cases, H_k^2 is selected randomly from $\mathcal{N}(H_k^1)$. Comparing the two lower panels it is worth noting that a shrinking of the neighborhoods leads to a concentration of the empirical distribution of relative deviations around zero per cent, i.e., to a more locally oriented behavior of the algorithm, but at the same time reduces the number of feasible moves in each iteration. Consequently, the risk of being stuck in a local minimum increases with shrinking neighborhoods. In the application presented by Chipman and Winker (2005), the use of neighborhoods defined as spheres of radius 8 with regard to the Hamming distance proved to be a good choice, although the quality of the results did not decrease dramatically when choosing spheres of radii 4 or 12 instead.

3.2 Local Updating

In the standard version of the algorithm described so far, in each iteration, x^c and $f(x^c)$ are given. Then, an element $x^n \in \mathcal{N}(x^n)$ is generated. Finally, $f(x^n)$ is calculated in order to obtain the difference $\Delta = f(x^n) - f(x^c)$ required to test the acceptance criterion. Consequently, the total complexity of a TA implementation is given by a constant times the total number of iterations times the complexity of a single evaluation of the objective function f. While the constant might be influenced by an efficient coding of the algorithm and the choice of appropriate hardware, the total number of iterations will typically depend on the complexity of the problem at hand. Of course, a reasonable choice of neighborhoods and the threshold sequence (see the following section) might help to reduce the number of iterations required to obtain a predefined quality of the results. Here, we will concentrate on the last argument in the complexity function of the algorithm, the evaluation of the objective function (however, see also Ferrall (2004)).

At first glance, the performance of the algorithm with respect to the objective function depends solely on an efficient calculation of the objective function for given x. In fact, often considerable performance gains can be obtained by searching for more efficient code for calculating the objective function. However, sometimes a different approach is also feasible. During each iteration step, the algorithm does not require $f(x^n)$ and $f(x^c)$, but solely the difference of the objective function values $\Delta = f(x^n) - f(x^c)$. Given x^n and x^c, in some cases, a direct calculation of Δ becomes possible at much lower computational cost than the complete evaluation of f.

For example, for the traveling salesman problem, each element $x \in \Omega$ represents a tour through all N cities of the given problem. In order to calculate the length of such a tour, the sum of N distances of pairs of cities has to be calculated. If the problem is small enough, all distance pairs ($N(N-1)$) can be calculated once and for all before starting the algorithm, for larger problem instances, they have to be calculated on the fly. A typical definition of neighborhood for traveling salesman tours consists in assuming that two

tours are neighbors if the second one can be obtained from the first one by exchanging the position of two cities. Doing so, only the distances for four pairs of cities change. The rest of the tour remains unchanged. Thus, instead of calculating the sum of N distances, we have to consider only four in order to obtain Δ. The speed up resulting from this local updating idea amounts to $4/N$, i.e., it becomes the more important the larger the problem instance grows.

A similar idea is used by Fang et al. (2003) in an application to uniform design problems. Making use of a new representation of the objective function, they can avoid to recalculate the whole objective function when moving from one candidate solution to a neighboring one. When moving from one solution to a neighboring one, only two elements of the design matrix are exchanged. The updating requires $2(n-2)$ updates, where n denotes the number of design points, i.e., represents a measure of the problem size. In contrast, a complete evaluation would require $\frac{n(n-1)}{2}k$ comparisons, where k denotes the number of columns of the design matrix. Thus, even if up to four or even slightly more elements are exchanged in a single iteration, a tremendous speed up results which is proportional to $\frac{1}{nk}$, i.e., the larger the design under consideration, the higher the efficiency gain. For the implementation presented in Fang et al. (2003), the actual speed up resulting from the local updating idea ranges from around 50% for rather small problem instances ($n = 8$, $k = 10$), increasing to 80% for $n = 18$ and $k = 30$ and reaching more than 90% for $n = 100$ and $k = 8$.

4 Threshold Sequence

The final crucial ingredient of any threshold accepting implementation is the threshold sequence. By considering two extreme cases, a first intuitive idea of its influence might be gained. First, if all threshold values are set equal to zero, in each iteration, the algorithm will only accept new solutions which are at least as good as the current one. Consequently, a threshold accepting implementation with a zero threshold sequence would perform like a classical greedy local search algorithm. In general, it would converge much faster than with positive threshold values, but will get stuck in a local minimum with high probability unless the problem is globally convex. Second, if all values of the threshold sequence are set to a very large value which happens to be larger than any possible difference of objective function values, the algorithm will act like a random walk through the search space as any generated candidate solution will be accepted. The performance of this degenerated threshold accepting implementation will be similar to a pure random search heuristic.

Obviously, an intermediate setting is selected for any reasonable threshold accepting. Unfortunately, not much is known about how to choose this sequence in a way to improve the performance of the algorithm. The convergence result for threshold accepting provided by Althöfer and Koschnik

(1991) states only the existence of an appropriate threshold sequence in order to obtain asymptotic convergence to the global optimum, but it does not provide any insights into the structure of the sequence. Consequently, the threshold sequence is often chosen in a rather ad hoc approach. Thereby, a linearly decreasing sequence appears to be preferred. The advantage of a linear threshold sequence consists in the fact, that for tuning purposes only the first value of the sequence has to be varied as it fixes the whole sequence. Existing experience with different functional forms for the threshold sequence suggests that the performance of the algorithm is quite robust with regard to the exact shape of the threshold sequence, while the size of the first threshold values has some impact. In fact, starting with too high threshold values makes the algorithm wandering around in the search space in a rather random fashion. In this case, computational resources are wasted. On the other hand, starting with a too small value for the threshold sequence, one risks to get stuck in a less favorable part of the search space. This trade–off has to be considered when conducting some tuning experiments with a threshold accepting implementation.

For discrete search spaces, a data driven method for the construction of the threshold sequence has been proposed by Winker and Fang (1997a). It is described in more detail in Winker (2001, p. 127f). It is based on the observation that for a finite (discrete) search space Ω, the set Δ of possible Δf is also finite (discrete). Obviously, for the algorithm, only the values of this set are relevant for the threshold sequence, as any value between two elements of the ordered set of possible Δf will have the same effect in the acceptance criterion. Although the size of Ω and, consequently, of Δ will exclude a complete evaluation even in the case of a finite search space, an empirical approximation to Δ can be obtained as follows. First, a large number of candidate solutions x_r^c is generated at random. Then, for each of these random designs a neighbor x_r^n is selected using the same neighborhood definition as for the optimisation procedure. For each resulting pair of designs, the difference of the objective function values between the larger and the smaller value is calculated $\Delta_r = |f(x_r^c) - f(x_r^n)|$. Ordering these values provides an approximation to the distribution of local relative changes of the objective function. Finally, taking into account the trade–off between too large or too small values of the threshold sequence at the beginning of the optimisation run, only a lower quantile of these sequence is actually employed as the threshold sequence. Typically, this lower quantile falls in the range of 10% to 50% for the applications considered in this contribution. Algorithm 2 provides the pseudo–code for this data driven generation of the threshold sequence.

Before describing the construction of the threshold sequence for the example of the optimal aggregation of time series, a last possible modification is introduced. Instead of using an absolute definition of the thresholds, a relative version can be employed. Consequently, the decision criterion becomes $f(x^n) < f(x^c)(1 + \tau_r)$ instead of $\Delta f = f(x^n) - f(x^c) < \tau_r$. The data driven construction of a threshold sequence can be performed as before by replacing

Algorithm 2 Pseudo–code for data driven generation of threshold sequence

1: Initialize n_R, lower quantile α, $n_D[n_R/\alpha]$
2: **for** $r = 1$ to n_D **do**
3: Choose (randomly) feasible solution x_r^c
4: Choose (randomly) neighbor solution $x_r^n \in \mathcal{N}(x_r^c)$
5: Calculate $\Delta_r = |f(x_r^c) - f(x_r^n)|$
6: **end for**
7: Sort $\Delta_1 \leq \Delta_2 \leq \ldots \leq \Delta_{n_D}$
8: Use $\Delta_{n_R}, \ldots, \Delta_1$ as threshold sequence

$\Delta_r = |f(x_r^c) - f(x_r^n)|$, e.g., by $\Delta_r = \mid f(x_r^c)/f(x_r^n) - 1 \mid$. The first advantage of this relative version of the threshold criterion is its independence from units of measurement. However, when employing the data driven method for constructing the threshold sequence, this advantage appears rather trivial. The second argument for employing the relative version comes into play, when the objective function takes values of widely differing orders of magnitude. Then, the threshold criterion exhibits some automatic scaling property. However, so far, there is no clear evidence for a superior performance of one or the other version of the criterion. This has to be left to future research.

As an example for the data driven generation of the threshold sequence, we refer again to the example of optimal aggregation introduced in subsection 3.1 (Chipman and Winker, 2005). The neighborhood definition uses the concept of the Hamming distance introduced before. As a final ingredient, a data generated threshold sequence is used which is obtained along the lines described in this section making use of the relative definition of the threshold criterion and a lower quantile $\alpha \in [0.3, 0.4]$.

Figure 2 shows a threshold sequence obtained by the data driven method for this application. The final values of the threshold sequence are equal to 1 since some of the simulated pairs of grouping matrices happen to belong to the same equivalence class of grouping matrices. Consequently, the threshold accepting algorithm degenerates to a classical local search heuristic during the last iterations of the algorithm. This feature of the automatically generated threshold sequence increases the probability to finish with a local optimum.

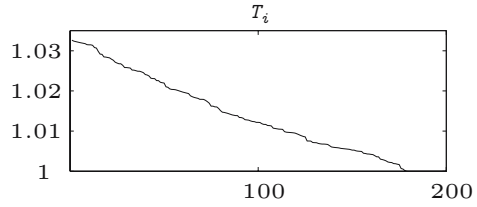

Fig. 2. A threshold sequence

Given the convergence result, this local optimum should be close to the global optimum and converge to it as the number of iterations of the algorithm tends to infinity.

5 Restart

Although it is not reported in many publications, most applications of optimisation heuristics use a restarting framework, i.e., the algorithm is rerun with different seeds for the random number generator or with different tuning parameters. Then, the presented results stem from the run with best performance. Some theoretical arguments on restart implementations can be found in Fox (1994). A heuristic argument in the context of genetic algorithms is provided by Farley and Jones (1994).

In this section, we will present some rational for this approach in the context of TA. However, it will turn out that it is essential that publications report the restarting framework and provide additional information besides the "best" result. In fact, stochastic search heuristics like threshold accepting can be interpreted as a stochastic mapping

$$TA : \Omega \rightarrow f_{min},\ f_{min} \sim D_{TA}(\mu, \sigma)\,, \tag{3}$$

where Ω is the search space and f_{min} the *random* realization of the minimum found by the algorithm for the given random number sequence. $D_{TA}(\mu, \sigma)$ denotes the distribution of f_{min} given the parameters used in the algorithm. Of course, this distribution is truncated from the left at the value of the global minimum $f_{min}^{glob} = \inf\{f(x)|x \in \Omega\}$. Consequently, D_{TA} will not be a normal distribution. It might be an interesting subject for future studies to analyze the properties of this distribution for different applications and different optimisation heuristics.

However, for practical purposes it might suffice to know about the existence of such a distribution. Furthermore, it might be obvious that an increase in the total number of iterations of a local search heuristic like threshold accepting should reduce the expected value μ of the distribution and – due to the left truncation – probably the standard deviation σ, too. Instead of using a parametric distributional assumption, we might use the empirical distribution obtained from a simulation study. For this purpose, the threshold accepting implementation is run several times with differing initializations (seeds) of the random number generator. For each run i, $i = 1, \ldots, N$, the minimum f_{min}^i is stored. Using the set $\{f_{min}^i | i = 1, \ldots, N\}$, it is possible to calculate the empirical mean and standard deviation for the best solution found by the algorithm or to provide empirical quantiles. It is also possible to report the minimum $\min\{f_{min}^i | i = 1, \ldots, N\}$ of all runs. In fact, this is the typical value provided in publications – sometimes accompanied by the number of runs N. However, from the interpretation of TA as a stochastic mapping, it becomes

evident that this value is not a robust statistic. Thus, it should not be the only information provided.

We recommend to provide at least the following information: The number of restarts N, the empirical mean and standard deviation or – alternatively – some quantiles of the empirical distribution. Furthermore, it should be reported for which parameter settings of the algorithm restarting has been considered.

Given these arguments about restarting, a further question has to be considered. Obviously, performing N restarts uses valuable computational resources. Instead, a smaller number of restarts – or a single run – with more iterations could be performed. Using more restarts provides a better approximation to the underlying distribution $D_{TA}(\mu_1, \sigma_1)$ for the given number of iterations. On the other hand, using less restarts and more iterations results in an approximation of lower quality to a different distribution $D_{TA}(\mu_2, \sigma_2)$ with a smaller expectation $\mu_2 < \mu_1$. In fact, given the convergence property of threshold accepting, D_{TA} will degenerate to a one point distribution in the global minimum with the number of iterations going to infinity. Unfortunately, not much is known about the rate of this convergence. Thus, it remains an empirical issue to decide about this trade–off.

To conclude this section, we present empirical findings for an implementation of threshold accepting to the well known traveling salesman problem. The problem instance with 442 points is described in more detail in Winker (2001, Chap. 8). For the analysis of a restarting situation, the following experimental setting is chosen. The threshold sequence is fixed for all runs to the same linear sequence. Then, the threshold accepting implementation is run with 100 000, 1 000 000 and 10 000 000 iterations. Obviously, in terms of computational resources, one run with 10 000 000 iterations corresponds to 10 runs with 1 000 000 or 100 runs with 100 000 iterations. In order to obtain good estimates of the lower percentiles, 100 runs were performed with the largest number of iterations. Consequently, the number of restarts with different random starting configurations was 1000 and 10 000 for 1 000 000 and 100 000 iterations, respectively. Table 1 summarizes the results obtained for a fixed threshold sequence, which is identical for all runs and numbers of iterations.

When considering these results, it turns out that the trade–off between more restarts with different seeds and a higher number of iterations per restart is in favour of the latter. As expected, both the mean and the lower percentiles become smaller as the number of iterations per run increases while holding the total use of computer resources (number of restart times number of iterations) constant. Given that users are typically interested in the very low percentiles of the distribution, it should be taken into account that the 1%–quantile for the runs with 10 000 000 iterations is estimated based on solely 100 observations. Consequently, this entry has to be interpreted with some care as it is estimated with less precision than other entries of the table.

The above analysis gives valuable information on the dependence of mean and percentiles on the number of iterations and restarts. However, it has not

Table 1. Restart threshold accepting

	Iterations per try		
	100 000	1 000 000	10 000 000
Restarts	10 000	1000	100
Mean	5317.07	5170.52	5138.22
SD	52.83	28.69	21.81
10%	5251.11	5135.45	5112.22
5%	5234.50	5124.83	5107.41
1%	5204.20	5109.90	5098.23

yet answered the practitioner's question whether it is preferable to perform a single run with a very high number of iterations, a few restarts with a moderate number of iterations, or many restarts with a low number of iterations. The outcomes of the experiment can also be used for this purpose. To this end, the results are grouped to 100 artificial sub–experiments each consisting of 100 tries with 100 000 iterations, 10 tries with 1 000 000 iterations and one try with 10 000 000 iterations. Now, only the overall best results obtained for each number of iterations are compared. Of course, considering this non–robust statistic is not a valid approach from a pure statistical perspective. However, it corresponds closely to the typical proceeding in real applications. Table 2 shows the number of times, the respective combinations of iterations per restart and number of restarts gave the best result. In the lower part the mean deviation from the best results for the three categories from the best results of these three categories are depicted.

If one has only computer resources for a total of 10 000 000 iterations, these results clearly indicate that one should neither perform many restarts with a very low number of iterations, nor only one huge run. Instead, the best expected performance is given by a choice which falls between the two extremes, i.e., restarting the threshold accepting a few times with a moderate number of iterations. For recent results for a uniform design problem see also Winker (2005).

Summarizing the arguments of this section, two aspects seem noteworthy. Firstly, the common practice to report only the best outcome of several restarts is not adequate. A minimum requirement is to report also the number

Table 2. Restart threshold accepting (comparison)

Iterations per try	100 000	1 000 000	10 000 000
Restarts	100	10	1
Times best in 100	0	65	35
Mean deviation from best	73.56	5.16	14.48

of restarts and some information on the empirical distribution D_{TA}, e.g., mean and standard deviations or quantiles. Secondly, the results indicate that mean value, standard deviation and low quantiles decrease, other things being equal, with an increasing number of iterations. Nevertheless, the combination of some restarts with a moderate number of iterations seems to be preferable in order to obtain high–quality results.

6 Conclusions

The concept of threshold accepting appears quite simple and yet powerful in its applications. By replacing a stochastic acceptance criterion by a deterministic one, it even reduces the complexity as compared to simulated annealing. Nevertheless, a more detailed presentation and discussion of the central ingredients of the algorithm highlights some aspects relevant for a successful implementation.

TA can be used to tackle highly complex optimisation problems which are not accessible by classical optimisation algorithms. The cost of implementation are small compared to more refined optimisation heuristics, e.g., population based approaches, or to tailor–made problem specific heuristics. The guidelines provided in this contribution might be helpful for the choice of settings and parameters resulting in a high quality outcome. Then, even if the TA implementation might not provide the global optimum, the best results obtained by this heuristic represent a benchmark which has to be beaten first by any potential challenger.

References

Aarts, E. H. L. and J. K. Lenstra: 1997, 'Local Search in Combinatorial Optimisation: Introduction'. In: E. Aarts and J. K. Lenstra (eds.): *Local Search in Combinatorial Optimisation*. Chichester, pp. 1–17, Wiley.

Althöfer, I. and K.-U. Koschnik: 1991, 'On the Convergence of Threshold Accepting'. *Applied Mathematics and Optimisation* **24**, 183–195.

Barr, R. S., B. L. Golden, J. P. Kelly, M. G. C. Resende, and W. R. Stewart: 1995, 'Designing and Reporting on Computational Experiments with Heuristic Methods'. *Journal of Heuristics* **1**, 9–32.

Brooks, C., S. P. Burke, and G. Persand: 2001, 'Benchmarks and the Accuracy of GARCH Model Estimation'. *International Journal of Forecasting* **17**(1), 45–56.

Chipman, J. S. and P. Winker: 2005, 'Optimal Aggregation of Linear Time Series Models'. *Computational Statistics and Data Analysis* **49**(2), 311–331.

Dueck, G. and T. Scheuer: 1990, 'Threshold Accepting: A General Purpose Algorithm Appearing Superior to Simulated Annealing'. *Journal of Computational Physics* **90**, 161–175.

Dueck, G. and P. Winker: 1992, 'New Concepts and Algorithms for Portfolio Choice'. *Applied Stochastic Models and Data Analysis* **8**, 159–178.

Dueck, G. and J. Wirsching: 1991, 'Threshold Accepting Algorithms for 0–1 Knapsack Problems'. In: H. Wacker and W. Zulehner (eds.): *Proceedings of the Fourth European Conference on Mathematics in Industry*. Stuttgart, pp. 225–262, B. G. Teubner.

Fang, K.-T., D. K. J. Lin, P. Winker, and Y. Zhang: 2000, 'Uniform Design: Theory and Application'. *Technometrics* **42**, 237–248.

Fang, K.-T., X. Lu, and P. Winker: 2003, 'Lower Bounds for Centered and Wrap–around L_2–discrepancies and Construction of Uniform Designs by Threshold Accepting'. *Journal of Complexity* **19**, 692–711.

Fang, K.-T., C.-X. Ma, and P. Winker: 2002, 'Centered L_2–Discrepancy of Random Sampling and Latin Hypercube Design, and Construction of Uniform Designs'. *Mathematics of Computation* **71**, 275–296.

Fang, K.-T., D. Maringer, Y. Tang, and P. Winker: 2005, 'Lower Bounds and Stochastic Optimisation Algorithms for Uniform Designs with Three or Four Levels'. *Mathematics of Computation* **75**, 859–878.

Farley, A. M. and S. Jones: 1994, 'Using a Genetic Algorithm to Determine an Index of Leading Economic Indicators'. *Computational Economics* **7**, 163–173.

Ferrall, C.: 2004, 'Solving Finite Mixture Models: Efficient Computation in Economics under Serial and Parallel Execution'. Technical report, Queen's University, Canada.

Fox, B. L.: 1994, 'Random Restarting versus Simulated Annealing'. *Computers, Mathematics and Applications* **27**(6), 33–35.

Gilli, M. and E. Këllezi: 2002a, 'The Threshold Accepting Heuristic for Index Tracking'. In: P. Pardalos and V. Tsitsiringos (eds.): *Financial Engineering, E-Commerce, and Supply Chain*, Applied Optimisation Series. Kluwer, pp. 1–18.

Gilli, M. and E. Këllezi: 2002b, 'A Global Optimisation Heuristic for Portfolio Choice with VaR and Expected Shortfall'. In: E. Kontoghiorghes, B. Rustem, and S. Siokos (eds.): *Computational Methods in Decision-making, Economics and Finance*. Kluwer, pp. 165–181.

Gilli, M. and P. Winker: 2003, 'A Global Optimisation Heuristic for Estimating Agent Based Models'. *Computational Statistics and Data Analysis* **42**(3), 299–312.

Hamming, R. W.: 1950, 'Error Detecting and Error Correcting Codes'. *Bell System Technical Journal* **29**, 147–160.

Hanafi, S., A. Freville, and A. E. Abdellaoui: 1996, 'Comparison of Heuristics for the 0–1 Multidimensional Knapsack Problem'. In: I. H. Osman and J. P. Kelly (eds.): *Meta–Heuristics: Theory and Applications*. Boston, MA, pp. 449–465, Kluwer.

Judd, K. D.: 1998, *Numerical Methods in Economics*. Cambridge, MA: MIT Press.

Kirkpatrick, S., C. Gelatt, and M. Vecchi: 1983, 'Optimisation by Simulated Annealing'. *Science* **220**, 671–680.

Maringer, D.: 2005, *Portfolio Management with Heuristic Optimisation*. Springer.

Osman, I. H. and G. Laporte: 1996, 'Metaheuristics: A Bibliography'. *Annals of Operations Research* **63**, 513–623.

Winker, P.: 1995, 'Identification of Multivariate AR–Models by Threshold Accepting'. *Computational Statistics and Data Analysis* **20**(9), 295–307.

Winker, P.: 2000, 'Optimized Multivariate Lag Structure Selection'. *Computational Economics* **16**, 87–103.

Winker, P.: 2001, *Optimisation Heuristics in Econometrics: Applications of Threshold Accepting*. Chichester: Wiley.

Winker, P.: 2005, 'The Stochastics of Threshold Accepting: Analysis of an Application to the Uniform Design Problem'. Technical Report 2005–003E, Staatswissenschaftliche Fakultät, Universität Erfurt.

Winker, P. and K.-T. Fang: 1997a, 'Application of Threshold Accepting to the Evaluation of the Discrepancy of a Set of Points'. *SIAM Journal on Numerical Analysis* **34**, 2028–2042.

Winker, P. and K.-T. Fang: 1997b, 'Optimal U–Type Designs'. In: H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof (eds.): *Monte Carlo and Quasi–Monte Carlo Methods 1996*. New York, pp. 436–448, Springer.

Winker, P. and M. Gilli: 2004, 'Applications of Optimisation Heuristics to Estimation and Modelling Problems'. *Computational Statistics & Data Analysis* **47**(2), 211–223.

The Autocorrelation Functions in SETARMA Models*

Alessandra Amendola, Marcella Niglio and Cosimo Vitale

Di.S.E.S., Università di Salerno

Summary. The dependence structure of a family of self exciting threshold autoregressive moving average (SETARMA) models, is investigated. An alternative representation for this class of models is proposed and the exact autocorrelation function is derived in the case of two regimes. Some practical implications of the theoretical results are analysed and discussed via several examples of SETARMA structures of fixed orders.

Key words: Threshold models, moments, autocorrelations, identification

1 Introduction

The family of linear autoregressive moving-average ARMA models can accommodate only a limited set of dynamic phenomena. Numerous generalisations of the ARMA models have been proposed with the intention of overcoming these limitations.

In this context, Priestley (1988), Tong (1990), Granger and Teräsvirta (1993), Tjøstheim (1994), Franses and Van Dijk (2000) and Fan and Yao (2003) have presented various nonlinear structures, highlighting their characteristic statistical features and proposing relevant applications.

In this paper, attention is focused on the class of threshold autoregressive (TAR) models, proposed by Tong between the end of the 1970's and the beginning of the 1980's, which have been generalised subsequently. Various features of the TAR models have been studied, such as the statistical properties of the generating process; the problems of estimating the coefficients of the models and of generating their forecasts have also been considered.

* Address for correspondence: Via Ponte Don Melillo, 84084, Fisciano (SA), Italy. E-mail:alamendola@unisa.it.

In spite of the intense interest in these models, some aspects of their structures require further investigation. (See Brockwell, Liu and Tweedie, 1992, Liu and Susko, 1992, De Gooijer, 1998, and more recently Ling and Tong, 2005).

Within the broad class of threshold models, attention has been focused on the self-exciting threshold autoregressive moving-average (SETARMA) model proposed in Tong (1983) and on its dependence structure.

Amendola et al. (2006) have derived the moments for this class of models by analytic methods. Their results have confirmed the well-known ability of the SETARMA models to capture the asymmetric properties of the distribution of the data. Moreover, it has been shown that the models can also account for the excess of kurtosis that characterises many financial times series.

The aim of the present paper is to derive an analytic expression of the global autocorrelation function of the SETARMA model. The model is locally linear, and an analysis of its autocorrelation can held in the investigation of the dependence structure of the process. The analysis of the autocorrelation functions of each regimes can also help in determining the orders of the respective autoregressive moving average models.

The structure of the paper is as follows. In Sect. 2, the SETARMA is introduced. In Sect. 3, the exact form of the moments of a SETARMA are presented briefly and, in Sect. 4, the autocorrelations of a two-regime SETARMA model are derived analytically. The generalisation to more than two regimes is straightforward. The theoretical results are confirmed through simulations. Some concluding remarks are made in the final section.

2 The SETARMA Model

2.1 The SETARMA Model

The Self-Exciting Threshold AutoRegressive Moving Average (SETARMA) model of order $(h; p_1, \ldots, p_h; q_1, \ldots, q_h)$, first presented by Tong (1983) and further mentioned in Tong (1990), is given by:

$$X_t = \sum_{i=1}^{h} \left[\phi_0^{(i)} + \sum_{j=1}^{p_i} \phi_j^{(i)} X_{t-j} + e_t - \sum_{w=1}^{q_i} \theta_w^{(i)} e_{t-w} \right] I(X_{t-d} \in R_i) \qquad (1)$$

where $e_t \sim i.i.d.(0, \sigma^2)$, $R_i = [r_{i-1}, r_i)$, for $i = 1, 2, \ldots, h$ and $\bigcup_{i=1}^{h} R_i = \mathbb{R}$, forms a partition of the real line such that $-\infty = r_0 < r_1 < r_2 < \ldots < r_h = +\infty$ with r_i the threshold values, d is the threshold delay, p_i and q_i are non negative integers referred to the AR order and MA order respectively, $\phi_j^{(i)}$ and $\theta_w^{(i)}$ are unknown parameters, with $j = 1, 2, \ldots, p_i$ and $w = 1, 2, \ldots, q_i$, and $I(\cdot)$ is the indicator function. The model is characterized by a piecewise linear structure which follows a linear ARMA model in each of the h regimes.

The attention of the present paper has been focused on a different SETARMA model with respect to the earlier version of Tong (1983), given as:

$$X_t = \sum_{i=1}^{h} \left[\phi_0^{(i)} + \sum_{j=1}^{p_i} \phi_j^{(i)} X_{t-j}^{(i)} + \sigma_i \left(e_t - \sum_{w=1}^{q_i} \theta_w^{(i)} e_{t-w} \right) \right] I(X_{t-d} \in R_i), \quad (2)$$

where $\sigma_i e_t \sim i.i.d.(0, \sigma_i^2)$, $0 < \sigma_i < \infty$ for $i = 1, \ldots, h$.

In particular the restriction that the regimes have a common error variance has been removed. Another aspect in which the present model differs from that of Tong is manifested when it switches from one regime to another. It can be observed that according to (1), the output of the i-th regime is a direct function of the process X_t, whereas, according to (2), it is related to the lagged values of X_t only via the innovation term e_t. The two representations (1) and (2) are equivalent in absence of any autoregressive component and (2) can be seen has a generalization of the model used in De Gooijer (1998) and Ling and Tong (2005).

In what follows, we shall consider a SETARMA$(2; p_1, p_2; q_1, q_2)$ model with a delay d and a threshold value r, such that $R_1 = [r, +\infty)$, $R_2 = (-\infty, r)$. This will reduce the complexity of the model and ease the burden of notation. The generalisation to the case of h regimes should be straightforward.

In a model with two regimes, the switching is regulated by the indicator process $I_{t-d} = I(X_{t-d} \in R_1)$ which, in view of its dichotomous nature, can be written as

$$I_{t-d} = \begin{cases} 1 & \text{if} \quad X_{t-d} \geq r, \\ 0 & \text{if} \quad X_{t-d} < r, \end{cases} \quad (3)$$

where $t = 1, 2, \ldots$, and $d > 0$.

In the case where $h = 2$, the model of (2) can be represented more compactly by

$$X_t = X_t^{(1)} I_{t-d} + X_t^{(2)} (1 - I_{t-d}) \equiv X_t^{(2)} + [X_t^{(1)} - X_t^{(2)}] I_{t-d}, \quad (4)$$

where, within their respective regimes, the outputs $X_t^{(i)} \sim \text{ARMA}(p_i, q_i); i = 1, 2$, are determined by linear autoregressive moving-average processes.

Since I_{t-d} in (4) plays a crucial role in the dynamic structure of X_t, its properties need to be examined more closely.

The Indicator Process I_{t-d}

The properties of the process $\{I_{t-d}\}$; $t = d+1, d+2, \ldots$ in (3), which controls the switching between the two regimes of the SETARMA model, reflect its dichotomous nature.

It is assumed that

[A1.] *The process I_{t-d} is second-order stationary and ergodic, and it has the following moments:*

$$\begin{aligned} &E(I_{t-d}) = P(I_{t-d} = 1) = P(X_{t-d} \geq r) = p, \\ &var(I_{t-d}) = p(1-p); \qquad t = d+1, d+2, \ldots, \\ &cov(I_{t-d}, I_{t-d-k}) = \gamma_I(k) = p_k - p^2, \end{aligned}$$

where

$$p_k = E(I_{t-d} I_{t-d-k}) = P(I_{t-d} = 1; I_{t-d-k} = 1) \qquad \text{for } k = 1, 2, \ldots. \tag{5}$$

It follows that the pair (I_{t-d}, I_{t-d-k}) has the joint probability distribution given in Table 1, where p and p_k must satisfy the inequality

$$\max(0, 2p-1) \leq p_k \leq p. \tag{6}$$

Moreover, there is

$$\sum_{k=0}^{\infty} |\gamma_I(k)| = p(1-p) + \sum_{k=1}^{\infty} |p_k - p^2| < +\infty, \tag{7}$$

where $\gamma_I(0) = p_0 - p^2 = p(1-p)$, since $p_0 = E(I_{t-d}^2) = p$. Figure 1 depicts the autocovariance $\gamma_I(k)$ as a function of the pair of values (p_k, p); and it reflects the inequality of (6).

The estimates of the probabilities p and p_k can be obtained taking advantage of the properties of the Bernoulli process I_{t-d}. In particular, given the assumption [A1.], estimates of p and p_k that are consistent in probability are given, respectively, by the following relative frequencies

$$\hat{p} = \frac{N(x_{t-d} \geq r)}{T-d}, \qquad \hat{p}_k = \frac{N(x_{t-d} \geq r, x_{t-d-k} \geq r)}{T-d}, \tag{8}$$

where $x_1, x_2, \ldots, x_T$ is a sample of length T generated by the process (2) and where $N(E)$ denotes the number of occurrences of the event E within the sample.

Table 1. Probability distribution of the variable (I_{t-d}, I_{t-d-k})

I_{t-d}/I_{t-d-k}	**0**	**1**	
0	$1-2p+p_k$	$p-p_k$	$1-p$
1	$p-p_k$	p_k	p
	$1-p$	p	1

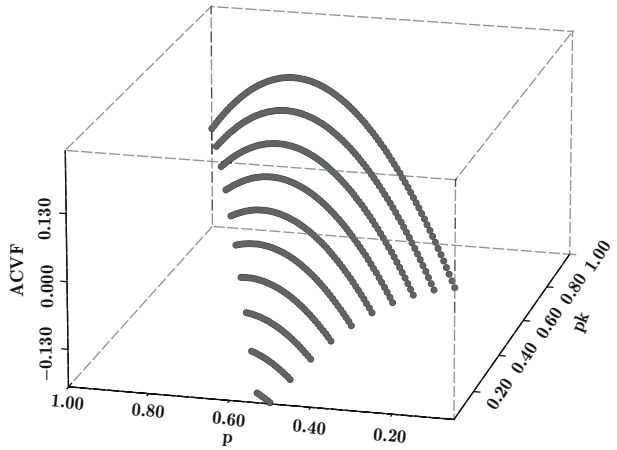

Fig. 1. Autocovariance function of the Bernoulli process I_t

Example 1. An example of probabilities p_k, with $k = 1, 2, \ldots$, that satisfy (6) and (7), is given by

$$p_k = p^2 + (p - p^2)^{k+1} \qquad k = 1, 2, \ldots \tag{9}$$

In this case, the probabilities of the distribution in Table 1 are given by

$$p - p_k = p(1-p)[1 - p^k(1-p^k)] \geq 0,$$

with $p(1-p) \in [0,1]$, $p^k(1-p^k) \in [0,1]$, and

$$1 - 2p + p_k = (1-p)^2 + p^{k+1}(1-p)^{k+1} \geq 0,$$

with $p^{k+1}(1-p)^{k+1} \in [0,\ (1/4)^{k+1}]$.

The autocovariance of I_t is

$$\gamma_I(k) = p_k - p^2 = p^{k+1}(1-p)^{k+1}.$$

Given $0 \leq p \leq 1$ and $0 \leq p(1-p) \leq 1/4$, it follows, by a convergence theorem of numeric series in geometric progression and ratio $|z| < 1$, that the sum of the autocovariance in (7) converges. That is

$$\sum_{k=0}^{\infty} |\gamma_I(k)| = p(1-p) \sum_{k=0}^{\infty} |p^k(1-p)^k| = \frac{p(1-p)}{1-p+p^2} < +\infty.$$

2.2 An Alternative Representation of the SETARMA Model

A SETARMA model is a generalisation of the ARMA model of Box and Jenkins (1976). Some of the results related to the ARMA model can be

extended to this class of nonlinear models. It is helpful, in pursuing these generalisations, to rewrite the model of (2), with $h = 2$, as follows:

$$X_t = \begin{cases} [\phi_0^{(1)} + \phi_{p_1}^{(1)}(B)X_t^{(i)} + \theta_{q_1}^{(1)}(B)\sigma_1 e_t]I(X_{t-d} \geq r), \\ [\phi_0^{(2)} + \phi_{p_2}^{(2)}(B)X_t^{(i)} + \theta_{q_2}^{(2)}(B)\sigma_2 e_t]I(X_{t-d} < r), \end{cases} \tag{10}$$

where $\phi_{p_i}^{(i)}(B) = \sum_{j=1}^{p_i} \phi_j^{(i)} B^j$ and $\theta_{q_i}^{(i)}(B) = 1 - \sum_{w=1}^{q_i} \theta_w^{(i)} B^w$, for $i = 1, 2$.

Two additional assumptions are required:

[A2.] *The polynomials* $\Phi^{(i)}(B) = 1 - \sum_{i=1}^{p_1} \phi_j^{(i)} B^j$ *and* $\theta_{q_i}^{(i)}(B) = 1 - \sum_{w=1}^{q_i} \theta_w^{(i)} B^w$ *have no roots in common, and all of their roots lie outside the unit circle*;

[A3.] *the joint process* $\mathbf{X_t} = (X_t^{(1)}, X_t^{(2)}, I_{t-d})$, *with* $X_t^{(i)} \sim \text{ARMA}(p_i, q_i)$ *for* $i = 1, 2$, *is strictly stationary, ergodic and invertible. (For a sufficient invertibility condition see Ling and Tong 2005, assumption 2.1)*;

The SETARMA$(2; p_1, p_2; q_1, q_2)$ model can be written alternatively as

$$X_t = \left[c_0^{(1)} + \sigma_1 \sum_{j=0}^{\infty} \psi_j^{(1)} B^j e_t \right] I_{t-d} + \left[c_0^{(2)} + \sigma_2 \sum_{j=0}^{\infty} \psi_j^{(2)} B^j e_t \right] (1 - I_{t-d}), \tag{11}$$

where

$$c_0^{(i)} = \frac{\phi_0^{(i)}}{1 - \sum_{j=1}^{p_i} \phi_j^{(i)}},$$

is the mean value of regime i, for $i = 1, 2$, and

$$\sum_{j=1}^{\infty} |\psi_j^{(i)}| < \infty,$$

with $\psi_0^{(i)} = 1$ and $i = 1, 2$.

The weights $\psi_j^{(i)}$ (for $i = 1, 2$ and $j = 0, 1, 2, \ldots$) are computed as

$$\psi_j^{(i)} = \begin{cases} 1 & \text{when} \quad j = 0, \\ \sum_{s=0}^{j-1} \psi_s^{(i)} \phi_{j-s}^{(i)} - \theta_j^{(i)} & \text{when} \quad j \geq 1, \end{cases} \tag{12}$$

with $\phi_j^{(i)} = 0$, for $j > p_i$, and $\theta_j^{(i)} = 0$, for $j > q_i$.

The model representation of (11) is used in deriving the theoretical results of the following sections.

3 The Moments of the SETARMA Model

A knowledge of the moments of a theoretical stochastic process enables us to understand the properties of its distribution and to assess the ability of the model to represent features that are observed in empirical data.

Therefore, Amendola et al. (2006) have derived the exact moments of order r and the central moments up to order four of the SETARMA model in (2), under the assumption that the errors e_t are Gaussian white noise, with $E(e_t) = 0$ and $E(e_t^2) = 1$.

In particular, using the local linearity of the SETARMA model and the properties of the Bernoulli process described in Sect. 2.1 (The Indicator Process I_{t-d}), they have demonstrated the following proposition:

Proposition 1. *Given* $X_t \sim \text{SETARMA}(2; p_1, p_2; q_1, q_2)$*, under the assumptions* [A2.] *and* [A3.] *and under the additional assumption that each regime admits at least moments of order* r*, it follows that the expected value of* X_t^r *is*

$$E(X_t^r) = \mu_r^{(1)} p + \mu_r^{(2)}(1-p), \tag{13}$$

where $\mu_r^{(i)} = E\left[\left(X_t^{(i)}\right)^r\right]$*, for* $i = 1, 2$.

The results in Proposition 1 indicate that the expected value of X_t in (11) is

$$E(X_t) = c_0^{(1)} p + c_0^{(2)}(1-p),$$

with

$$c_0^{(i)} = \frac{\phi_0^{(i)}}{1 - \sum_{j=1}^{p_i} \phi_j^{(i)}},$$

for $i = 1, 2$; and that $E(X_t) = 0$ if $c_0^{(1)} = c_0^{(2)} = 0$. The latter implies that $\phi_0^{(i)} = 0$, for $i = 1, 2$.

The result (13) can also be used to evaluate the variance of the SETARMA process. Thus

Corollary 1. *Under the hypotheses of Proposition 1, the process* $X_t \sim \text{SETARMA}(2; p_1, p_2; q_1, q_2)$*, has the variance*

$$var(X_t) = p\Psi_{(1)}^2\sigma_1^2 + (1-p)\Psi_{(2)}^2\sigma_2^2 + p(1-p)(c_0^{(1)} - c_0^{(2)})^2, \tag{14}$$

where $\Psi_{(i)}^2\sigma_i^2$ *(for* $i = 1, 2$*) is the variance of regime* i *in* (11)*, with* $\Psi_{(i)}^2 = \sum_{j=0}^{\infty}\left(\psi_j^{(i)}\right)^2 < \infty$.

The starting point for this result is

$$\begin{aligned} var(X_t) = var\left[I_{t-d}X_t^{(1)}\right] + var\left[(1 - I_{t-d})X_t^{(2)}\right] \\ + 2cov\left[I_{t-d}X_t^{(1)}, (1 - I_{t-d})X_t^{(2)}\right], \end{aligned} \tag{15}$$

where the three terms on the right of (15) are

$$\begin{aligned}
&\text{(a)} \quad var\left[I_{t-d}X_t^{(1)}\right] = p\gamma_1(0) + p(1-p)(c_0^{(1)})^2,\\
&\text{(b)} \quad var\left[(1-I_{t-d})X_t^{(2)}\right] = (1-p)\gamma_2(0) + p(1-p)(c_0^{(2)})^2,\\
&\text{(c)} \quad cov\left[I_{t-d}X_t^{(1)}, (1-I_{t-d})X_t^{(2)}\right] = -p(1-p)c_0^{(1)}c_0^{(2)}.
\end{aligned}$$

The study of moments assumes an important role when the distribution of the process is under analysis. In this context, it is appropriate to investigate the shape of the distribution through the third and fourth central moments of X_t in (11).

Corollary 2. *Under the hypotheses of Proposition 1, the third central unconditional moment of X_t is given by*

$$\begin{aligned}
E[(X_t-\mu)^3] = (c_0^{(1)} - c_0^{(2)}) \Big[&\bar{\mu}_{3I}(c_0^{(1)} - c_0^{(2)})^2 \\
&+ 3Var(I_t)(\sigma_1^2\Psi_{(1)}^2 - \sigma_2^2\Psi_{(2)}^2)\Big],
\end{aligned} \tag{16}$$

with $Var(I_t) = p(1-p)$ and $\bar{\mu}_{3I} = p(1-p)(1-2p)$.

The result (16) shows that the third central moment of the SETARMA model equals zero if the intercepts of the two regimes—or equivalently the mean values of the two regimes $c_0^{(1)}$ and $c_0^{(2)}$ in model (11) – are zeros. This highlights the relevance of the values assumed by $\phi_0^{(1)}$ and $\phi_0^{(2)}$ when the ability of the model to represent the skewness of the data is under analysis.

Corollary 3. *Under the hypotheses of Proposition 1, the fourth central unconditional moment of the* SETARMA *model is given by*

$$\begin{aligned}
\bar{\mu}_4 &= E[(X_t-\mu)^4] \\
&= \bar{\mu}_{4I}(c_0^{(1)} - c_0^{(2)})^4 + 3\left[p\sigma_1^4(\Psi_{(1)}^2)^2 + (1-p)\sigma_2^4(\Psi_{(2)}^2)^2\right] \\
&\quad + 6Var(I_t)(c_0^{(1)} - c_0^{(2)})^2\left[p\sigma_1^2\Psi_{(1)}^2 + (1-p)\sigma_2^2\Psi_{(2)}^2\right],
\end{aligned} \tag{17}$$

where $\bar{\mu}_{4I} = E[(I_t-p)^4] = p(1-p)[1-3p(1-p)]$.

The result in Corollary 3 shows that, when $c_0^{(1)} = c_0^{(2)} = 0$, the fourth central moment of model (11) is proportional to the weighted mean of the variances of the two regimes whose parameters contribute to the exact form derived for $\bar{\mu}_4$. The results (14), (16) and (17) relate to the unconditional variance, and the third and fourth central moments respectively. They can be used in computing the indexes of skewness (18) and of excess kurtosis (19) as follows:

$$\gamma_{AS} = \frac{E[(X_t-\mu)^3]}{E[(X_t-\mu)^2]^{3/2}}, \tag{18}$$

$$\gamma_K = \frac{E[(X_t - \mu)^4]}{E[(X_t - \mu)^2]^2} - 3, \tag{19}$$

where it can be readily observed that, when the intercepts, $\phi_0^{(1)}$ and $\phi_0^{(2)}$, are zeros, (or equivalently $c_0^{(1)} = c_0^{(2)} = 0$,) the skewness index $\gamma_{AS} = 0$ and the data generated by the corresponding SETARMA process with Gaussian innovations have a symmetric distribution.

4 The Autocorrelation

The autocorrelation coefficient $\rho(k)$ of a second-order stationary process

$$\rho(k) = \frac{cov(X_t, X_{t\pm k})}{[Var(X_t)Var(X_{t\pm k})]^{1/2}} \qquad k = 0, 1, 2, \ldots, N, \tag{20}$$

is an important tool in the analysis of linear time series. It reveals the linear dependence among the random variables; and it is used in the approach of Box and Jenkins (1976) in identifying the orders of ARMA models.

When nonlinear models are involved, the autocorrelation coefficient $\rho(k)$ provides insufficient information for investigating fully the relationship amongst the variables $X_t; t = 1, 2, \ldots, N$. However, it can give useful indications of how close the generating mechanism of $\{X_t\}$ is to linearity. For this purpose, Tong (1990) has proposed an *index of linearity* based on the squared autocorrelation. More recently, Nielsen and Madsen (2001) have generalised some traditional tools, such as the global and partial autocorrelation functions, for use in identifying nonlinear models.

When $X_t \sim$ SETARMA$(2; p_1, p_2; q_1, q_2)$, the autocorrelation coefficient $\rho(k)$ can be used to investigate whether a particular sequence of values have all been generated by the same regime.

For a start, the stationarity of X_t allows us to write the denominator of (20) as

$$[Var(X_t)Var(X_{t\pm k})]^{1/2} = Var(X_t),$$

where, referring to model (11), $Var(X_t)$ is given in (14).

The numerator of (20) requires more detailed investigation as the following proposition indicates.

Proposition 2. *Given* $X_t \sim$ SETARMA$(2; p_1, p_2; q_1, q_2)$ *and under the assumptions* [A2.] *and* [A3.], *which allow the SETARMA model to be written in the alternative form of* (11), *the autocovariance of* X_t *at lag* k, *with* $k = 0, 1, \ldots, N$, *is*

$$\gamma(k) = \sum_{j=0}^{\infty} \Big[p_k \sigma_1^2 \psi_j^{(1)} \psi_{k+j}^{(1)} + (1 - 2p + p_k)\sigma_2^2 \psi_j^{(2)} \psi_{k+j}^{(2)} + (p - p_k)\sigma_1\sigma_2 \times (\psi_j^{(1)} \psi_{k+j}^{(2)} + \psi_{k+j}^{(1)} \psi_j^{(2)}) \Big] + (p_k - p^2)(c_0^{(1)} - c_0^{(2)})^2, \tag{21}$$

with $p_k = E(I_{t-d} I_{t-d-k})$.

Proof. The stationarity assumption implies the symmetry of the autocovariance of X_t, such that $cov(X_t, X_{t-k}) = cov(X_t, X_{t+k})$; so the proof can be limited to the case of $\gamma(k) = cov(X_t, X_{t-k})$.

Starting with the definition of $\gamma(k)$, we have

$$\begin{aligned}\gamma(k) = cov\left[I_{t-d}X_t^{(1)}, I_{t-d-k}X_{t-k}^{(1)}\right] \\ +cov\left[I_{t-d}X_t^{(1)}, (1-I_{t-d-k})X_{t-k}^{(2)}\right] \\ +cov\left[(1-I_{t-d})X_t^{(2)}, I_{t-d-k}X_{t-k}^{(1)}\right] \\ +cov\left[(1-I_{t-d})X_t^{(2)}, (1-I_{t-d-k})X_{t-k}^{(2)}\right].\end{aligned} \tag{22}$$

The terms of (22) are given by

$$\begin{aligned}&\text{(a)}\quad cov\left[I_{t-d}X_t^{(1)}, I_{t-d-k}X_{t-k}^{(1)}\right] \\ &\qquad = p_k\gamma_1(k) + (p_k - p^2)(c_0^{(1)})^2, \\ &\text{(b)}\quad cov\left[I_{t-d}X_t^{(1)}, (1-I_{t-d-k})X_{t-k}^{(2)}\right] \\ &\qquad = (p-p_k)\gamma_{12}(k) - (p_k - p^2)c_0^{(1)}c_0^{(2)}, \\ &\text{(c)}\quad cov\left[(1-I_{t-d})X_t^{(2)}, I_{t-d-k}X_{t-k}^{(1)}\right] \\ &\qquad = (p-p_k)\gamma_{21}(k) - (p_k - p^2)c_0^{(1)}c_0^{(2)}, \\ &\text{(d)}\quad cov\left[(1-I_{t-d})X_t^{(2)}, (1-I_{t-d-k})X_{t-k}^{(2)}\right] \\ &\qquad = (1-2p+p_k)\gamma_2(k) + (p_k - p^2)(c_0^{(2)})^2,\end{aligned}$$

with $E[(1-I_{t-d})(1-I_{t-d-k})] = 1-2p+p_k$ and $E[(1-I_{t-d})I_{t-d-k}] = p-p_k$.

Placing the results (a)–(d) in (22) gives

$$\begin{aligned}\gamma(k) = p_k\gamma_1(k) + (1-2p+p_k)\gamma_2(k) + (p-p_k)(\gamma_{12}(k) + \gamma_{21}(k)) \\ +(p_k - p^2)(c_0^{(1)} - c_0^{(2)})^2,\end{aligned} \tag{23}$$

where

$$\gamma_i(k) = \sigma_i^2 \sum_{j=0}^{\infty} \psi_j^{(i)}\psi_{k+j}^{(i)}, \quad \text{for} \quad i = 1, 2,$$

$$\gamma_{12}(k) = \sigma_1\sigma_2 \sum_{j=0}^{\infty} \psi_j^{(1)}\psi_{k+j}^{(2)},$$

$$\gamma_{21}(k) = \sigma_1\sigma_2 \sum_{j=0}^{\infty} \psi_{k+j}^{(1)}\psi_j^{(2)},$$

which enable (21) to be derived.

In order to show how the autocovariance of the SETARMA model can be computed, we consider some examples.

Example 2. The simplest models such as SETARMA(2; 0, 0; 1,0) and SETARMA(2; 0 ,0; 0,1), which are described as SETMA models by De Gooijer (1998), have the following autocovariances:

$$\gamma(k) = \operatorname{cov}(X_t, X_{t-k}) = \begin{cases} \sigma_1^2 \theta_1^{(1)} p_k & \text{if} \quad k = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

and

$$\gamma(k) = \operatorname{cov}(X_t, X_{t-k}) = \begin{cases} \sigma_2^2 \theta_1^{(2)} (1 - 2p + p_k) & \text{if} \quad k = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

where σ_i^2 is the variance of the error in regime i $(i = 1, 2)$.

It is interesting to note that, in example 2, the SETMA autocovariance cannot be distinguished from that of a MA model, which has significant consequences for model selection.

A different autocovariance structure is obtained when the autoregressive component is of order one.

Example 3. Given $X_t \sim$ SETARMA(2;1,0;0,0) , $\gamma(k)$ in (21) is

$$\gamma(k) = p_k \frac{\sigma_1^2 \left(\phi_1^{(1)}\right)^k}{1 - \left(\phi_1^{(1)}\right)^2} + (p - p_k)\sigma_1\sigma_2 \left(\phi_1^{(1)}\right)^k + (p_k - p^2) \left(\frac{\phi_0^{(1)}}{1 - \phi_1^{(1)}} - \phi_0^{(2)} \right)^2 . \tag{26}$$

If both regimes have an autoregressive component of order 1, then the SETARMA(2; 1,1; 0,0) model becomes:

$$X_t = \left(\phi_0^{(1)} + \phi_1^{(1)} X_{t-1}^{(1)} + e_t \sigma_1\right) I_{t-d} + \left(\phi_0^{(2)} + \phi_1^{(2)} X_{t-1}^{(2)} + e_t \sigma_2\right) (1 - I_{t-d}),$$

and so

$$\begin{aligned} \gamma(k) = p_k & \frac{\sigma_1^2 \left(\phi_1^{(1)}\right)^k}{1 - \left(\phi_1^{(1)}\right)^2} + (1 - 2p + p_k) \frac{\sigma_2^2 \left(\phi_1^{(2)}\right)^k}{1 - \left(\phi_1^{(2)}\right)^2} \\ & + (p - p_k)\sigma_1\sigma_2 \frac{\left(\phi_1^{(1)}\right)^k + \left(\phi_1^{(2)}\right)^k}{1 - \left(\phi_1^{(1)}\phi_1^{(2)}\right)} \\ & + (p_k - p^2) \left(\frac{\phi_0^{(1)}}{1 - \phi_1^{(1)}} - \frac{\phi_0^{(2)}}{1 - \phi_1^{(2)}} \right)^2 . \end{aligned} \tag{27}$$

More general results are obtained when a SETARMA(2; 1,0; 0,1) is considered whose autocovariances are given, in the case of $k = 1$, by

$$\begin{aligned}\gamma(1) = p_1 \frac{\sigma_1^2 \phi_1^{(1)}}{1-(\phi_1^{(1)})^2} - (1-2p+p_1)\sigma_2^2\theta_1^{(2)} \\ + (p-p_1)\sigma_1\sigma_2\phi_1^{(1)}\left[1-\theta_1^{(2)}\left(\phi_1^{(1)}-1\right)\right] \\ + (p_1-p^2)\left(\frac{\phi_0^{(1)}}{1-\phi_1^{(1)}} - \phi_0^{(2)}\right)^2\end{aligned} \tag{28}$$

and, when $k > 1$, by

$$\begin{aligned}\gamma(k) = p_k \frac{\sigma_1^2(\phi_1^{(1)})^k}{1-(\phi_1^{(1)})^2} + \ (p-p_k)\sigma_1\sigma_2\left(\phi_1^{(1)}\right)^k\left(1-\theta_1^{(2)}\phi_1^{(1)}\right) \\ + (p_k-p^2)\left(\frac{\phi_0^{(1)}}{1-\phi_1^{(1)}} - \phi_0^{(2)}\right)^2.\end{aligned} \tag{29}$$

By placing the results (14) and (21) in (20), the autocorrelation $\rho(k)$ for model (11) is obtained, which assumes different forms when the orders p_i and q_i $(i = 1, 2)$ of the two regimes are selected. Based on these results, the autocorrelations of some SETARMA models are presented in the following section.

4.1 The Autocorrelation of SETARMA Models

Given a SETARMA(2; 1,1; 1,1) model with no intercepts, the numerator of the autocorrelation (20) is such that for $k = 1$:

$$\gamma_i(1) = \frac{\sigma_i^2\phi_1^{(i)}}{1-\left(\phi_1^{(i)}\right)^2}\left[1-\phi_1^{(i)}\theta_1^{(i)}\left(\phi_1^{(i)}-\theta_1^{(i)}\right)\right], \quad \text{for} \quad i = 1, 2,$$

$$\gamma_{12}(1) = \frac{\sigma_1\sigma_2}{1-\phi_1^{(1)}\phi_1^{(2)}}\left[-\phi_1^{(2)}\theta_1^{(1)} + \left(\phi_1^{(1)}-\theta_1^{(1)}\right)\left(1-\phi_1^{(1)}\theta_1^{(2)}\right)\right],$$

$$\gamma_{21}(1) = \frac{\sigma_1\sigma_2}{1-\phi_1^{(1)}\phi_1^{(2)}}\left[-\phi_1^{(1)}\theta_1^{(2)} + \left(\phi_1^{(2)}-\theta_1^{(2)}\right)\left(1-\phi_1^{(2)}\theta_1^{(1)}\right)\right],$$

whereas, more generally, when $k > 1$

$$\begin{aligned}\gamma_i(k) &= cov\left[X_t^{(i)}X_{t-k}^{(i)}\right] \\ &= \sigma_i^2\left(\phi_1^{(i)}\right)^{k-1}\frac{(1-\phi_1^{(i)}\theta_1^{(i)})(\phi_1^{(i)}-\theta_1^{(i)})}{1-\left(\phi_1^{(i)}\right)^2}, \quad \text{for} \quad i = 1, 2,\end{aligned}$$

$$\gamma_{12}(k) = cov\left[X_t^{(1)} X_{t-k}^{(2)}\right]$$
$$= \frac{\sigma_1\sigma_2}{1-\phi_1^{(1)}\phi_1^{(2)}}\left[\left(\phi_1^{(1)}\right)^{k-1}\left(\phi_1^{(1)}-\theta_1^{(1)}\right)\left(1-\phi_1^{(1)}\theta_1^{(2)}\right)\right],$$
$$\gamma_{21}(k) = cov\left[X_{t-k}^{(1)} X_t^{(2)}\right]$$
$$= \frac{\sigma_1\sigma_2}{1-\phi_1^{(1)}\phi_1^{(2)}}\left[\left(\phi_1^{(2)}\right)^{k-1}\left(\phi_1^{(2)}-\theta_1^{(2)}\right)\left(1-\phi_1^{(2)}\theta_1^{(1)}\right)\right],$$

Finally the variance at its denominator is

$$Var(X_t) = p\sigma_1\frac{1+\left(\theta_1^{(1)}\right)^2-2\phi_1^{(1)}\theta_1^{(1)}}{1-\left(\phi_1^{(1)}\right)^2}+(1-p)\sigma_2\frac{1+\left(\theta_1^{(2)}\right)^2-2\phi_1^{(2)}\theta_1^{(2)}}{1-\left(\phi_1^{(2)}\right)^2}.$$

From the previous results, it can be shown that autocorrelation $\rho(k)$ of the SETARMA(2; 1,1; 1,1) model under analysis tend to zero as k increases:

$$\lim_{k\to\infty}\rho(k) = 0,$$

and so the autocorrelation between X_t and $X_{t\pm k}$ is zero as the temporal lag $|k|$ grows.

These results can be clearly observed in frame (b) of Fig. 2, where the correlogram of the series generated from a SETARMA(2; 1,1; 1,1) model is portrayed, with X_t given by

$$X_t = \begin{cases} 0.6X_{t-1}^{(1)} + e_t^{(1)} - 0.4e_{t-1}^{(1)} & X_{t-1} \geq 0, \\ -0.6X_{t-1}^{(2)} + e_t^{(2)} + 0.4e_{t-1}^{(2)} & X_{t-1} < 0, \end{cases} \tag{30}$$

where $e_t^{(1)} = e_t$, $e_t^{(2)} = 0.5e_t$ and $\{e_t\}$ is a white-noise process with $e_t \sim N(0,1)$, for $t = 1, 2, \ldots, 10000$.

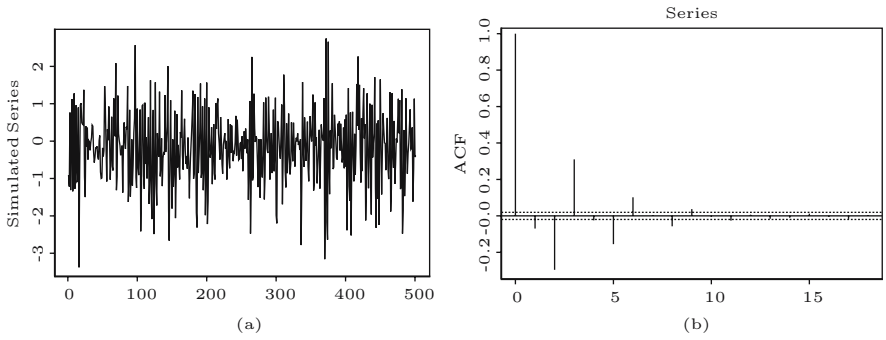

Fig. 2. Frame (a): Sample of the generated series of length 500; Frame (b): Correlogram of the generated series

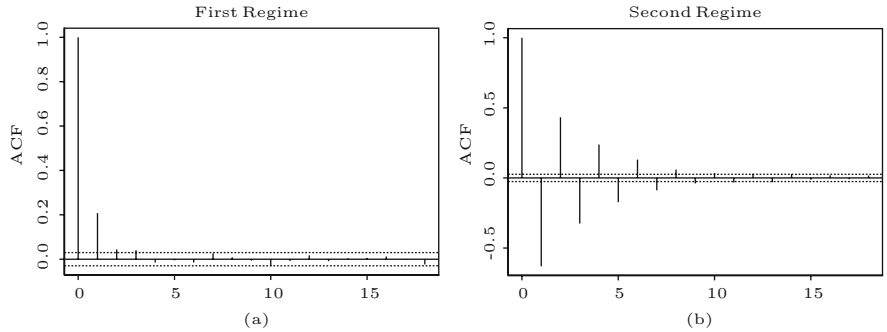

Fig. 3. Correlograms of the generated data which belong to the first and second regime [frames (a) and (b) respectively]

The sample autocorrelation, which decreases exponentially to zero, gives useful information about the dependence structure of the series under analysis. As Tong (1990) has indicated, this can be considered a first step in investigating the dependence among the X_t's which needs to be evaluated with more sophisticated instruments in order to avoid model misspecification. Further it is informative for the identification of the order of the two regimes when the threshold delay d and the threshold value r are known and therefore the correlograms of the values which belong to each regime can be observed—see frames (a) and (b) in Fig. 3.

5 Conclusions

In time series analysis, the autocorrelation function is a tool for studying the dependence structure of the data, which can also suggest how they should be modelled. This has provided the motive for deriving the exact form of the SETARMA autocorrelation function $\rho(k)$ where by the dependence structure of this family of processes can be investigated. The results obtained also highlight the way in which diagnostic tools, based on $\rho(k)$, can be used to determine the extent to which the generating process departs from the linearity, which can be helpful in identifying a SETARMA.

Acknowledgements

The authors are grateful to D.S.G. Pollock and the four anonymous referees for their valuable comments and suggestions.

References

Amendola A., Niglio M., Vitale C., (2006), The moments of SETARMA models, *Statistics & Probability Letters*, **76**, 625–633.

Brännäs K., De Gooijer J.G., (2004), Asymmetries in conditional mean and variance: Modelling stock returns by asMA-asQGARCH, *Journal of Forecasting*, **23**, 155–171.

Box G.E.P., Jenkins G.M., (1976), *Time series analysis, forecasting and control*, Holden-Day, San Francisco.

Brock W.A., De Lima P.J.F., (1996), Nonlinear time series complexity theory and finance, in *Handbook of Statistics*, **14**, Elsevier.

Brockwell P.J., Liu J., Tweedie R., (1992), On the extence of stationary threshold autoregressive moving-average processes, *Journal of Time Series Analysis*, **13**, 95–107.

Cont R., (2001), Empirical properties of asset returns: stylized facts and statistical issue, *Quantitative Finance*, **1**, 223–236.

De Gooijer J., (1998), On threshold moving-average models, *Journal of Time Series Analysis*, **19**, 1–18.

Engle R.F., (1982) Autoregrassive conditional heteroskedasticity with estimates of the variance for U.K. inflation, *Econometrica*, **50**, 987–1008.

Fan J, Yao Q., (2003), *Nonlinear time series: parametric and nonparametric methods*, Springer, New York.

Franses P.H., Van Dijk D., (2000), *Non-linear time series models in empirical finance*, Cambridge University Press, Cambridge.

Granger C.W.J., Teräsvirta T., (1993), *Modelling nonlinear relationships*, Oxford University Press, Oxford.

Ling S., Tong H., (2005), Testing for a linear MA model against threshold MA models, *The Annals of Statistics*, **33**, 2529–2552.

Liu J., Li W.K., Li C.W., (1997), On a threshold autoregression with conditional heteroskedasticity, *Journal of Statistical Planning & Inference*, **62**, 279–300.

Liu J., Susko E., (1992), On strict stationary and ergodicity of a nonlinear ARMA model, *Journal of Applied Probabilyty*, **29**, 363–373.

Nielsen H.A., Madsen H., (2001), A generalization of some classical time series tools, *Computational Statistics & Data Analysis*, **37**, 13–31.

Pagan A., (1996), The econometrics of financial market, *Journal of Empirical Finance*, **3**, 15–102.

Priestley M. B., (1988), *Non-Linear and Non-Stationary Time Series Analysis*, Academic Press, San Diego.

Tjøstheim D., (1994), Non-linear time series: a selective review, *Scandinavian Journal of Statistics*, **21**, 97–130.

Tong H., (1983), *Threshold models in nonlinear time series analysis*, Springer-Verlag, New York.

Tong H., (1990), *Non-linear time series: A dynamical system approach*, Clarendon Press, Oxford.

Trend Estimation and De-Trending

Stephen Pollock

Department of Economics, Queen Mary College, University of London,
Mile End Road, London E1 4NS, UK

Summary. An account is given of a variety of linear filters which can be used for extracting trends from economic time series and for generating de-trended series. A family of rational square-wave filters is described which enable designated frequency ranges to be selected or rejected. Their use is advocated in preference to other filters which are commonly used in quantitative economic analysis.

Key words: Linear filters, economic time series, extracting trends

1 Introduction: The Variety of Linear Filters

Whenever we form a linear combination of successive elements of a discrete-time signal $y(t) = \{y_t; t = 0, \pm 1, \pm 2, \ldots\}$, we are performing an operation which is described as linear filtering. Such an operation can be represented by the equation

$$x(t) = \psi(L)y(t) = \sum_j \psi_j y(t-j), \tag{1}$$

wherein

$$\psi(L) = \big\{ \cdots + \psi_{-1}L^{-1} + \psi_0 I + \psi_1 L + \cdots \big\}, \tag{2}$$

is described as the filter.

The effect of the operation it to modify the signal $y(t)$ by altering the amplitudes of its cyclical components and by advancing or delaying them in time. These modifications are described, respectively, as the gain effect and the phase effect of the filter. The gain effect is familiar through the example of the frequency-specific amplification of sound recordings which can be achieved with ordinary domestic sound systems. A phase effect in the form of a time delay is bound to accompany any signal processing that takes place in real time.

In quantitative economic analysis, filters are used for smoothing data series, which is a matter of attenuating or even discarding the high-frequency

components of the series and preserving the low-frequency components. The converse operation, which is also common, is to extract and discard the low-frequency trend components so as to leave a stationary sequence of residuals, from which the dynamics of short-term economic relationships can be estimated more easily.

As it stands, the expression under (2) represents a Laurent series comprising an indefinite number of terms in powers of the lag operator L and its inverse $L^{-1} = F$ whose effects on the sequence $y(t)$ are described by the equations $Ly(t) = y(t-1)$ and $L^{-1}y(t) = Fy(t) = y(t+1)$.

In practice, $\psi(L)$ often represents a finite polynomial in positive powers of L, which is described as a one-sided moving-average operator. Such a filter can only impose delays upon the components of $y(t)$.

Alternatively, the expression $\psi(L)$ might stand for the series expansion of a rational function $\delta(L)/\gamma(L)$; in which case the series is liable to comprise an indefinite number of ascending powers of L, beginning with $L^0 = I$. Such a filter is realised via a process of feedback, which may be represented by the equation

$$\gamma(L)x(t) = \delta(L)y(t), \tag{3}$$

or, more explicitly, by

$$\begin{aligned} x(t) = \delta_0 y(t) + \delta_1 y(t-1) + \cdots + \delta_d y(t-d) \\ -\gamma_1 x(t-1) - \cdots - \gamma_g x(t-g). \end{aligned} \tag{4}$$

Once more, the filter can only impose time delays upon the components of $x(t)$; and, because the filter takes a rational form, there are bound to be different delays at the various frequencies.

Occasionally, a two-sided symmetric filter in the form of

$$\psi(L) = \delta(F)\delta(L) = \psi_0 I + \psi_1(F + L) + \cdots + \psi_d(F^d + L^d) \tag{5}$$

is employed in smoothing the data or in eliminating its seasonal components. The advantage of such a filter is the absence of a phase effect. That is to say, no delay is imposed on any of the components of the signal. The so-called Cramér–Wold factorisation which sets $\psi(L) = \delta(F)\delta(L)$, and which must be available for any properly-designed filter, provides a straightforward way of explaining the absence of a phase effect. For the factorisation enables the transformation of (1) to be broken down into two operations:

$$\text{(i)} \quad z(t) = \delta(L)y(t) \qquad \text{and} \qquad \text{(ii)} \quad x(t) = \delta(F)z(t). \tag{6}$$

The first operation, which runs in real time, imposes time delays on every component of $x(t)$. The second operation, which works in reversed time, imposes an equivalent reverse-time delay on each component. The reverse-time delays, which are advances in other words, serve to eliminate the corresponding real-time delays.

The processed sequence $x(t)$ may be generated via a single application of the two-sided filter $\psi(L)$ to the signal $y(t)$, or it may be generated in two operations via the successive applications of $\delta(L)$ to $y(t)$ and $\delta(F)$ to $z(t) = \delta(L)y(t)$. The question of which of these techniques has been used to generate $y(t)$ in a particular instance should be a matter of indifference.

The final species of linear filter that may be used in the processing of economic time series is a symmetric two-sided rational filter of the form

$$\psi(L) = \frac{\delta(F)\delta(L)}{\gamma(F)\gamma(L)}. \tag{7}$$

Such a filter must, of necessity, be applied in two separate passes running forwards and backwards in time and described, respectively, by the equations

$$\text{(i)} \quad \gamma(L)z(t) = \delta(L)y(t) \qquad \text{and} \qquad \text{(ii)} \quad \gamma(F)x(t) = \delta(F)z(t). \tag{8}$$

Such filters represent a most effective way of processing economic data in pursuance of a wide range of objectives.

The essential aspects of linear filtering are recounted in numerous texts devoted to signal processing. Two that are worthy of mention are by Haykin (1989) and by Oppenheim and Schafer (1989). The text of Pollock (1999) bridges the gap between signal processing and time-series analysis.

In this paper, we shall concentrate on the dual objectives of estimating economic trends and of de-trending data series. However, before we present the methods that we wish to advocate, it seems appropriate to provide a critical account of some of the methods that are in common use.

2 Differencing Filters

In quantitative economics, the traditional means of reducing a time series to stationarity has been to take as many differences of the series as are necessary to eliminate the trend and to generate a series that has a convergent autocovariance function. A sequence of d such operations can be represented by the equation

$$x(t) = (I - L)^d y(t). \tag{9}$$

This approach to trend-elimination has a number of disadvantages, which can prejudice the chances of using the processed data successfully in estimating economic relationships.

The first of the deleterious effects of the difference operator, which is easily emended, is that it induces a phase lag. Thus, when it is applied to data observed quarterly, the operator induces a time lag of one-and-a-half months. To compensate for the effect, the differenced data may be shifted forwards in time to the points that lie midway between the observations. When applied twice, the operator induces a lag of three months. In that case, the appropriate

recourse in avoiding a phase lag is to apply the operator both in real time and in reversed time. The resulting filter is

$$(I - F)(I - L) = -F + 2I - L, \tag{10}$$

which is a symmetric two-sided filter with no phase effect.

As Fig. 1 shows, this filter serves to attenuate the amplitude of the components of $y(t)$ over a wide range of frequencies. It also serves to increase the amplitude of the high-frequency components. If the intention is only to remove the trend from the data, then the amplitude of these components should not be altered. In order not to affect the high-frequency components, the filter coefficients must be scaled by a factor of 0.25.

To understand this result, one should consider the transfer-function of the resulting filter, which is obtained by replacing the lag operator L by the complex argument z^{-1} to give

$$\psi_D(z) = \frac{1}{4}\big(-z + 2 - z^{-1}\big). \tag{11}$$

The effect of the filter upon the component of the highest observable frequency, which is the so-called Nyquist frequency of $\omega = \pi$, is revealed by setting $z = \exp\{\mathrm{i}\pi\}$, which creates the filter's frequency-response function. This is

$$\begin{aligned}\psi_D(e^{i\pi}) &= \frac{1}{4}\Big\{2 - (e^{\mathrm{i}\pi} + e^{-\mathrm{i}\pi})\Big\} \\ &= \frac{1}{4}\Big\{2 - 2\cos(\pi)\Big\} = 1.\end{aligned} \tag{12}$$

Thus, the gain of the filter, which is the factor by which the amplitude of a cyclical component is altered, is unity at the frequency $\omega = \pi$, which is what is required.

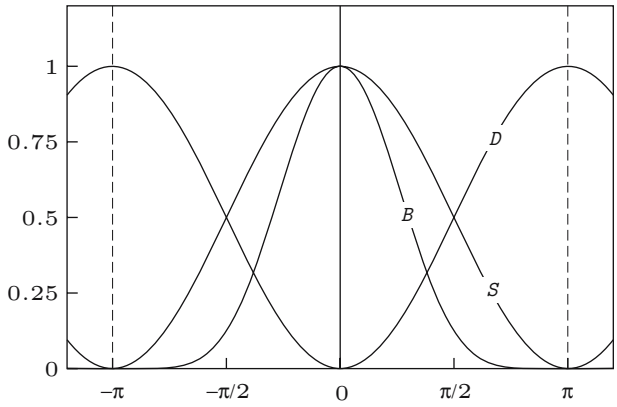

Fig. 1. The frequency-response functions of the lowpass filter $\psi_S(z) = \frac{1}{4}(z+2+z^{-1})$, the highpass filter $\psi_D(z) = \frac{1}{4}(-z + 2 - z^{-1})$ and the binomial filter $\psi_B(z) = \frac{1}{64}(1 + z)^3(1 + z^{-1})^3$

The condition that has been fulfilled by the filter may be expressed most succinctly by writing $|\psi_D(-1)| = 1$, where the vertical lines denote the operation defined by

$$|\psi(z)| = \sqrt{\psi(z)\psi(z^{-1})}, \tag{13}$$

which, in the case where $z = \exp\{i\omega\}$, amounts to taking the complex modulus. In that case, z is located on the unit circle; and, when it is expressed as a function of ω, $|\psi(\exp\{\mathrm{i}\omega\})|$ becomes the so-called amplitude-response function, which indicates the absolute value of the filter gain at each frequency.

In the case of the phase-neutral differencing filter of (10), as in the case of any other phase-free filter, the condition $\psi(z) = \psi(z^{-1})$ is fulfilled. This condition implies that the transfer function $\psi(z) = |\psi(z)|$ is a non-negative real-valued function. Therefore, the operation of finding the modulus is redundant. In general, however, the transfer function is a complex-valued function $\psi(z) = |\psi(z)| \exp\{\mathrm{i}\theta(\omega)\}$ whose argument $\theta(\omega)$, evaluated at a particular frequency, corresponds to the phase shift at that frequency.

Observe that the differencing filter also obeys the condition $|\psi_D(1)| = 0$. This indicates that the gain of the filter is zero at zero frequency, which corresponds to the fact that it annihilates a linear trend, which may be construed as a zero-frequency component.

The adjunct of the highpass trend-removing filter $\psi_D(z)$ is a complementary lowpass trend-estimation or smoothing filter defined by

$$\psi_S(z) = 1 - \psi_D(z) = \frac{1}{4}\big(z + 2 + z^{-1}\big). \tag{14}$$

As can be seen from Fig. 1, the two filters $\psi_S(z)$ and $\psi_D(z)$ bear a relation of symmetry, with is to say that, when they are considered as functions on the interval $[0, \pi]$, they represent reflections of each other about a vertical axis drawn through the frequency value of $\omega = \pi/2$. The symmetry condition can be expressed succinctly via the equations $\psi_S(-z) = \psi_D(z)$ and $\psi_D(-z) = \psi_S(z)$.

The differencing filter $\psi_D(z) = \frac{1}{4}(1 - z)(1 - z^{-1})$ and its complement $\psi_S(z) = \frac{1}{4}(1 + z)(1 + z^{-1})$ can be generalised in a straightforward manner to generate higher-order filters. Thus, we may define a binomial lowpass filter via the equation

$$\psi_B(z) = \frac{1}{4^n}(1 + z)^n(1 + z^{-1})^n. \tag{15}$$

This represents a symmetric two-sided filter whose coefficients are equal to the ordinates of the binomial probability function $b(2n; p = \frac{1}{2}, q = \frac{1}{2})$. The gain or frequency response of this filter is depicted in Fig. 1 for the case where $2n = 6$.

An n increases, the profile of the coefficients of the binomial filter tends increasingly to resemble that of a Gaussian normal probability density function. The same is true of the profile of the frequency-response function defined over the interval $[-\pi, \pi]$, which is the Fourier transform of the sequence of

coefficients. In this connection, one might recall that the Fourier (integral) transform of a Gaussian distribution is itself a Gaussian distribution. As n increases, the span of the filter coefficients widens. At the same time, the dispersion of the frequency-response function diminishes, with the effect that the filter passes an ever-diminishing range of low-frequency components.

It is clear that, for the family of binomial filters, the symmetry of the relationship between the highpass and lowpass filters prevails only in the case of $n = 1$. Thus, if $\psi_C(z) = 1 - \psi_B(z)$, then, in general, $\psi_C(z) \neq \psi_B(-z)$. This is to be expected from the characterisation that we have given above.

It remains to conclude this section by demonstrating the effect that the simple differencing filter of (10) is liable to have on a typical economic time series. An example is provided by a series of monthly measurements on the U.S. money stock from January 1960 to December 1970. Over the period in question, the stock appears to grow at an accelerating rate.

Figure 2 shows the effect of fitting a polynomial of degree five in the temporal index t to the logarithms of the data. This constitutes a rough-and-ready means of estimating the trend.

The periodogram of the residuals from the polynomial regression is displayed in Fig. 3. Here, there is evidence of a strong seasonal component at

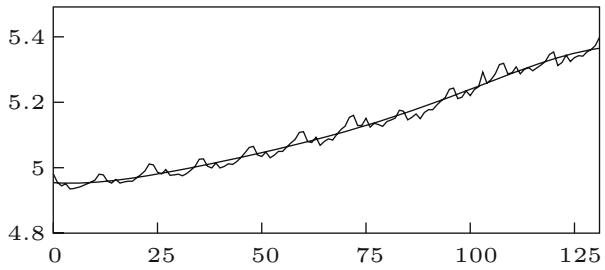

Fig. 2. The logarithms of 132 monthly observations on the U.S. money stock with an interpolated polynomial time trend of degree 5

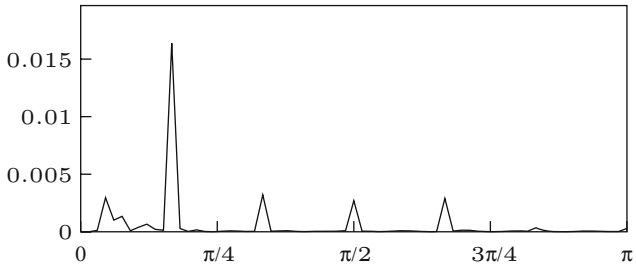

Fig. 3. The periodogram of the residuals from fitting a 5th degree polynomial time trend to the logarithms of the U.S. money stock

the frequency of $\omega = \pi/6$. Components of a lesser amplitude are also evident at the harmonic frequencies of $\omega = \pi/3$, $\pi/2$, $2\pi/3$, and there is a barely perceptible component at the frequency of $\omega = 5\pi/6$.

Apart from these components, which are evidently related to an annual cycle in the money stock, there is a substantial low-frequency component, which spreads over a range of adjacent frequencies and which attains its maximum amplitude at a frequency that corresponds to a period of roughly four years. This component belongs to the trend; and the fact that it is evident in the periodogram of the residuals is an indication of the inadequacy of the polynomial as a means of estimating the trend.

Figure 4 shows the periodogram of the logarithmic money-stock sequence after is has been subjected to the differencing filter of (10). As might be expected, the effect of the filter has been to remove the low-frequency trend components. However, it also has an effect which spreads into the mid and high-frequency ranges. In summary, we might say that the differencing filter has destroyed or distorted much of the information that would be of economic interest. In particular, the pattern of the seasonal effect has been corrupted. This distortion is liable to prejudice our ability to build effective forecasting models that are designed to take account of the seasonal fluctuations.

One might be tempted to use the lowpass binomial filter, defined under (15), as a means of extracting the trend. However, as Fig. 1 indicates, even with a filter order of 6, there would be substantial leakage from the seasonal components into the estimated trend; and we should need to deseasonalise the data before applying the filter.

In the ensuing sections, we shall describe alternative procedures for trend extraction and trend estimation. The first of these procedures, which is the subject of the next section, is greatly superior to the differencing procedure. Nevertheless, it is still subject to a variety of criticisms. The procedure of the ultimate section is the one which we shall recommend.

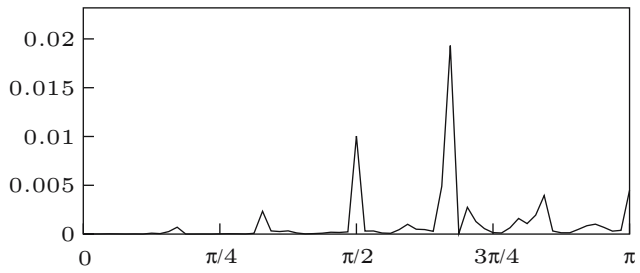

Fig. 4. The periodogram of a sequence obtained by applying the second-order differencing filter to the logarithms of the U.S. money stock

3 Notch Filters

The binomial filter $\psi_B(z)$, which we have described in the previous section, might be proposed as a means of extracting the low-frequency components of an economic time series, thereby estimating the trend. The complementary filter, which would then serve to generate the de-trended series, would take the form of $\psi_C(z) = 1 - \psi_B(z)$.

Such filters, however, would be of limited use. In order to ensure that a sufficiently restricted range of low-frequency components are passed by the binomial filter, a large value of n would be required. This would entail a filter with numerous coefficients and a wide time span. When a two-sided filter of $2n + 1$ coefficients reaches the end of the data sample, there is a problem of overhang. Either the final n sample elements must remain unprocessed, or else n forecast values must be generated in order to allow the most recent data to be processed. The forecasts, which could be provided by an ARIMA model, for example, might be of doubtful accuracy.

In applied economics, attention is liable to be focussed on the most recent values of a data series; and therefore a wide-span symmetric filter, such as the binomial filter, is at a severe disadvantage. It transpires that methods are available for constructing lowpass filters which require far fewer parameters.

To describe such methods, let us review the original highpass differencing filter of (10). Such a filter achieves the effect of annihilating a trend component by placing a zero of the function $\psi(z)$ on the unit circle at the point $z = 1$, which corresponds to a frequency value of $\omega = 0$. Higher-order differencing filters are obtained by placing more than one zero at this location. However, the effect of the zeros is likely to be felt over the entire frequency range with the deleterious consequences that we have already illustrated with a practical example.

In order to limit the effects of a zero of the filter, the natural recourse is to place a pole in the denominator of the filter's transfer function located at a point in the complex plane near to the zero. The pole should have a modulus that is slightly less than unity. The effect will be that, at any frequencies remote from the target frequency of $\omega = 0$, the pole and the zero will virtually cancel, leaving the frequency response close to unity. However, at frequencies close to $\omega = 0$, the effect of the zero, which is on the unit circle, will greatly outweigh the effect of the pole, which is inside it, and a narrow notch will be cut in the frequency response of the transfer function.

The device that we have described is called a notch filter. It is commonly used in electrical engineering to eliminate unwanted components, which are sometimes found in the recordings of sensitive electrical transducers and which are caused by the inductance of the alternating current of the mains electrical supply. In that case, the zero of the transfer function is placed, not at $z = 1$, but at some point on the unit circle whose argument corresponds to the mains frequency Also, the pole and the zero must be accompanied by their complex conjugates.

The poles in the denominator of the electrical notch filter are commonly placed in alignment with the corresponding zeros. However, the notch can be widened by placing the pole in a slightly different alignment. Such a recourse is appropriate when the mains frequency is unstable. Considerations of symmetry may then dictate that there should be a double zero on the unit circle flanked by two poles. If μ denotes a zero and κ denotes a pole, then this prescription would be met by setting

$$\mu_1, \mu_2 = e^{\mathrm{i}\omega} \qquad \text{and} \qquad \kappa_1, \kappa_2 = \rho e^{\mathrm{i}\omega \pm \epsilon} \quad \text{with} \quad 0 < \rho < 1, \tag{16}$$

where ω denotes the target frequency and ϵ denotes a small offset. The accompanying conjugate values are obtained by reversing the sign of the imaginary number i.

The concept of a notch filter with offset poles leads directly to the idea of a rational trend-removal filter of the form

$$\frac{\delta(z^{-1})}{\gamma(z^{-1})} = \frac{(1 - z^{-1})^2}{(1 - \kappa z^{-1})(1 - \kappa^* z^{-1})}, \tag{17}$$

where $\kappa = \rho \exp\{\mathrm{i}\epsilon\}$ is a pole which may be specified in terms of its modulus ρ and its argument ϵ, and where $\kappa^* = \rho \exp\{-\mathrm{i}\epsilon\}$ is its conjugate. To generate a phase-neutral filter, this function must be compounded with the function $\delta(z)/\gamma(z)$, which corresponds to the same filter applied in reversed time. Although only two parameters ρ and ϵ are involved, the search for an appropriate specification for the filter is liable to be difficult and time-consuming in the absence of a guiding design formula.

A notch filter, which has acquired considerable popularity amongst economists, and which depends on only one parameter, is given by the formula

$$\psi_N(z) = \frac{\delta(z)\delta(z^{-1})}{\gamma(z)\gamma(z^{-1})} = \frac{(1 - z)^2(1 - z^{-1})^2}{(1 - z)^2(1 - z^{-1})^2 + \lambda^{-1}}. \tag{18}$$

The placement of its poles and zeros within the complex plane is illustrated in Fig. 5. The complement of the filter, which is specified by

$$\psi_P(z) = 1 - \psi_N(z) = \frac{\lambda^{-1}}{(1 - z)^2(1 - z^{-1})^2 + \lambda^{-1}} \tag{19}$$

is know to economists as the Hodrick–Prescott smoothing filter.

The filter was presented originally by Hodrick and Prescott (1980) in a widely circulated discussion paper. The paper was published as recently as (1997). Examples of the use of this filter have been provided by Kydland and Prescott (1990), King and Rebelo (1993) and by Cogley and Nason (1995).

The Hodrick–Prescott filter has an interesting heuristic. It transpires that it is the optimal estimator of the trajectory of a second-order random walk observed with error. Its single adjustable parameter λ^{-1} corresponds to the signal-to-noise ratio, which is the ratio of the variance of the white-noise

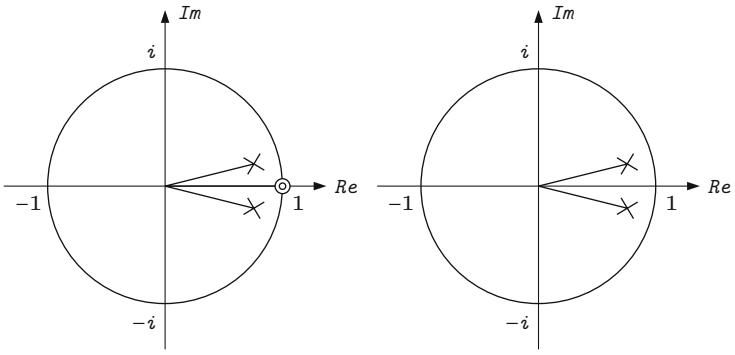

Fig. 5. The pole–zero diagram of the real-time components of the notch filter ψ_N (left) and of the Hodrick–Prescott filter $\psi_P = 1 - \psi_N$ (right) in the case where $\lambda = 64$. The poles are marked by crosses and, in the case of the notch filter, the double zero at $z = 1$ is marked by concentric circles

process that drives the random walk and the variance of the error that obscures its observations. It is usual to describe λ as the smoothing parameter.

The filter is also closely related to the Reinsch (1976) smoothing spline, which is used extensively in industrial design. With the appropriate choice of the smoothing parameter, the latter represents the optimal estimator of the underlying trajectory of an integrated Wiener process observed with error.

The effect of increasing the value of λ in the formula for the smoothing filter is to reduce the range of the low-frequency components that are passed by the filter. The converse effect upon the notch filter is to reduce the width of the notch that impedes the passage of these components. These two effects are illustrated in Figs. 6 and 7, which depict the frequency-response functions of the two filters. Figure 8 shows the trajectory of the poles of the filter as a function of the value of λ.

In order to implement either the smoothing filter or the notch filter, it is necessary to factorise their common denominator to obtain an expression for

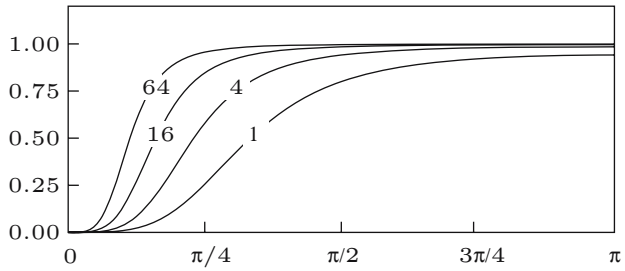

Fig. 6. The frequency-response function of the notch filter ψ_N for various values of the smoothing parameter λ

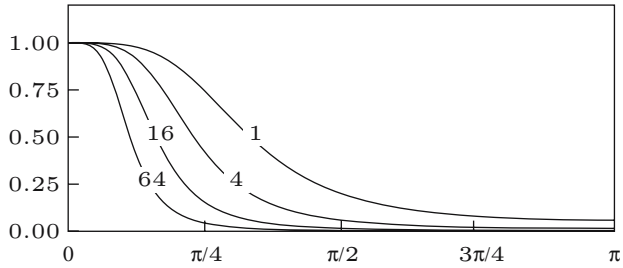

Fig. 7. The frequency-response function of the Hodrick–Prescott smoothing filter ψ_P for various values of the smoothing parameter λ

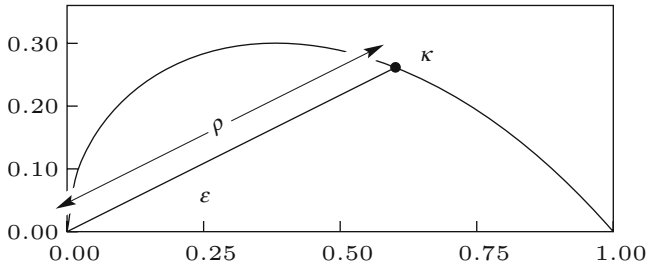

Fig. 8. The trajectory in the complex plane of a pole of the norch filter ψ_N. The pole approaches $z = 1$ as $\lambda^{-1} \to 0$

$\gamma(z)$. Since $z^2\gamma(z)\gamma(z^{-1})$ is a polynomial of degree four, one can, in principle, find analytic expressions for the poles which are in terms of the smoothing parameter λ. Alternatively, one may apply the iterative procedures which are used in the obtaining the Cramér–World factorisation of a Laurent polynomial. This is, in fact, how Fig. 8 has been constructed.

The Hodrick–Prescott smoothing filter has been subjected to criticisms from several sources. In particular, it has been claimed—by Harvey and Jaeger (1993) amongst others—that thoughtless de-trending using the filter can lead investigators to detect spurious cyclical behaviour in economic data. The claim can only be interpreted to mean that, sometimes, the notch filter will pass cyclical components which ought to be impeded and attributed to the trend. One might say, in other words, that in such circumstances, the trend has been given a form which is too inflexible. This problem, which cannot be regarded as a general characteristic of the filter, arises from a mismatch of the chosen value of the smoothing parameter with the characteristics of the data series. However, it must be admitted that it is often difficult to find an appropriate value for the parameter.

A more serious shortcoming of the filter concerns the gradation between the stopband, which is the frequency range which is impeded by the filter, and the passband which is the frequency range where the components of a series are unaffected by the filter. This gradation may be too gentle for some

purposes, in which case there can be no appropriate choice of value for the smoothing parameter.

In order to construct a frequency-selective filter which is accurately attuned to the characteristics of the data, and which can discriminate adequately between the trend and the residue, a more sophisticated methodology may be called for. We shall attempt to provide this in the ensuing sections of the paper.

4 Rational Square-Wave Filters

In the terminology of digital signal processing, an ideal frequency-selective filter is one for which the frequency response is unity over a certain range of frequencies, described as the passband, and zero over the remaining frequencies, which constitute the stopband. In a lowpass filter ψ_L, the passband covers a frequency interval $[0, \omega_c]$ ranging from zero to a cut-off point. In the complementary highpass filter ψ_H, it is the stopband which stands on this interval. Thus

$$|\psi_L(e^{i\omega})| = \begin{cases} 1, & \text{if } \omega < \omega_c \\ 0, & \text{if } \omega > \omega_c \end{cases} \qquad \text{and} \qquad |\psi_H(e^{i\omega})| = \begin{cases} 0, & \text{if } \omega < \omega_c \\ 1, & \text{if } \omega > \omega_c. \end{cases} \tag{20}$$

In this section, we shall derive a pair of complementary filters that fulfil this specification approximately for a cut-off frequency of $\omega_c = \pi/2$. Once we have designed these prototype filters, we shall be able to apply a transformation that shifts the cut-off point from $\omega = \pi/2$ to any other point $\omega_c \in [0, \pi]$.

The idealised conditions of (20), which define a periodic square wave, are impossible to fulfil in practice. In fact, the Fourier transform of the square wave is an indefinite sequence of coefficients defined over the positive and negative integers; and, in constructing a practical moving-average filter, only a limited number of central coefficients can be taken. In such a filter, the sharp disjunction between the passband and the stopband, which characterises the ideal filter, is replaced by a gradual transition. The cost of a more rapid transition is bound to be an increased number of coefficients.

A preliminary step in designing a pair of complementary filters is to draw up a list of specifications that can be fulfilled in practice. We shall be guided by the following conditions:

$$\begin{aligned}
&\text{(i)} && \psi_L(z) + \psi_H(z) = 1, && \textit{Complementarity} \\
&\text{(ii)} && \psi_L(-z) = \psi_H(z), \quad \psi_H(-z) = \psi_L(z), && \textit{Symmetry} \\
&\text{(iii)} && \psi_L(z^{-1}) = \psi_L(z), \quad \psi_H(z^{-1}) = \psi_H(z), && \textit{Phase-Neutrality} \\
&\text{(iv)} && |\psi_L(1)| = 1, \quad |\psi_L(-1)| = 0, && \textit{Lowpass Conditions} \\
&\text{(v)} && |\psi_H(1)| = 0, \quad |\psi_H(-1)| = 1. && \textit{Highpass Conditions}
\end{aligned} \tag{21}$$

There is no reference here to the rate of the transition from the passband to the stopband. In fact, the condition under (iv) and (v) refer only to the end points of the frequency range $[0, \pi]$, which are the furthest points from the cut-off.

Observe that the symmetry condition $\psi_L(-z) = \psi_H(z)$ under (ii) necessitates placing the cut-off frequency at $\omega_c = \pi/2$. The condition implies that, when it is reflected about the axis of $\omega_c = \pi/2$, the frequency response of the lowpass filter becomes the frequency response of the highpass filter. This feature is illustrated by Fig. 10.

It will be found that all of the conditions of (21) are fulfilled by the highpass differencing filter ψ_D, defined under (11), in conjunction with the complementary lowpass smoothing filter $\psi_S = 1 - \psi_D$, defined under (14). However, we have already rejected ψ_D and ψ_S on the grounds that their transitions between the passband to the stopband are too gradual.

In order to minimise the problem of spectral leakage whilst maintaining a transition that is as rapid as possible, we now propose to fulfil the conditions of (21) via a pair of rational functions that take the forms of

$$\psi_L(z) = \frac{\delta_L(z)\delta_L(z^{-1})}{\gamma(z)\gamma(z^{-1})} \qquad \text{and} \qquad \psi_H(z) = \frac{\delta_H(z)\delta_H(z^{-1})}{\gamma(z)\gamma(z^{-1})}. \tag{22}$$

The condition of phase neutrality under (iii) is automatically satisfied by these forms. We propose to satisfy the lowpass and highpass conditions under (iv) and (v) by specifying that

$$\delta_L(z) = (1+z)^n \qquad \text{and} \qquad \delta_H(z) = (1-z)^n. \tag{23}$$

Similar specifications are also to be found in the binomial filter ψ_B of (15) and in the notch filter ψ_N of (18).

Given the specifications under (22), it follows that the symmetry condition of (ii) will be satisfied if and only if every root of $\gamma(z) = 0$ is a purely imaginary number. It follows from (i) that the polynomial $\gamma(z)$ must fulfil the condition that

$$\gamma(z)\gamma(z^{-1}) = \delta_L(z)\delta_L(z^{-1}) + \delta_H(z)\delta_H(z^{-1}). \tag{24}$$

On putting the specifications of (23) and (24) into (22), we find that

$$\begin{aligned} \psi_L(z) &= \frac{(1+z)^n(1+z^{-1})^n}{(1+z)^n(1+z^{-1})^n + (1-z)^n(1-z^{-1})^n} \\ &= \frac{1}{1 + \left(i\dfrac{1-z}{1+z}\right)^{2n}} \end{aligned} \tag{25}$$

and that

$$\psi_H(z) = \frac{(1-z)^n(1-z^{-1})^n}{(1+z)^n(1+z^{-1})^n + (1-z)^n(1-z^{-1})^n} \tag{26}$$
$$= \frac{1}{1+\left(i\dfrac{1+z}{1-z}\right)^{2n}}.$$

These will be recognised as instances of the Butterworth filter, which is familiar in electrical engineering—see, for example, Roberts and Mullis (1987).

The Butterworth filter, in common with the Hodrick–Prescott filter can also be derived by applying the Wiener–Kolmogorov theory of signal extraction to an appropriate statistical model. In that context, the filter represents a device for obtaining the minimum-mean-square-error estimate of the component in question. See Kolmogorov (1941) and Wiener (1950) for the original expositions of the theory and Whittle (1983) for a modern account.

A defining characteristic of the Wiener–Kolmogorov filters is the condition of complementarity of (21) (i). On that basis, we might also regard the complementary binomial filters $\psi_D(z)$ and $\psi_S(z)$ of (11) and (15), respectively, as Wiener–Kolmogorov filters; but they are unusual in being represented by polynomials of finite degree, whereas filters of this class are more commonly represented by rational functions.

Since $\delta_L(z)$ and $\delta_H(z)$ are now completely specified, it follows that $\gamma(z)$ can be determined via the Cramér–Wold factorisation of the polynomial of the RHS of (24). However, it is relatively straightforward to obtain analytic expressions for the roots of the equation $\gamma(z)\gamma(z^{-1}) = 0$. The roots come in reciprocal pairs; and, once they are available, they may be assigned unequivocally to the factors $\gamma(z)$ and $\gamma(z^{-1})$. Those roots which lie outside the unit circle belong to $\gamma(z)$ whilst their reciprocals, which lie inside the unit circle, belong to $\gamma(z^{-1})$. Therefore, consider the equation

$$(1+z)^n(1+z^{-1})^n + (1-z)^n(1-z^{-1})^n = 0, \tag{27}$$

which is equivalent to the equation

$$1+\left(i\frac{1-z}{1+z}\right)^{2n} = 0. \tag{28}$$

Solving the latter for

$$s = i\frac{1-z}{1+z} \tag{29}$$

is a matter of finding the $2n$ roots of -1. These are given by

$$s = \exp\left\{\frac{i\pi j}{2n}\right\}, \quad \text{where} \quad j = 1, 3, 5, \ldots, 4n-1, \tag{30}$$
$$\text{or} \quad j = 2k-1;\ k = 1, \ldots, 2n.$$

The roots correspond to a set of $2n$ points which are equally spaced around the circumference of the unit circle. The radii, which join the points to the

centre, are separated by angles of π/n; and the first of the radii makes an angle of $\pi/(2n)$ with the horizontal real axis.

The inverse of the function $s = s(z)$ is the function

$$z = \frac{i - s}{i + s} = \frac{i(s + s^*)}{2 - i(s - s^*)}. \tag{31}$$

Here, the final expression comes from multiplying top and bottom of the second expression by $s^* - i = (i + s)^*$, where s^* denotes the conjugate of the complex number s, and from noting that $ss^* = 1$. On substituting the expression for s from (29), it is found that the solutions of (28) are given, in terms of z, by

$$z_k = i\frac{\cos\{\pi(2k-1)/2n\}}{1 + \sin\{\pi(2k-1)/2n\}}, \quad \text{where} \quad k = 1, \ldots, 2n. \tag{32}$$

The roots of $\gamma(z^{-1}) = 0$ are generated when $k = 1, \ldots, n$. Those of $\gamma(z) = 0$ are generated when $k = n + 1, \ldots, 2n$.

Figure 9 shows the disposition in the complex plane of the poles and zeros of the prototype lowpass filter $\psi(z)_L$ for the case where $n = 6$, whilst Fig. 10 shows the gain of this filter together with that of the complementary filter $\psi(z)_H$.

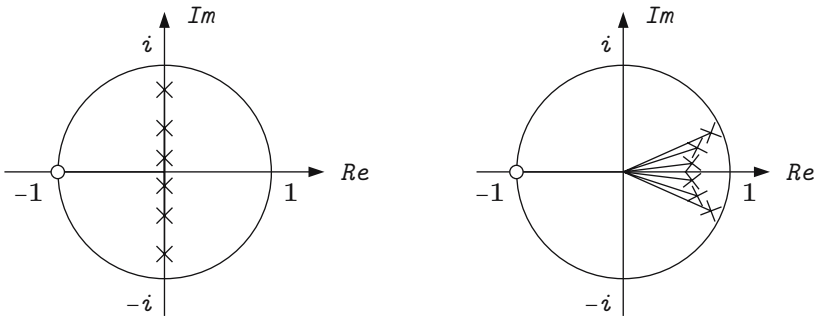

Fig. 9. The pole–zero diagrams of the lowpass square-wave filters for $n = 6$ when the cut-off is at $\omega = \pi/2$ (left) and at $\omega = \pi/8$

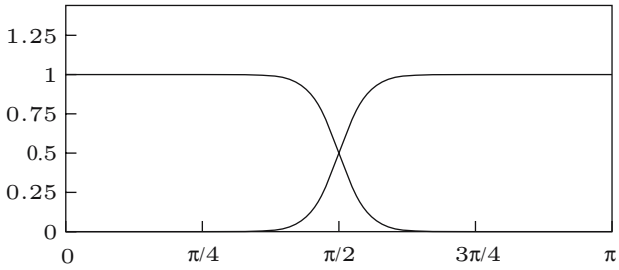

Fig. 10. The frequency-responses of the prototype square-wave filters with $n = 6$ and with a cut-off at $\omega = \pi/2$

5 Frequency Transformations

The object of the filter $\psi_L(z)$ is to remove from a time series a set of trend components whose frequencies range from $\omega = 0$ to a cut-off value of $\omega = \omega_c$. The prototype version of the filter has a cut-off at the frequency $\omega = \pi/2$. In order to convert the prototype filter to one that will serve the purpose, a means must be found for mapping the frequency interval $[0, \pi/2]$ into the interval $[0, \omega_c]$. This can be achieved by replacing the argument z, wherever it occurs in the filter formula, by the argument

$$g(z) = \frac{z - \alpha}{1 - \alpha z}, \tag{33}$$

where $\alpha = \alpha(\omega_c)$ is an appropriately specified parameter.

The function $g(z)$ fulfils the following conditions:

$$\begin{aligned}
&\text{(i)} && g(z)g(z^{-1}) = 1, \\
&\text{(ii)} && g(z) = z \quad \text{if} \quad \alpha = 0, \\
&\text{(iii)} && g(1) = 1 \quad \text{and} \quad g(-1) = -1, \\
&\text{(iv)} && \text{Arg}\big\{g(z)\big\} \geq \text{Arg}\{z\} \quad \text{if} \quad \alpha > 1, \\
&\text{(v)} && \text{Arg}\big\{g(z)\big\} \leq \text{Arg}\{z\} \quad \text{if} \quad \alpha < 1.
\end{aligned} \tag{34}$$

The conditions (i) and (ii) indicate that, if $g(z) \neq z$, then the modulus of the function is invariably unity. Thus, as z encircles the origin, $g = g(z)$ travels around the unit circle. The conditions of (iii) indicate that, if $z = e^{i\omega}$ travels around the unit circle, then g and z will coincide when $\omega = 0$ and when $\omega = \pi$—which are the values that bound the positive frequency range over which the transfer function of the filter is defined. Finally, conditions (iv) and (v) indicate that, if $g \neq z$, then g either leads z uniformly or lags behind it as the two travel around the unit circle from $z = 1$ to $z = -1$.

The value of α is completely determined by any pair of corresponding values for g and z. Thus, from (33), it follows that

$$\begin{aligned}
\alpha &= \frac{z - g}{1 - gz} \\
&= \frac{g^{1/2}z^{-1/2} - g^{-1/2}z^{1/2}}{g^{1/2}z^{1/2} - g^{-1/2}z^{-1/2}}.
\end{aligned} \tag{35}$$

Imagine that the cut-off of a prototype filter is at $\omega = \theta$ and that it is desired to shift it to $\omega = \kappa$. Then $z = e^{i\kappa}$ and $g = e^{i\theta}$ will be corresponding values; and the appropriate way of shifting the frequency would be to replace

the argument z within the filter formula by the function $g(z)$ wherein the parameter α is specified by

$$\begin{aligned}\alpha &= \frac{e^{i(\theta-\kappa)/2} - e^{-i(\theta-\kappa)/2}}{e^{i(\theta+\kappa)/2} - e^{-i(\theta+\kappa)/2}} \\ &= \frac{\sin\{(\theta-\kappa)/2\}}{\sin\{(\theta+\kappa)/2\}}.\end{aligned} \tag{36}$$

To find an explicit form for the transformed filter, we may begin by observing that, when $g(z)$ is defined by (33), we have

$$\frac{1-g(z)}{1+g(z)} = \left\{\frac{1+\alpha}{1-\alpha}\right\}\left\{\frac{1-z}{1+z}\right\}. \tag{37}$$

Here there is

$$\begin{aligned}\frac{1+\alpha}{1-\alpha} &= \frac{\sin\{(\theta+\kappa)/2\} + \sin\{(\theta-\kappa)/2\}}{\sin\{(\theta+\kappa)/2\} - \sin\{(\theta-\kappa)/2\}} \\ &= \frac{\sin(\theta/2)\cos(\kappa/2)}{\cos(\theta/2)\sin(\kappa/2)}.\end{aligned} \tag{38}$$

In the prototype filter, we are setting $\theta = \pi/2$ and, in the transformed filter, we are setting $\kappa = \omega_c$, which is the cut-off frequency. The result of these choices is that

$$\frac{1+\alpha}{1-\alpha} = \frac{1}{\tan(\omega_c/2)}. \tag{39}$$

It follows that the lowpass filter with a cut-off at ω_c takes the form of

$$\begin{aligned}\psi_L(z) &= \frac{1}{1+\lambda\left(i\dfrac{1-z}{1+z}\right)^{2n}} \\ &= \frac{(1+z)^n(1+z^{-1})^n}{(1+z)^n(1+z^{-1})^n + \lambda(1-z)^n(1-z^{-1})^n},\end{aligned} \tag{40}$$

where $\lambda = \{1/\tan(\omega_c)\}^{2n}$. The same reasoning shows that the highpass filter with a cut-off at ω_c takes the form of

$$\begin{aligned}\psi_H(z) &= \frac{1}{1+\dfrac{1}{\lambda}\left(i\dfrac{1+z}{1-z}\right)^{2n}} \\ &= \frac{\lambda(1-z)^n(1-z^{-1})^n}{(1+z)^n(1+z^{-1})^n + \lambda(1-z)^n(1-z^{-1})^n}.\end{aligned} \tag{41}$$

In applying the frequency transformation to the prototype filter, we are also concerned with finding revised values for the poles. The conditions

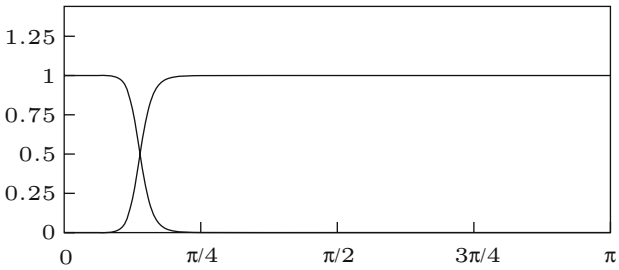

Fig. 11. The frequency-responses of the square-wave filters with $n = 6$ and with a cut-off at $\omega = \pi/8$

under (iii) indicate that the locations of the zeros will not be affected by the transformation. Only the poles will be altered. Consider, therefore, the generic factor within the denominator of the prototype. This is $z - i\rho$, where $i\rho$ is one of the poles specified under (30). Replacing z by $g(z)$ and setting the result to zero gives the following condition:

$$\frac{z - \alpha}{1 - \alpha z} - i\rho = 0. \tag{42}$$

This indicates that the pole at $z = \rho$ will be replaced by a pole at

$$z = \frac{\alpha + i\rho}{1 + i\rho\alpha} = \frac{\alpha(1 - \rho^2) + i\rho(1 - \alpha^2)}{1 - \rho^2\alpha^2}, \tag{43}$$

where the final expression comes from multiplying top and bottom of its predecessor by $1 - i\rho\alpha$.

Figure 11, displays the pole-zero diagram of the prototype filter and of a filter with a cut-off frequency of $\pi/8$. It also suggests that one of the effects of a frequency transformation may be to bring some of poles closer to the perimeter of the unit circle. This can lead to stability problems in implementing the filter, and it is liable to prolong the transient effects of ill-chosen start-up conditions.

6 Implementing the Filters

The classical signal-extraction filters are intended to be applied to lengthy data sets. The task of adapting them to limited samples often causes difficulties and perplexity. The problems arise from not knowing how to supply the initial conditions with which to start a recursive filtering process. By choosing inappropriate starting values for the forwards or the backwards pass, one can generate a so-called transient effect, which is liable, in fact, to affect all of the processed values.

Of course, when the values of interest are remote from either end of a long sample, one can trust that they will be barely affected by the start-up conditions. However, in many applications, such as in the processing of

economic data, the sample is short and the interest is concentrated at the upper end where the most recent observations are to be found.

One approach to the problem of the start-up conditions relies upon the ability to extend the sample by forecasting and backcasting. The additional exta-sample values can be used in a run-up to the filtering process wherein the filter is stabilised by providing it with a plausible history, if it is working in the direction of time, of with a plausible future, if it is working in reversed time. Sometimes, very lengthy extrapolations are called for—see Burman (1980), for example.

The approach that we shall adopt in this paper is to avoid the start-up problem altogether by deriving specialised finite-sample versions of the filters on the basis of the statistical theory of conditional expectations.

Some of the more successful methods for treating the problem of the start-up conditions that have been proposed have arisen within the context of the Kalman filter and the associated smoothing algorithms—see Ansley and Kohn (1985), De Jong (1991), and Durbin and Koopman (2001), for example. The context of the Kalman filter is a wide one; and it seems that the necessary results can be obtained more easily by restricting the context.

Let us begin, therefore, by considering a specific model for which the square-wave filter would represent the optimal device for extracting the signal, given a sample of infinite length. The model is represented by the equation

$$\begin{aligned} y(t) &= \xi(t) + \eta(t) \\ &= \frac{(1+L)^n}{(1-L)^2}\nu(t) + (1-L)^{n-2}\varepsilon(t), \end{aligned} \tag{44}$$

where $\nu(t)$ and $\varepsilon(t)$ are statistically independent sequences generated by normal white-noise processes. This can be rewritten as

$$\begin{aligned} (1-L)^2 y(t) &= (1+L)^n \nu(t) + (1-L)^n \varepsilon(t) \\ &= \zeta(t) + \kappa(t), \end{aligned} \tag{45}$$

where $\zeta(t) = (1-L)^2\xi(t) = (1+L)^n\nu(t)$ and $\kappa(t) = (1-L)^2\eta(t) = (1-L)^n\varepsilon(t)$ both follow noninvertible moving-average processes.

The statistical theory of signal extraction, as expounded by Whittle (1983), for example, indicates that the lowpass filter $\psi_L(z)$ of (40) will generate the minimum mean-square-error estimate of the sequence $\xi(t)$, provided that the smoothing parameter has the value of $\lambda = \sigma_\varepsilon^2/\sigma_\nu^2$. The theory also indicates that the Hodrick–Prescott filter will generate the optimal estimate in the case where $\xi(t)$ is a second-order random walk and $\eta(t)$ is a white-noise process:

$$\begin{aligned} y(t) &= \xi(t) + \eta(t) \\ &= \frac{1}{(1-L)^2}\nu(t) + \eta(t). \end{aligned} \tag{46}$$

Now imagine that there are T observations of the process $y(t)$ of (44), which run from $t = 0$, to $t = T - 1$. These are gathered in a vector

$$y = \xi + \eta. \tag{47}$$

To find the finite-sample the counterpart of (45), we need to represent the second-order difference operator $(1 - L)^2$ in the form of a matrix. The matrix that finds the differences $d_2, \ldots, d_{T-1}$ of the data points $y_0, y_1, y_2, \ldots, y_{T-1}$ is in the form of

$$Q' = \begin{bmatrix} 1 & -2 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & 0 & \ldots & -2 & 1 \end{bmatrix}. \tag{48}$$

Premultiplying (47) by this matrix gives

$$\begin{aligned} d = Q'y &= Q'\xi + Q'\eta \\ &= \zeta + \kappa, \end{aligned} \tag{49}$$

where $\zeta = Q'\xi$ and $\kappa = Q'\eta$. The first and second moments of the vector ζ may be denoted by

$$E(\zeta) = 0 \qquad \text{and} \qquad D(\zeta) = \sigma_\nu^2 M, \tag{50}$$

and those of κ by

$$\begin{aligned} E(\kappa) = 0 \qquad \text{and} \qquad D(\kappa) &= Q'D(\eta)Q \\ &= \sigma_\varepsilon^2 Q'\Sigma Q, \end{aligned} \tag{51}$$

where both M and $Q'\Sigma Q$ are symmetric Toeplitz matrices with $2n+1$ nonzero diagonal bands. The generating functions for the coefficients of these matrices are, respectively, $\delta_L(z)\delta_L(z^{-1})$ and $\delta_H(z)\delta_H(z^{-1})$, where $\delta_L(z)$ and $\delta_H(z)$ are the polynomials defined in (23).

The optimal predictor z of the twice-differenced signal vector $\zeta = Q'\xi$ is given by the following conditional expectation:

$$\begin{aligned} E(\zeta|d) &= E(\zeta) + C(\zeta, d)D^{-1}(d)\big\{d - E(d)\big\} \\ &= M(M + \lambda Q'\Sigma Q)^{-1}d = z, \end{aligned} \tag{52}$$

where $\lambda = \sigma_\varepsilon^2/\sigma_\nu^2$. The optimal predictor k of the twice-differenced noise vector $\kappa = Q'\eta$ is given, likewise, by

$$\begin{aligned} E(\kappa|d) &= E(\kappa) + C(\kappa, d)D^{-1}(d)\big\{d - E(d)\big\} \\ &= \lambda Q'\Sigma Q(M + \lambda Q'\Sigma Q)^{-1}d = k. \end{aligned} \tag{53}$$

It may be confirmed that $z + k = d$.

The estimates are calculated, first, by solving the equation

$$(M + \lambda Q'\Sigma Q)g = d \tag{54}$$

for the value of g and, thereafter, by finding

$$z = Mg \qquad \text{and} \qquad k = \lambda Q'\Sigma Qg. \tag{55}$$

The solution of (54) is found via a Cholesky factorisation which sets $M + \lambda Q'\Sigma Q = GG'$, where G is a lower-triangular matrix. The system $GG'g = d$ may be cast in the form of $Gh = d$ and solved for h. Then $G'g = h$ can be solved for g.

There is a straightforward correspondence between the finite-sample implementations of the filter and the formulations that assume an infinite sample. In terms of the lag-operator polynomials, (54) would be rendered as

$$\begin{aligned} \gamma(F)\gamma(L)g(t) &= d(t), \qquad \text{where} \\ \gamma(F)\gamma(L) &= \delta_L(F)\delta_L(L) + \lambda\delta_H(F)\delta_H(L). \end{aligned} \tag{56}$$

The process of solving (54) via a Cholesky decomposition corresponds to the application of the filter in separate passes running forwards and backwards in time respectively:

$$\text{(i)} \quad \gamma(L)f(t) = d(t) \qquad \text{(ii)} \quad \gamma(F)g(t) = f(t). \tag{57}$$

The coefficients of successive rows of the Cholesky factor G converge upon the values of the coefficients of $\gamma(z)$; and, at some point, it may become appropriate to use the latter instead. This will save computer time and computer memory.

The two equations under (55) correspond respectively to

$$z(t) = \delta_L(F)\delta_L(L)g(t) \qquad \text{and} \qquad k(t) = \delta_H(F)\delta_H(L)q(t). \tag{58}$$

Our object is to recover from z an estimate x of the trend vector ξ. This would be conceived, ordinarily, as a matter of integrating the vector z twice via a simple recursion which depends upon two initial conditions. The difficulty is in discovering the appropriate initial conditions with which to begin the recursion.

We can circumvent the problem of the initial conditions by seeking the solution to the following problem:

$$\text{Minimise} \qquad (y - x)'\Sigma^{-1}(y - x) \qquad \text{Subject to} \qquad Q'x = z. \tag{59}$$

The problem is addressed by evaluating the Lagrangean function

$$L(x, \mu) = (y - x)'\Sigma^{-1}(y - x) + 2\mu'(Q'x - z). \tag{60}$$

By differentiating the function with respect to x and setting the result to zero, we obtain the condition

$$\Sigma^{-1}(y - x) - Q\mu = 0. \tag{61}$$

Premultiplying by $Q'\Sigma$ gives

$$Q'(y - x) = Q'\Sigma Q\mu. \tag{62}$$

But, from (54) and (55), it follows that

$$\begin{aligned} Q'(y - x) &= d - z \\ &= \lambda Q'\Sigma Qg, \end{aligned} \tag{63}$$

whence we get

$$\begin{aligned} \mu &= (Q'\Sigma Q)^{-1}Q'(y - x) \\ &= \lambda g. \end{aligned} \tag{64}$$

Putting the final expression for μ into (61) gives

$$x = y - \lambda\Sigma Qg. \tag{65}$$

This is our solution to the problem of estimating the trend vector ξ. Notice that there is no need to find the value of z explicitly, since the value of x can be expressed more directly in terms of $g = \Sigma^{-1}z$.

It is notable that there is a criterion function which will enable us to derive the equation of the trend estimation filter in a single step. The function is

$$L(x) = (y - x)'\Sigma^{-1}(y - x) + \lambda x'QM^{-1}Q'x, \tag{66}$$

wherein $\lambda = \sigma_\varepsilon^2/\sigma_\nu^2$ as before. This is minimised by the value specified in (65). The criterion function becomes intelligible when we allude to the assumptions that $y \sim N(\xi, \sigma_\varepsilon^2\Sigma)$ and that $Q'\xi = \zeta \sim N(0, \sigma_\nu^2 M)$; for then it plainly resembles a combination of two independent chi-square variates.

The effect of the square-wave filter is illustrated in Figs. 12–14 which depict the detrending of the logarithmic series of the U.S. money stock. It is notable that, in contrast to periodogram of Fig. 3, which relates to the the residuals from fitting a polynomial trend, the periodogram of Fig. 14 shows virtually no power in the range of frequencies below that of the principal seasonal frequency.

We should point out that our derivation and the main features of our algorithm are equally applicable to the task of implementing the Hodrick–Prescott (H–P) filter and the Reinsch smoothing spline. In the case of the H–P filter, we need only replace the matrices Σ and M in the equations above by the matrices I and $Q'Q$ respectively. Then (52) becomes

$$(I + \lambda Q'Q)^{-1}d = z, \tag{67}$$

whilst (65), which provides the estimate of the signal or trend, becomes

$$x = y - \lambda Qz. \tag{68}$$

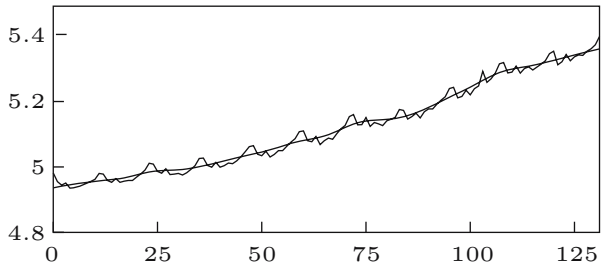

Fig. 12. The data on the U.S. money stock with an interpolated trend estimated by a lowpass square-wave filter with $n = 6$ and a cut off at $\omega = \pi/8$

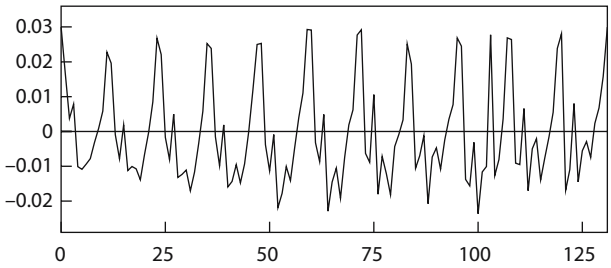

Fig. 13. The residual sequence obtained by detrending the logarithm of the money stock data with a square-wave filter

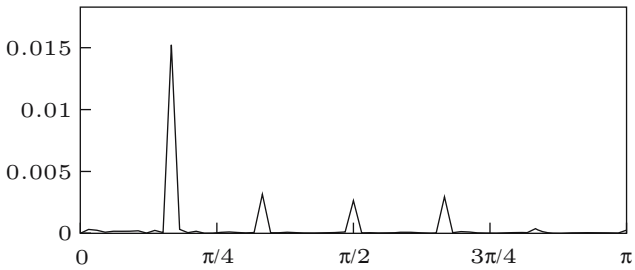

Fig. 14. The periodogram of the residuals from detrending the logarithm of the U.S. money stock data

References

Ansley, C.F., and R. Kohn, (1985), Estimation, Filtering and Smoothing in State Space Models with Incompletely Specified Initial Conditions, *The Annals of Statistics,* **13,** 1286–1316.

Burman, J.P., (1980), Seasonal Adjustment by Signal Extraction, *Journal of the Royal Statistical Society, Series A,* **143,** 321–337.

Cogley, T., and J.M. Nason, (1995), Effects of the Hodrick–Prescott Filter on Trend and Difference Stationary Time Series, Implications for Business Cycle Research, *Journal of Economic Dynamics and Control,* **19,** 253–278.

De Jong, P., (1991), The Diffuse Kalman Filter, *The Annals of Statistics,* **19,** 1073–1083.

Durbin, J., and S.J. Koopman, (2001), *Time Series Analysis by State Space Methods,* Oxford University Press.

Harvey, A.C., and A. Jaeger, (1993), Detrending, Stylised Facts and the Business Cycle, *Journal of Applied Econometrics,* **8,** 231–247.

Haykin, S., (1989), *Modern Filters, Macmillan Publishing Company,* New York.

Hodrick, R.J., and E.C Prescott, (1980), *Postwar U.S. Business Cycles: An Empirical Investigation,* Working Paper, Carnegie–Mellon University, Pittsburgh, Pennsylvania.

Hodrick R.J., and Prescott, E.C., (1997), Postwar U.S. business bycles: An Empirical Investigation, Journal of Money, Credit and Banking, **29,** 1–16.

King, R.G., and S.G. Rebelo, (1993), Low Frequency Filtering and Real Business Cycles, *Journal of Economic Dynamics and Control,* **17,** 207–231.

Kolmogorov, A.N., (1941), Interpolation and Extrapolation. *Bulletin de l'academie des sciences de U.S.S.R.,* Ser. Math., **5,** 3–14.

Kydland, F.E., and C. Prescott, (1990), Business Cycles: Real Facts and a Monetary Myth, *Federal Reserve Bank of Minneapolis Quarterly Review,* **14,** 3–18.

Oppenheim A.V. and R.W. Schafer, (1989), *Discrete-Time Signal Processing,* Prentice-Hall, Englewood Cliffs, New Jersey.

Pollock, D.S.G., (1999), *Time-Series Analysis, Signal Processing and Dynamics,* The Academic Press, London.

Reinsch, C.H., (1976), Smoothing by Spline Functions, *Numerische Mathematik,* **10,** 177–183.

Roberts, R.A., and C.T. Mullis, (1987), *Digital Signal Processing,* Addison Wesley, Reading, Massachusetts.

Whittle, P., (1983), *Prediction and Regulation by Linear Least-Square Methods, Second Revised Edition,* Basil Blackwell, Oxford.

Wiener, N., (1950), *Extrapolation, Interpolation and Smoothing of Stationary Time Series,* MIT Technology Press, John Wiley and Sons, New York.

Non-Dyadic Wavelet Analysis

Stephen Pollock and Iolanda Lo Cascio

Department of Economics, Queen Mary College, University of London,
Mile End Road, London E1 4NS, UK

Summary. The conventional dyadic multiresolution analysis constructs a succession of frequency intervals in the form of $(\pi/2^j, \pi/2^{j-1}); j = 1, 2, \ldots, n$ of which the bandwidths are halved repeatedly in the descent from high frequencies to low frequencies. Whereas this scheme provides an excellent framework for encoding and transmitting signals with a high degree of data compression, it is less appropriate to statistical data analysis. A non-dyadic mixed-radix wavelet analysis which allows the wave bands to be defined more flexibly than in the case of a conventional dyadic analysis is described. The wavelets that form the basis vectors for the wave bands are derived from the Fourier transforms of a variety of functions that specify the frequency responses of the filters corresponding to the sequences of wavelet coefficients.

Key words: Wavelet analysis, signal extraction

1 Introduction: Dyadic and Non-Dyadic Wavelet Analysis

The statistical analysis of time series can be pursued either in the time domain or in the frequency domain, or in both. A time-domain analysis will reveal the sequence of events within the data, so long as the events do not coincide. A frequency-domain analysis, which describes the data in terms of sinusoidal functions, will reveal its component sequences, whenever they subsist in separate frequency bands. The analyses in both domains are commonly based on the assumption of stationarity. If the assumption is not satisfied, then, often, a transformation can be applied to the data to make them resemble a stationary series. For a stationary series, the results that are revealed in one domain can be transformed readily into equivalent results in the other domain.

The revolution in statistical Fourier analysis that occurred in the middle of the twentieth century established the equivalence of the two domains under the weak assumption of statistical stationarity. Previously, it had seemed that

frequency-domain analysis was fully applicable only to strictly periodic functions of a piecewise continuous nature. However, the additional flexibility of statistical Fourier analysis is not sufficient to cope with phenomena that are truly evolving through time. A sufficient flexibility to deal with evolutionary phenomena can be achieved by combining the time domain and the frequency domain in a so-called wavelet analysis.

The replacement of classical Fourier analysis by wave packet analysis occurred in the realms of quantum mechanics many years ago when Schrödinger's time-dependent wave equation became the model for all sorts of electromagnetic phenomena. (See Dirac 1958, for example.) This was when the dual wave-particle analogy of light superseded the classical wave analogy that had displaced the ancient corpuscular theory. It is only recently, at the end of the twentieth century, that formalisms that are similar to those of quantum mechanics have penetrated statistical time-series analysis. The result has been the new and rapidly growing field of wavelet analysis.

The common form of dyadic wavelet analysis entails a partition of the time-frequency plane of the sort that is depicted in Fig. 1, which relates to the wavelet analysis of a sample of $T = 2^7 = 128$ points. The wavelets are functions of continuous time that reside in a succession of horizontal frequency bands. Each band contains a succession of wavelets, distributed over time, of which the centres lie in the cells that partition the band. Within a given band, the wavelets have a common frequency content and a common temporal dispersion, but their amplitude, which is their vertical scale, is free to vary. As we proceed down the frequency scale from one band to the next, the bandwidth of the frequencies is halved and the temporal dispersion of the wavelets, which is reflected in the width of the cells, is doubled.

The wavelet bands are created by a recursive process of subdivision. In the first round, the frequency range is divided in two. The upper band $[\pi/2, \pi]$ is populated by $T/2$ wavelets, separated, one from the next, by two sampling intervals, and the lower band $[0, \pi/2]$ is populated by the same number of scaling functions in a similar sequence. Thus, there are as many functions as there are data points. In the next round, the lower half of the frequency

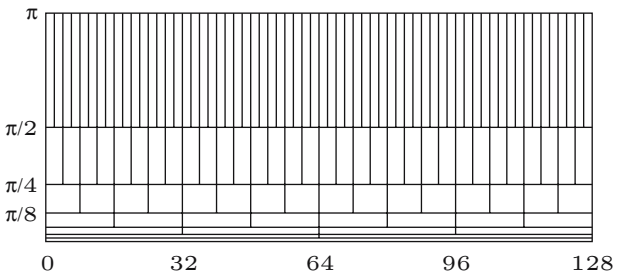

Fig. 1. The partitioning of the time–frequency plane according to a dyadic multiresolution analysis of a data sequence of $T = 128 = 2^7$ points

range is subdivided into an upper band $[\pi/4, \pi/2]$ of wavelets and a lower band $[0, \pi/4]$ of scaling functions, with both bands containing $T/4$ functions, separated by four intervals. The process can be repeated such that, in the jth round, the jth band is divided into an upper band of wavelets and a lower band of scaling functions, with $T/2^j$ functions in each. If that number is no longer divisible by 2, then the process must terminate. However, if $T = 2^n$, as is the case for Fig. 1, then it can be continued through n rounds until the nth band contains a single wavelet, and there is a constant function to accompany it in place of a scaling function.

The object of the wavelet analysis is to associate an amplitude coefficient to each of the wavelets. The variation in the amplitude coefficients enables a wavelet analysis to reflect the changing structure of a non-stationary time series. By contrast, the amplitude coefficients that are associated with the sinusoidal basis functions of a Fourier analysis remain constant throughout the sample. Accounts of wavelet analysis, which place it within the context of Fourier analysis, have been given by Newland (1993) and by Boggess and Narcowich (2001). Other accessible accounts have been given by Burrus, Gopinath and Guo (1998) and by Misiti, Misiti, Oppenheim and Poggi (1997) in the user's guide to the MATLAB Wavelets Toolbox.

The wavelets that are employed within the dyadic scheme are usually designed to be mutually orthogonal. They can be selected from a wide range of wavelet families. The most commonly employed wavelets are from the Daubechies (1988), (1992) family. Figures 2 and 3 display the level-1 D4

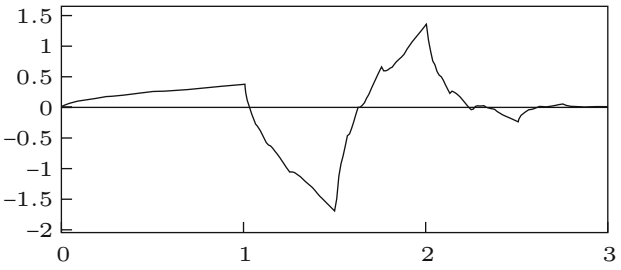

Fig. 2. The Daubechies D4 wavelet function calculated via a recursive method

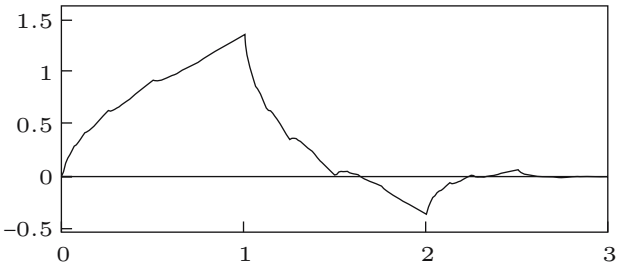

Fig. 3. The Daubechies D4 scaling function calculated via a recursive method

Daubechies wavelet and scaling function, which are generated on the first division of the time-frequency plane, and which span the upper and the lower halves of the frequency range $[0, \pi]$, respectively. These are highly localised continuous functions of a fractal nature that have finite supports with a width of three sampling intervals. The Daubechies wavelets have no available analytic forms, and they are not readily available in sampled versions. They are defined, in effect, by the associated dilation coefficients. These express a wavelet in one frequency band and a scaling function in the band below—which has the same width and which stretches to zero—as a linear combination of the more densely packed and less dispersed scaling functions that form a basis for the two bands in combination.

The fact that the Daubechies wavelets are know only via their dilation coefficients is no impediment to the discrete wavelet transform. This transform generates the amplitude coefficients associated with the wavelet decomposition of a data sequence; and it is accomplished via the pyramid algorithm of Mallat (1989). The continuous-time wavelets are, in reality, a shadowy accompaniment—and, in some ways, an inessential one—of a discrete-time analysis that can be recognised as an application of the techniques of multirate filtering, which are nowadays prevalent in communications engineering. (For examples, see Vaidyanathan 1993, Strang and Nguyen 1997 and Vetterli and Kovacević 1995.) In this perspective, the dilation coefficients of the wavelets and of the associated scaling functions are nothing but the coefficients of a pair of quadrature mirror filters that are applied in successive iterations of the pyramid algorithm. This uncommon relationship between the continuous-time and the discrete-time aspects of the analysis is undoubtedly the cause of many conceptual difficulties.

The Daubechies–Mallat paradigm has been very successful in application to a wide range of signal processing problems, particularly in audio-acoustic analysis and in the analysis of digitised picture images, which are two-dimensional signals in other words. There are at least two reasons for this success. The first concerns the efficiency of the pyramid algorithm, which is ideal for rapid processing in real time. The second reason lies in the Daubechies wavelets themselves. Their restricted supports are a feature that greatly assists the computations. This feature, allied to the sharp peaks of the wavelets, also assists in the detection of edges and boundaries in images.

The system of Daubechies and Mallat is not suited to all aspects of statistical signal extraction. For a start, the Daubechies wavelets might not be the appropriate ones to select. Their disjunct nature can contrast with the smoother and more persistent motions that underlie the data. The non-availability of their discretely sampled versions may prove to be an impediment; and the asymmetric nature of the associated dilation coefficients might conflict with the requirement, which is commonly imposed upon digital filters, that there should be no phase effects. (The absence of phase effects is important when, for example, wavelets are used as an adjunct to transfer-function modelling, as in the investigations of Ramsey and Lampart

1998 and of Nason and Sapatinas 2002.) A more fundamental difficulty lies in the nature of the dyadic decomposition. In statistical analyses, the structures to be investigated are unlikely to fall neatly into dyadic time and frequency bands, such as those of Fig. 1; and the frequency bands need to be placed wherever the phenomena of interest happen to be located.

For an example of a statistical data series that requires a more flexible form of wavelet analysis, we might consider the familiar monthly airline passenger data of Box and Jenkins (1976), depicted in Fig. 4, which comprises $T = 144 = 3^2 \times 2^4$ data points. The detrended series, which is obtained by taking the residuals from fitting a quadratic function to the logarithms of the data, is shown in Fig. 5. The detrended data manifest a clear pattern of seasonality, which is slowly evolving in a manner that is readily intelligible if one thinks of the development of air travel over the period in question—the summer peak in air travel was increasing relative to the winter peak throughout the period. The components of the seasonal pattern lie in and around a set of harmonically related frequencies $\{\pi j/6; j = 1, \ldots, 6\}$. This can be seen in Fig. 6, which displays the periodogram of the seasonal fluctuations.

In order to capture the evolving seasonal pattern, one might apply a wavelet analysis to some narrow bands surrounding the seasonal frequencies. To isolate bands extending for 5 degrees on either side of the seasonal frequencies, (excepting the frequency of π, where there is nothing above,) one

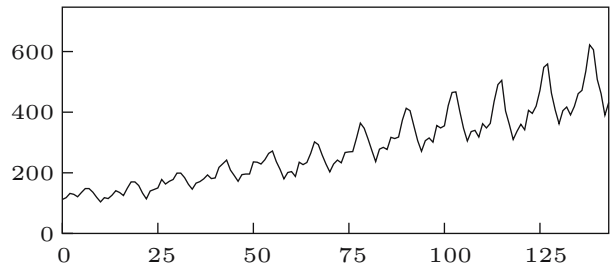

Fig. 4. International airline passengers: monthly totals (thousands of passengers) January 1949–December 1960: 144 observations

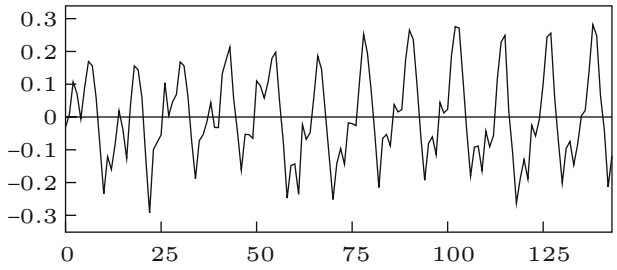

Fig. 5. The seasonal fluctuation in the airline passenger series, represented by the residuals from fitting a quadratic function to the logarithms of the series

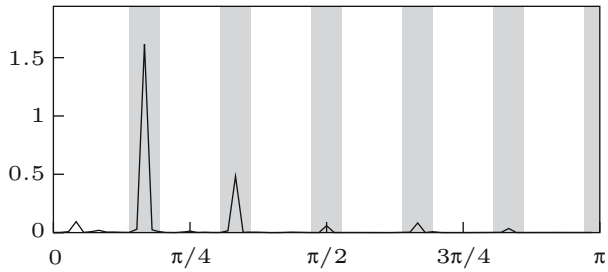

Fig. 6. The periodogram of the seasonal fluctuations in the airline passenger series

must begin by dividing the frequency range in $36 = 3^2 \times 2^2$ equal bands. The requisite wavelets will be obtained by dilating the first-level wavelet by a factor of 3 as well as by the dyadic factor of 2. These bands are indicated on Fig. 6. The other choices for the bandwidths would be 6 degrees, $7\frac{1}{2}$ degrees 10, degrees and 15 degrees—the latter affording no interstices between the bands.

2 The Aims of the Paper

The intention of this paper is to provide the framework for a flexible method of wavelet analysis that is appropriate to nonstationary data that have been generated by evolving structures that fall within non-dyadic frequency bands. For this purpose, we have to consider collections of wavelets and filters that are related to each other by dilation factors in addition to the factor of 2. At the same time, we shall endeavour to accommodate samples of all sizes, thereby relieving the restriction that $T = 2^n$, which is necessary for a complete dyadic decomposition.

We shall use the so-called Shannon wavelet as a prototype, since it is readily amenable to dilations by arbitrary factors. Since the Shannon wavelets are defined by a simple analytic function, their sampled versions are readily available; and their ordinates constitute the coefficients of symmetric digital filters that have no phase effects.

Thus, in the case of the Shannon wavelets, the connection between the continuous-time analysis and the discrete-time analysis is uniquely straightforward: the sampled ordinates of the wavelets and scaling functions constitute the filter coefficients of the discrete-time analysis, which are also the coefficients of the dilation relationships. The orthogonality conditions that affect the Shannon wavelets are easy to demonstrate. The conditions are important in a statistical analysis, since they enable the testing of hypotheses to proceed on the basis of simple chi-square statistics.

The disadvantage of the Shannon wavelets is in their wide dispersion. They have an infinite support, which is the entire real line. However, they can be

adapted to the analysis of a finite data sequence of T points by wrapping their sampled coefficients around a circle of circumference T and by adding the coincident coefficients. The wrapping is achieved by sampling the corresponding energy functions in the frequency domain at regular intervals. The wavelet coefficients in the time domain may be obtained by applying the discrete Fourier transform to the square roots of the ordinates sampled from the energy functions.

The band limitation of the energy functions enhances the efficiency of computations performed in the frequency domain, which entail simple multiplications or modulations. At the same time, it prejudices the efficiency of computations performed in the time domain, which entail the circular convolutions of sequences of length T. For this reason, we choose to conduct our filtering operations in the frequency domain. The mixed-radix fast Fourier transform of Pollock (1999) may be used to carry the data into the frequency domain; and it may be used, in its inverse mode, to carry the products of the filtering operations back to the time domain.

Despite the availability of these techniques for dealing with finite samples, the wide dispersion of the Shannon wavelets remains one of their significant disadvantages. Therefore, we must also look for wavelets of lesser dispersion. It is true that the Daubechies wavelets that have finite supports can be adapted to a non-dyadic analysis. Nevertheless, we choose to look elsewhere for our wavelets. Our recourse will be to derive the wavelets from energy functions specified in the frequency domain. By increasing the dispersion of these frequency-domain functions, we succeed in decreasing the dispersion of the corresponding wavelets in the time domain.

Much of what transpires in this paper may be regarded as an attempt to preserve the salient properties of the Shannon wavelets while reducing their dispersion in the time domain. In particular, we shall endeavour to maintain the conditions of sequential orthogonality between wavelets in the same band that are manifest amongst the Shannon wavelets. We shall also preserve the symmetry of the wavelets. The cost of doing so is that we must forego the conditions of orthogonality between wavelets in adjacent bands. However, the mutual orthogonality between wavelets in non-adjacent bands will be preserved. The latter conditions are appropriate to the analysis of spectral structures that are separated by intervening dead spaces. The seasonal structures within the airline passenger data, revealed by Fig. 6, provide a case in point.

Before embarking on our own endeavours, we should make some reference to related work. First, it should be acknowledged that a considerable amount of work has been done already in pursuit of a non-dyadic wavelet analysis. The objective can be described as that of partitioning the time–frequency plane in ways that differ from that of the standard dyadic analysis, represented in Fig. 1, and of generating the wavelets to accompany the various schemes.

A program for generalising the standard dyadic analysis has led to the so-called wavelet packet analysis, of which Wickerhauser (1994) is one of the

principal exponents. An extensive account has also been provided by Percival and Walden (2000). The essential aim, at the outset, is to decompose the frequency interval $[0, \pi]$ into 2^j equal intervals. Thereafter, a rich variety of strategies are available.

An alternative approach has been developed under the rubric of M-band wavelet analysis. This uses a particular type of filter bank architecture to create M equal subdivisions of each of the octave bands of a dyadic analysis. Seminal contributions have been made by Gopinath and Burrus (1993) and by Steffen, Heller, Gopinath and Burrus (1993). The work of Vaidyanathan (1990), (1993) on filter banks has also been influential in this connection.

Next, there is the matter of the uses of wavelets in statistical analysis. Here, the developments have been far too diverse and extensive for us to give a reasonable list of citations. However, it is appropriate to draw attention to a special issue of the *Philosophical Transactions of the Royal Society of London* that has been devoted to the area. Amongst other pieces, it contains an article by Ramsey (1999), which deals with application of wavelets to financial matters, and a survey by Nason and von Sachs (1999), which covers a wide range of statistical issues.

3 The Shannon Wavelets

The Shannon wavelet, which is also known as the sinc function, arises from an attempt to derive a time-localised function from an ordinary trigonometrical function. It is the result of applying a hyperbolic taper to the sine wave to give

$$\operatorname{sinc}(\omega t) = \frac{\sin(\omega t)}{\pi t}. \tag{1}$$

Woodward (1953) was responsible for naming the sinc function. It has been called the Shannon function in recognition of its central *role* in the Shannon–Nyquist sampling theory—see, for example, Shannon and Weaver (1964) or Boggess and Narcowich (2001).

The Figs. 7–9 plot the functions

$$\begin{aligned} \phi_{(0)}(t) &= \frac{\sin(\pi t)}{\pi t}, \\ \phi_{(1)}(t) &= \frac{\sin(\pi t/2)}{\pi t}, \\ \psi_{(1)}(t) &= \frac{\cos(\pi t)\sin(\pi t/2)}{\pi t}, \end{aligned} \tag{2}$$

both for $t \in \mathcal{R}$, which is the real line, and for $t \in \mathcal{I} = \{0, \pm 1, \pm 2, \ldots\}$, which is the set of integers representing the points at which the data are sampled. Here, $\phi_{(0)}(t)$ is the fundamental scaling function, whereas $\phi_{(1)}(t)$ is the scaling function at level 1 and $\psi_{(1)}(t)$ is the level-1 wavelet.

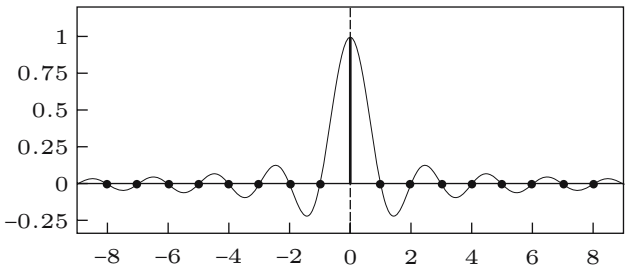

Fig. 7. The scaling function $\phi_{(0)}(t)$

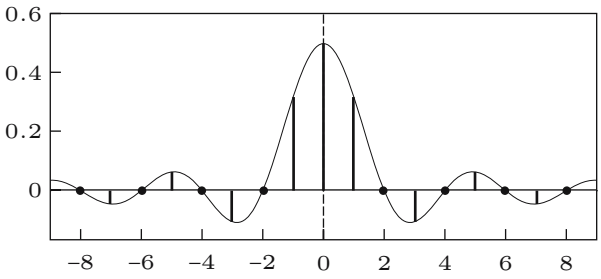

Fig. 8. The scaling function $\phi_{(1)}(t) = \phi_{(0)}(t/2)$

These time-domain functions with $t \in \mathcal{R}$ are the Fourier transforms of the following square-wave or boxcar functions defined in the frequency domain:

$$\phi_{(0)}(\omega) = \begin{cases} 1, & \text{if } |\omega| \in (0, \pi); \\ 1/2, & \text{if } \omega = \pm\pi, \\ 0, & \text{otherwise} \end{cases}$$

$$\phi_{(1)}(\omega) = \begin{cases} 1, & \text{if } |\omega| \in (0, \pi/2); \\ 1/\sqrt{2}, & \text{if } \omega = \pm\pi/2, \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$\psi_{(1)}(\omega) = \begin{cases} 1, & \text{if } |\omega| \in (\pi/2, \pi); \\ 1/\sqrt{2}, & \text{if } \omega = \pm\pi/2, \\ 1/2, & \text{if } \omega = \pm\pi, \\ 0, & \text{otherwise} \end{cases}$$

Here and elsewhere, we are using the same symbols to denote the time-domain functions and the frequency-domain functions that are their Fourier transforms. The arguments of the functions alone will serve to make the distinctions.

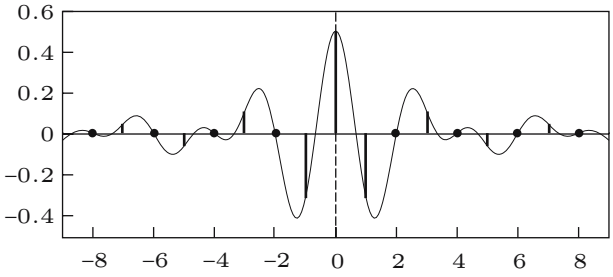

Fig. 9. The wavelet function $\psi_{(1)}(t) = \cos(\pi t)\phi_{(0)}(t/2)$

Within the frequency interval $[-\pi, \pi]$ on the real line, the points $\pm\pi$ and $\pm\pi/2$ constitute a set of measure zero. Therefore, any finite values can be attributed to the ordinates of the functions at these points without affecting the values of their transforms, which are the functions of (2). It is when the frequency-domain functions are sampled at a finite set of points, including the points in question, that it becomes crucial to adhere to the precise specifications of (3).

When $t \in \mathcal{I}$, the time-domain functions of (2) become sequences that correspond to periodic functions in the frequency domain, with a period of 2π radians. These functions are derived by superimposing copies of the aperiodic functions of (3) displaced successively by 2π radians in both positive and negative directions. Thus, for example, the periodic function derived from $\phi_{(0)}(\omega)$ is

$$\tilde{\phi}_{(0)}(\omega) = \sum_{j=-\infty}^{\infty} \phi_{(0)}(\omega + 2\pi j). \tag{4}$$

which is just a constant function with a value of unity.

We are defining the periodic functions in terms of the closed intervals $[(2j-1)\pi, (2j+1)\pi]; j \in \mathcal{I}$, such that adjacent intervals have a common endpoint. This is subject to the proviso that only half the value of the ordinate at the common endpoint is attributed to each interval. An alternative recourse, to which we resort elsewhere, is to define the periodic functions in terms of the non-overlapping half-open intervals such as $[2\pi[j-1], 2\pi j)$ and to attribute to the included endpoint the full value of its ordinate.

The time-domain sequences also constitute the coefficients of the ideal frequency-selective filters of which the above-mentioned periodic functions constitute the frequency responses. Given that the frequency responses are real-valued in consequence of the symmetry of the time-domain sequences, they can also be described as the amplitude responses or the gain functions of the filters. In the case of the Shannon wavelet, the periodic frequency functions also represent the energy spectra of the wavelets.

The fundamental scaling function $\phi_{(0)}(t)$ with $t \in \mathcal{I}$, which is depicted in Fig. 7, is nothing but the unit impulse sequence. Therefore, the set of sequences

$\{\phi_{(0)}(t-k); t, k \in \mathcal{I}\}$, obtained by integer displacements k of $\phi_{(0)}(t)$, constitute the ordinary Cartesian basis in the time domain for the set of all real-valued time series.

The level-1 scaling function $\phi_{(1)}(t) = \phi_{(0)}(t/2)$ of Fig. 8 is derived from the level 0 function by a dilation that entails doubling its temporal dispersion. The level 1 wavelet function $\psi_{(1)}(t)$ of Fig. 9 is derived from $\phi_{(1)}(t)$ by a process of frequency shifting, which involves multiplying the latter by $\cos(\pi t)$, which is $(-1)^t$ when $t \in \mathcal{I}$, which carries the function into the upper half of the frequency range.

The set of displaced scaling sequences $\{\phi_{(1)}(t-2k); t, k \in \mathcal{I}\}$, which are separated from each other by multiples of two points, provides a basis for the space of all sequences that are band limited to the frequency range $(0, \pi/2)$. The corresponding set of wavelet sequences $\{\psi_{(1)}(t-2k); t, k \in \mathcal{I}\}$, which is, in effect, a version of the scaling set that has undergone a frequency translation, provides a basis for the upper frequency range $(\pi/2, \pi)$. From the fact that, with the exclusion of the boundary points, the two ranges are non-overlapping, it follows that the two basis sets are mutually orthogonal (since sinusoids at different frequencies are mutually orthogonal.) Therefore, the two sets together span the full range $(0, \pi)$.

The elements within the basis sets are also mutually orthogonal. To see this, consider the fact that the boxcar frequency-response functions are idempotent. When multiplied by themselves they do not change, albeit that, with the resulting change of units, they come to represent the energy spectra of the wavelets. The time-domain operation corresponding to this frequency-domain multiplication is autoconvolution. The time-domain functions are real and symmetric, so their autoconvolution is the same as their autocorrelation. Therefore, the discrete wavelet sequences are their own autocorrelation functions. (We should say that, in this context, we are talking of autocorrelations where, in strict parlance, a statistician might talk of autocovariances.)

On inspecting the graphs of these functions, we see that there are zeros at the points indexed by $k = 2t$, which correspond to the conditions of orthogonality. We may describe the mutual orthogonality of the displaced wavelets as sequential orthogonality. Orthogonality conditions that extend across frequency bands may be described as lateral orthogonality.

To represent these relationships algebraically, we may consider a wavelet and its transform denoted by $\psi(t) \longleftrightarrow \psi(\omega)$. The autoconvolution of the wavelet gives the autocorrelation function $\xi^{\psi}(t) = \psi(t) * \psi(-t) = \psi(t) * \psi(t)$, where the second equality is in consequence of the symmetry of the wavelet. The corresponding operation in the frequency domain gives the modulation product $\xi^{\psi}(\omega) = \psi(\omega)\psi(-\omega) = \{\psi(\omega)\}^2$, where the second equality is in consequence of the fact that the Fourier transform of a real-valued symmetric sequence is also real-valued and symmetric. Thus, there is

$$\xi^{\psi}(t) = \psi(t) * \psi(t) \longleftrightarrow \xi^{\psi}(\omega) = \{\psi(\omega)\}^2, \tag{5}$$

where $\xi^{\psi}(\omega)$ is the energy spectrum of the wavelet. The peculiar feature of the Shannon wavelet is that $\psi(t) = \xi^{\psi}(t)$, for all t. The corresponding boxcar function has $\psi(\omega) = \xi^{\psi}(\omega)$, everywhere except at the points of discontinuity.

The conventional dyadic multiresolution wavelet analysis, represented by Fig. 1, is concerned with a succession of frequency intervals in the form of $(\pi/2^j, \pi/2^{j-1}); j = 1, 2, \ldots, n$, of which the bandwidths are halved repeatedly in the descent from high frequencies to low frequencies. By the jth round, there will be j wavelet bands and one accompanying scaling-function band.

By applying the scheme described by Mallat (1989), known as the pyramid algorithm, to the discrete versions of the functions, $\phi_{(1)}(t)$ and $\psi_{(1)}(t)$, sets of wavelet sequences can be generated that span these bands. The generic set at level j, denoted by $\{\psi_{(j)}(t - 2^j k); t, k \in \mathcal{I}\}$, contains mutually orthogonal sequences that are separated by multiples of 2^j points, and it is accompanied by a set of scaling sequences $\{\phi_{(j)}(t - 2^j k); t, k \in \mathcal{I}\}$ that span the lower frequency band $[0, \pi/2^j)$. (Here, as before, t is the index of the sequence, whereas k is the index of its displacement relative to the other wavelet sequences within the same band.)

A dyadic wave-packet analysis extends this scheme so that, by the jth round, there are 2^j bands of equal width spanning the intervals $([\ell - 1]\pi/2^j, \ell\pi/2^j); \ell = 1, \ldots, 2^j$. Each such band is spanned by a set of orthogonal functions $\{\psi_{(\ell/2^j)}(t - 2^j k); t, k \in \mathcal{I}\}$ which are separated by multiples of 2^j points. The first and the second of these bands—counting in terms of rising frequencies, which reverses the dyadic convention—are spanned by the functions $\{\psi_{(1/2^j)}(t - 2^j k) = \phi_{(j)}(t - 2^j k)\}$ and $\{\psi_{(2/2^j)}(t - 2^j k) = \psi_{(j)}(t - 2^j k)\}$ respectively, which are also found in the dyadic multiresolution wavelet analysis.

In order to generalise such schemes, we need to consider dividing the frequency range by other prime numbers and their products. For this purpose, we must consider the function defined in the frequency domain by

$$\psi(\omega) = \begin{cases} 1, & \text{if } |\omega| \in (\alpha, \beta); \\ 1/\sqrt{2}, & \text{if } \omega = \pm\alpha, \pm\beta, \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

In case it is required to divide the range into p equal intervals, there will be $\alpha = \pi(j-1)/p$ and $\beta = \pi j/p; j = 1, \ldots, p$. The corresponding time-domain function is

$$\begin{aligned} \psi(t) = \frac{1}{\pi t}\{\sin(\beta t) - \sin(\alpha t)\} &= \frac{2}{\pi t}\cos\{(\alpha + \beta)t/2\}\sin\{(\beta - \alpha)t/2\} \\ &= \frac{2}{\pi t}\cos(\gamma t)\sin(\delta t), \end{aligned} \tag{7}$$

where $\gamma = (\alpha + \beta)/2$ is the centre of the pass band and $\delta = (\beta - \alpha)/2$ is half its width. The equality, which follows from the identity $\sin(A + B) - \sin(A - B) = 2\cos A \sin B$, suggests two interpretations. On the LHS is the difference

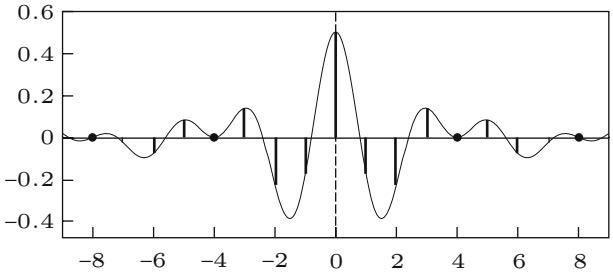

Fig. 10. A wavelet within a frequency band of width $\pi/2$ running from $3\pi/8$ to $7\pi/8$

between the coefficients of two lowpass filters with cut-off frequencies of β and α respectively. On the RHS is the result of shifting a lowpass filter with a cut-off frequency of δ so that its centre is moved from $\omega = 0$ to $\omega = \gamma$.

The process of frequency shifting is best understood by taking account of both positive and negative frequencies when considering the lowpass filter. Then, the pass band covers the interval $(-\delta, \delta)$. To convert to the bandpass filter, two copies of the pass band are made that are shifted so that their new centres lie at $-\gamma$ and γ. The pass bands have twice the width that one might expect. In the limiting case, the copies are shifted to the centres $-\pi$ and π. There they coincide, and we have $\psi(t) = 2\cos(\pi t)\sin(\delta t)/\pi t$. To reconcile this with formula for $\psi_{(1)}(t)$ of (2), wherein $\delta = \pi/2$, we must divide by 2.

We shall show, in Sect. 7, that, when the interval $[0, \pi]$ is partitioned by a sequence of p frequency bands of equal width, an orthogonal basis can be obtained for each band by displacing its wavelets successively by p elements at a time. We shall also show that, when such a band of width π/p is shifted in frequency by an arbitrary amount, the conditions of orthogonality will be maintained amongst wavelets that are separated by $2p$ elements.

This fact, which does not appear to have been recognised previously, can be exploited in fitting pass bands around localised frequency structures that do not fall within the divisions of an even grid. For the present, we shall do no more than illustrate the fact with Fig. 10, which shows the effect of translating the Shannon scaling function $\phi_{(1)}(t)$ of width $\pi/2$ up the frequency scale by an arbitrary amount. It can be see that there are orthogonality conditions affecting wavelets at displacements that are multiples of 4 points.

4 Compound Filters

The algorithms of wavelet analysis owe their efficiency to the manner in which the various bandpass filters can be constructed from elementary component filters. The resulting filters may be described as compound filters. The manner in which the filters are formed is expressed more readily in the frequency

domain than in the time domain. The subsequent translation of the compound filters from the frequency domain to the time domain is straightforward.

Figure 11 represents, in graphical terms, the construction of the second-level scaling function $\phi(2\omega)\phi(\omega)$ and wavelet $\psi(2\omega)\phi(\omega)$. These are shown in the third row of the diagram. The fourth row of the diagram shows the remaining wave-packet functions, which come from dividing the domain of the (level-1) wavelet $\psi(\omega)$ in half. The functions, which are defined over the real line, have a period of 2π. Therefore, they extend beyond the interval $[-\pi, \pi]$ which covers only the central segment. The serrated edges in the diagram are to indicate the severance of the segment from the rest of the function.

To represent the construction algebraically, we may use $\psi_{j/N}(\omega)$ to denote the jth filter in a sequence of N filters that divide the frequency range into equal bands, running from low frequency to high frequency. Then, the level-1 scaling function is $\phi_{(1)}(\omega) = \psi_{1/2}(\omega)$ and the level-1 wavelet function is $\psi_{(1)}(\omega) = \psi_{2/2}(\omega)$. The second-level scaling function is $\phi_{(2)}(\omega) = \phi_{1/4}(\omega)$, whereas the second-level wavelet is $\psi_{(2)}(\omega) = \psi_{2/4}(\omega)$. The algebra for the second-level functions is as follows:

$$\begin{aligned}
\phi_{(2)}(\omega) = \phi_{1/4}(\omega) &= \phi_{(1)}(2\omega)\phi_{(1)}(\omega) = \psi_{1/2}(2\omega)\psi_{1/2}(\omega), \\
\psi_{(2)}(\omega) = \psi_{2/4}(\omega) &= \psi_{(1)}(2\omega)\phi_{(1)}(\omega) = \psi_{2/2}(2\omega)\psi_{1/2}(\omega), \\
\psi_{3/4}(\omega) &= \psi_{(1)}(2\omega)\psi_{(1)}(\omega) = \psi_{2/2}(2\omega)\psi_{2/2}(\omega), \\
\psi_{4/4}(\omega) &= \phi_{(1)}(2\omega)\psi_{(1)}(\omega) = \psi_{1/2}(2\omega)\psi_{2/2}(\omega).
\end{aligned} \tag{8}$$

The formulae for the filters at the $(j+1)$th level of an ordinary dyadic analysis, of the kind depicted in Fig. 1, are

$$\begin{aligned}
\phi_{(j+1)}(\omega) &= \phi_{(j)}(2\omega)\phi_{(1)}(\omega) = \phi_{(1)}(2^j\omega)\phi_{(j)}(\omega), \\
\psi_{(j+1)}(\omega) &= \psi_{(j)}(2\omega)\phi_{(1)}(\omega) = \psi_{(1)}(2^j\omega)\phi_{(j)}(\omega).
\end{aligned} \tag{9}$$

The equalities can be established via recursive expansions of the formulae. Regardless of which of the forms are taken, we get

$$\phi_{(j+1)}(\omega) = \prod_{i=0}^{j} \phi_{(1)}(2^i\omega) \quad \text{and} \quad \psi_{(j+1)}(\omega) = \psi_{(1)}(\omega)\prod_{i=1}^{j} \phi_{(1)}(2^i\omega). \tag{10}$$

The formulae of (9) can be translated into the time domain. A modulation in the frequency domain corresponds to a convolution in the time domain. Raising the frequency value of any function $\psi_{(k)}(\omega)$ by a factor of n entails interpolating $n-1$ zeros between the elements of the corresponding time-domain sequence $\psi_{(k)}(t)$ to give a sequence that may be denoted by $\psi_{(k)}(t \uparrow n)$. Thus, it can be seen that

$$\begin{aligned}
\phi_{(j+1)}(t) &= \phi_{(j)}(t \uparrow 2) * \phi_{(1)}(t) = \phi_{(1)}(t \uparrow 2^j) * \phi_{(j)}(t), \\
\psi_{(j+1)}(t) &= \psi_{(j)}(t \uparrow 2) * \phi_{(1)}(t) = \psi_{(1)}(t \uparrow 2^j) * \phi_{(j)}(t).
\end{aligned} \tag{11}$$

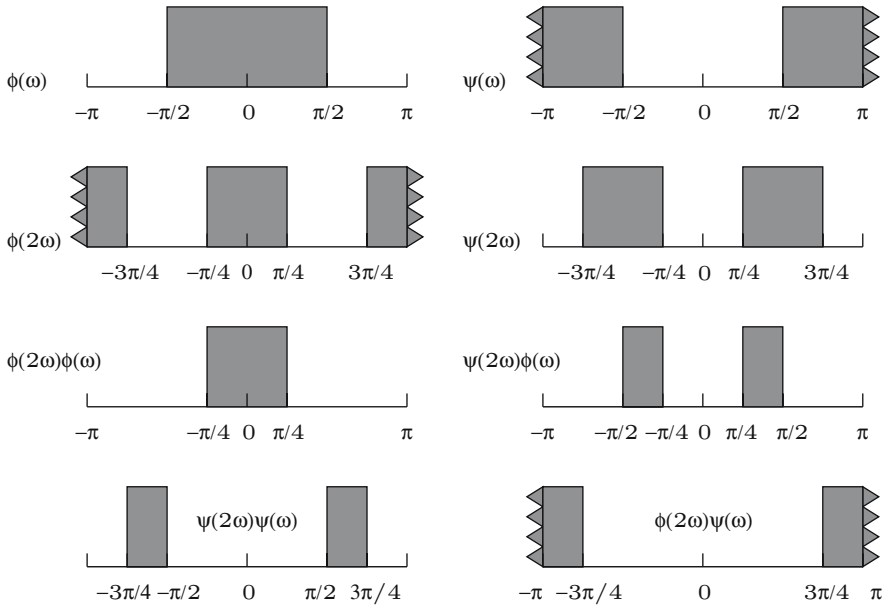

Fig. 11. The formation of second-level wavelets and scaling functions illustrated in terms of their frequency-response functions

As they stand, these time-domain formulae are not practical: the sequences $\phi_{(1)}(t)$ and $\psi_{(1)}(t)$ of the ordinates of the Shannon functions are infinite and they converge none too rapidly. The practical finite-sample versions of the formulae will be derived in the next section.

Figure 12 shows how the dyadic scheme for forming compound filters can be extended through successive rounds; and it portrays the subdivision of the

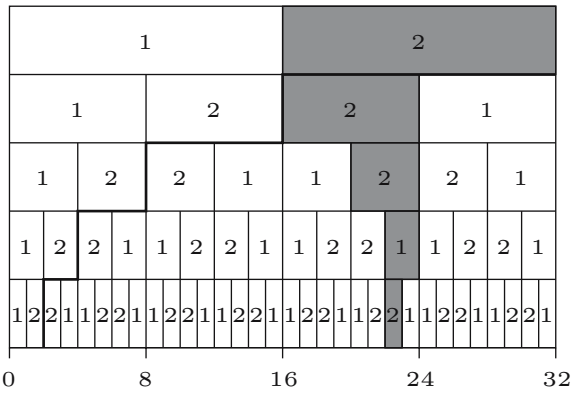

Fig. 12. The scheme for constructing compound filters in the dyadic case. The diagram highlights the construction of the filter $\psi_{23/32}(\omega)$

wavelets bands to create a set of bands of equal width that cover the entire frequency range. The figure represents five successive rounds; and it highlights the construction of the bandpass filter which is the 23rd in a succession of 32 filters with pass bands of ascending frequency. In these terms, the filter is

$$\psi_{23/32}(\omega) = \psi_{2/2}(16\omega)\psi_{1/2}(8\omega)\psi_{2/2}(4\omega)\psi_{2/2}(2\omega)\psi_{2/2}(\omega). \tag{12}$$

The bold lines in Fig. 12, which create a flight of steps descending from right to left, relate to the pyramid algorithm of the ordinary dyadic multiresolution analysis. In the jth round, the algorithm separates into two components a filtered sequence that is associated with frequency interval $(0, \pi/2^{j-1})$. From the high-frequency component are derived the amplitude coefficients of the wavelets of the jth level. The low-frequency component is passed to the next round for further subdivision.

To see how the dyadic scheme may be generalised, consider the case where the positive frequency range $[0, \pi]$ is already divided into n equal intervals, by virtue of n bandpass filters denoted $\psi_{1/n}(\omega), \ldots, \psi_{n/n}(\omega)$. The objective is to subdivide each interval into p sub intervals, where p is a prime number.

Imagine that there also exists a set of p bandpass filters, $\psi_{1/p}(\omega)$, $\ldots, \psi_{p/p}(\omega)$, that partition the interval $[0, \pi]$ into p equal parts. Amongst the latter, the ideal specification of the generic bandpass filter is

$$\psi_{j/p}(\omega) = \begin{cases} 1, & \text{if } |\omega| \in \mathcal{I}_j, \\ 1/2 & \text{if } |\omega| = (j-1)\pi/p, j\pi/p, \\ 0, & \text{if } |\omega| \in \mathcal{I}_j^c, \end{cases} \tag{13}$$

where the open interval $\mathcal{I}_j = ([j-1]\pi/p, j\pi/p)$ is the jth of the p subdivisions of $[0, \pi]$, and where $\mathcal{I}_j^c$ is the complement within $[0, \pi]$ of the closed interval $\mathcal{I}_j \cup \{(j-1)\pi/p, j\pi/p\}$ that includes the endpoints. But the function $\psi_{j/p}(\omega)$ is symmetric such that $\psi_{j/p}(\omega - \pi) = \psi_{j/p}(\pi - \omega)$. It also has a period of 2π such that $\psi_{j/p}(\omega - \pi) = \psi_{j/p}(\omega + \pi)$. The two conditions imply that $\psi_{j/p}(\pi + \omega) = \psi_{j/p}(\pi - \omega)$. It follows that

$$\psi_{j/p}(\pi + \omega) = \begin{cases} 1, & \text{if } |\omega| \in \mathcal{I}_{p+1-j}, \\ 0, & \text{if } |\omega| \in \mathcal{I}_{p+1-j}^c, \end{cases} \tag{14}$$

where $\mathcal{I}_{p+1-j}^c$ in the jth interval in the reverse sequence $\{\mathcal{I}_p, \mathcal{I}_{p-1}, \ldots, \mathcal{I}_1\}$.

To subdivide the first of the n intervals, which is $(0, \pi/n)$, into p parts, the filters $\psi_{1/p}(n\omega), \ldots, \psi_{p/p}(n\omega)$ are used, in which the argument ω has been multiplied by n. These have the same effect on the first interval as the original filters have on the interval $[0, \pi]$. To subdivide the second of the n intervals, which is $(\pi/n, 2\pi/n)$, the filters $\psi_{p/p}(n\omega), \ldots, \psi_{1/p}(n\omega)$ are used, which are in reversed order. For, in this case, $\omega \in (\pi/n, 2\pi/n)$ gives $n\omega = \pi + \lambda$ with $\lambda \in (0, \pi)$; and, therefore, the conditions of (14) apply.

Now we may recognise that the 2π periodicity of $\psi_{j/p}(\omega)$ implies that, amongst the n intervals that are to be sub divided, all odd-numbered intervals

may be treated in the manner of the first interval, whereas all even-numbered intervals may be treated in the manner of the second interval.

The generic compound filter, which has a pass band on the jth interval out of np intervals, is specified by

$$\psi_{j/pn}(\omega) = \psi_{k/p}(n\omega)\psi_{\ell/n}(\omega), \tag{15}$$

where

$$\ell = (j \text{ div } p) + 1 \quad \text{and} \quad k = \begin{cases} (j \bmod p), & \text{if } \ell \text{ is odd;} \\ p + 1 - (j \bmod p), & \text{if } \ell \text{ is even.} \end{cases} \tag{16}$$

Here, $(j \text{ div } p)$ is the quotient of the division of j by p and $(j \bmod p)$ is the remainder. (Reference to the first two rows of Figs. 13–15 will help in verifying this formula.)

Given a succession of prime factors, some of which may be repeated, the formula of (15) may be used recursively to construct compound filters of a correspondingly composite nature. However, whereas the prime factorisation of the sample size $T = p_1 p_2 \cdots p_q$ is unique, the order of the factors is arbitrary. By permuting the order, one can find a variety of compositions that amount to the same bandpass filter.

Figures 13–15 represent three ways of constructing the filter $\psi_{22/30}$ from elementary components, which are from the sets $\{\psi_{j/2}, j = 1, 2\}$, $\{\psi_{k/3}, k = 1, 2, 3\}$ and $\{\psi_{\ell/5}, \ell = 1, 2, \ldots, 5\}$. There are altogether 6 ways in which the filter may be constructed; but it seems reasonable, to opt for the construction, represented by Fig. 13, that takes the prime factors in order of descending magnitude. In practice, the filters are represented, in the frequency domain, by the ordinates of their frequency response functions, sampled at equal intervals; and the ordering of the factors by declining magnitude will serve to minimise the number of multiplications entailed in the process of compounding the filters. This is reflected in the fact that, compared with the other figures, Fig. 13 has the least highlighted area.

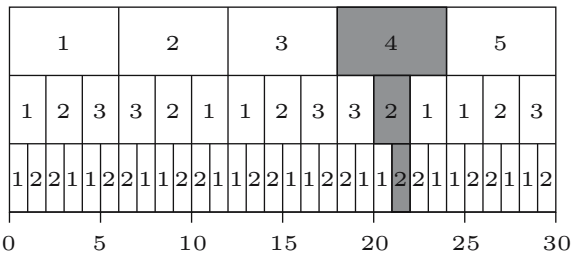

Fig. 13. The 22nd bandpass filter out of 30 factorised as $\psi_{22/30}(\omega) = \psi_{2/2}(15\omega)\psi_{2/3}(5\omega)\psi_{4/5}(\omega)$

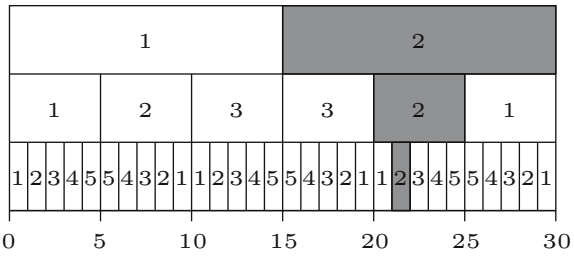

Fig. 14. The 22nd bandpass filter out of 30 factorised as $\psi_{22/30}(\omega) = \psi_{2/5}(6\omega)\psi_{2/3}(2\omega)\psi_{2/2}(\omega)$

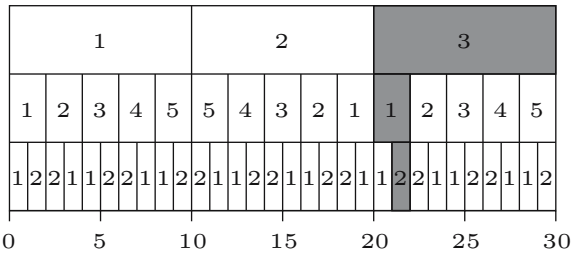

Fig. 15. The 22nd bandpass filter out of 30 factorised as $\psi_{22/30}(\omega) = \psi_{2/2}(15\omega)\psi_{1/5}(3\omega)\psi_{3/3}(\omega)$

Raising the frequency value of $\psi_{k/p}(\omega)$ by a factor of n entails interpolating $n-1$ zeros between every point of the corresponding time-domain sequence $\psi_{k/p}(t)$. The following expression indicates the correspondence between the equivalent operations of compounding the filter in the time domain and the frequency domain:

$$\psi_{j/np}(t) = \{\psi_{k/p}(t) \uparrow n\} * \psi_{\ell/n}(t) \longleftrightarrow \psi_{j/np}(\omega) = \psi_{k/p}(n\omega)\psi_{\ell/n}(\omega). \quad (17)$$

When creating a compound filter via convolutions in the time domain, the prime factors should be taken in ascending order of magnitude.

In this section, we have described a method of generating the wavelets by compounding a series of filters. In theory, the process can be pursued either in the time domain, via convolutions, or in the frequency domain, via modulations. In the case of the Shannon wavelets, the frequency domain specifications at all levels use the same rectangular template, which is mapped onto the appropriate frequency intervals. Therefore, is makes no difference whether the wavelets are produced via the compounding process or directly from the template, appropriately scaled and located in frequency.

For other wavelet specifications, a choice must be made. Either they are generated by a process of compounding, which is generally pursued in the time domain using a fixed set of dilation coefficients as the template, or else they are generated in the frequency domain using a fixed energy-function template.

The results of the two choices may be quite different. It is the latter choice that we shall make in the remainder of this paper.

Example 1. The Daubechies D4 wavelet and the scaling function of Figs. 2 and 3 relate to a dyadic analysis that proceeds in the time domain on the basis of a set of four dilation coefficients. The dilation coefficients for the scaling function are $p_0 = (1 + \sqrt{3})/4$, $p_1 = (3 + \sqrt{3})/4$, $p_2 = (3 - \sqrt{3})/4$ and $p_3 = (\sqrt{3} - 1)/4$. The coefficients that are used in creating the wavelets from the scaling functions are $q_0 = p_3$, $q_1 = -p_3$, $q_2 = p_1$ and $q_3 = -p_0$. The sequences $p_{(1)}(t) = \{p_t\}$ and $q_{(1)}(t) = \{q_t\}$ may be compared to the sequences of Shannon coefficients $\phi_{(1)}(t)$ and $\psi_{(1)}(t)$ respectively. On that basis, the following equations can be defined, which correspond to those of (11):

$$p_{(j+1)}(t) = p_{(j)}(t \uparrow 2) * p_{(1)}(t) = p_{(1)}(t \uparrow 2^j) * p_{(j)}(t), \tag{18}$$
$$q_{(j+1)}(t) = p_{(j)}(t \uparrow 2) * q_{(1)}(t) = q_{(1)}(t \uparrow 2^j) * p_{(j)}(t).$$

The first of the alternative forms, which entails the interpolation of a zero between each of the coefficients of $p_{(j)}(t)$, corresponds to the recursive system that has been used in generating Figs. 2 and 3. That is to say, the diagrams have been created by proceeding through a number of iterations and then mapping the resulting coefficients, which number $2^{j+1} + 2^j - 2$ at the jth iteration, into the interval $[0, 3]$. The difference between the wavelets and the scaling functions lies solely in the starting values. This algorithm provides a way of seeking the fixed-point solution to the following dilation equation that defines the D4 scaling function $\phi_D(t)$ with $t \in \mathcal{R}$:

$$\phi_D(t) = p_0\phi_D(2t) + p_1\phi_D(2t-1) + p_2\phi_D(2t-2) + p_3\phi_D(2t-3). \tag{19}$$

The second of the forms is implicated in the pyramid algorithm of Mallat (1989). In this case, the difference between the wavelets and the scaling functions lies solely in the final iteration.

5 Adapting to Finite Samples

The wavelet sequences corresponding to the ideal bandpass filters are defined on the entire set of integers $\{t = 0, \pm 1, \pm 2, \ldots\}$ whereas, in practice, a discrete wavelet analysis concerns a sample of T data points. This disparity can be overcome, in theory, by creating a periodic extension of the data that replicates the sample in all intervals of the duration T that precede and follow it. By this means, the data value at a point $t \notin \{0, 1, \ldots, T-1\}$, which lies outside the sample, is provided by $y_t = y_{\{t \bmod T\}}$, where $(t \bmod T)$ lies within the sample. With the periodic extension available, one can think of multiplying the filter coefficients point by point with the data.

As an alternative to extending the data, one can think of creating a finite sequence of filter coefficients by wrapping the infinite sequence $\psi(t) = \{\psi_t\}$ around a circle of circumference T and adding the overlying coefficients to give

$$\psi_t^\circ = \sum_{k=-\infty}^{\infty} \psi_{\{t+kT\}} \quad \text{for} \quad t = 0, 1, \ldots, T-1. \tag{20}$$

The inner product of the resulting coefficients $\psi_0^\circ, \ldots, \psi_{T-1}^\circ$ with the sample points $y_0, \ldots, y_{T-1}$ will be identical to that of the original coefficients with the extended data. To show this, let $\tilde{y}(t) = \{\tilde{y}_t = y_{\{t \bmod T\}}\}$ denote the infinite sequence that is the periodic extension of $y_0, \ldots, y_{T-1}$. Then,

$$\begin{aligned} \sum_{t=-\infty}^{\infty} \psi_t \tilde{y}_t &= \sum_{k=-\infty}^{\infty} \left\{ \sum_{t=0}^{T-1} \psi_{\{t+kT\}} \tilde{y}_{\{t+kT\}} \right\} \\ &= \sum_{t=0}^{T-1} y_t \left\{ \sum_{k=-\infty}^{\infty} \psi_{\{t+kT\}} \right\} = \sum_{t=0}^{T-1} y_t \psi_t^\circ. \end{aligned} \tag{21}$$

Here, the first equality, which is the result of cutting the sequence $\{\psi_t \tilde{y}_t\}$ into segments of length T, is true in any circumstance, whilst the second equality uses the fact that $\tilde{y}_{\{t+kT\}} = y_{\{t \bmod T\}} = y_t$. The final equality invokes the definition of ψ_t°.

In fact, the process of wrapping the filter coefficients should be conducted in the frequency domain, where it is simple and efficient, rather than in the time domain, where it entails the summation of infinite series. We shall elucidate these matters while demonstrating the use of the discrete Fourier transform in performing a wavelets analysis.

To begin, let us consider the z-transforms of the filter sequence and the data sequence:

$$\psi(z) = \sum_{t=-\infty}^{\infty} \psi_t z^t \qquad \text{and} \qquad y(z) = \sum_{t=0}^{T-1} y_t z^t. \tag{22}$$

Setting $z = \exp\{-\mathrm{i}\omega\}$ in $\psi(z)$ creates a continuous periodic function in the frequency domain of period 2π, denoted by $\psi(\omega)$, which, by virtue of the discrete-time Fourier transform, corresponds one-to-one with the doubly infinite time-domain sequence of filter coefficients.

Setting $z = z_j = \exp\{-\mathrm{i}2\pi j/T\}; j = 0, 1, \ldots, T-1$, is tantamount to sampling the (piecewise) continuous function $\psi(\omega)$ at T points within the frequency range of $\omega \in [0, 2\pi)$. (Given that the data sample is defined on a set of positive integers, it is appropriate to replace the symmetric interval $[-\pi, \pi]$, considered hitherto, in which the endpoints are associated with half the values of their ordinates, by the positive frequency interval $[0, 2\pi)$, which excludes the endpoint on the right and attributesthe full value of the ordinate at zero

frequency to the left endpoint.) The powers of z_j now form a T-periodic sequence, with the result that

$$\begin{aligned}\psi(z_j) &= \sum_{t=-\infty}^{\infty} \psi_t z_j^t \qquad (23)\\ &= \left\{\sum_{k=-\infty}^{\infty} \psi_{kT}\right\} + \left\{\sum_{k=-\infty}^{\infty} \psi_{(kT+1)}\right\} z_j + \cdots + \left\{\sum_{k=-\infty}^{\infty} \psi_{(kT+T-1)}\right\} z_j^{T-1}\\ &= \psi_0^\circ + \psi_1^\circ z_j + \cdots + \psi_{T-1}^\circ z_j^{T-1} = \psi^\circ(z_j).\end{aligned}$$

There is now a one-to-one correspondence, via the discrete Fourier transform, between the values $\psi(z_j); j = 0, 1, \ldots, T-1$, sampled from $\psi(\omega)$ at intervals of $2\pi/T$, and the coefficients $\psi_0^\circ, \ldots, \psi_{T-1}^\circ$ of the circular wrapping of $\psi(t)$. Setting $z = z_j = \exp\{-\mathrm{i}2\pi j/T\}; j = 0, 1, \ldots, T-1$, within $y(z)$ creates the discrete Fourier transform of the data sequence, which is commensurate with the square roots of the ordinates sampled from the energy function.

To elucidate the correspondence between operations in the two domains, we may replace z in (22) by a circulant matrix $K = [e_1, \ldots, e_{T-1}, e_0]$, which is formed from the identity matrix $I_T = [e_0, e_1, \ldots, e_{T-1}]$ of order T by moving the leading vector to the end of the array. Since $K^q = K^{T+q}$, the powers of the matrix form a T-periodic sequence, as do the powers of $z = \exp\{-\mathrm{i}2\pi j/T\}$. (A full account of the algebra of circulant matrices has been provided by Pollock 2002.)

The matrix K is amenable to a spectral factorisation of the form $K = \bar{U}DU$, where

$$\begin{aligned}U &= T^{-1/2}W = T^{-1/2}[\exp\{-\mathrm{i}2\pi tj/T\}; t, j = 0, \ldots, T-1] \quad \text{and}\\ \bar{U} &= T^{-1/2}\bar{W} = T^{-1/2}[\exp\{\mathrm{i}2\pi tj/T\}; t, j = 0, \ldots, T-1] \qquad (24)\end{aligned}$$

are unitary matrices such that $U\bar{U} = \bar{U}U = I_T$, and where

$$D = \mathrm{diag}\{1, \exp\{-\mathrm{i}2\pi/T\}, \ldots, \exp\{-\mathrm{i}2\pi(T-1)/T\}\} \qquad (25)$$

is a diagonal matrix whose elements are the T roots of unity, which are found on the circumference of the unit circle in the complex plane.

Using $K = \bar{U}DU$ in place of z in (22) creates the following circulant matrices:

$$\Psi^\circ = \psi^\circ(K) = \bar{U}\psi^\circ(D)U \qquad \text{and} \qquad Y = y(K) = \bar{U}y(D)U. \qquad (26)$$

The multiplication of two circulant matrices generates the circular convolution of their elements. Thus the product

$$\Psi^\circ Y = \{\bar{U}\psi^\circ(D)U\}\{\bar{U}y(D)U\} = \bar{U}\psi^\circ(D)y(D)U. \qquad (27)$$

is a matrix in which the leading vector contains the elements of the circular convolution of $\{\psi_0^\circ, \ldots, \psi_{T-1}^\circ\}$ and $\{y_0, \ldots, y_{T-1}\}$, of which the inner product of (21) is the first element.

The leading vector of $\Psi^\circ Y$ can be isolated by postmultiplying this matrix by $e_0 = [1, 0, \ldots, 0]'$. But $Ue_0 = T^{-1/2}We_0 = T^{-1/2}h$, where $h = [1, 1, \ldots, 1]'$ is the summation vector. Therefore,

$$\Psi^\circ Y e_0 = T^{-1}\bar{W}\{\psi^\circ(D)y(D)h\}, \tag{28}$$

where $\psi^\circ(D)y(D)h$ is a vector whose elements are the products of the diagonal elements of $\psi^\circ(D)$ and $y(D)$. Equation (28) corresponds to the usual matrix representation of an inverse discrete Fourier transform, which maps a vector from the frequency domain into a vector of the time domain.

Observe that (27) also establishes the correspondence between the operation of cyclical convolution in the time domain, represented by the product of the matrices on the LHS, and the operation of modulation in the frequency domain, represented by the pairwise multiplication of the elements of two diagonal matrices. The correspondence can be represented by writing $\Psi^\circ Y \longleftrightarrow \psi^\circ(D)y(D)$. Using such notation, we can represent the finite-sample version of (17) by

$$\psi^\circ_{j/np}(K) = \psi^\circ_{k/p}(K^n)\psi^\circ_{\ell/n}(K) \longleftrightarrow \psi^\circ_{j/np}(D) = \psi^\circ_{k/p}(D^n)\psi^\circ_{\ell/n}(D). \tag{29}$$

If $\alpha(z)$ is a polynomial of degree $T-1$ and, if n is a factor of T, then $\alpha(K^n) = \bar{U}\alpha(D^n)U$ is a circulant matrix of order T in which there are T/n nonzero bands, with $n-1$ bands of zeros lying between one nonzero band and the next. The generic nonzero coefficient, which is on the tth nonzero subdiagonal band, is $\alpha^\circ_t = \sum_{j=0}^{T/n} \alpha_{\{t+jn\}}$. The jth diagonal element of the matrix D^n, which is entailed in the spectral factorisation of $\alpha(K^n)$, takes the values $\exp\{-\mathrm{i}2\pi nj/T\}; j = 0, 1, \ldots, T-1$. Compared to the corresponding elements of D, its frequency values have been increased by a factor of n.

In the case of a piecewise continuous energy function $\xi(\omega) = |\psi(\omega)|^2$, defined on the interval $[-\pi, \pi]$, one can afford to ignore the endpoints of the interval together with any points of discontinuity within the interval. These constitute a set of measure zero in the context of the remaining frequency values. When such points are taken in the context of a sample of T frequency values, they can no longer be ignored, as the example at the the end of this section indicates.

The method of coping with finite samples via a periodic extension of the data is also a feature of a discrete Fourier analysis. It requires the data to be free of an overall trend. Otherwise, there will be a radical disjunction in passing from the end of one replication of the sample to the begining of the next. Such disjunctions will affect all of the Fourier coefficients. However, the effect upon the coefficients of a wavelet analysis will be limited to the extent that the wavelets are localised in time. A disadvantage of the Shannon wavelets is that they are widely dispersed; and, in the next section, we shall be developing wavelets that are more localised.

Example 2 (The Wrapped Shannon Wavelet). Consider a set of frequency-domain ordinates sampled from a boxcar energy function, defined over the

interval $[-\pi, \pi]$, at the points $\omega_j = 2\pi j/T; j = 1 - T/2, \ldots, 0, \ldots, T/2$, where T is even:

$$\xi_j^\circ = \begin{cases} 1, & \text{if } j \in \{1-d, \ldots, d-1\}, \\ 1/2, & \text{if } j = \pm d, \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

Here, $d < T/2$ is the index of the point of discontinuity. The (inverse) Fourier transform of these ordinates constitutes the autocorrelation function of the wrapped Shannon wavelet. The transform of the square roots of the ordinates is the wavelet itself.

The z-transform of the energy sequence is $\xi^\circ(z) = \{S^+(z) + S^-(z)\}/2$, wherein

$$\begin{aligned} S^+(z) &= z^{-d} + \cdots + z^{-1} + 1 + z + \cdots + z^d \quad \text{and} \\ S^-(z) &= z^{1-d} + \cdots + z^{-1} + 1 + z + \cdots + z^{d-1}. \end{aligned} \tag{31}$$

Setting $z = e^{-\mathrm{i}\omega_1 t}$, where $\omega_1 = 2\pi/T$, and using the formula for the partial sum of a geometric progression, gives the following Dirichlet kernels:

$$S^+(t) = \frac{\sin\{\omega_1 t(d + 1/2)\}}{\sin(\omega_1 t/2)}, \qquad S^-(t) = \frac{\sin\{\omega_1 t(d - 1/2)\}}{\sin(\omega_1 t/2)}. \tag{32}$$

But $\sin(A+B)+\sin(A-B) = 2\sin A\cos B$, so, with $A = \omega_1 td$ and $B = \omega_1 t/2$, we have

$$\xi^\circ(t) = \frac{1}{2T}\{S^+(t) + S^-(t)\} = \frac{\cos(\omega_1 t/2)\sin(d\omega_1 t)}{T\sin(\omega_1 t/2)}, \tag{33}$$

This expression gives the values of the circular autocorrelation function of the wrapped wavelet at the points $t = 1, \ldots, T-1$. The value at $t = 0$ is $\xi_0^\circ = 2d/T$, which comes from setting $z = 1$ in the expressions for $S^+(z)$ and $S^-(z)$ of (31).

If $d = T/4$, such that the points of discontinuity are at $\pm\pi/2$, as in the specification of $\phi_{(0)}(t)$ under (3), then $\sin(d\omega_1 t) = \sin(\pi t/2)$ and $\xi^\circ(2t) = 0$ for $t = 1, \ldots, T-1$. This confirms that the relevant conditions of sequential orthogonality are indeed fulfilled by the wrapped wavelet.

The technique of frequency shifting may be applied to the formula of (33). Let g be the index that marks the centre of the pass band. Then, the autocorrelation function of the wrapped wavelet corresponding to a bandpass filter with lower and upper cut-off points of $a = g - d$ and $b = g + d$ is given by

$$\xi^\circ(t) = 2\cos(g\omega_1 t)\frac{\cos(\omega_1 t/2)\sin(d\omega_1 t)}{T\sin(\omega_1 t/2)}. \tag{34}$$

To find the wavelets themselves, we transform a set of frequency-domain coefficients that are the square roots of those of the energy function. For the

wavelet corresponding to the ideal lowpass filter with a cut-off at $j = \pm d$, we have

$$T\phi^{\circ}(t) = \frac{\sin\{\omega_1 t(d - 1/2)\}}{\sin(\omega_1 t/2)} + \sqrt{2}\cos(d\omega_1). \tag{35}$$

For the wavelet corresponding to the ideal bandpass filter with a cut-off points at $j = \pm a, \pm b$, there is

$$\begin{aligned} T\psi^{\circ}(t) = 2\cos\{(a + b)\omega_1 t/2\}\frac{\sin\{(b - a - 1)\omega_1 t/2\}}{\sin(\omega_1 t/2)} \\ +\sqrt{2}\{\cos(a\omega_1) + \cos(b\omega_1)\}. \end{aligned} \tag{36}$$

6 Conditions of Sequential Orthogonality in the Dyadic Case

The advantage of the Shannon wavelets is that they provide us with a ready-made orthogonal bases for the frequency bands that accompany a multiresolution analysis or a wave packet analysis. We have illustrated this feature, in Sect. 3, with the case of the infinite wavelet and scaling function sequences that correspond to the first level of a dyadic analysis.

The conditions of sequential orthogonality also prevail in the case of the wrapped wavelet sequence. This may be demonstrated with reference to the autocorrelation functions of (33) and (34). The only restriction is that the bandwidth $2\delta = \beta - \alpha$ must divide the frequency range $[0, \pi)$ an integral number of times, say q times. In that case, the orthogonal basis of each of the bands that partition the range will be formed by displacing the corresponding Shannon wavelet by q elements at a time.

In this section, we shall look for the general conditions that are necessary to ensure that the displaced wavelet sequences are mutually orthogonal. The conditions of orthogonality will be stated in terms of the frequency-domain energy function and its square root, which is the Fourier transform of the time-domain wavelet function.

To avoid unnecessary complexity, we shall deal in terms of the continuous frequency-domain function rather the sampled version, which has been the subject of Sect. 5. Except in cases where the energy function has an absolute discontinuity or a saltus, as in the case of the boxcar function associated with Shannon wavelets, the results can be applied without hesitation to the sampled function.

We may begin, in this section, by considering the first of the prime numbers which is $q = 2$, which is the case of the dyadic wavelets. This is the only even prime number; and, therefore, it demands special treatment. In the next section, we shall deal with the case where q is any other prime number, beginning with the triadic case, where $q = 3$. This is a prototype for all other cases.

Ignoring subscripts, let $\xi(t) \longleftrightarrow \xi(\omega)$ denote the autocorrelation function, which may belong equally to a scaling function or to a wavelet, together with its Fourier transform, which is the corresponding energy spectrum. Then, the condition of orthogonality is that

$$\xi(2t) = \begin{cases} \xi_0, & \text{if } t = 0; \\ 0, & \text{if } t \neq 0, \end{cases} \tag{37}$$

which is to say that $\xi(2t) = \xi_0\delta(t)$, where $\delta(t)$ is the unit impulse function in the time domain. The transform of the impulse function is a constant function in the frequency domain: $\delta(t) \longleftrightarrow 1$. To see what this implies for the energy spectrum, define $\lambda = 2\omega$ and use the change of variable technique to give

$$\begin{aligned} \xi(2t) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \xi(\omega)e^{\mathrm{i}\omega(2t)}d\omega \qquad (38)\\ &= \frac{1}{4\pi}\int_{-2\pi}^{2\pi} \xi(\lambda/2)e^{\mathrm{i}\lambda t}d\lambda \\ &= \frac{1}{4\pi}\int_{-2\pi}^{-\pi} \xi(\lambda/2)e^{\mathrm{i}\lambda t}d\lambda + \frac{1}{4\pi}\int_{-\pi}^{\pi} \xi(\lambda/2)e^{\mathrm{i}\lambda t}d\lambda + \frac{1}{4\pi}\int_{\pi}^{2\pi} \xi(\lambda/2)e^{\mathrm{i}\lambda t}d\lambda. \end{aligned}$$

But the Fourier transform of the sequence $\xi(2t)$ is a periodic function with one cycle in 2π radians. Therefore, the first integral must be translated to the interval $[0, \pi]$, by adding 2π to the argument, whereas the third integral must be translated to the interval $[-\pi, 0]$, by subtracting 2π from the argument. After their translation, the first and the third integrands combine to form the segment of the function $\xi(\pi + \lambda/2)$ that falls in the interval $[-\pi, \pi]$. The consequence is that

$$\xi(2t) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \{\xi(\lambda/2) + \xi(\pi + \lambda/2)\}e^{\mathrm{i}\lambda t}d\lambda. \tag{39}$$

This relationship can be denoted by $\xi(2t) \longleftrightarrow \frac{1}{2}\{\xi(\lambda/2) + \xi(\pi + \lambda/2)\}$. The necessary condition for the orthogonality of the displaced wavelet sequences is that the Fourier transform on the RHS is a constant function. In that case, the argument $\lambda/2$ can be replaced by ω, and the condition becomes

$$\{\xi(\omega) + \xi(\pi + \omega)\} = c, \tag{40}$$

where c is a constant.

It will be observed that, if $\xi(\omega) = \xi_{1/2}(\omega)$ stands for energy spectrum of the dyadic scaling function, then $\xi(\omega + \pi) = \xi_{2/2}(\omega)$ will be the energy spectrum of the wavelet. The condition $\xi_{1/2}(\omega) + \xi_{2/2}(\omega) = 1$, which actually prevails, corresponds to the conservation of energy. Pairs of filters for which the squared gains satisfy the condition are called quadrature mirror filters.

Example 3 (The Triangular Energy Function). Consider the periodic energy functions defined over the frequency interval $[-\pi, \pi]$ by

$$\xi_{1/2}(\omega) = \begin{cases} 1 - |\omega|/\pi, & \text{if } |\omega| \in [0, \pi); \\ 0, & \text{if } \omega = \pm\pi, \end{cases}$$
$$\xi_{2/2}(\omega) = \begin{cases} |\omega|/\pi, & \text{if } |\omega| \in [0, \pi/2); \\ 1/2, & \text{if } \omega = \pm\pi, \end{cases} \tag{41}$$

Here $\xi_{1/2}(\omega)$ is a triangle that results from the autoconvolution in the frequency domain of the box function $\phi_{(1)}(\omega)$ of (3), whilst $\xi_{2/2}(\omega)$ is a version translated by π radians. It is manifest that these functions obey the condition of (40), since $\xi_{1/2}(\omega) + \xi_{2/2}(\omega) = 1$.

The Fourier transforms are given by

$$\xi_{1/2}(t) = \left\{ \frac{\sin(\pi t/2)}{\pi t} \right\}^2, \tag{42}$$
$$\xi_{2/2}(t) = \cos(\pi t)\xi_{1/2}(t).$$

Here, $\xi_{1/2}(t)$ is the square of the sinc function, whereas $\xi_{2/2}(t)$ is the result of a frequency shifting operation applied to $\xi_{1/2}(t)$.

Example 4 (The Chamfered Box). A generalisation of the function $\xi_{1/2}(\omega)$ of (41), which also obeys the condition of (40), is one that can be described as a chamfered box or a split triangle, and which is defined by

$$\xi_{1/2}(\omega) = \begin{cases} 1, & \text{if } |\omega| \in (0, \pi/2 - \epsilon), \\ 1 - \dfrac{|\omega + \epsilon - \pi/2|}{2\epsilon}, & \text{if } |\omega| = (\pi/2 - \epsilon, \pi/2 + \epsilon), \\ 0, & \text{otherwise.} \end{cases} \tag{43}$$

Setting $\epsilon = \pi/2$ reduces this to the triangular function of (41). Also subsumed under the sampled version of the present function is the sampled version of the boxcar energy function, in which the problem caused by the discontinuity at the cut-off point is handled, in effect, by chamfering the edge. (When the edge of the box is chamfered in the slightest degree, the two function values at the point of discontinuity, which are zero and unity, will coincide at a value of one half.)

A function that has the same Fourier transform as the chamfered box can be formed from the difference of two triangle functions. The first triangle is defined in the frequency domain by

$$\Lambda_1(\omega) = \begin{cases} \dfrac{1}{2}\left(\dfrac{\pi}{2\epsilon} + 1\right) - \dfrac{|\omega|}{2\epsilon}, & \text{if } |\omega| \in (0, \pi/2 + \epsilon), \\ 0, & \text{otherwise.} \end{cases} \tag{44}$$

The Fourier transform is

$$\Lambda_1(t) = \left\{ \frac{\sin\{(\pi/2 + \epsilon)t\}}{\pi t} \right\}^2, \tag{45}$$

The second triangle is defined by

$$\Lambda_2(\omega) = \begin{cases} \frac{1}{2}\left(\frac{\pi}{2\epsilon} - 1\right) - \frac{|\omega|}{2\epsilon}, & \text{if } |\omega| \in (0, \pi/2 - \epsilon), \\ 0, & \text{otherwise.} \end{cases} \tag{46}$$

The Fourier transform for this one is

$$\Lambda_2(t) = \left\{\frac{\sin\{(\pi/2 - \epsilon)t\}}{\pi t}\right\}^2. \tag{47}$$

The Fourier transform of the function of $\xi_{1/2}(\omega)$ of (43) is

$$\xi_{1/2}(t) = \Lambda_1(t) - \Lambda_2(t). \tag{48}$$

In the example above, the autocorrelation functions fulfil the orthogonality condition of (40) by virtue of their anti-symmetry in the vicinity of the cut-off values $\omega_c = \pm\pi/2$. For a sine wave the condition of anti-symmetry is expressed in the identity $\sin(-\omega) = -\sin(\omega)$. For the energy functions, the points of symmetry have the coordinates $(\omega_c, 0.5)$ and the conditions of anti-symmetry, which prevail in the intervals $(\omega_c - \epsilon, \omega_c + \epsilon)$, where $\epsilon \leq \pi/2$, are expressed in the identity

$$0.5 - \xi(\omega_c - \omega) = \xi(\omega_c + \omega) - 0.5 \quad \text{for} \quad \omega \in (-\epsilon, \epsilon) \tag{49}$$

We may describe this as the condition of sigmoid anti-symmetry, or of S-symmetry for short. The terminology is suggested by the following example which uses an ordinary cosine in constructing the autocorrelation function.

Example 5 (The Cosine Bell). The cosine bell, with a period of 2π, is defined in the frequency domain by

$$\xi_{1/2}(\omega) = 0.5\{1 + \cos(\omega)\}. \tag{50}$$

It is S-symmetric about the point $\pi/2$ in the frequency interval $(0, \pi)$ and about the point $-\pi/2$ in the frequency interval $(-\pi, 0)$. The function is not band-limited in frequency domain—but it is band-limited in the time domain. The Fourier transform of the continuous periodic function is the three-point sequence $\{0.25, 0.5, 0.25\}$, which can be recognised as the autocorrelation function of the discrete Haar scaling function. The transform of the continuous periodic function $\xi_{2/2}(\omega) = 0.5\{1 - \cos(\omega)\}$ is the sequence $\{-0.25, 0.5, -0.25\}$, which can be recognised as the autocorrelation function of the discrete Haar wavelet.

The Haar wavelet, which is the one with the minimum temporal dispersion, is defined, in discrete terms, on two points by

$$\psi(t) = \begin{cases} 0.5, & \text{if } t = 0, \\ -0.5, & \text{if } t = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{51}$$

The accompanying scaling function is

$$\phi(t) = \begin{cases} 0.5, & \text{if } t = 0, 1, \\ 0, & \text{otherwise.} \end{cases} \tag{52}$$

Now consider a sequence of ordinates sampled from the energy function $\xi_{2/2}(\omega) = 0.5\{1 - \cos(\omega)\}$ at the Fourier frequencies $\omega_j = 2\pi/T; j = 0, 1, \ldots, T-1$, which extend over the interval $[0, 2\pi)$. The sequence will be real-valued and even such that $\xi_{2/2}(\omega_j) = \xi_{2/2}(\omega_{T-j})$. The Fourier transform of these ordinates will, likewise, be a real-valued even sequence of T points of the form $\{0.5, -0.25, 0, \ldots, 0, -0.25\}$. This can be envisaged either as a single cycle of a periodic function or as a set of points distributed evenly around a circle of circumference T. The sequence constitutes the circular autocorrelation function of a wrapped Haar wavelet.

The Haar wavelet is not an even function. To derive a wavelet that is real and even and which has the same autocorrelation function as the wrapped Haar wavelet, we must transform into the time domain the square roots of the ordinates sampled from the cosine bell energy function. An example of such a wavelet is provided by Fig. 16.

Example 6 (The Split Cosine Bell). A derivative of the cosine bell, which is band-limited in the frequency domain, is provided by the split cosine bell. This has a horizontal segment interpolated at the apex of the bell which, consequently, must show a more rapid transition in the vicinities of $\pm\pi/2$.

$$\xi_{1/2}(\omega) = \begin{cases} 1, & \text{if } |\omega| \in (0, \pi/2 - \epsilon); \\ 0.5\left[1 + \cos\left\{\dfrac{\pi}{2\epsilon}|\omega + \epsilon - \pi/2|\right\}\right], & \text{if } |\omega| \in (\pi/2 - \epsilon, \pi/2 + \epsilon), \\ 0, & \text{otherwise} \end{cases} \tag{53}$$

Setting $\epsilon = \pi/2$ reduces this to the cosine bell of (50).

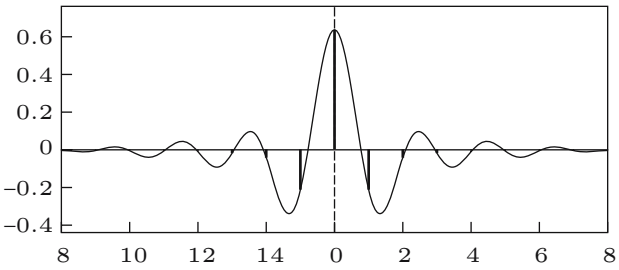

Fig. 16. A circulant wavelet sequence on 16 points corresponding to a cosine bell energy function

Observe that, if ϵ divides π an even number of times, then the split cosine bell can be expressed as a sum of cosine bells, each of width 4ϵ, at displacements relative to each other that are multiples of 2ϵ. In that case, the Fourier transform of the function has a particularly simple analytic expression.

The split cosine bell has been advocated by Bloomfield (1976) as a means of truncating and tapering the time-domain coefficients of an ideal bandpass filter to create a practical FIR filter. Its advantage over the chamfered box in this connection lies in the fact that it possesses a first derivative that is continuous everywhere. Its avoidance of discontinuities reduces the spectral leakage. It is therefore to be expected that a wavelet derived from a cosine bell energy function will have a lesser temporal dispersion than one that has been derived from the corresponding chamfered box.

Example 7 (The Butterworth Function). Another family of energy functions from which the wavelets may be derived is provided by the function that defines the frequency response of a digital Butterworth filter with a cut-off point at $\omega = \pi/2$:

$$\begin{aligned} \xi_{1/2}(\omega) &= (1 + \{\tan(\omega/2)\}^{2n})^{-1}, \\ \xi_{2/2}(\omega) &= 1 - \xi_{1/2}(\omega). \end{aligned} \tag{54}$$

When $n = 1$, the Butterworth function is the square of a cosine. Increasing the value of n increases the rate of the transition between the pass band and the stop band of the filter, such that the function converges to the boxcar function $\phi_{(1)}(\omega)$ of (3)—see Pollock (1999), for example.

In the context of the Butterworth digital filter, the integer parameter n represents the degree of a polynomial operator. In the present context, there is no reason why n should be restricted to take integer values. It will be found, for example, that, when $n = 0.65$, the Butterworth function provides a close approximation to the triangular energy function of (41). This is shown in Fig. 23 together with the effects of other values of the parameter.

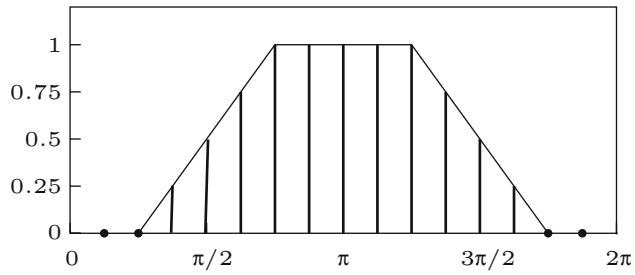

Fig. 17. A sampled energy function of 16 points in the shape of a chamfered box

The Butterworth function, which satisfies the condition of S-symmetry, appears to be preferable to the split cosine bell. The relative merits of various

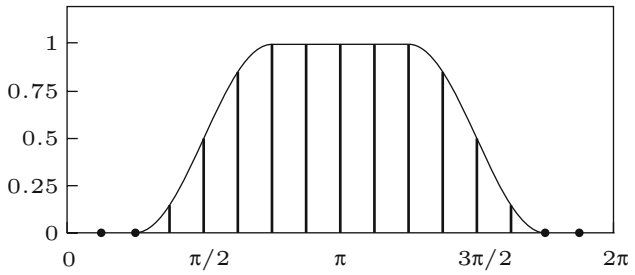

Fig. 18. A sampled energy function of 16 points in the shape of a split cosine bell

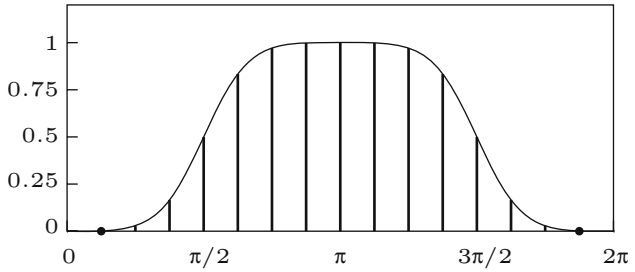

Fig. 19. A sampled energy function of 16 points defined by a Butterworth function

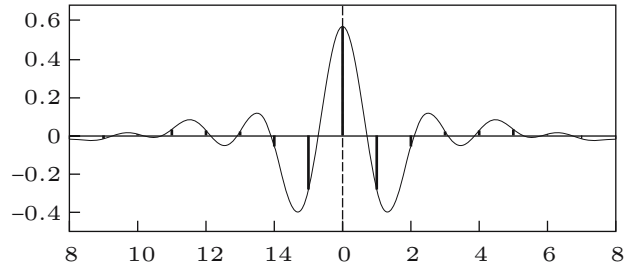

Fig. 20. A circulant wavelet sequence on 16 points corresponding to the energy function of Fig. 17, which is in the shape of a chamfered box

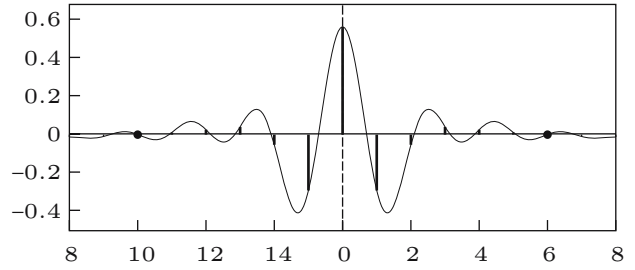

Fig. 21. A circulant wavelet sequence on 16 points corresponding to the split cosine bell energy function of Fig. 18

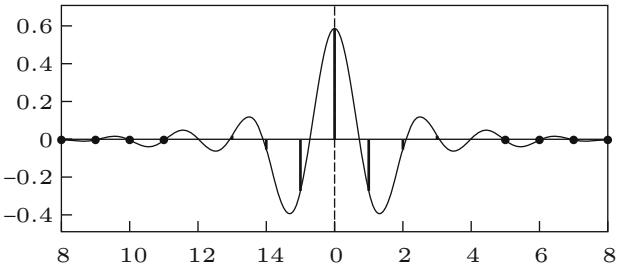

Fig. 22. A circulant wavelet sequence on 16 points corresponding to the Butterworth energy function with $n = 2$ of Fig. 19

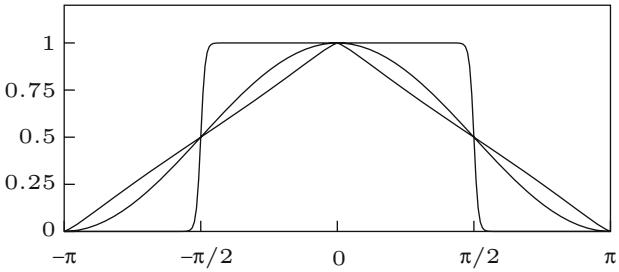

Fig. 23. The Butterworth function with the parameter values $n = 0.62$ (the triangle), $n = 1$ (the bell) and $n = 20$ (the boxcar)

families of wavelets proposed in this section can be assessed with reference to Figs. 17–22, which show the energy functions together with the wavelets that are derived from them.

A remarkable feature of the Butterworth wavelet is that, beyond a short distance from the central point, where $t = 0$, the ordinates are virtually zeros. The virtual zeros are indicated in Fig. 22 by black dots, the first of which corresponds to a value of $\psi(t = 6) = -0.00096$. Moreover, such values are reduced as T increases and as the wavelet is wrapped around a widening circle.

One might recall the fact that, for a non-circulant wavelet on a finite support, the condition of sequential orthogonality necessitates an even number of points—see, for example, Percival and Walden (2000, p.69). This precludes the symmetry of the coefficients about a central value. Nevertheless, the Butterworth wavelet, which satisfies the orthogonality conditions, has virtually a finite support and is also symmetric.

7 Conditions of Orthogonality in the Non-dyadic Case

We shall now consider the general case where the wavelets subsist in q bands within the frequency interval $[0, \pi]$, where q is a prime number. We shall begin

by considering the triadic case where $q = 3$. This serves as a prototype for all other cases. First, it is necessary to indicate the manner in which the triadic wavelets may be constructed from various S-symmetric energy functions, such as those that have been considered in the previous section.

Consider an energy function $B(\omega)$ defined on the interval $[-\pi, \pi]$ that corresponds to a dyadic scaling function or, equally, to a half-band lowpass filter with a nominal cut-off frequency of $\pi/2$. This function can be mapped, via a compression of the frequency axis, onto an interval of length $2\pi/3$. The effect is achieved by multiplying the frequency argument by a factor of 3 to give $B(3\omega); \omega \in [-\pi/3, \pi/3]$.

To construct the triadic lowpass wavelets, for which the nominal range of the energy function is the interval $(-\pi/3, \pi/3)$, copies of $B(3\omega)$ are placed at the centres $-\pi/6$ and $\pi/6$. The result is the function defined on the interval $[-\pi, \pi]$ by

$$\xi_{1/3}(\omega) = \begin{cases} B(3\omega + \pi/6) + B(3\omega - \pi/6), & \text{if } \omega \in [-\pi/2, \pi/2], \\ 0, & \text{otherwise.} \end{cases} \tag{55}$$

Figure 24a shows the manner in which the copied functions are fused together.

To construct the triadic bandpass wavelet, for which the nominal pass band is the interval $(\pi/3, 2\pi/3)$, the two copies of $B(3\omega)$ are translated to centres at $\pi/2$ and $-\pi/2$ and combined to give

$$\xi_{2/3}(\omega) = \begin{cases} B(3\omega + \pi/2), & \text{if } \omega \in [-5\pi/6, -\pi/6], \\ B(3\omega - \pi/2), & \text{if } \omega \in [5\pi/6, \pi/6], \\ 0, & \text{otherwise.} \end{cases} \tag{56}$$

The result is shown in Fig. 24b.

In the case of the triadic highpass wavelet, for which the nominal pass band is the interval $(2\pi/3, \pi)$, the two copies of $B(3\omega)$ are translated to centres as $5\pi/6$ and $-5\pi/6$ to give

$$\xi_{3/3}(\omega) = \begin{cases} B(3\omega + 5\pi/6) + B(3\omega - 5\pi/6), & \text{if } \omega \in [-\pi/2, \pi/2], \\ 0, & \text{otherwise.} \end{cases} \tag{57}$$

This can also be obtained simply by translating the centre of $\xi_{1/3}(\omega)$ from $\omega = 0$ to $\omega = \pm\pi$. The feature becomes fully apparent only when the interval $[-\pi, \pi]$ is wrapped around the circle such that π and $-\pi$ coincide at the point diametrically opposite the point where $\omega = 0$. The result is shown in Fig. 24c in terms of the linear interval.

Let $\xi(t) \longleftrightarrow \xi(\omega)$ denote the autocorrelation function of any one of the triadic wavelets together with the energy function, which is its Fourier transform. Then, the relevant condition of sequential orthogonality is that $\xi(3t) = 0$ if $t \neq 0$. Define $\lambda = 3\omega$. Then,

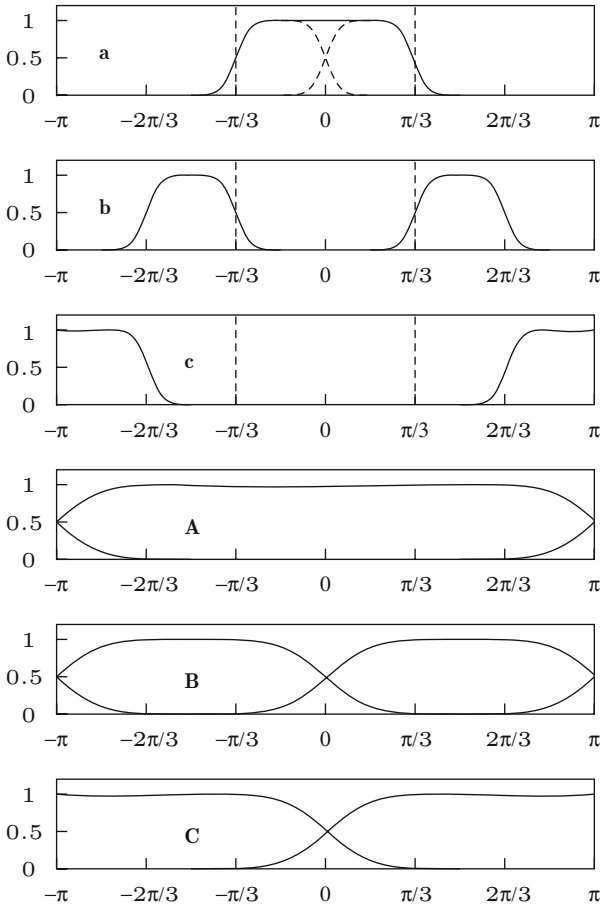

Fig. 24. The triadic energy functions, Figs. a–c. The segments of the latter, which are demarcated by the dotted lines and which are each of length $2\pi/3$, are dilated by a factor of 3 and overlaid on the interval $[-\pi, \pi]$ to form Figs. A–C

$$\xi(3t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \xi(\omega) e^{\mathrm{i}\omega(3t)} d\omega \tag{58}$$
$$= \frac{1}{6\pi}\int_{-3\pi}^{3\pi} \xi(\lambda/3) e^{\mathrm{i}\lambda t} d\lambda$$

$$= \frac{1}{6\pi}\left\{\int_{-3\pi}^{-\pi} \xi(\lambda/3) e^{\mathrm{i}\lambda t} d\lambda + \int_{-\pi}^{\pi} \xi(\lambda/3) e^{\mathrm{i}\lambda t} d\lambda + \int_{\pi}^{3\pi} \xi(\lambda/3) e^{\mathrm{i}\lambda t} d\lambda\right\}$$
$$= \frac{1}{6\pi}\int_{-\pi}^{-\pi} \sum_{j=-1}^{1} \xi([2\pi j + \lambda]/3) e^{\mathrm{i}\lambda t} d\lambda.$$

It follows that

$$\xi(3t) \longleftrightarrow \sum_{j=-1}^{1} \frac{1}{3}\xi([2\pi j + \lambda]/3) = \delta(\omega). \tag{59}$$

The condition of sequential orthogonality is that $\delta(\omega)$ must be a constant function such that its Fourier transform is the unit impulse. Figs. 24(A–C) show how the segments of the three energy functions that are demarcated by the dotted lines are dilated and overlaid on the interval $[-\pi, \pi]$. In each case, adding the segments produces the constant function $\delta(\omega) = c$. The overlaying of the segments occurs when each is wrapped around the same circle of circumference 2π. In fact, the segments need not be separated one from another. They can be wrapped around the circle in one continuous strip.

The condition of lateral orthogonality cannot be satisfied by wavelets in adjacent frequency bands. This a consequence of the spectral leakage from each band into the neighbouring bands. However, in the present triadic specification, which interpolates a third band between the lowpass and highpass bands and which limits the extent of the leakage on either side to one half of the nominal bandwidth, conditions of lateral orthogonality prevail between non-adjacent bands.

Now let us consider a bandpass filter with a nominal width of $\pi/3$ centered, in the positive frequency range, on some point $\theta \in [\pi/6, 5\pi/6]$ that lies between the centres of the lowpass and the highpass filters. The energy function of the filter is specified over the interval $[-\pi, \pi]$ by

$$\xi_{\theta/3}(\omega) = \begin{cases} B(3\omega + \theta) & \text{if } \omega \in [\theta - \pi/3, \theta + \pi/3], \\ B(3\omega - \theta), & \text{if } \omega \in [-\pi/3 - \theta, \pi/3 - \theta], \\ 0, & \text{otherwise.} \end{cases} \tag{60}$$

It can be shown that, regardless of the actual value taken by θ within the designated range, the condition $\xi_{\theta/3}(6t) = 0$ prevails for all $t \neq 0$, which is to say that wavelets within the band that are separated by multiples of 6 points are mutually orthogonal.

To demonstrate this, we must consider the decomposition $\xi_{\theta/3}(\omega) = \xi^{+}_{\theta/3}(\omega) + \xi^{-}_{\theta/3}(\omega)$, where $\xi^{+}_{\theta/3}(\omega)$ has a zero segment in place of the segment of $B(3\omega - \theta)$, and where $\xi^{-}_{\theta/3}(\omega)$ has a zero segment in place of the segment of $B(3\omega + \theta)$. Since θ is arbitrary, the interaction of $\xi^{+}_{\theta/3}(\omega)$ and $\xi^{-}_{\theta/3}(\omega)$ is undetermined, and we must treat the two functions separately.

In order that $\xi^{+}_{\theta/3}(\omega)$ should generate a uniform function over the interval $[-\pi, \pi]$ and beyond, it must be dilated by a factor of 6 before being wrapped around the circle. Then, the pass band, which has a (nominal) width of $\pi/3$, will acquire a width of 2π, which is sufficient to encompass the circle with a band of constant height. Equivalent conditions apply to $\xi^{-}_{\theta/3}(\omega)$. The upshot is that the wavelets that lie within a pass band of width $\pi/3$, located at an arbitrary centre, are mutually orthogonal when separated by multiples of 6 points.

The generalisations of the analysis of this section from $q = 3$ to cases of other integers is immediate. For the case where the interval $[0, \pi]$ is divided in $q > 2$ bands of equal width, the condition for the sequential orthogonality of wavelets separated by q points is that

$$\xi(qt) \longleftrightarrow \sum_{j=(1-q)/2}^{(q-1)/2} \frac{1}{q}\xi([2\pi j + \lambda]/q) = c. \tag{61}$$

For a band of width π/q, with $q \geq 2$, centred on an arbitrary point θ within $[\pi/2q, \pi - \pi/2q]$, the proof that wavelets separated by multiples $2q$ points are mutually orthogonal in indicated by the proof for the case where $q = 3$. We shall conclude the paper with an example that shows of how these conditions can be used in the analysis of the finite data sequence that was described in the introduction.

Example 8. Figure 25 shows the time–frequency plane for 144 data points, partitioned in a manner that is appropriate to the analysis of the monthly airline passenger data of Fig. 5. The bands that have been highlighted cover the spectral structure of the seasonal fluctuations that is revealed by the periodogram of Fig. 6. On either side of the the seasonal frequencies $\{\pi j/6; j = 1, \ldots, 5\}$, there are adjacent bands of $7\frac{1}{2}$ degrees in width. Altogether, there are 24 bands of equal width dividing the frequency range, and the time span of the sample is divided into six sections, each of which spans a two-year period.

With this partitioning, it should be possible to reveal the evolution of the seasonal pattern by showing the progression of the amplitude coefficients of the wavelets within the highlighted bands. In testing the statistical null hypothesis of temporal homogeneity, which is liable to be rejected, it is helpful to have wavelets that are mutually orthogonal.

The interstices between the highlighted bands are effective in ensuring the lateral orthogonality of the wavelets, whenever they are derived from one of the templates that have been provided in Sect. 6. However, the wavelets in the contiguous bands that fall on either side of the frequencies $\{\pi j/6; j =$

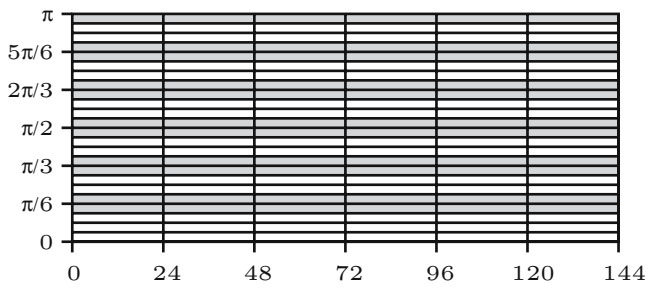

Fig. 25. The time–frequency plane for 144 data points partitioned with 24 frequency intervals and 6 time periods

$1, \ldots, 5\}$ will not be mutually orthogonal. This problem can be overcome by combining these bands. The combined bands will be populated by twice as many wavelets as the original narrower bands. However, the distances that separate orthogonal wavelets will remain the same at 24 points.

References

Bloomfield, P., (1976), *Fourier Analysis of Time Series: An Introduction,* John Wiley and Sons, New York.

Boggess, A., and F.J. Narcowich, (2001), *A First Course in Wavelets with Fourier Analysis,* Prentice-Hall, Upper Saddle River, New Jersey.

Box, G.E.P., and G.M. Jenkins, (1976), *Time Series Analysis: Forecasting and Control,* Revised Edition, Holden-Day, San Francisco.

Burrus, C.S, R.A. Gopinath and H. Guo, (1998), *Introduction to Wavelets and Wavelet Transforms: A Primer,* Prentice-Hall, Upper Saddle River, New Jersey.

Daubechies, I., (1988), Orthonormal Bases of Compactly Supported Wavelets, *Communications in Pure and Applied Mathematics,* 41, 909–996.

Daubechies, I., (1992), *Ten Lectures on Wavelets,* Society for Industrial and Applied Mathematics, Philadelphia.

Dirac, P.A.M., (1958), *The Principles of Quantum Mechanics, Fourth Edition,* Oxford University Press, Oxford.

Gopinath, R.A., and C.S. Burrus, (1993), Wavelet Transforms and Filter Banks. In C.K. Chui, editor, *Wavelets: A Tutorial in Theory and Applications,* 603–655, Academic Press, San Diego. Volume 2 of *Wavelet Analysis and its Applications.*

Mallat, S.G., (1989), A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 11, 674–693.

Misiti, M., Y. Misiti, G. Oppenheim and J-M. Poggi, (1997), *Wavelet Toolbox for Use with MATLAB,* The MathWorks Inc., Natick, Massachusetts.

Nason, G.P., and R. von Sachs, (1999), Wavelets in Time Series Analysis, *Philosophical Transactions of the Royal Society of London, Series A,* 357, 2511–2526.

Nason, G.P., and T. Sapatinas, (2002), Wavelet Packet Transfer Function Modelling of Nonstationary Time Series, *Statistics and Computing,* 12, 45–56.

Newland, D.E., (1993), *An Introduction to Random Vibrations, Spectral Analysis and Wavelets: 3rd edition,* Longman Scientific and Technical, Harlow, Essex.

Percival, D.B., and A.T. Walden (2000), *Wavelet Methods for Time Series Analysis,* Cambridge University Press, Cambridge.

Pollock, D.S.G., (1999), *Time-Series Analysis, Signal Processing and Dynamics,* Academic Press, London.

Pollock, D.S.G., (2002), Circulant Matrices and Time-Series Analysis, *The International Journal of Mathematical Education in Science and Technology,* 33, 213–230.

Ramsey, J.B., and C. Lampart, (1998), The Decomposition of Economic Relationships by Time Scale Using Wavelets: Expenditure and Income, *Studies in Nonlinear Dynamics and Econometrics,* 3, 23–42.

Ramsey, J.B., (1999), The Contribution of Wavelets to the Analysis of Economics and Financial Data, *Philosophical Transactions of the Royal Society of London, Series A,* 357, 2593–2606.

Shannon, C.E., and W. Weaver, (1964), *The Mathematical Theory of Communication,* The University of Illinois Press, Urbana, Illinois.

Steffen, P., P.N. Heller, R.A. Gopinath, and C.S. Burrus, (1993), Theory of Regular M-band Wavelets, *IEEE Transactions on Signal Processing,* 41, 3497–3511.

Strang, G., and T. Nguyen, (1997), *Wavelets and Filter Banks,* Prentice-Hall, Upper Saddle River, New Jersey.

Vaidyanathan, P.P., (1990), Multirate Digital Filters, Filter Banks, Polyphase Networks and Applications: A Tutorial, *Proceedings of the IEEE,* 78, 56–93.

Vaidyanathan, P.P., (1993), Multirate Systems and Filter Banks, *Proceedings of the IEEE,* 78, 56–93.

Vetterli, M., and J. Kovacević (1995), *Wavelets and Subband Coding,* Prentice-Hall, Upper Saddle River, New Jersey.

Wickerhauser, V.M., (1994), *Adapted Wavelet Analysis from Theory to Software,* A.K. Peters, Natick, Massachusetts.

Woodward, P.M., (1953), *Probability and Information Theory with Applications to Radar,* McGraw-Hill Book Company, New York.

Measuring Core Inflation by Multivariate Structural Time Series Models*

Tommaso Proietti

Dipartimento S.E.F. e ME. Q., University of Rome "Tor Vergata"

Summary. The measurement of core inflation can be carried out by optimal signal extraction techniques based on the multivariate local level model, by imposing suitable restrictions on its parameters. The various restrictions correspond to several specialisations of the model: the core inflation measure becomes the optimal estimate of the common trend in a multivariate time series of inflation rates for a variety of goods and services, or it becomes a minimum variance linear combination of the inflation rates, or it represents the component generated by the common disturbances in a dynamic error component formulation of the multivariate local level model. Particular attention is given to the characterisation of the optimal weighting functions and to the design of signal extraction filters that can be viewed as two sided exponentially weighted moving averages applied to a cross-sectional average of individual inflation rates. An empirical application relative to U.S. monthly inflation rates for 8 expenditure categories is proposed.

Key words: Common trends, dynamic factor analysis, homogeneity, exponential smoothing, Wiener–Kolmogorov filter

1 Introduction

Core inflation measures are considered to be more appropriate for the assessment of the trend movements in aggregate prices than is the official aggregate inflation rate. It is usually thought that the raw inflation rate, obtained as the percentage change in the consumer price index (CPI, henceforth) over a given horizon, is too noisy to provide a good indication of the inflationary pressures in the economy.

* Address for Correspondence: Via Columbia 2, 00133, Rome. E-mail: tommaso.proietti@uniroma2.it. The author wishes to thank Stephen Pollock and two anonymous referees for their very helpful suggestions.

Like many other key concept in economics, there is no consensus on the notion of core inflation, despite the fact that quasi-official measures are routinely produced by statistical agencies. This is because the notion serves a variety of purposes. Nevertheless, an increasing number of indices of core inflation are being produced in a variety of ways.

As a consequence, a large body of literature has been devoted to core inflation. An excellent review is Wynne (1999), who makes a basic distinction between methods which use only sectional information, and those which also use the time dimension. Another useful distinction is between aggregate or disaggregate approaches.

The most popular measures fall within the disaggregate approach, using only the cross-sectional distribution of inflation rates at a given point in time. They aim at reducing the influence of items that are presumed to be more volatile, such as food and energy. Other measures exclude mortgage interest costs, and some also exclude the changes in indirect taxes.

Bryan and Cecchetti (1994) (see also Bryan, Cecchetti and Wiggins II, 1997) argue that the systematic exclusion of specific items, such as food and energy, is arbitrary, and, after remarking that the distribution of relative price changes exhibits skewness and kurtosis, propose to use the median or the trimmed mean of the cross-sectional distribution.

Cross-sectional measures, using only contemporaneous price information, are not subject to revision as new temporal observations become available, and this is often, although mistakenly, seen as an advantage. The corresponding core inflation measures tend to be rather rough and do not provide clear signals of the underlying inflation. We show here that measures that are constructed via a time series approach are better behaved.

Other approaches arise in the structural vector autoregressive (VAR) framework, starting from the seminal work of Quah and Vahey (1995), who, within a bivariate stationary VAR model of real output growth and inflation, defined core inflation as that component of measured inflation which has no long run effect on real output.

This paper considers the measurement of core inflation in an unobserved components framework; in particular, the focus will be on dynamic models that take a stochastic approach to the measurement of inflation, such as those introduced in Selvanathan and Prasada Rao (1994, Chap. 4). We propose and illustrate measures of core inflation that arise when standard signal extraction principles are applied to restricted versions of a workhorse model, which is the multivariate local level model (MLLM, henceforth).

The parametric restrictions are introduced in order to accommodate several important cases: the first is when the core inflation measure is the optimal estimate of the common trend in a multivariate time series of inflation rates for a variety of goods and services. In an alternative formulation it is provided by the minimum variance linear combination of the inflation rates. In another it arises as the component generated by the common disturbances in a dynamic error component formulation of the MLLM.

Particular attention is devoted to the characterisation of the Wiener–Kolmogorov optimal weighting functions and to the design of signal extraction filters that can be viewed as a two sided exponentially weighted moving averages applied to a contemporaneously aggregated inflation series.

The paper is organised as follows. Section 2 deals with aggregate measures of core inflation and their limitations. The MLLM and its main characteristics are presented in Sect. 3. Signal extraction for the unrestricted MLLM is considered in Sect. 4. Section 5 introduces several measures of core inflation that can be derived from the MLLM under suitable restrictions of its parameters. In particular, we entertain three classes of restrictions, namely the common trend, the homogeneity, and the dynamic error components restrictions. Section 6 derives the signal extraction filters for the dynamic factor model proposed by Bryan and Cecchetti (1994); and it compares them with those derived from the MLLM. Inference and testing for the MLLM and its restricted versions are the topic of Sect. 7). Finally, in Sect. 8, the measures considered in the paper are illustrated with reference to a set of U.S. time series concerning the monthly inflation rates for 8 expenditure categories.

2 Aggregate Measure of Core Inflation

Statistical agencies publish regularly two basic descriptive measures of inflation that are built upon a consumer or retail price index. The first is the percentage change over the previous month. The second is the percent change with respect to the same month of the previous year.

Unfortunately, neither index is a satisfactory measure of trend inflation. The first turns out to be very volatile, as it is illustrated by the upper panel of Fig. 1, which displays the monthly inflation rates for the for U.S. consumer price index (city average, source: Bureau of Labor Statistics) for the sample period 1993.1–2003.8. By contrast, the annual changes in relative prices are much smoother (see the lower panel of Fig. 1), but, being based on an asymmetric filter, they suffer from a phase shift in the signal, which affects the timing of the turning points in inflation. Furthermore, if the consumer price index is strictly non seasonal, then the series of yearly inflation rates is non invertible at the seasonal frequencies. With p_t representing the price index series, and with $y_t = \ln(p_t/p_{t-1})$, the yearly inflation rate is approximately $S(L)y_t$, where $S(L) = 1 + L + L^2 + \cdots + L^{11}$, which is a one sided filter with zeros at the seasonal frequencies.

One approach, which is followed by statistical agencies, is to reduce the volatility of inflation by discarding the goods or services that are presumed to be more volatile, such as food and energy. The monthly and yearly inflation rates constructed from the CPI excluding *Food and Energy* are indeed characterised by reduced variability, as Fig. 1 shows; yet they are far from satisfactory and they can be criticised on several grounds, not the least of which is their lack of smoothness.

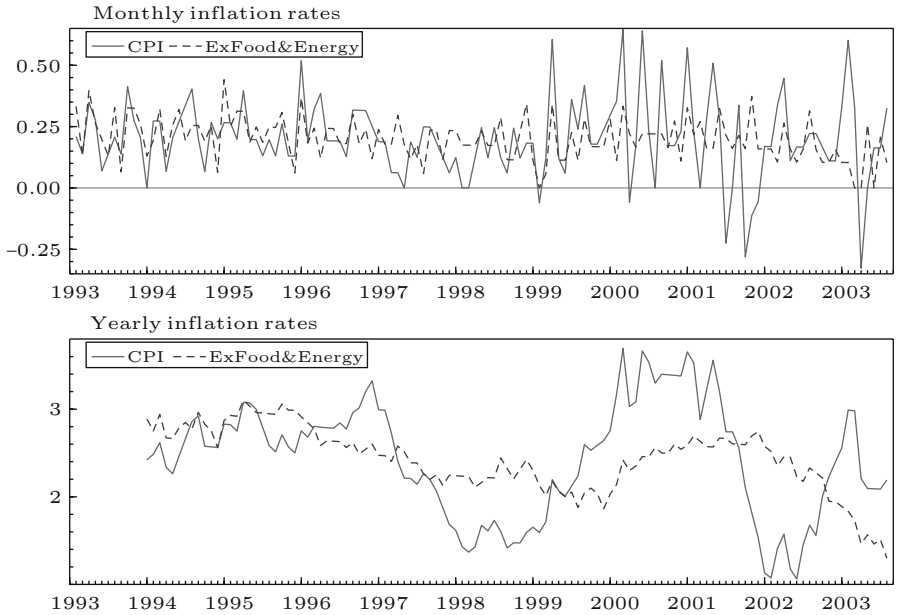

Fig. 1. U.S. CPI Total and Excluding Food & Energy, 1993.1–2003.8. Monthly and yearly inflation rates

3 The Multivariate Local Level Model

The measures of core inflation proposed in this paper arise from applying optimal signal extraction techniques derived from various restricted versions of a multivariate times series model. The model in question is the multivariate generalisation of the local level model (MLLM), according to which a multivariate time series can be decomposed into a trend component, represented by a multivariate random walk, and a white noise (WN) component. Letting $\mathbf{y}_t$ denote an $N \times 1$ vector time series referring to the monthly changes in the prices of N consumer goods and services,

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t, \qquad t = 1, 2, \ldots, T, \qquad \boldsymbol{\epsilon}_t \sim \mathrm{WN}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon), \\ \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t-1} + \boldsymbol{\eta}_t, \qquad\qquad\qquad\quad \boldsymbol{\eta}_t \sim \mathrm{WN}(\mathbf{0}, \boldsymbol{\Sigma}_\eta). \end{aligned} \tag{1}$$

The disturbances $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$ are assumed to be mutually uncorrelated and uncorrelated to $\boldsymbol{\mu}_0$.

Before considering restricted versions of the model, we review its main features both in the time and the frequency domain (see Harvey, 1989, for more details). The model assumes that the monthly inflation rates are integrated of order one (prices are integrated of order two). This assumption can actually be tested. In Sect. 7 we review the locally best invariant test of the hypothesis that monthly inflation rates are stationary versus the alternative

that they are I(1). Taking first differences, we can reexpress model (1) in its stationary form:

$$\Delta \mathbf{y}_t = \boldsymbol{\eta}_t + \Delta \boldsymbol{\epsilon}_t.$$

The crosscovariance matrices of $\Delta \mathbf{y}_t$, $\boldsymbol{\Gamma}_\Delta(\tau) = \mathrm{E}(\Delta \mathbf{y}_t \Delta \mathbf{y}'_{t-\tau})$ are then

$$\begin{aligned} \boldsymbol{\Gamma}_\Delta(0) &= \boldsymbol{\Sigma}_\eta + 2\boldsymbol{\Sigma}_\epsilon, \\ \boldsymbol{\Gamma}_\Delta(1) &= \boldsymbol{\Gamma}_\Delta(-1)' = -\boldsymbol{\Sigma}_\epsilon, \\ \boldsymbol{\Gamma}_\Delta(\tau) &= \mathbf{0}, \quad |\tau| > 1. \end{aligned}$$

Notice that the autocovariance at lag 1 is negative (semi)definite and symmetric: $\boldsymbol{\Gamma}_\Delta(1) = \boldsymbol{\Gamma}_\Delta(1)' = \boldsymbol{\Gamma}_\Delta(-1)$. This symmetry property implies that the multivariate spectrum is real-valued. Denoting by $\mathbf{F}(\lambda)$ the spectral density of $\Delta \mathbf{y}_t$ at the frequency λ, we have $\mathbf{F}(\lambda) = (2\pi)^{-1}[\boldsymbol{\Sigma}_\eta + 2(1-\cos\lambda)\boldsymbol{\Sigma}_\epsilon]$. The autocovariance generating function (ACGF) of $\Delta \mathbf{y}_t$ is

$$\mathbf{G}(L) = \boldsymbol{\Sigma}_\eta + |1-L|^2 \boldsymbol{\Sigma}_\epsilon. \tag{2}$$

The reduced form of the MLLM is a multivariate IMA(1,1) model:

$$\Delta \mathbf{y}_t = \boldsymbol{\xi}_t + \boldsymbol{\Theta} \boldsymbol{\xi}_{t-1}.$$

Equating (2) to the ACGF of the vector MA(1) representation for $\Delta \mathbf{y}_t$, it is possible to show that the parameterisation (1) is related to the reduced form parameters via:

$$\boldsymbol{\Sigma}_\eta = (\mathbf{I} + \boldsymbol{\Theta})\boldsymbol{\Sigma}_\xi (\mathbf{I} + \boldsymbol{\Theta}'), \quad \boldsymbol{\Sigma}_\varepsilon = -\boldsymbol{\Theta}\boldsymbol{\Sigma}_\xi = -\boldsymbol{\Sigma}_\xi \boldsymbol{\Theta}'.$$

The structural form has $N(N+1)$ parameters, whereas the unrestricted vector IMA(1,1) model has $N^2 + N(N+1)/2$. In fact, $\boldsymbol{\Sigma}_\varepsilon = -\boldsymbol{\Theta}\boldsymbol{\Sigma}_\xi = -\boldsymbol{\Sigma}_\xi \boldsymbol{\Theta}'$ imposes $N(N-1)/2$ restrictions.

4 Signal Extraction

Assuming a doubly infinite sample, the minimum mean square linear estimator (MMSLE) of the underlying level component is

$$\tilde{\boldsymbol{\mu}}_t = \mathbf{W}(L)\mathbf{y}_t,$$

with weighting matrix polynomial

$$\mathbf{W}(L) = |1-L|^2 \mathbf{G}_\mu(L)\mathbf{G}(L)^{-1} = \boldsymbol{\Sigma}_\eta \left(\boldsymbol{\Sigma}_\eta + |1-L|^2 \boldsymbol{\Sigma}_\epsilon\right)^{-1},$$

where $\mathbf{G}_\mu(L)$ is the pseudo ACGF of the trend component and we have defined $|1-L|^2 = (1-L)(1-L^{-1})$. This results from the application of the Wiener–Kolmogorov (WK, henceforth) filtering formulae given in Whittle (1983), which apply also to the nonstationary case (Bell, 1984).

The matrix polynomial $\mathbf{W}(L)$ performs two-sided exponential smoothing

$$\mathbf{W}(L) = (\mathbf{I} + \boldsymbol{\Theta})\boldsymbol{\Sigma}_\xi(\mathbf{I} + \boldsymbol{\Theta}')(\mathbf{I} + \boldsymbol{\Theta}' L^{-1})^{-1}\boldsymbol{\Sigma}_\xi^{-1}(\mathbf{I} + \boldsymbol{\Theta} L)^{-1}$$

and it has $\mathbf{W}(1) = \mathbf{I}_N$, which generalises to the multivariate case the unit sum property of the weights for the extraction of the trend component.

The filtered or real time estimator of the trend is an exponentially weighted average of current and past observations:

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{t|t} &= (\mathbf{I} + \boldsymbol{\Theta})(\mathbf{I} + \boldsymbol{\Theta} L)^{-1}\mathbf{y}_t \\ &= (\mathbf{I} + \boldsymbol{\Theta}) \textstyle\sum_{j=0}^{\infty}(-\boldsymbol{\Theta})^j \mathbf{y}_{t-j}.\end{aligned}$$

5 Measures of Core Inflation Derived from the MLLM

In this section we explore that have to be imposed on the WK filter to make it yield univariate summary measures of tendency of the form:

$$\tilde{\mu}_t = \mathbf{w}(L)'\mathbf{y}_t, \quad \mathbf{w}(L) = q(L)\mathbf{w} \tag{3}$$

where $q(L)$ is a univariate symmetric two-sided filter and $\mathbf{w}$ is a static vector of cross-sectional weights. Purely static measures arise when $q(L) = 1$. The signal extraction filters of (3) are the basis of the measurement of core inflation, when $\mathbf{y}_t$ represents N inflation rates that have to be combined in a single measure.

The cross-sectional weights can be model based or they can originate from a priori knowledge. It is instructive to look at the various ways that they can originate and at their different various meanings.

5.1 Aggregate Measures (Known Weights)

The first core inflation measures arise from the contemporaneous aggregation of the multivariate trend component in (1) using known weights. The MLLM is invariant under contemporaneous aggregation; thus, $\mathbf{w}'\mathbf{y}_t$, where $\mathbf{w}$ is a vector of known weights (e.g. expenditure shares in the core inflation example), follows a univariate local level model.

The aggregated time series, $\mathbf{w}'\mathbf{y}_t$, has thus a local level model representation, and the miminum-mean-square linear estimator of the trend component, $\mathbf{w}'\boldsymbol{\mu}_t$, based on a doubly infinite sample, has the above structure (3), with:

$$q(L) = \frac{1}{1 + q^{-1}|1 - L|^2} = \frac{(1+\theta)^2}{|1 + \theta L|^2} \tag{4}$$

and $q = \mathbf{w}'\boldsymbol{\Sigma}_\eta\mathbf{w}/\mathbf{w}'\boldsymbol{\Sigma}_\epsilon\mathbf{w}$, and $\theta = [\sqrt{(q^2 + 4q)} - 2 - q]/2$, $-1 < \theta \leq 0$.

Alternatively, we could fit a univariate local level model to the aggregate time series. The corresponding core inflation measure is given by (4), but q would be estimated directly, rather than obtained from the aggregation of the covariance matrix of the disturbances of the multivariate specification.

5.2 Common Trend

Common trends arise when $\text{rank}(\mathbf{\Sigma}_\eta) = K < N$, so that

$$\mathbf{\Sigma}_\eta = \mathbf{Z}\mathbf{\Sigma}_{\eta^\dagger}\mathbf{Z}'$$

where $\mathbf{Z}$ is $N \times K$ and $\mathbf{\Sigma}_{\eta^\dagger}$ is a full rank $K \times K$ matrix.

When there is a single common trend, $K = 1$, driving the $\boldsymbol{\mu}_t$'s in (1), we can write:

$$\mathbf{y}_t = \mathbf{z}\mu_t + \boldsymbol{\mu}_0 + \boldsymbol{\epsilon}_t,$$

where $\mathbf{z}$ is a $N \times 1$ vector of loadings and $\mu_t = \mu_{t-1} + \eta_t, \eta_t \sim \text{WN}(0, \sigma_\eta^2)$.

The WK filter for μ_t, assuming a doubly infinite sample, takes the form (3) with

$$\mathbf{w} = (\mathbf{z}'\mathbf{\Sigma}_\epsilon^{-1}\mathbf{z})^{-1}\mathbf{\Sigma}_\epsilon^{-1}\mathbf{z}, \tag{5}$$

and $q(L)$ given by (4), where the signal–noise ratio is given by

$$q = \sigma_\eta^2(\mathbf{z}'\mathbf{\Sigma}_\epsilon^{-1}\mathbf{z}).$$

If $\mathbf{\Sigma}_\epsilon$ is diagonal (i.e. if it represents the idiosyncratic noise) and $\mathbf{z}$ is a constant vector (the common trend enters the series with the same coefficient) then the cross-sectional weights (5) applied to $\mathbf{y}_t$ produce a weighted average $\bar{y}_t = \mathbf{w}'\mathbf{y}_t$, in which the more noisy series are downweighted. The application of the univariate two sided filter $q(L)$, which is a bidirectional exponentially weighted average, to $\bar{y}_t$ yields the estimated component.

The expression (5) assumes that $\mathbf{\Sigma}_\epsilon$ is full rank; if its rank is $N - 1$ then $q(L) = 1$ and $\mathbf{w} = (\mathbf{v}'\mathbf{z})^{-1}\mathbf{v}$, where $\mathbf{v}$ is the eigenvector corresponding to the zero eigenvalue of $\mathbf{\Sigma}_\epsilon$. Hence, in this special case, the filter is fully static.

5.3 Homogeneity

The MLLM is said to be *homogeneous* if the covariance matrices of the disturbances are proportional (see Enns et al., 1982, and Harvey, 1989, Chap. 8):

$$\mathbf{\Sigma}_\eta = q\mathbf{\Sigma}_\epsilon.$$

Here, q denotes the proportionality factor.

Under homogeneity, the model is a seemingly unrelated IMA(1,1) process $\Delta\mathbf{y}_t = \boldsymbol{\xi}_t + \theta\boldsymbol{\xi}_{t-1}$, with scalar MA parameter, $\theta = [\sqrt{(q^2 + 4q} - 2 - q]/2$, taking values in [-1,0], and $\boldsymbol{\xi}_t \sim \text{WN}(\mathbf{0}, \mathbf{\Sigma}_\xi), \mathbf{\Sigma}_\xi = -\mathbf{\Sigma}_\epsilon/\theta$.

The trend extraction filter is scalar and can be applied to each series in turn:

$$\tilde{\boldsymbol{\mu}}_t = \frac{1}{1 + q^{-1}|1 - L|^2}\mathbf{y}_t.$$

As a matter of fact, the Kalman filter and smoother become decoupled, and inferences are particularly simplified (see also Sect. 7).

Consider forming a linear combination of the trend component $\boldsymbol{\mu}_t$: $\bar{\mu}_t = \mathbf{w}'\boldsymbol{\mu}_t$. Obviously,

$$\tilde{\bar{\mu}}_{t|\infty} = \frac{1}{1+q^{-1}|1-L|^2}\mathbf{w}'\mathbf{y}_t.$$

If $\mathbf{w}$ is known (as in the case of expenditure shares), then the summary measure coincides with that arising from the aggregate approach. The difference, however, lies with the signal–noise ratio, which is estimated more efficiently if the model is homogeneous. Again, the measure of core inflation is a static weighted average, with given weights, of the individual trends characterising each of the series.

Consider now the alternative strategy of forming a measure of the type (3) by means of a static linear combination of the estimated trends, $\tilde{\mu}_t = \mathbf{w}'\tilde{\boldsymbol{\mu}}_t$, with weights

$$\mathbf{w} = \frac{\boldsymbol{\Sigma}_\eta^{-1}\mathbf{z}}{\mathbf{z}'\boldsymbol{\Sigma}_\eta^{-1}\mathbf{z}}. \tag{6}$$

It is easy to show that the weights $\mathbf{w}$ produce the linear combination $\mathbf{w}'\boldsymbol{\mu}_t$ which minimises the variance $\mathbf{w}'\boldsymbol{\Sigma}_\eta\mathbf{w}$ under the constraint $\mathbf{w}'\mathbf{z} = 1$, where $\mathbf{z}$ is an arbitrary vector. Hence, these weights provide the smoothest (i.e. the least variable) component that preserves the level ($\mathbf{w}'\mathbf{i} = 1$), where $\mathbf{i}$ is an $N \times 1$ vector with unit elements, $\mathbf{i} = [1, \ldots, 1]'$.

Another interpretation of (6) is that $\mathbf{w}'\boldsymbol{\mu}_t$ is the GLS estimates of μ_t in the model $\boldsymbol{\mu}_t = \mathbf{z}\mu_t + \boldsymbol{\mu}_t^*$, considered as a fixed effect; $\mathbf{w}'\mathbf{y}_t$ are known as Bartlett scores in factor analysis (see Anderson, 1984, p. 575). Notice, however, that here $\mathbf{z}$ is a known vector, that has to be specified a priori (e.g. we may look for weights summing up to unity, in which circumstance, $\mathbf{z} = \mathbf{i}$). It is not a necessary feature of the true model.

5.4 Dynamic Error Components

Suppose that the level disturbances have the following error components structure (Marshall, 1992):

$$\boldsymbol{\eta}_t = \mathbf{z}\eta_t + \boldsymbol{\eta}_t^*, \quad \eta_t \sim \mathrm{WN}(0, \sigma_\eta^2), \quad \boldsymbol{\eta}_t^* \sim \mathrm{WN}(\mathbf{0}, \mathbf{N}_\eta),$$

where η_t is disturbance that is common to all the trends and $\boldsymbol{\eta}_t^*$ is the idiosyncratic disturbance (typically, but not necessarily, $\mathbf{N}_\eta$ is a diagonal matrix). Correspondingly, $\boldsymbol{\Sigma}_\eta = \sigma_\eta^2\mathbf{z}\mathbf{z}' + \mathbf{N}_\eta$, and the trends can be rewritten as

$$\mu_t = \mathbf{z}\mu_t + \boldsymbol{\mu}_t^*, \;\; \Delta\mu_t = \eta_t, \;\; \Delta\boldsymbol{\mu}_t^* = \boldsymbol{\eta}_t^*.$$

In general, the WK filter for μ_t does not admit the representation (3), as we have:

$$\tilde{\mu}_t = \frac{\sigma_\eta^2}{1+\sigma_\eta^2\mathbf{z}'\mathbf{A}(L)^{-1}\mathbf{z}}\mathbf{z}'\mathbf{A}(L)^{-1}\mathbf{y}_t, \quad \mathbf{A}(L) = \mathbf{N}_\eta + |1-L|^2\boldsymbol{\Sigma}_\epsilon.$$

However, if $\mathbf{N}_\eta = q^*\boldsymbol{\Sigma}_\epsilon$, then the WK filter takes the form (3) with

$$q(L) = \frac{\sigma_\eta^2 \mathbf{z}'\mathbf{N}_\eta^{-1}\mathbf{z}}{\sigma_\eta^2 \mathbf{z}'\mathbf{N}_\eta^{-1}\mathbf{z} + 1 + q^{-1}|1-L|^2} \tag{7}$$

and

$$\mathbf{w} = \frac{\mathbf{N}_\eta^{-1}\mathbf{z}}{\mathbf{z}'\mathbf{N}_\eta^{-1}\mathbf{z}} = \frac{\boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{z}}{\mathbf{z}'\boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{z}}.$$

This type of homogeneity may arise, for instance, when the idiosyncratic trend disturbances are a fraction of the irregular component, that is $\boldsymbol{\eta}_t^* = \rho\boldsymbol{\epsilon}_t$, $\rho^2 = q^*$. This makes the overall trend and irregular components correlated, but μ_t would still be uncorrelated with $\boldsymbol{\epsilon}_t$.

Now let us consider the case when $\boldsymbol{\epsilon}_t$ has the same error components structure: $\boldsymbol{\epsilon}_t = \mathbf{z}\epsilon_t + \boldsymbol{\epsilon}_t^*$, with $\epsilon_t \sim \mathrm{WN}(0, \sigma_\epsilon^2)$, $\boldsymbol{\epsilon}_t^* \sim \mathrm{WN}(\mathbf{0}, \mathbf{N}_\epsilon)$, so that

$$\boldsymbol{\Sigma}_\epsilon = \sigma_\epsilon^2 \mathbf{z}\mathbf{z}' + \mathbf{N}_\epsilon.$$

The WK filter for μ_t is now

$$\tilde{\mu}_t = \frac{\sigma_\eta^2}{1 + \sigma_\eta^2 \mathbf{z}'\mathbf{A}(L)^{-1*}\mathbf{z}}\mathbf{z}'\mathbf{A}(L)^{-1*}\mathbf{y}_t, \quad \mathbf{A}(L)^* = \mathbf{N}_\eta + |1-L|^2\mathbf{N}_\epsilon,$$

and, under the homogeneity condition $\mathbf{N}_\eta = q^*\mathbf{N}_\epsilon$, produces exactly the same filter as the previous case, with q^* replacing q in (7).

Gathering the components driven by the common disturbances into $\varsigma_t = \mu_t + \epsilon_t$, and writing $\mathbf{y}_t = \mathbf{z}\varsigma_t + \boldsymbol{\mu}_t^* + \boldsymbol{\epsilon}_t^*$, the MMSLE of ς_t is

$$\tilde{\varsigma}_t = \frac{c(L)}{1 + c(L)\mathbf{z}'\mathbf{A}(L)^{-1*}\mathbf{z}}\mathbf{z}'\mathbf{A}(L)^{-1*}\mathbf{y}_t, \quad c(L) = \sigma_\eta^2(1+q^{-1}|1-L|^2), \;\; q = \sigma_\eta^2/\sigma_\epsilon^2.$$

Moreover, if $\mathbf{N}_\eta = q\mathbf{N}_\epsilon$, with the same q, $\sigma_\eta^2\mathbf{A}(L) = c(L)\mathbf{N}_\eta$, then $\tilde{\varsigma}_t$ is extracted by a static linear combination with weights

$$\mathbf{w} = \frac{\mathbf{N}_\eta^{-1}\mathbf{z}}{\sigma_\eta^{-2} + \mathbf{z}'\mathbf{N}_\eta^{-1}\mathbf{z}}.$$

Notice that this last case arises when the system is homogeneous $\boldsymbol{\Sigma}_\eta = q\boldsymbol{\Sigma}_\epsilon$, and $\boldsymbol{\Sigma}_\epsilon$ has an error component structure. Otherwise, if $\mathbf{N}_\eta = \sigma_{\eta^*}^2\mathbf{I}, \mathbf{N}_\epsilon = \sigma_{\epsilon^*}^2\mathbf{I}$, then the filter for ς_t is as in (3) with $\mathbf{w} = \mathbf{z}$ and

$$q(L) = \frac{\sigma_\eta^2(1+q^{-1}|1-L|^2)}{\sigma_{\eta*}^2(1+q^{-1*}|1-L|^2)}, q^* = \sigma_{\eta*}^2/\sigma_{\epsilon^*}^2$$

6 Dynamic Factor Models

This section discusses the signal extraction filters that are optimal for a class of dynamic factor models proposed by Stock and Watson (1991) for the purpose of constructing a model-based index of coincident indicators for the U.S. econonomy.

This class has been adopted by Bryan and Cecchetti (1994) for the measurement of core inflation, and it applies to a vector of monthly inflation rates, $\mathbf{y}_t$, which are expressed as as follows:

$$\begin{array}{lll} \mathbf{y}_t & = \mathbf{z}\mu_t + \boldsymbol{\mu}_t^*, & \\ \varphi(L)\mu_t & = \eta_t, & \eta_t \sim \mathrm{WN}(0, \sigma_\eta^2) \\ \mathbf{D}(L)\boldsymbol{\mu}_t^* & = \boldsymbol{\eta}_t^*, & \boldsymbol{\eta}_t^* \sim \mathrm{WN}(0, \mathbf{N}_\eta) \end{array} \tag{8}$$

where $\mathbf{D}(L) = \mathrm{diag}\{d_i(L), i = 1, \ldots, N\}$, and $\varphi(L)$ and $d_i(L)$ are AR scalar polynomials, possibly nonstationary, $\mathbf{N}_\eta = \mathrm{diag}\{\sigma_{i*}^2, i = 1, \ldots, N\}$, and η_t is uncorrelated with $\boldsymbol{\eta}_t^*$ at all leads and lags.

The autocovariance generating function of μ_t and the cross-covariance generating function of $\mathbf{y}_t$ are respectively:

$$g_\mu(L) = \frac{\sigma_\eta^2}{|\varphi(L)|^2}, \qquad \boldsymbol{\Gamma}_y(L) = g_\mu(L)\mathbf{z}\mathbf{z}' + \mathbf{M}(L),$$

where we have written

$$\mathbf{M}(L) = \mathbf{D}(L)^{-1}\mathbf{N}_\eta\mathbf{D}(L^{-1})^{-1} = \mathrm{diag}\{\sigma_{i*}^2|d_i(L)|^{-2}, i = 1, \ldots, N\}.$$

Moreover, the cross-covariance generating function between μ_t and $\mathbf{y}_t$ is simply $g_\mu(L)\mathbf{z}'$. Hence, the WK signal extraction formula is:

$$\begin{aligned} \tilde{\mu}_t &= g_\mu(L)\mathbf{z}'[\boldsymbol{\Gamma}_y(L)]^{-1}\mathbf{y}_t \\ &= \left[g_\mu(L)^{-1} + \mathbf{z}'\mathbf{M}(L)^{-1}\mathbf{z}\right]^{-1}\mathbf{z}'\mathbf{M}(L)^{-1}\mathbf{y}_t \\ &= \left[\sigma_\eta^{-2}|\varphi(L)|^2 + \textstyle\sum_i |d_i(L)|^2\sigma_{i*}^{-2}\right]^{-1}\sum_{i=1}^N |d_i(L)|^2\sigma_{i*}^{-2}y_{it}. \end{aligned}$$

When $\varphi(L) = d_i(L), i = 1, \ldots, N$, which is a seemingly unrelated time series equations (SUTSE) system, the WK specialises as follows:

$$\tilde{\mu}_t = \left[\sigma_\eta^{-2} + \sum_{i=1}^N \frac{1}{\sigma_{i*}^2}\right]^{-1} \sum_{i=1}^N \frac{1}{\sigma_{i*}^2}y_{it} = [\sigma_\eta^{-2} + \mathbf{z}'\mathbf{N}_\eta^{-1}\mathbf{z}]^{-1}\mathbf{z}'\mathbf{N}_\eta^{-1}\mathbf{y}_t.$$

Hence, $\tilde{\mu}_t$ results only from the contemporaneous aggregation of the individual time series, with weights that do not sum to unity, although they are still proportional to the reciprocal of the specific variances. If $\varphi(L) = \Delta$, then the dynamic factor model (8) is a special case of the MLLM (1), with no irregular component.

7 Inference and Testing

The state space methodology provides a means of computing the minimum-mean-square linear estimators (MMSLE) of the core inflation component, μ_t, and of any latent variable in the model. In finite samples the MMSLE of $\boldsymbol{\mu}_t$ in (1) is computed by Kalman filter and the associated smoothing algorithm (see Durbin and Koopman, 2001).

The Kalman filter (KF) is started off at $t = 2$ with $\tilde{\boldsymbol{\mu}}_{2|1} = \mathbf{y}_1$ and $\mathbf{P}_{2|1} = \boldsymbol{\Sigma}_\epsilon + \boldsymbol{\Sigma}_\eta$ computes for $t = 2, \ldots, T$:

$$\begin{aligned} \boldsymbol{\nu}_t &= \mathbf{y}_t - \tilde{\boldsymbol{\mu}}_{t|t-1}, & \mathbf{F}_t &= \mathbf{P}_{t|t-1} + \boldsymbol{\Sigma}_\epsilon \\ & & \mathbf{K}_t &= \mathbf{P}_{t|t-1}\mathbf{F}_t^{-1}, \\ \tilde{\boldsymbol{\mu}}_{t+1|t} &= \tilde{\boldsymbol{\mu}}_{t|t-1} + \mathbf{K}_t\boldsymbol{\nu}_t, & \mathbf{P}_{t+1|t} &= \mathbf{P}_{t|t-1} + \boldsymbol{\Sigma}_\eta - \mathbf{K}_t\mathbf{F}_t\mathbf{K}_t'. \end{aligned}$$

Denoting the information up to time t by $\mathbf{Y}_t = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t\}$, the above quantities have the following interpretation: $\boldsymbol{\nu}_t = \mathbf{y}_t - \mathrm{E}(\mathbf{y}_t|\mathbf{Y}_{t-1}), \mathbf{F}_t = \mathrm{Var}(\mathbf{y}_t|\mathbf{Y}_{t-1})$, $\tilde{\boldsymbol{\mu}}_{t|t-1} = \mathrm{E}(\boldsymbol{\mu}_t|\mathbf{Y}_{t-1}), \mathbf{P}_{t|t-1} = \mathrm{Var}(\boldsymbol{\mu}_t|\mathbf{Y}_{t-1})$ $\tilde{\boldsymbol{\mu}}_t = \mathrm{E}(\boldsymbol{\mu}_t|\mathbf{Y}_t)$, $\mathbf{P}_t = \mathrm{Var}(\boldsymbol{\mu}_t|\mathbf{Y}_t)$.

The Kalman filter performs the prediction error decomposition of the likelihood function. The latter can be maximised using a quasi-Newton numerical optimisation method.

When the model is homogeneous, inferences are made easier by the fact that the Kalman filter and smoother become decoupled. In fact, at $t = 2$, $\tilde{\boldsymbol{\mu}}_{2|1} = \mathbf{y}_1$ and $\mathbf{P}_{2|1} = (q+2)\boldsymbol{\Sigma}_\epsilon = p_{2|1}\boldsymbol{\Sigma}_\epsilon$, where we have written $p_{2|1} = (q+1)$. Now, consider the KF quantities that are independent of the data:

$$\begin{aligned} \mathbf{F}_t &= \mathbf{P}_{t|t-1} + \boldsymbol{\Sigma}_\epsilon, \quad \mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{F}_t^{-1}, \\ \mathbf{P}_{t+1|t} &= \mathbf{P}_{t|t-1} + q\boldsymbol{\Sigma}_\epsilon - \mathbf{P}_{t|t-1}\mathbf{F}_t^{-1}\mathbf{P}_{t|t-1}; \end{aligned}$$

$\mathbf{F}_t$ and $\mathbf{P}_{t+1|t}$ will be proportional to $\boldsymbol{\Sigma}_\epsilon$: $\mathbf{F}_t = f_t\boldsymbol{\Sigma}_\epsilon$, $\mathbf{P}_{t|t-1} = p_{t|t-1}\boldsymbol{\Sigma}_\epsilon$; also, $\mathbf{K}_t = p_{t|t-1}/(1 + p_{t|t-1})\mathbf{I}_N$, where the scalar quantities are delivered by the univariate KF for the LLM with signal to noise ratio q.

Hence, the innovations and inferences about the states can be from N univariate KFs. Correspondingly, it can be shown that $\boldsymbol{\Sigma}_\epsilon$ can be concentrated out the likelihood function, and the concentrated likelihood can be maximised with respect to q (see Harvey, 1989, p. 439).

7.1 Homogeneity Tests

The Lagrange multiplier test of the homogeneity restriction, $\mathrm{H}_0 : \boldsymbol{\Sigma}_\eta = q\boldsymbol{\Sigma}_\epsilon$, was derived in the frequency domain by Fernandez and Harvey (1990). The frequency domain log-likelihood function is built on the stationary representation of the local level model, $\Delta\mathbf{y}_t = \boldsymbol{\eta}_t + \Delta\boldsymbol{\epsilon}_t$ and it takes the form:

$$\mathcal{L}(\boldsymbol{\psi}) = -\frac{NT^*}{2}\ln 2\pi - \frac{1}{2}\sum_{j=0}^{T^*-1}\left\{\ln|\mathbf{G}_j| + 2\pi \cdot \mathrm{trace}\left[\mathbf{G}_j^{-1}\mathbf{I}^*(\lambda_j)\right]\right\},$$

where $T^* = T - 1$, $\boldsymbol{\psi}$ is a vector containing the $p = N(N+1)$ unknown parameters of the disturbance covariance matrices, G_j is the spectral generating function at frequency $\lambda_j = 2\pi j/T^*$, $\mathbf{G}_j = \boldsymbol{\Sigma}_\eta + 2(1-\cos\lambda_j)\boldsymbol{\Sigma}_\epsilon$ is the spectral generating function of the MLLM evaluated at the Fourier frequency λ_j and $\mathbf{I}^*(\lambda_j)$ is the (real part of) multivariate sample spectrum at the same frequency.

The LM test of the restriction $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ takes the form

$$LM = \mathsf{D}\mathcal{L}(\boldsymbol{\psi}_0)\mathcal{I}(\boldsymbol{\psi}_0)^{-1}\mathsf{D}\mathcal{L}(\boldsymbol{\psi}_0)' \tag{9}$$

where $\mathsf{D}\mathcal{L}(\boldsymbol{\psi}_0)$ is $1 \times p$ vector containing the partial derivatives with respect to the parameters evaluated at the null and $\mathcal{I}(\boldsymbol{\psi}_0)$ is the information matrix evaluated at $\boldsymbol{\psi}_0$. Expressions for $\mathsf{D}\mathcal{L}(\boldsymbol{\psi}_0)$ and $\mathcal{I}(\boldsymbol{\psi}_0)$ are given in Fernandez and Harvey (1990).

The unrestricted local level model has $N(N+1)$ parameters, whereas the homogenous model has $N(N+1)/2+1$, so the test statistic (9) is asymptotically distributed as a χ^2 random variable with $N(N+1)/2-1$ degrees of freedom.

The homogeneous dynamic error component model further restricts $\boldsymbol{\Sigma}_\epsilon = \sigma_{\bar{\epsilon}}\mathbf{i}\mathbf{i}' + \mathbf{N}_\epsilon$, and when the disturbances $\boldsymbol{\epsilon}_t^*$ are fully idiosyncratic, the model has only $N+2$ parameters. This restriction can be tested using (9), which gives a χ^2 test with $N(N+1) - N - 2$ degrees of freedom.

7.2 Testing for a Multivariate RW and for Common Trends

Nyblom and Harvey (2003, NH henceforth) have developed the locally best invariant test of the hypothesis $H_0 : \boldsymbol{\Sigma}_\eta = \mathbf{0}$ against the homogenous alternative $H_1 : \boldsymbol{\Sigma}_\eta = q\boldsymbol{\Sigma}_\epsilon$. The test statistic is

$$\xi_N = \mathrm{tr}[\hat{\boldsymbol{\Gamma}}^{-1}\mathbf{S}],$$

where

$$\mathbf{S} = \frac{1}{T^2}\sum_{t=1}^{T}\left[\sum_{i=1}^{t}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'\right]$$

$$\hat{\boldsymbol{\Gamma}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})'$$

and has rejection region $\xi_N > c$. Under the null hypothesis, the limiting distribution of ξ_N is Cramèr-von Mises with N degrees of freedom. Although the test maximises the power against a homogeneous alternative, it is consistent for any $\boldsymbol{\Sigma}_\eta$.

A non parametric adjustment, along the lines of the KPSS test, is required when ϵ_t is serially correlated and heteroscedastic. This is obtained by replacing $\hat{\mathbf{\Gamma}}$ with

$$\hat{\mathbf{\Gamma}}_l = \hat{\mathbf{\Gamma}}(0) + \sum_{\tau=1}^{l} \left(1 - \frac{\tau}{l+1}\right) [\hat{\mathbf{\Gamma}}(\tau) + \hat{\mathbf{\Gamma}}(\tau)']$$

where $\hat{\mathbf{\Gamma}}(\tau)$ is the autocovariance of $\mathbf{y}_t$ at lag τ.

When the test is applied to the linear transformation $\mathbf{A}'\mathbf{y}_t$, where $\mathbf{A}$ is a known $r \times N$ matrix, testing the stationarity of $\mathbf{A}'\mathbf{y}_t$ amounts to testing the null that there are r cointegrating relationships. If $\mathbf{A}'\mathbf{\Sigma}_\eta\mathbf{A} = 0$, then we can rewrite $\mathbf{y}_t = \mathbf{Z}\boldsymbol{\mu}_t + \boldsymbol{\mu}_0 + \boldsymbol{\epsilon}_t$, with $\mathbf{A}'\mathbf{Z} = \mathbf{0}$, so that $\mathbf{A}'\mathbf{y}_t = \mathbf{A}'\boldsymbol{\mu}_0 + \mathbf{A}'\boldsymbol{\epsilon}_t$.

The test statistics for this hypothesis is

$$\xi_r(\mathbf{A}) = \mathrm{tr}[(\mathbf{A}'\hat{\mathbf{\Gamma}}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}\mathbf{A}],$$

and its limiting distribution is Cramèr-von Mises with r degrees of freedom.

NH also consider the test of the null hypothesis that there are k common trends, versus the alternative that there are more.

$$\mathrm{H}_0 : \mathrm{rank}(\mathbf{\Sigma}_\eta) = k, \quad \text{vs.} \quad \mathrm{H}_1 : \mathrm{rank}(\mathbf{\Sigma}_\eta) > k.$$

The test statistic is based on the sum of the $N - k$ smallest eigenvalues of $\hat{\mathbf{\Gamma}}^{-1}\mathbf{S}$,

$$\zeta(k, N) = \lambda_{k+1} + \ldots + \lambda_N$$

The significance points of $\zeta(k, N)$ are tabulated in NH (2003) for a set of (k, N) pairs.

8 Illustration

Our illustrative example deals with extracting a measure of core inflation from a multivariate time series consisting of the monthly inflation rates for 8 expenditure categories.

The series are listed in table 1 and refer to the U.S. city average for the sample period 1993.1–2003.8 (source: Bureau of Labor Statistics). The relative importance of the components of the U.S, inflation rate in building up the U.S. inflation rate, i.e. their CPI weights, is reported in the second column of table 3. Figure 2 displays the eight series $y_{it}, i = 1, 2, \ldots, 8$.

Fitting the univariate local level model to $\mathbf{w}'\mathbf{y}_t$, where $\mathbf{w}$ is the vector containing the CPI weights reproduced in the second column of table 3, that is $\mathbf{w}'\mathbf{y}_t = \mu_t + \epsilon_t, \epsilon_t \sim \mathrm{WN}(0, \sigma_\epsilon^2)$, $\mu_t = \mu_{t-1} + \eta_t, \eta_t \sim \mathrm{WN}(0, q\sigma_\epsilon^2)$, gives the maximum likelihood estimate $\tilde{q} = 0$.

The estimated signal to noise ratio implies that CPI monthly inflation is stationary and that the corresponding aggregate core inflation measure is represented by the time average of the CPI all items monthly inflation rates, that is $\tilde{\mu} = T^{-1}\sum_t \mathbf{w}'\mathbf{y}_t$.

Table 1. Description of the series and their CPI weights

Expenditure group	*CPI weights*
1. Food and beverages	0.162
2. Housing	0.400
3. Apparel	0.045
4. Transportation	0.176
5. Medical care	0.058
6. Recreation	0.059
7. Education and communication	0.053
8. Other goods and services	0.048

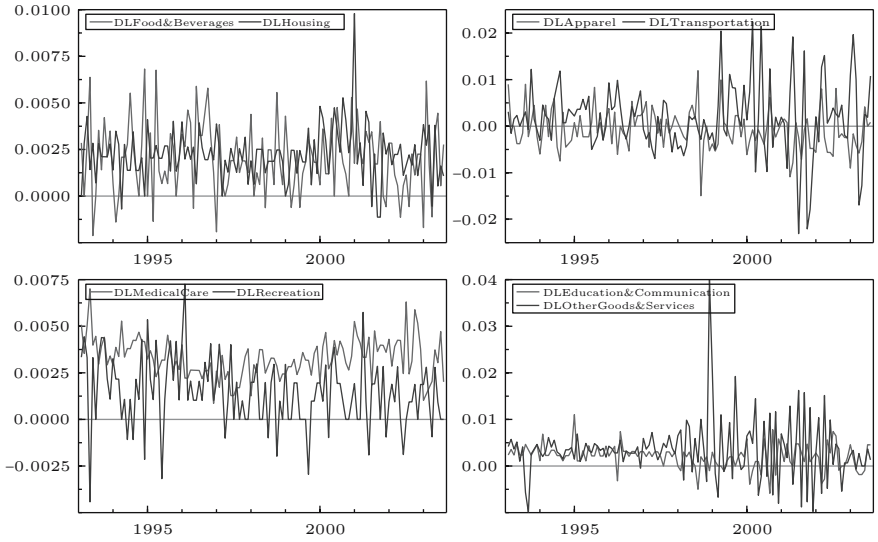

Fig. 2. U.S. CPI, 1993.1–2003.8. Monthly relative price changes for the eight expenditure categories

Table 2. Nyblom and Harvey (2003) stationarity test, cointegration test and common trend test

Truncation lag (l)	NH	NH-coint	CT(1)
0	3.334	2.510	1.112
1	2.876	2.306	0.982
2	2.638	2.134	0.894
5	2.214	1.787	0.720
10	1.682	1.346	0.605
5% crit. value	2.116	1.903	0.637

Table 3. Core inflation measures: weights defining $\bar{\mu}_t = w'\mu_t$

Expenditure group	CPI weights	MV weights	DECM weights
Food and beverages	0.162	0.119	0.114
Housing	0.400	0.174	0.215
Apparel	0.045	0.025	0.023
Transportation	0.176	−0.005	0.007
Medical care	0.058	0.435	0.435
Recreation	0.059	0.146	0.126
Education and communication	0.053	0.079	0.068
Other goods and services	0.048	0.026	0.010

8.1 Stationarity and Common Trends

The issue concerning the stationarity of CPI monthly inflation can be also be handled within a multivariate framework. Assuming that $\mathbf{y}_t = (y_{1t}, \ldots, y_{8t})'$ is modelled as in (1), we can use the NH statistic ξ_N to test $\mathrm{H}_0 : \boldsymbol{\Sigma}_\eta = \mathbf{0}$ versus the alternative that $\boldsymbol{\Sigma}_\eta = q\boldsymbol{\Sigma}_\epsilon$.

The values of the NH statistic are reported in the second column of table 2 for various values of the truncation lag l used in computing the Newey-West nonparametric correction for autocorrelation and heteroscedasticity. They lead to reject the null that $\mathbf{y}_t$ is stationarity for values of l up to 5.

The third column reports the values of the NH-cointegration test. The latter tests the null hypothesis that $\mathbf{A}'\mathbf{y}_t$ is stationary, where $\mathbf{A}$ is chosen such that $\mathbf{A}'\mathbf{i} = \mathbf{0}$, which corresponds to the hypothesis that there is a single common trend which enters each of the series with the same loading; this is also known as the balanced growth hypothesis: as the series share the same common trend, the difference between any pair is stationary. This hypothesis is clearly rejected for low values of the truncation parameter, up to $l = 2$.

Finally, CT(1) is the statistic for testing the null hypothesis that a single common trend is present (the 5% critical value has been obtained by simulation), based on the statistic $\zeta(1, 8) = \sum_{i=2}^{8} \lambda_i(\mathbf{S}^{-1}\boldsymbol{C})$, where $\lambda_i(\cdot)$ is the i-th ordered eigenvalue of the matrix in argument.

Taken together, the results of the NH-coint and CT(1) tests do not suggest the presence of a single common trend driving the eight CPI monthly inflation rates. Nevertheless, if the MLLM is estimated with a single common trend (see Sect. 5.2) then the estimated vector of loadings is

$$\tilde{\mathbf{z}}' = [1.000, 0.038, 0.069, 0.081, 0.016, 0.086, 0.023, -0.021],$$

and the estimated trend disturbance variance is $\tilde{\sigma}^2_\eta = 0.00065282$. Estimation of the MLLM with common trends was carried out in Stamp 6 by Koopman et al. (2000). The other computations and inferences were performed in Ox 3 by Doornik (1999). It can be seen that the series *Food and beverages* plays a dominant role in the definition of the common trend.

8.2 Homogeneous MLLM

Maximum likelihood estimation of the local level model (1) with the homogeneity restriction is particularly straightforward, since the innovations and inferences about the states can be obtained by running N univariate Kalman filters (this is known as *decoupling*). The matrix $\mathbf{\Sigma}_\epsilon$ can be concentrated out the likelihood function and the concentrated likelihood can be maximised with respect to the signal-noise ratio q.

The estimation results are the following: $\hat{q} = 0.0046$, and

$$\tilde{\mathbf{\Sigma}}_\epsilon = \begin{bmatrix} 0.04 & -0.10 & 0.04 & 0.07 & 0.02 & -0.04 & -0.03 & -0.01 \\ -0.00 & 0.02 & 0.04 & 0.23 & 0.05 & 0.13 & 0.07 & -0.10 \\ 0.00 & 0.00 & 0.17 & 0.07 & -0.01 & -0.00 & -0.08 & -0.08 \\ 0.01 & 0.02 & 0.02 & 0.55 & 0.04 & 0.10 & -0.22 & 0.02 \\ 0.00 & 0.00 & -0.00 & 0.00 & 0.01 & -0.14 & -0.02 & -0.07 \\ -0.00 & 0.00 & -0.00 & 0.01 & -0.00 & 0.03 & -0.06 & -0.04 \\ -0.00 & 0.00 & -0.01 & -0.04 & -0.00 & -0.00 & 0.06 & -0.14 \\ -0.00 & -0.01 & -0.02 & 0.01 & -0.00 & -0.00 & -0.02 & 0.39 \end{bmatrix}$$

where in the upper triangle we report the correlations, which are usually very low.

The frequency domain test for homogeneity (Fernandez and Harvey, 1990) takes the value 31.275 on 35 degrees of freedom and therefore it is not significant (the p-value is 0.65). This suggests that the homogenous specification is a good starting point for building up core inflation measures.

8.3 Homogeneous Dynamic Error Components Model

Within the homogeneous model of the previous subsection we considered the error component structure $\mathbf{\Sigma}_\epsilon = \sigma_\epsilon^2 \mathbf{i}\mathbf{i}' + \mathbf{N}_\epsilon$, in which there is a common disturbance linking the trends and the irregular component; $\mathbf{N}_\epsilon$ was specified as a diagonal matrix.

When estimated by maximum likelihood, the signal–noise ratio is close to that of the homogenous case, $\hat{q} = 0.0043$; moreover, the common irregular disturbance variance is estimated $\hat{\sigma}_{\bar{\epsilon}}^2 = \times 10^{-7}$ and

$$b\hat{N}_\epsilon = \mathrm{diag}(0.035, 0.019, 0.173, 0.546, 0.009, 0.032, 0.059, 0.391).$$

However, the DECM restriction, $H_0 : \mathbf{\Sigma}_\epsilon = \sigma_{\bar{\epsilon}}^2 \mathbf{i}\mathbf{i}' + \mathbf{N}_\epsilon, \mathbf{\Sigma}_\eta = q\mathbf{\Sigma}_\epsilon$, is strongly rejected, with the LM test taking the value 153.43 on 60 degrees of freedom.

8.4 Core Inflation Measures

Bearing in mind the empirical results of the previous sections, we now discuss three measures of core inflation obtained from the multivariate MLLM.

The first is derived from the homogeneous local level model and is defined as $\mathbf{w}'\tilde{\boldsymbol{\mu}}_{t|T}$, where $\tilde{\boldsymbol{\mu}}_{t|T}$ are the smoothed estimates of the trends and w is the vector of CPI weights, equal to the budget share of the expenditure groups. This is reproduced along with the 95% confidence interval in the first panel of Fig. 3.

The second measure uses the minimum variance (MV) weights $w = \hat{\boldsymbol{\Sigma}}_{\epsilon}^{-1}\mathbf{i}/(\mathbf{i}'\hat{\boldsymbol{\Sigma}}_{\epsilon}^{-1}\mathbf{i})$, reproduced in the third column of table 3. *Housing* and *Transportation* result heavily downweighted (the MV weight is negative for the latter). The corresponding core inflation measure, displayed in the right upper panel of Fig. 3, is much smoother than the previous, and characterised by lower estimation error variance.

The last measure of core inflation is derived from the dynamic error component local level model with homogeneity and is defined as $\mathbf{w}'\tilde{\boldsymbol{\mu}}_{t|T}$, where $\mathbf{w} = \mathbf{N}_{\eta}^{-1}\mathbf{i}/(\mathbf{i}'\mathbf{N}_{\eta}^{-1}\mathbf{i})$ where $\mathbf{N}_{\eta}$ is a diagonal matrix. Although the DECM restriction was strongly rejected, the weights and the corresponding core inflation measure agree very closely with the minimum variance one.

The overall conclusion is that the point estimates of the three core inflation measures agree very closely.

For comparison purposes, in the last panel of Fig. 3 we display the core inflation measure estimated using the structural VAR approach by Quah and

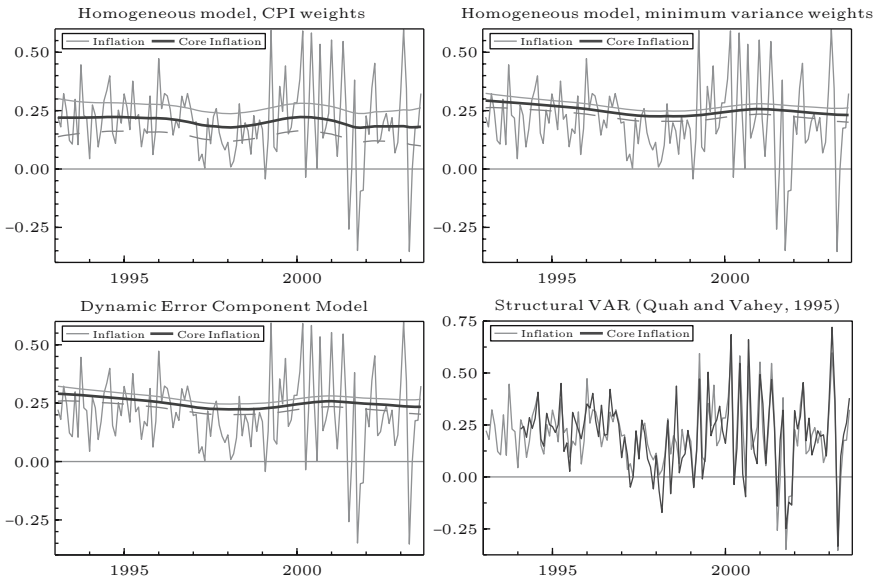

Fig. 3. U.S. CPI, 1993.1–2000.12. Core inflation measures derived from a multivariate local level model with homogeneity and variance components restrictions

Vahey (1995, QV henceforth). A bivariate vector autoregressive (VAR) model was estimated for the series $\mathbf{u}_t = [\Delta y_t, \Delta x_t]'$,

$$\boldsymbol{\Phi}(L)\mathbf{u}_t = \boldsymbol{\beta} + \boldsymbol{\xi}_t, \quad \boldsymbol{\Phi}(L) = \mathbf{I} - \boldsymbol{\Phi}_1 L - \cdots - \boldsymbol{\Phi}_p L^p,$$

where y_t is the monthly inflation rate, computed using the CPI total, and x_t is the logarithm of the industrial production index (source: Federal Reserve Board, sample period: 1993.1–2003.8). The VAR lag length which minimises the Akaike information criterion resulted $p = 11$, which is close to the value adopted by QV in their original paper. QV define core inflation as the component of inflation that can be attributed to nominal disturbances that have no long run impact on output. Their identification proceeds as follows: the structural disturbances, $\boldsymbol{\zeta}_t = [\zeta_{1t}, \zeta_{2t}]'$, are defined as linear transformations of the time series innovations, $\xi_t = \mathbf{B}\boldsymbol{\zeta}_t$, where $\mathbf{B} = \{b_{ij}; i, j = 1, 2\}$ is a full rank matrix such that $\boldsymbol{\Phi}(1)^{-1}\mathbf{B}$ is upper triangular (i.e. the nominal disturbance ζ_{1t} has no permanent effect on x_t). Correspondingly, the core inflation measure is:

$$m_t = [\varphi_{11}(L)b_{11} + \varphi_{12}(L)b_{21}]\zeta_{1t},$$

where $\boldsymbol{\Phi}(L)^{-1} = \{\varphi_{ij}(L); i, j = 1, 2\}$.

Several differences arise with the measures extracted from the MLLM. The QV measure tracks actual inflation very closely. The plot clearly show that m_t is indeed very volatile. This lack of smoothness can be partly attributed to the fact that this measure is based on a one sided filter. On the other hand, it may be argued that the industrial production index is not fully adequate to capture core inflation.

9 Conclusions

The paper has illustrated how core inflation measures can be derived from optimal signal extraction principles based on the multivariate local level model. The approach is purely statistical, in that a coherent statistical representation of the dynamic features of the series the model is sought, along with sensible ways of synthesizing the dynamics of a multivariate time series in a single indicator of underlying inflation. The advantage over indices excluding particular items, such as food and energy, is that maximum likelihood estimation of the parameters of the model indicate what items have to be downweighted in the estimation of core inflation.

Two main directions for future research can be envisaged: the first is enlarging the cross-sectional dimension by using more disaggregate price data. The second is to provide more economic content to the measurement by including in the model a Phillips' type relationship featuring among the inflation determinants measures of monetary growth, the output gap and inflation expectations.

References

Anderson, T. W. *An Introduction to Multivariate Statistical Analysis.* Second Edition, John Wiley & Sons, 1984.

Bell, W. Signal Extraction for Nonstationary Time Series. *The Annals of Statistics*, 12, 646–664, 1984.

Bryan M. F. and S. G. Cecchetti. Measuring Core Inflation. In N. G. Mankiw, editor, *Monetary Policy.* Chicago: University of Chicago Press, 1994.

Bryan M. F., S. G. Cecchetti and R. L. Wiggins II. Efficient Inflation Estimation. *NBER Working Paper* n. 6183. , Cambridge, MA., 1997.

Doornik, J. A. *Ox: An Object-Oriented Matrix Programming Language.* London: Timberlake Consultants Ltd., 1999.

Durbin, J. and S. J. Koopman. *Time Series Analysis by State Space Methods.* Oxford University Press, Oxford, 2001.

Enns, P. G., J. A. Machak, W. A. Spivey and W. J. Wrobleski. Forecasting applications of an adaptive multiple exponential smoothing model. *Management Science*, 28, 1035–1044, 1982.

Fernandez F. J. and A. C. Harvey. Seemingly unrelated time series equations and a test for homogeneity. *Journal of Business and Economic Statistics*, 8, 71–81, 1990.

Harvey A. C. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, Cambridge, 1989.

Harvey A. C. and S. J. Koopman. Signal extraction and the formulation of unobserved components models. *Econometrics Journal*, 3, 84–107, 2000.

Koopman S. J. and A. C. Harvey. Computing observation weights for signal extraction and filtering. *Journal of Economic Dynamics and Control*, 27, 1317–33, 2003.

Koopman S. J., A. C. Harvey, J. A. Doornik and N. Shephard. *STAMP 6: Structural Time Series Analysis Modeller and Predictor.* London: Timberlake Consultants Ltd., 2000.

Marshall P. Estimating Time-Dependent Means in Dynamic Models for Cross-sections of Time Series. *Empirical Economics*, 17, 25–33, 1992.

Nyblom J. and A. C. Harvey. Tests of Common Stochastic Trends. *Econometric Theory*, 16, 76–99, 2003.

Quah D. and S. Vahey. Measuring Core Inflation. *Economic Journal*, 105, 1130–1144, 1995.

Selvanathan, E. A. and D. S. Prasada Rao. *Index Numbers. A Stochastic Approach.* Macmillan, London, 1994.

Stock J. H. and M. W. Watson. *A Probability Model of Coincident Economic Indicators.* In K. Lahiri and G. H. Moore, editors, *Leading Economic Indicators: New Approaches and Forecasting Records.* Cambridge University Press, New York, pp. 63–85, 1991.

Whittle P. *Prediction and Regulation by Least Squares Methods.* 2nd ed. Oxford: Blackwell, 1983.

Wynne M. A. Core Inflation: A Review of some Conceptual Issues. *ECB Working Paper* No. 5.5. European Central Bank, Frankfurt am Main, 1999.

Part III

Financial Modelling

Part III

Financial Modelling

Random Portfolios for Performance Measurement

Patrick Burns

Burns Statistics

Summary. Random portfolios—portfolios that obey constraints but ignore utility—are shown to measure investment skill effectively. Problems are highlighted regarding performance measurement using information ratios relative to a benchmark. Random portfolios can also form the basis of investment mandates—this allows active fund managers more freedom to implement their ideas, and provides the investor more flexibility to gain utility. The computation of random portfolios is briefly discussed.

Key words: Random portfolios, performance measurement, investment mandates, investment opportunity, Monte Carlo simulation

1 Introduction

The accurate assessment of the skill of fund managers is quite obviously of great value. It is also well known to be a very difficult task. A variety of techniques, some quite clever, have been devised. Some methods measure individual managers, others a class of managers. A few references are (Kosowski et al., 2001), (Muralidhar, 2001), (Engstrom, 2004), (Ding and Wermers, 2004). There are also (Ferson and Khang, 2002) and (Grinblatt and Titman, 1993).

More accurate performance measurement allows a quicker determination of whether or not a fund manager has skill. It can also provide a more fair method of compensating fund managers for their contribution to the investor.

A perfect measure would be to look at all portfolios that the fund manager might have held, and compare their realized utility to the realized utility of the fund under question. The portfolios that the manager might have held are those that satisfy the constraints that the manager uses.

For practical reasons we take a random sample from the set of portfolios satisfying the constraints to use in the comparisons. We are free to use whatever measure (or measures) of quality that we like, and we will have a statistical statement of the significance of the quality of the fund. This

procedure eliminates much of the noise that results from assessing a fund's outperformance relative to an index (or to peers).

Random portfolios have other uses as well, such as evaluating trading strategies as discussed in (Burns, 2006).

R (R Development Core Team, 2005) was used for computations and graphs for this paper. Random portfolios and optimisations were done with the POP Portfolio Construction Suite (Burns Statistics, 2005).

2 Generating Random Portfolios

It is worth noting that naive approaches to generating random portfolios typically do not work well. For example, permuting the weights of an actual portfolio seldom yields a portfolio with realistic properties. In particular the volatility of such portfolios is generally quite large. Real portfolios are not a haphazard collection of assets.

Generating portfolios of some number of equally weighted assets is an easy approach. While this can produce a distribution that is much better than nothing, it is not what a fund manager will do. This technique will produce portfolios that a fund manager would not hold because they are too volatile. It also fails to allow portfolios that fund managers would hold—equal weighting is a significant limitation.

A set of constraints is required in order to produce random portfolios that are believable. One approach to generating random portfolios when there are constraints is to use the rejection method—produce a series of random portfolios and reject all of those that violate at least one constraint. This is not an effective method in practice. The probability of a portfolio being accepted is generally extremely small when realistic constraints are in place.

(Dawson and Young, 2003) outline a mathematical algorithm for generating random portfolios when there are only linear constraints. Perhaps the most important constraint when generating random portfolios is a limit on the volatility, which is not linear. Integer constraints such as limits on the number of assets in the portfolio and limits on the number of assets to trade are also often desired in practice. Thus an assumption of only linear constraints still only produces an approximation to the real situation.

Genetic algorithms provide a practical means of generating random portfolios. Genetic algorithms are generally attributed to (Holland, 1975), but others contributed as well—see (Fogel, 2006). The use of genetic and related algorithms in finance is discussed in (Maringer, 2005). The original formulation of a genetic algorithm was quite inefficient, but advances have been made. The basic idea is that at any one time there is a population of candidate solutions. Two of these solutions are selected as parents that produce child solutions—the children combine, in some random way, features of the parents. Better solutions have the best chance of surviving in the population.

In the case of random portfolios the genetic algorithm merely needs a number that describes how much a solution violates the constraints. Once we find a solution that violates none of the constraints, we are done (with generating one random portfolio). Since other solutions in the population will be correlated with our selected portfolio, we need to start over with a completely new population to generate an additional random portfolio.

In practice this approach with genetic algorithms (and many other types of algorithm) will produce random portfolios that usually have at least one constraint that is close to binding. That is, the portfolios are not uniformly spread over the space in which the constraints are met—the portfolios are concentrated on the boundary.

There are two attitudes to this. One is that the portfolios should be uniform. The other is that actual portfolios will also be close to binding on some of the constraints—thus being concentrated on the boundary may actually be closer to what is really wanted. After all, if none of the constraints were likely to be binding when building a portfolio, then there would be no point in having the constraints.

The (Dawson and Young, 2003) algorithm includes a method of thinning solutions near edges. Their technique could possibly be adapted to use with genetic and other algorithms to produce more uniform sets of random portfolios. The first step is to decide if a portfolio is too close to one or more constraints. If it is, then the portfolio could be dropped. While this is not going to be an especially efficient algorithm, it may well be practical since it is generally quite quick to generate a portfolio.

Alternatively, a portfolio near the boundary could be moved. Their approach is to select another portfolio that has already been generated and move towards that portfolio. This works because the second portfolio is unlikely to be tight on the same constraints as the first portfolio, and—in their framework—the feasible region is convex. In most practical applications, where there is a limit on the number of assets in the portfolios, the feasible region is not convex. The technique needs modification for these cases.

In the examples used here, no adjustments have been made for concentrations on the boundary.

3 Management Against a Benchmark

Currently a great amount of performance analysis is relative to a benchmark. Sometimes this is done because it is deemed reasonable, but other times for lack of an alternative. A good discussion of the use and abuse of benchmarks is (Siegel, 2003). (Kothari and Warner, 2001) use random portfolios to examine problems with benchmarks—they get results quite similar to those in this section.

In this section (and the next) we use a dataset of the daily returns of an unsystematic collection of 191 large-cap and small-cap US equities. The data

start at the beginning of 1996 and end after the third quarter of 2004. Results are reported for each quarter except the first two. The first two quarters are excluded so that all results are out of sample—in some operations the variance matrix is estimated with the previous two quarters of data.

One thousand random portfolios were created from this universe with the constraints that no more than 100 names were in a portfolio, no short values were allowed, the maximum weight of any asset was 10%, and the sum of the 8 largest weights was no more than 40%. In some figures the first 500 random portfolios are compared to the second 500 in order to indicate the significance of any pattern that might appear.

Three artificial benchmarks were created. The first is the equal weighting of the assets. The other two have weights that were randomly generated. These latter two are referred to as the "unequally weighted benchmarks". Note that the randomness is only in the selection of the weights of the assets, and these weights are held fixed throughout time.

3.1 Outperforming the Benchmark

The information ratios of the random portfolios were calculated relative to the benchmark that has equal weight in each stock. (An information ratio is the annualized return in excess of the benchmark divided by the annualized standard deviation of the differences in returns—an excess return derived from a regression rather than subtraction is more desirable (see (Siegel, 2003)) but for simplicity is not done here.) Figure 1 shows the probability that the random portfolios have a positive information ratio against this benchmark for each

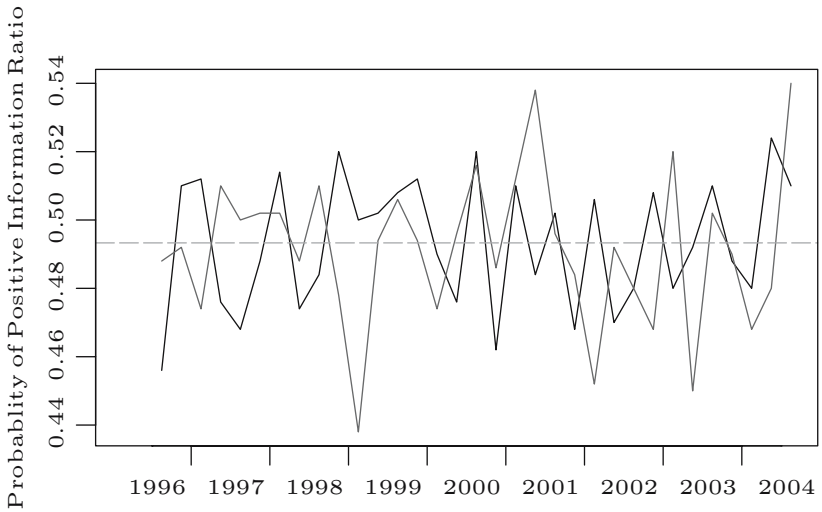

Fig. 1. The empirical frequency of a positive information ratio by quarter relative to the equally weighted benchmark. Each line represents 500 random portfolios

quarter. One line corresponds to the first 500 random portfolios and the other line to the second 500.

We might have expected the fraction of portfolios that outperform the equally weighted benchmark to be closer to 50%. (The average probability is indicated by the horizontal line.) The p-value is 0.006 for the test that positive and negative information ratios are equally likely. Note though that the benchmark is outside the constraints that we have put on the random portfolios—the portfolios can have at most 100 constituents while the benchmark has 191.

While there is a slight tendency for the equally weighted benchmark to outperform, there seems to be no systematic difference between quarters.

We now look at two benchmarks with (randomly generated) unequal weights. The mean weight is about 0.5%, and the maximum weight in each benchmark is slightly over 2.5%. Figures 2 and 3 show the probability of a positive information ratio. In these plots there are undeniable differences between quarters. In some quarters there is a strong tendency for the benchmark to outperform the random portfolios, in others a strong tendency for the benchmark to underperform. In most quarters the two sets of random portfolios have almost identical fractions of outperformance. There is no consistency of outperformance between the two benchmarks.

On reflection this result should not be so surprising—though the extent of the effect may be. A benchmark will be hard to beat during periods when the most heavily weighted assets in the benchmark happen to do well. Likewise, when the assets with large weights in the benchmark do relatively poorly, then the benchmark will be easy to beat.

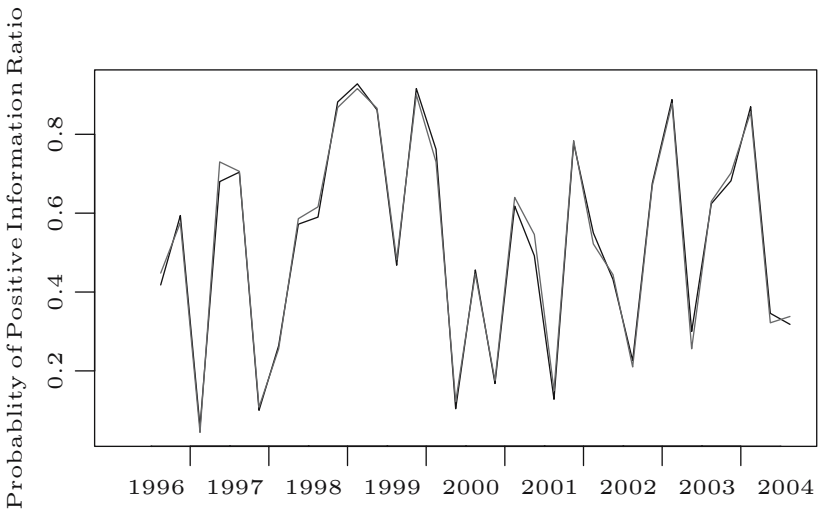

Fig. 2. The empirical frequency of a positive information ratio by quarter relative to the first unequally weighted benchmark. Each line represents 500 random portfolios

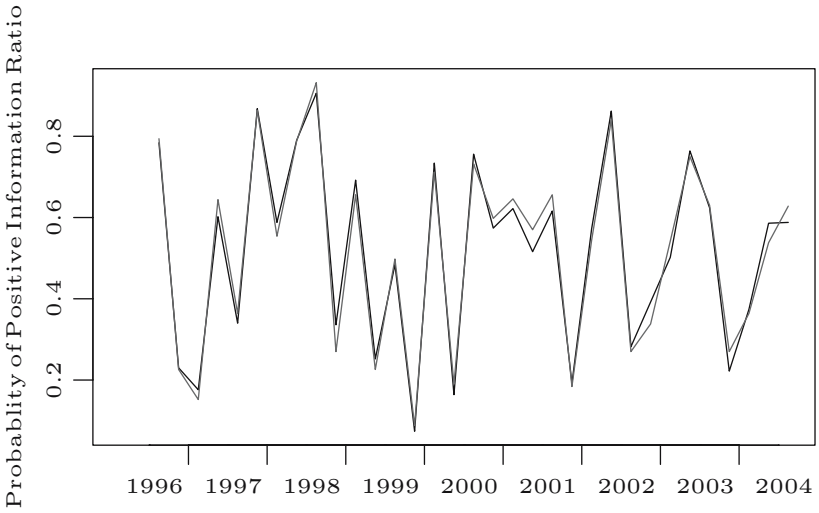

Fig. 3. The empirical frequency of a positive information ratio by quarter relative to the second unequally weighted benchmark. Each line represents 500 random portfolios

Figure 4 shows the quarterly returns of each of the three benchmarks plotted against each other. The three benchmarks are obviously highly correlated. This seems contradictory since the probabilities of outperforming the benchmarks didn't appear to be related. The explanation is illustrated by Fig. 5. This shows the returns of the three benchmarks and the probability of outperformance for each quarter. Even slight differences in return between the benchmarks cause dramatic differences in the probability of outperformance. That is, random portfolios provide a very sensitive measure of performance.

Clearly the more unequal the weights in a benchmark, the more extreme the swings will be in the probability of outperforming. In this regard, the random benchmarks that are used here are not at all extreme compared to many indices that are used in practice as benchmarks.

Table 1 shows a history of U.S. mutual fund outperformance relative to the "best fitting" benchmark of each fund. The data in this table were computed by Craig Israelsen using the Morningstar database. (Israelsen, 2003) alludes to the method of choosing the benchmark for each fund.

There are two histories for the S&P 500—one with all of the available funds, and one containing only the funds that were live in all of the years. This was to explore the possibility of survival bias. Survival bias appears to be minimal.

The pattern of outperformance of the S&P 500 by the funds is quite similar to that for the unequally weighted benchmarks as exhibited in Figs. 2

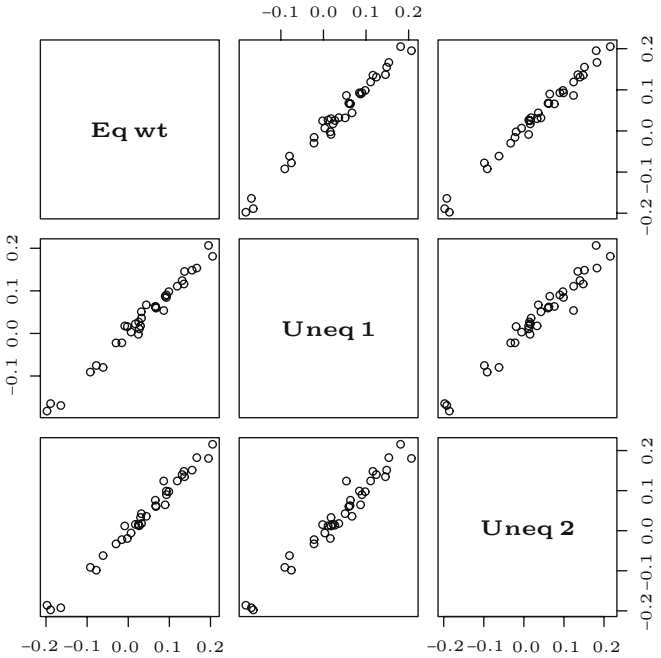

Fig. 4. Scatterplots of quarterly returns of the three hypothetical benchmarks

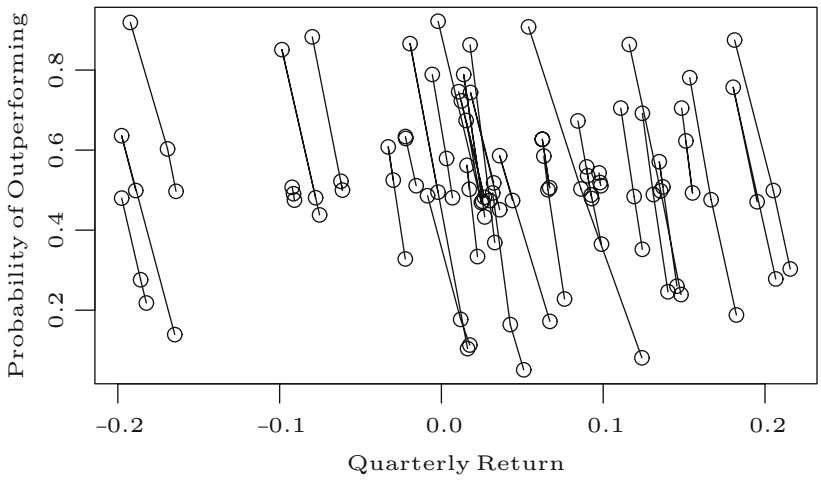

Fig. 5. The empirical frequency of outperformance by quarterly return for the three hypothetical benchmarks

Table 1. US mutual funds outperforming benchmarks. Source: Craig Israelsen

	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
S&P 500 return	1.31	37.53	22.94	33.35	28.57	21.04	−9.10	−11.88	−22.09	28.67
# of funds (all)	100	115	126	136	155	173	183	204	212	212
% funds outperform	30	9.6	24.6	14.0	24.5	27.2	72.7	60.8	56.6	13.2
95% conf. int.	(21,40)	(5,16)	(17,33)	(9,21)	(18,32)	(21,34)	(66,79)	(54,68)	(50,63)	(6,20)
# of funds (full history)	100	100	100	100	100	100	100	100	100	100
% funds outperform	30	9	22	13	16	23	74	63	59	12
95% conf. int.	(21,40)	(4,16)	(14,31)	(7,21)	(9,25)	(15,32)	(64,82)	(53,72)	(49,69)	(6,20)
S&P Midcap 400 ret	−3.59	30.92	19.18	32.24	19.11	14.72	17.72	−0.60	−14.53	35.59
# of funds	40	48	52	63	77	95	101	113	121	121
% funds outperform	55.0	41.7	50.0	23.8	27.3	53.7	33.7	39.8	38.0	28.1
95% conf. int.	(38,71)	(28,57)	(36,64)	(14,36)	(18,39)	(43,64)	(25,44)	(31,49)	(29,47)	(20,37)
Russell 2000 return	−1.82	28.44	16.49	22.36	−2.55	21.26	−3.02	2.49	−20.48	47.25
# of funds	26	34	40	50	69	84	97	111	115	115
% funds outperform	61.5	58.8	75.0	56.0	60.9	63.1	71.1	56.8	67.8	41.7
95% conf. int.	(41,80)	(41,75)	(59,87)	(41,70)	(48,72)	(52,73)	(61,80)	(47,66)	(58,76)	(33,51)

and 3—some years a large fraction of funds underperform and other years a large fraction outperform.

Interpreting this data in the way that it is often used, we infer that managers were, in general, bad during the 90's, then they suddenly became very good for three years starting in 2000, then returned to being bad in 2003. This is clearly a ridiculous inference, but nonetheless is often done.

The outperformance of funds relative to the other two benchmarks, while not completely stable, is much less variable. The S&P Midcap 400 almost always beats more than half of the funds that track it, while the Russell 2000 is almost always beat by more than half the funds that track it. There are several possibilities:

- The fund managers that track the S&P Midcap are inept, and the fund managers that track the Russell 2000 are quite skillful.
- The S&P Midcap has been hard to beat and the Russell 2000 has been easy to beat.
- The volatilities of the funds are substantially different from the benchmark volatility.
- The outperformance is an artifact of the way that benchmarks have been assigned to funds.

We don't have enough information to decide among these. Random portfolios could help inform us.

Some would argue—given the evidence we've just seen—that benchmarks should be equally weighted indices. Even if this were accepted as practical (see (Siegel, 2003) for some reasons why it isn't), it still doesn't solve the issue of accurately measuring skill. Figure 6 shows the probability of the random portfolios having an information ratio greater than two relative to the equally weighted benchmark. There are definite systematic differences by quarter—sometimes a large information ratio is easier to achieve than at other times.

3.2 Information Ratios and Opportunity

Figure 6 implies that the distribution of information ratios changes from quarter to quarter. Information ratios are not purely a measure of skill, but rather are a combination of skill and opportunity. (Statman and Scheid, 2005) focuses on this topic.

Imagine a case where all of the assets in the universe happen to have the same return over a time period. Portfolios will vary from each other during the period and hence have non-zero tracking error relative to the index. However, all portfolios will end the period with the same return—all information ratios will be zero.

Figure 7 shows the standard deviation of the information ratios of the random portfolios for each quarter and each of the three benchmarks. The naive assumption is that the standard deviations should all be 2. (A one-year

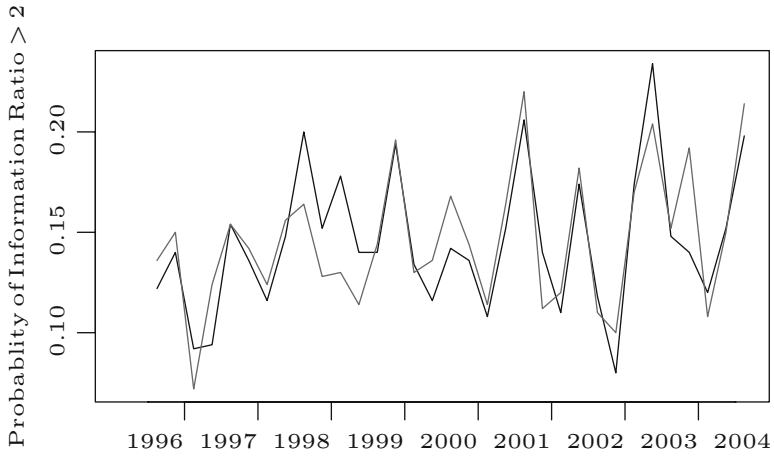

Fig. 6. The empirical frequency of an information ratio greater than two relative to the equally weighted benchmark. Each line represents 500 random portfolios

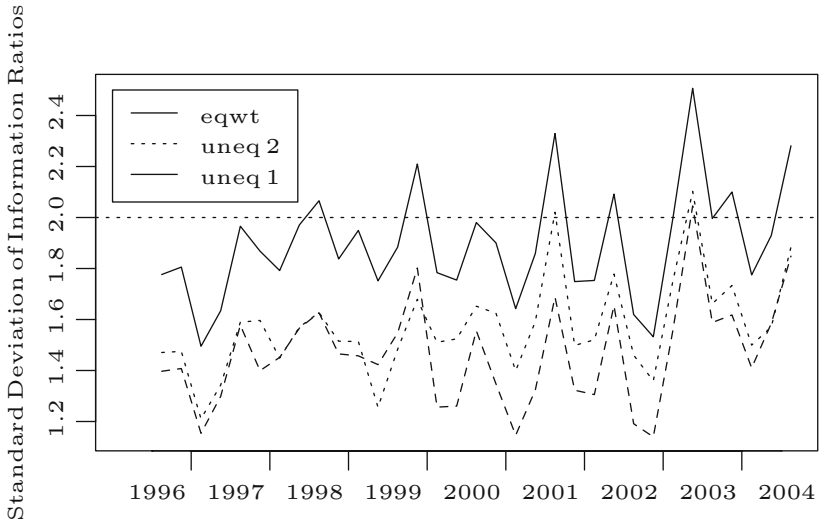

Fig. 7. Standard deviations of the information ratios of random portfolios by quarter

information ratio "should" have standard deviation 1, in which case annualized quarterly ratios will have a variance of 4 since the annual ratio is the mean of the four quarterly ratios.) The plot exhibits definite differences between quarters and between benchmarks.

Certainly the cross-sectional spread of the full-period returns has an effect on the standard deviation of information ratios. The volatility over time of the assets will also have an effect. Figure 8 shows an experiment of varying these. The data are from the first quarter of 2004. Each point in the figure has

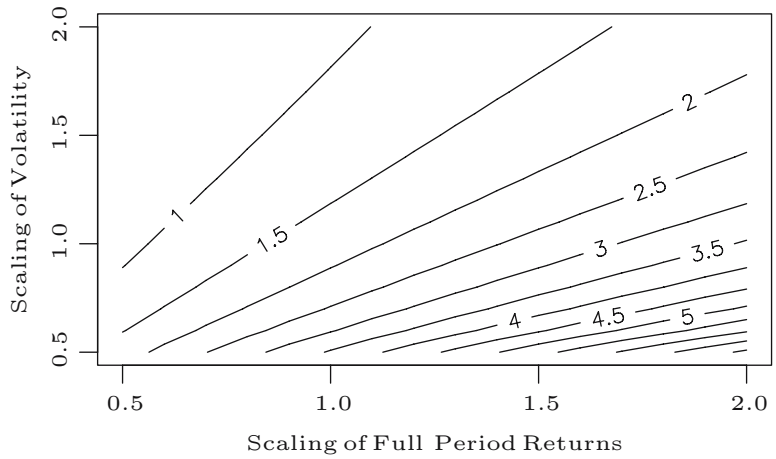

Fig. 8. The standard deviation of information ratios as volatility and returns are artificially varied (using data from the first quarter of 2004)

had the volatility of each asset multiplied by a value and the returns for the period multiplied by a value. (The return of an asset is adjusted by adding the same value to each of the daily returns for the asset. The volatility is adjusted by scaling the deviation of the returns around the mean return for the asset.)

The point at (1, 1) corresponds to the real data—there the standard deviation of the information ratios (relative to the equally weighted benchmark) is about 1.8. The points that are at 2 on the horizontal axis have twice the spread of returns as the real data (a stock that really had a 3% return gets a 6% return, and a stock with a -1% return gets a -2% return). The points that are at 0.5 on the vertical axis have half of the volatility as the real data for all of the assets. The point at (2, 0.5) has a standard deviation of information ratios that is about 7.

Figure 8 shows that the cross-sectional spread of asset returns is very important to the opportunity to achieve a large information ratio. The spread of returns has a bigger impact as the volatility of the assets decreases. Obviously in reality there is a connection between the volatility of the individual assets and the cross-sectional spread of returns, but there is no reason to suppose that they are in lock step.

3.3 Measuring Skill via Information Ratios

In order to study the ability to measure skill, a set of 100 "managers" was created. At the beginning of each quarter each manager performs a portfolio optimisation. The managers all use the same variance matrix, but each has a unique vector of expected returns. The variance matrix is estimated from the previous two quarters using a statistical factor model. The expected returns

in the optimisation are based on the actual returns that are realized in the quarter (since this is looking at future data, it is not a strategy that real fund managers have available to them). The expected returns for the stocks are random normals with mean equal to 0.1 times the realized mean daily return for the asset. The standard deviation for the random normals is 0.1 times the standard deviation of the realized daily returns for the asset. The objective of the optimisation was to maximize the information ratio—the absolute ratio, not relative to any benchmark.

A common approach to testing for skill is to compute the information ratio of the fund relative to its benchmark. The test is then to see if this information ratio is too large given the null hypothesis that the true value is zero. There are at least two approaches to the test. One is to feed the information ratios for the individual periods—33 quarters in the current case—to a t-test. More common is to calculate the information ratio for the whole period and use the fact that the standard deviation is theoretically known, then use the normal distribution. The statistics and p-values from these two approaches should be similar. Figure 9 shows the p-values from the normal test for the 100 "managers" for the information ratio based on the first unequally weighted benchmark for the full time period. The skill of the managers shows up by quite a large number having p-values close to zero.

Another view is in Fig. 10 which shows the number of hypothetical managers with significant p-values as each quarter is observed—an additional point on the x-axis becomes available as each quarter is completed. There are a couple of aspects to this plot that are worrisome. The number of significant managers is much more variable when only a few quarters have been

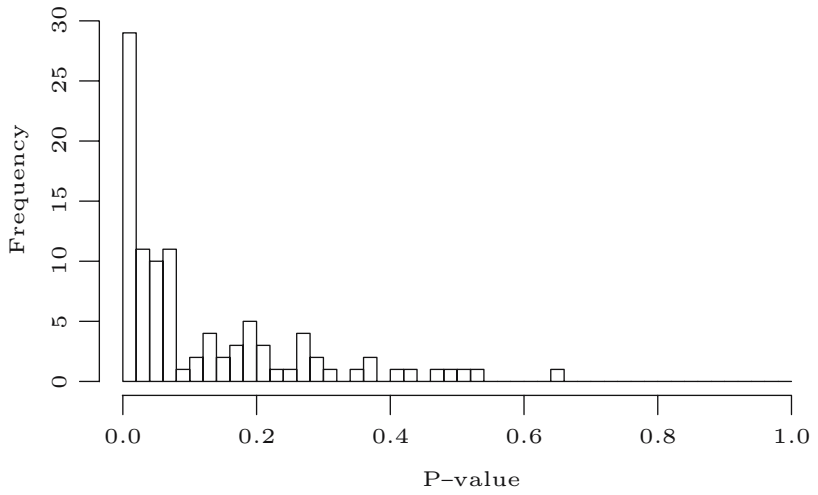

Fig. 9. P-values of the 100 hypothetical managers based on the information ratio relative to unequally weighted benchmark 1 over 33 quarters using the normal test

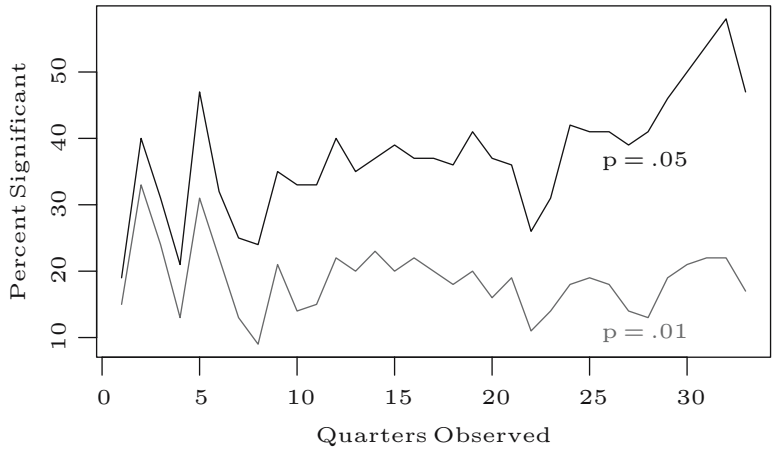

Fig. 10. Percent of hypothetical managers with significant p-values from the normal test over time for random benchmark 1

observed. While the number of managers that are significant at the 5% level grows reasonably steadily as we would expect, the number that are significant at 1% seems to stagnate.

We have seen that the assumption of known standard deviation in the normal test is actually violated. Figure 11 shows the distribution of normal test p-values using information ratios relative to the equally weighted benchmark when there is no skill. Each of the "no skill managers" selects one of the 1000 random portfolios at random each quarter—100,000 such managers were created. If theory were correct, then the distribution in the plot would be

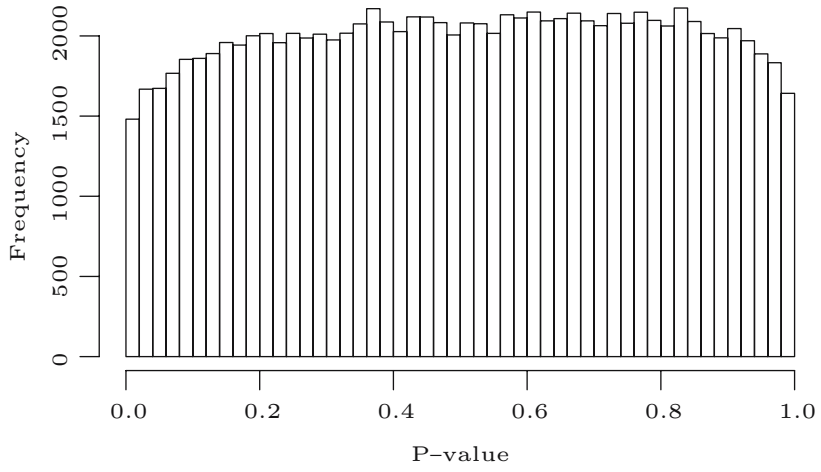

Fig. 11. Distribution of p-values from the normal test of information ratios relative to the equally weighted benchmark on portfolios with zero skill

uniform (that is, flat). The distribution does not have enough mass in the tails, near 0 and 1. This implies that it is harder (in this case) than it should be to prove fund managers either skilled or unskilled. The deviation from the uniform distribution will be time, benchmark and universe dependent.

4 Measuring Skill with Random Portfolios

We've already seen that assessing the skill of fund managers with information ratios has severe problems.

A second commonly used method is to rank a fund relative to similar funds. This has problems of its own. It supposes that all funds within the category are doing the same thing. For instance, it isn't entirely obvious how differences in volatility should be taken into account, and seemingly small differences in the universe that is used could have a major impact. (Surz, 2006) has a fuller criticism of this form of performance measurement as well as arguments for using random portfolios.

Even if all of the funds in a category used precisely the same universe, had the same volatility and so on, we still wouldn't know if the top-ranked managers had skill. It could be that no manager in the category has skill and that the top-ranked managers are merely the luckiest.

(Kacperczyk et al., 2006) use a rather unique form of performance measurement. This looks at the published positions of a fund at a point in time, and then compares the fund's subsequent return to the return of the published portfolio. Outperformance relative to published portfolios is found to be persistent. Using random portfolios to assess the significance of the outperformance would be easy and quite accurate.

Random portfolios provide an opportunity to measure skill more effectively than the methods just discussed. First, we take a statistical detour.

4.1 Combining p-values

In using random portfolios to measure skill, it will be necessary to combine p-values from different periods of time. A key assumption of combining p-values is that they need to be statistically independent. In our context as long as the tests are for non-overlapping periods of time, this will be true to a practical extent, if not absolutely true.

One way of combining p-values is called Stouffer's method. In this technique the individual p-values are transformed into the quantiles of a standard normal. The p-value of the average of the quantiles is then found. In R the command to do this is:

```
pnorm(sum(qnorm(x)) / sqrt(length(x)))
```

where x is the vector of individual p-values.

Stouffer's method easily admits the use of weights for the individual p-values—for example, if not all of the time periods were the same length.

A weighted sum of the quantiles is performed, and then standardized by its standard deviation—the square root of the sum of squared weights.

When using Stouffer's method, we do not want any of the individual p-values to be either 0 or 1. We need to use centered p-values:

$$p_{centered} = \frac{n_x + .5}{N + 1}$$

where N is the number of random portfolios and n_x is the number of portfolios that are as extreme or more extreme than the observed fund. Stouffer's method is used to combine p-values in what follows. See (Burns, 2004) for a discussion of why Fisher's method of combining p-values is inappropriate.

4.2 Tests with the Example Data

Figure 9 shows a test of skill using information ratios relative to a benchmark. Here we use the same data to test skill based on the mean-variance utility using random portfolios.

The first step is to decide what specific utility is to be computed. In the case of mean-variance utility we need to specify the risk aversion parameter. We then compute the utility achieved within each quarter by each random portfolio and by each manager. The utility of a manager within a quarter is compared to the utilities of the random portfolios—this provides a p-value for that manager in that quarter. Finally, we combine these p-values to derive a p-value for the whole period for each manager.

Figure 12 plots the p-values based on random portfolio tests using mean-variance utility with risk aversion 2. This has many more very small p-values

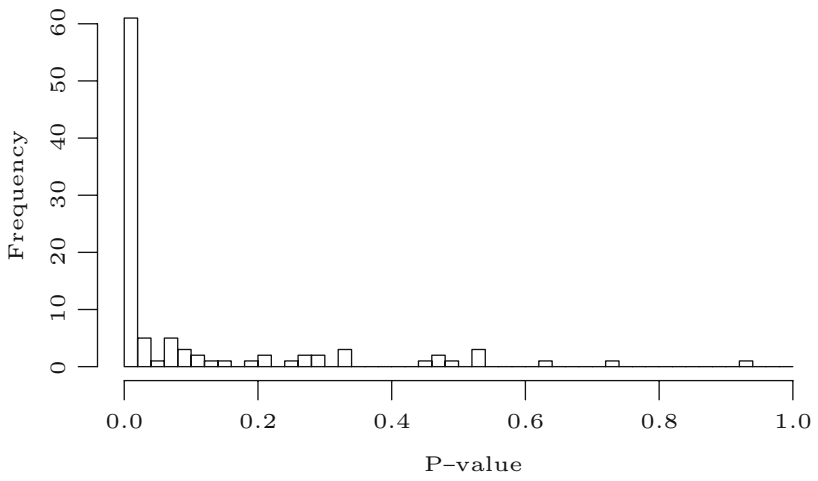

Fig. 12. P-values of the 100 hypothetical managers based on random portfolios using mean-variance utility with risk aversion 2 (over 33 quarters)

Table 2. Counts (out of 100) of the number of hypothetical managers achieving significance levels in the various forms of tests over 33 quarters

test	< 0.05	< 0.01	< 0.001
random portfolio, risk aversion = 2	67	57	37
random portfolio, risk aversion = 1	68	56	35
random portfolio, risk aversion = 0.5	67	52	34
random portfolio, risk aversion = 0	66	51	33
information ratio, equal wt benchmark	35	12	0
information ratio, random benchmark 1	47	17	1
information ratio, random benchmark 2	43	14	0

than Fig. 9. Table 2 shows the number of hypothetical managers that achieved various significance levels for different forms of the tests. The tests using random portfolios clearly have more power than those using information ratios. About a third of the random portfolio tests achieve a p-value less than 0.001, while only one manager in one of the information ratio tests achieves this.

Figure 13 shows the number of hypothetical managers with significant p-values as the number of quarters observed increases. (As is also the case with Fig.10 the managers do not change their behavior as time goes on.) This plot shows the number of significant p-values growing rather steadily. The problems that p-values based on information ratios seemed to have are not in evidence in this plot.

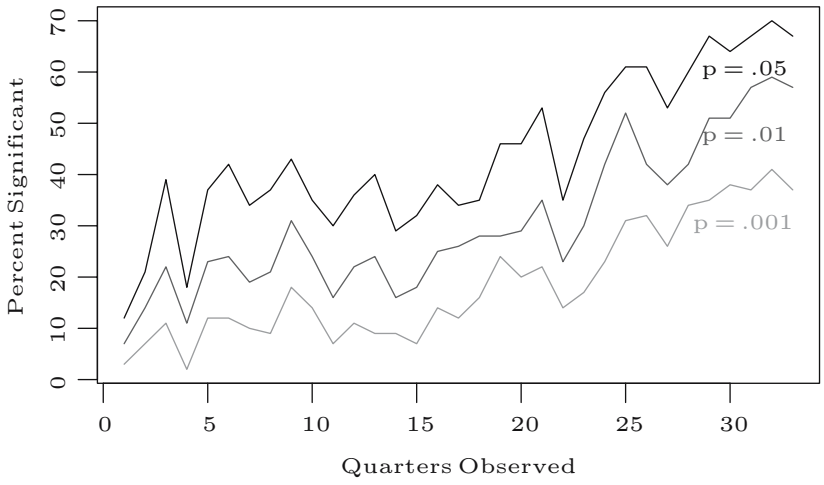

Fig. 13. Percent of hypothetical managers with significant p-values over time using random portfolios with risk aversion 2

5 Investment Mandates

Mandates are the contracts that tell fund managers what they should do with the investor's money. Mandates should be created so that the investor maximizes the usefulness of the entire portfolio. At present this goal is probably not realized very well.

5.1 Tracking Error Should be Maximized

Currently fund managers are often expected to have a relatively small tracking error to their benchmark. If there were no opportunity to invest passively in the benchmark, then this could be a rational approach. But is this the right approach when passive investment is possible?

If both passive and active funds are held, then the total portfolio is enhanced from lower volatility when the correlation between the passive and active portions decreases (assuming the expected return and volatility of the active fund do not change).

We can see what minimizing correlation means for the tracking error by some minor manipulation of its definition. We will denote the active fund by A and the benchmark by B, other notation should be self-explanatory.

$$\mathrm{TE}_{\mathrm{B}}^2(A) = \mathrm{Var}\{A - B\} = \mathrm{Var}\{A\} + \mathrm{Var}\{B\} - 2\mathrm{Cov}\{A, B\} \tag{1}$$

Putting the covariance term alone on the left side and transforming to correlation gives us

$$\mathrm{Cor}\{A, B\} = \frac{\mathrm{Cov}\{A, B\}}{\sqrt{\mathrm{Var}\{A\}}\sqrt{\mathrm{Var}\{B\}}} = \frac{\mathrm{Var}\{A\} + \mathrm{Var}\{B\} - \mathrm{TE}_{\mathrm{B}}^2(A)}{2\sqrt{\mathrm{Var}\{A\}}\sqrt{\mathrm{Var}\{B\}}}. \tag{2}$$

Holding the variance of the active fund constant, the correlation between the active fund and the benchmark is minimized when the (squared) tracking error is maximized.

This directly contradicts (Kahn, 2000), cited by (Waring and Siegel, 2003). Who is right?

5.2 What is Risk?

The argument we've just seen says that tracking errors are ideally large, while (Kahn, 2000) argues that tracking errors should be small. The discrepancy boils down to the definition of risk. The argument in which tracking errors should be large takes the risk to be the mean-variance utility of the entire portfolio—the active part plus the passive part. The argument in which tracking errors should be small takes risk to be the deviation from the benchmark.

Optimal behavior is vastly different depending on which is the more realistic definition of risk.

Calling risk the deviation from the benchmark is the appropriate choice when the benchmark is the liabilities of the fund. If there is no deviation from the benchmark, then the fund carries no risk. For example if the fund needs to deliver x times the value of the S&P 500 in 10 years, then this situation applies with the benchmark equal to the S&P 500.

Alternatively if the benchmark is the S&P 500 but it could reasonably have been some other index of U.S. equities, then exactly reproducing the S&P 500 is not going to be a zero risk solution. This is the more common case.

However, using the absolute utility of the portfolio (where we want to maximize tracking error) is also wrong—it ignores the liabilities altogether, as if we knew nothing about them.

There has been some work on evaluating policies when the liabilities are known only with uncertainty—see, for instance, (Board and Sutcliffe, 2005). A lot of work, however, assumes that liabilities are known, which is almost always not true. One way of thinking about uncertain liabilities is that it is a generalization of a dual benchmark optimisation. So perhaps an approximate answer can be obtained by performing an optimisation with several benchmarks.

My (uneducated) guess is that using the absolute utility is almost always closer to the right answer than using the active utility.

(Muralidhar, 2001) on p. 157 speaks of an example where the actively managed portfolio had a lower asset-liability risk (in a certain sense) than the benchmark portfolio. This is obviously a case where deviation from the benchmark should not be considered to be the risk.

Traditionally there has been another reason to prefer small tracking errors: small tracking errors enhance the ability to declare skill when information ratios are used. Consider an extreme case. Two fund managers outperform an index by 3%, their funds each have the same volatility, but one has a tracking error of 1% while the other has a tracking error of 10%. From a global perspective the two fund managers are equivalent—they have the same return and the same volatility. But in terms of proving skill via the information ratio relative to the index, the first fund manager would be judged to have skill while the second could not be.

6 Random Portfolio Mandates

Random portfolios can be used as the basis of mandates. The investor specifies the constraints that the fund manager is to obey; the manager is judged, and possibly paid, based on the fund's performance relative to random portfolios that obey the constraints. This process gives fund managers the freedom to shape their portfolios the way that they see fit, and provides investors with an accurate measure of the value to them of a fund manager.

In a traditional mandate the investor and fund manager agree on a benchmark and a tracking error allowance. With a random portfolio mandate, it is

the constraints that need to be agreed upon. Of course each party will have views on the constraints.

In general fund managers want constraints to be loose so that they have a lot of freedom, and the random portfolios are allowed to do stupid things. The investor wants to set the constraints so that the fund manager is likely to add as much value as possible. This tends to favor relatively tight constraints on such things as volatility.

While there is a natural tension between the fund manager and the investor, there is also quite a lot of room for cooperation. It is in the interests of both that the fund manager is given enough freedom to capitalize on good investment ideas.

6.1 An Example Mandate

Here we briefly outline what a random portfolio mandate might look like. The items in our mandate are:

- **The evaluation period is 6 months.**
  The frequency of evaluation needs to take the fund manager's strategy into account. Obviously an evaluation over 1 week when the manager is looking at time horizons on the order of 3 to 6 months will be pure noise. A manager that typically holds positions for less than a day could be evaluated very frequently, but the evaluation need not be especially frequent.
- **The universe of assets is the constituents of the S&P 500 at the beginning of the period.**
  To keep things simple, the universe is fixed throughout the period regardless of constituent changes in the index itself. An alternative would be to allow new constituents into the universe, in which case the random portfolios would be given the opportunity to trade into the new assets.
- **The number of assets in the portfolio is to be between 50 and 100, inclusive.**
  These numbers reflect the desire by the fund manager to hold 100 names or slightly fewer, while the lower bound ensures that the fund never gets too concentrated.
  If it is found that the size of the portfolio has a material effect on the distribution of utility, then the random portfolios can be generated with sizes that characterize the actual sizes that the portfolios are likely to be. (In this case the range of allowable sizes would probably be reduced.)
- **The positions are to be long only.**
- **The maximum weight of any asset will be 5%.**
  This seems like a straightforward constraint, but isn't—there could be numerous interpretations of what it means. One practical choice is that a position can be no more than 5% at the point when it is created or added to.

- **The volatility of the fund will be no more than 150% of the volatility of the minimum variance portfolio that satisfies the remaining constraints.**
  This clearly needs more careful definition. Not just any volatility will do—it has to be agreed. One choice would be to provide a specific variance matrix of the universe of assets. An equivalent approach is to provide the specification of how the variance matrix is to be produced. For example, use the default arguments of the POP function `factor.model.stat` with 4 years of daily log returns.

6.2 Operational Issues

In the example mandate, volatility is constrained statically—only information available at the beginning of the period is used. While this avoids the problem of the fund manager unintentionally breaching the mandate because of changes during the period, it doesn't necessarily state how the investor would like the fund manager to behave. The investor may desire the fund manager to control the volatility of the fund throughout the period using updated information. While slightly more involved, the random portfolios can have trading requirements imposed upon them during the period. However, if the fund manager is being judged based on a utility that includes volatility as a component, then the fund manager should already be taking changes in the volatility environment into account in the best interests of the investor.

The evaluation criterion can be at least as useful in shaping the fund manager's behavior as the constraints. The criterion can be anything that can be computed using information that is available at the end of the period—we are not limited to any particular measures such as the return or a mean-variance utility. For example, the criterion might include the skewness of the daily log returns during the evaluation period, and the correlation with some proxy of the rest of the investor's portfolio.

The fund manager may be at a disadvantage (or advantage) relative to the random portfolios if they are allowed unlimited turnover. If a portfolio is already in place, then it is reasonable for the random portfolios to be generated so that there is a maximum amount of trading from the portfolio that exists at the start of the period.

Proposed revisions may arise about the form of the mandate. For example the fund manager may come to think that a particular constraint is not the best approach. With the use of random portfolios the fund manager can demonstrate to the investor the effect of changing the constraint. The mandate can be revised from period to period as more is learned.

6.3 Performance Fees

Performance fees can easily be based on random portfolios from a mandate. As stated earlier, the criterion used to measure success can be specialized to fit

the particular situation. As long as the criterion is a close match to the actual utility of the investor, then the interests of the investor and fund manager are aligned when a performance fee is used.

It is probably sensible to reward overlapping time periods—for example, to have a quarterly, yearly and three-yearly component of the fee. This should help to reduce the fund manager's ability to game the performance fee.

The starting point for a performance fee based on random portfolios is likely to be the average utility of the random portfolios. The fund manager should be paid for utility that is delivered above the base. How much is paid for an increment in utility is, of course, up to the investor. From the investor's point of view the payment should be for how much utility is delivered, not how difficult it is to deliver.

As in (Waring and Siegel, 2003) the investor is (or should be) doing a portfolio optimisation. The investor is selecting weights for a variety of active and passive funds. The investor may assign different utility functions to different fund managers (for instance the investor could vary what the managers should have a small correlation to), but it is sensible for the investor to pay the same amount to each fund manager for an equivalent increase in utility. More weight should be given to managers who have the ability to deliver a lot of utility.

7 Summary

Random portfolios have been shown to be of use in two respects—measuring skill and forming investment mandates.

The measurement of skill with random portfolios avoids some of the noise that is introduced when performance is measured relative to a benchmark. This means that knowledge of skill can be more precise. An accurate assessment of skill with random portfolios requires a knowledge of both the returns of the fund and the constraints that the fund obeys. Less accurate assessment can be done where the constraints are not known specifically. Even in this case, though, the results will still be dramatically better than comparing to a benchmark or to peers. The statistical statements that result are distribution-free—there are no assumptions on the distribution of returns or measures of utility.

Mandates that are based on random portfolios allow fund managers to play to their strengths because they need not be tied to a benchmark. This also allows more flexibility for the investor to shape the behavior of the fund managers to best advantage.

Other uses of random portfolios include the assessment of the opportunity set available to a fund manager with a given strategy.

Acknowledgements

I thank Craig Israelsen for providing the data in Table 1. Larry Siegel and Barton Waring helped with a discussion that clarified some of my thinking. Comments from two anonymous referees improved the presentation.

References

Board, J. and C. Sutcliffe: 2005, 'Joined-Up Pensions Policy in the UK: An Asset-Liability Model for Simultaneously Determining the Asset Allocation and Contribution Rate'. Technical report, ICMA Centre, University of Reading.

Burns, P.: 2004, 'Performance Measurement via Random Portfolios'. Working paper, Burns Statistics, http://www.burns-stat.com/.

Burns, P.: 2006, 'Random Portfolios for Evaluating Trading Strategies'. Working paper, Burns Statistics, http://www.burns-stat.com/.

Burns Statistics: 2005, 'POP Portfolio Construction User's Manual'. http://www.burns-stat.com.

Dawson, R. and R. Young: 2003, 'Near-uniformly Distributed, Stochastically Generated Portfolios'. In: S. Satchell and A. Scowcroft (eds.): *Advances in Portfolio Construction and Implementation*. Butterworth–Heinemann.

Ding, B. and R. Wermers: 2004, 'Mutual Fund Stars: The Performance and Behavior of U.S. Fund Managers'. Technical report, SSRN, http://papers.ssrn.com.

Engstrom, S.: 2004, 'Does Active Portfolio Management Create Value? An Evaluation of Fund Managers' Decisions'. Technical Report 553, Stockholm School of Economics, http://swopec.hhs.se/hastef/.

Ferson, W. and K. Khang: 2002, 'Conditional Performance Measurement using Portfolio Weights: Evidence for Pension Funds'. *Journal of Financial Economics* **65**, 249–282.

Fogel, D. B.: 2006, *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. IEEE Press—Wiley Interscience, 3rd edition.

Grinblatt, M. and S. Titman: 1993, 'Performance Measurement without Benchmarks: An Examination of Mutual Fund Returns'. *Journal of Business* **66**, 47–68.

Holland, J. H.: 1975, *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

Israelsen, C. L.: 2003, 'Relatively Speaking'. *Financial Planning Magazine*.

Kacperczyk, M., C. Sialm, and L. Zheng: 2006, 'Unobserved Actions of Mutual Funds'. Technical report, SSRN, http://papers.ssrn.com.

Kahn, R. N.: 2000, 'Most Pension Plans Need More Enhanced Indexing'. *Investment Guides, Institutional Investor*.

Kosowski, R., A. Timmermann, H. White, and R. Wermers: 2001, 'Can Mutual Fund Stars Really Pick Stocks? New Evidence from a Bootstrap Analysis'. Working paper, http://papers/ssrn.com.

Kothari, S. P. and J. B. Warner: 2001, 'Evaluating Mutual Fund Performance'. *The Journal of Finance* **56**, 1985–2010.

Maringer, D.: 2005, *Portfolio Management with Heuristic Optimisation*. Springer.

Muralidhar, A. S.: 2001, *Innovations in Pension Fund Management*. Stanford University Press.

R Development Core Team: 2005, 'R: A Language and Environment for Statistical Computing'. R Foundation for Statistical Computing, http://www.r-project.org. ISBN 3-900051-07-0.

Siegel, L. B.: 2003, *Benchmarks and Investment Management*. CFA Institute (for hardcopy). Full text available online at http://www.qwafafew.org/?q=filestore/download/120.

Statman, M. and J. Scheid: 2005, 'Measuring the Benefits of Diversification and the Performance of Money Managers'. *Journal of Investment Consulting* **8**(1), 18–28.

Surz, R. J.: 2006, 'A Fresh Look at Investment Performance Evaluation: Unifying Best Practices to Improve Timeliness and Reliability'. *Journal of Portfolio Management Summer*.

Waring, M. B. and L. B. Siegel: 2003, 'The Dimensions of Active Management'. *The Journal of Portfolio Management* **29**(3), 35–51.

Real Options with Random Controls, Rare Events, and Risk-to-Ruin

Nicos Koussis, Spiros H. Martzoukos and Lenos Trigeorgis

Department of Public and Business Administration, University of Cyprus

Summary. Situations involving real investment options in the presence of multiple sources of jump risk, and controls are analyzed. Randomly arriving jumps include also the special cases of jump-to-ruin on the underlying asset, or on the contingent claim. Management has available impulse-type controls with random outcome. The analytic solutions when available, and a Markov-Chain numerical approach for solving more general investment decision problems are demonstrated.

Key words: Flexibility, real options, multi-class jump-diffusion processes, catastrophic risks, controls with random outcome, Markov-Chains

1 Introduction

Real investment options often involve multiple sources of jump risk and managerial controls. Randomly arriving value jumps may come from multiple sources representing political, technological or other risks. We also examine important special cases involving catastrophic jumps-to-ruin: one is a threat to the underlying asset, and the other is a threat to the contingent claim. Management has available impulse-type controls (with random outcome) that capture managerial actions aiming at enhancing the value of investment opportunities. This control approach can be applied in cases where firms, before making a capital-intensive investment decision (to bring a product to market, etc.), can invest to improve its attributes and enhance its market appeal or lower its cost of production. This can be done via R&D or by adopting existing technological innovations. In the case of European options we demonstrate the analytic solutions, and provide a Markov-Chain numerical solution framework to handle more general problems of simple or sequential investment decisions involving potential early exercise. We provide a synthesis of the random controls (impulse controls with random outcome) discussed in Martzoukos (2000) with the general jump-diffusion approach in the presence

of multiple sources of jump risk of Martzoukos and Trigeorgis (2002), and include the special cases of catastrophic jumps.

We first present the general framework before controls are introduced, and provide the partial differential equation (or, to be more precise the partial integro-differential equation, PIDE) that such a claim must follow. Such equations are generally hard to handle, but analytic solutions for special cases involving European options (in the spirit of Merton, 1976) are obtained, and the impact of catastrophic jumps is examined. We demonstrate that the two classes of jumps-to-ruin assumptions result in significantly different solutions in the case of put options. We then consider the case of superimposing impulse controls with random outcome for the stochastic process governing the underlying asset. The analytic solutions we present are useful in terms of real option valuation and for the study of optimal investment decisions. The analytic solutions also provide a benchmark for testing the accuracy of the numerical Markov-chain solution framework proposed in the last section.

2 General Framework: European Options with Multiple Types of Jumps

Stochastic processes with discontinuous (Poisson-type) events have been extensively studied in the context of option pricing by Merton (1976) and others (e.g., Ball and Torous, 1985, Amin, 1993). This literature has mostly focused on the case of a single source of discontinuity (information arrival). Notable exceptions are Jones (1984), who studied hedging of financial (European) options involving two classes of jumps, Martzoukos and Trigeorgis (2002) who studied complex real options in the presence of multiple sources of jumps. Kou (2002) and Kou and Wang (2004) study financial options with a single source of jumps that can take values from two different probability distributions – this is in effect equivalent to a jump-diffusion with two sources of jumps. In this paper we make the assumption that jump risk is not priced. For issues relating to pricing of the jump risk see Bates (1991), Bardham and Chao (1996), Chan (1999), Henderson and Hobson (2003), and Martzoukos (2003).

In our general framework we assume the existence of multiple $(N + 2)$ sources of jumps. The N classes are non-catastrophic risks affecting the underlying asset, the $(N+1)^{\text{th}}$ class is a catastrophic risk that affects the underlying asset, and the $(N + 2)^{\text{th}}$ class is a catastrophic risk that affects the contingent claim. The underlying asset S is assumed to follow a continuous-time stochastic process of the form:

$$\frac{dS}{S} = \mu\, dt + \sigma\, dZ + \sum_{i=1}^{N+1} (k_i dq_i). \tag{1}$$

Here μ is the instantaneous drift and σ the instantaneous standard deviation of continuous returns, dZ is an increment to a standard Wiener process, dq_i is a jump counter that takes a value 1 with probability $\lambda_i dt$ or a value 0 with probability $(1-\lambda_i)dt$, λ_i is the (annual) frequency of a jump of type i, and k_i is the jump size for each event class i. Summation is over the $N+1$ classes (types) of rare events involving the underlying asset. The N classes represent non-catastrophic risks, and the $(N+1)^{\text{th}}$ class involves a catastrophic risk that causes the underlying asset value S to jump to zero. Due to the impact of the N non-catastrophic jump risks, and before the catastrophic $(N+1)^{\text{th}}$ class jump-to-ruin occurs, the actual trend of the underlying (asset) value process equals $\mu + \sum_{i=1}^{N+1}(\lambda_i \bar{k}_i)$, involving a term $\lambda\overline{k} \equiv \lambda E[k]$ for each event class i that affects the underlying asset, with $E[.]$ denoting the expectations operator. Under risk-neutral valuation, the underlying asset S follows the process:

$$\frac{dS}{S} = (r-\delta^*)\,dt + \sigma\,dZ + \sum_{i=1}^{N+1}(k_i dq_i). \tag{1a}$$

Following Merton (1976), we assume the jump risk to be diversifiable (and hence not priced) and that an intertemporal capital asset pricing model holds (Merton, 1973). Thus, we do not need to invoke the standard replication and continuous-trading arguments of Black and Scholes (1973). The risk-neutral drift above differs from the riskless rate r by δ^*, where $\delta^* = \delta + \sum_{i=1}^{N+1}(\lambda_i \bar{k}_i)$. The parameter δ may represent any form of "dividend yield" or opportunity cost (e.g., in McDonald and Siegel, 1984, δ may be a deviation from the equilibrium required rate of return, while in Brennan, 1991, δ is a convenience yield). In the presence of random jumps, the deterministic drift component in general includes a "*compensation*" term, $\lambda\overline{k} \equiv \lambda E[k]$, for each event class. This term is also present in the jump-diffusion model of Merton (1976) where the underlying asset is traded, in order to ensure that the expected return of the asset equals the required (risk-neutral) return $r - \delta$ (even when $\overline{k} \neq 0$). For a non-traded real option, the compensation term may be absent (e.g., Dixit and Pindyck, 1994, pp. 170–172). In this case the rare events may not only affect volatility but may also affect the expected growth rate (via $\lambda\overline{k}$). We similarly assume that the $(N+1)^{\text{th}}$ class jump is due to technical uncertainty (uncorrelated to market movement) and is thus not compensated. In what follows we assume for simplicity that jumps are not related to market events but only to asset-specific technical uncertainties, so that all the compensation terms are absent (thus $\delta^* = \delta$). Generally, several of the N classes may have market-related risks requiring that compensation terms be added in the following equations – this without any effect on the results and insights presented herein, so the general notation δ^* may be retained.

The stochastic differential (1a) can alternatively be expressed in integral form as:

$$\ln[S(T) - \ln[S(0)] = \int_0^T [r - \delta^* - .5\sigma^2]\, dt + \int_0^T \sigma\, dZ(t) + \sum_{i=1}^{N} \sum_{q=1}^{n_i} \ln(1 + k_{i,q}) \tag{2}$$

$$\text{if for jump class } i = N + 1 \text{ and } q = 0,$$

or

$$S(T) = 0, \text{if for jump class } i = N + 1 \text{ and } q \neq 0. \tag{2a}$$

Equation (2) holds in the absence of any realization of jumps-to-ruin, while (2a) holds when there has been a realization of a jump-to-ruin on the underlying asset. Note that realizations of jump-to-ruin on the contingent claim do not affect the underlying asset. In the above equation, the nested summation is over the realizations for all N classes of jumps of size $k_{i,q}$, with $n = (n_1, \ldots, n_N)$ an $N-$element vector with each element being the number of realized i-class jump occurrences. For each (independent) *event class*, we assume that the distribution of jump size, $1 + k_i$, is log-normal. That is, $\ln(1 + k_i) \sim \mathbf{N}(\gamma_i - .5\sigma_i^2,\, \sigma_i^2)$, with $\mathbf{N}(.,.)$ denoting the normal density function with mean $\gamma_i - .5\sigma_i^2$ and variance σ_i^2, and $E[k_i] \equiv \bar{k}_i = \exp(\gamma_i) - 1$. The value of an option on claim F contingent on underlying asset S is characterized by the following partial integro-differential equation (PIDE):

$$\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 F}{\partial S^2} + [r - \delta^*] S \frac{\partial F}{\partial S} - \frac{\partial F}{\partial t} - rF + \sum_{i=1}^{N} \{\lambda_i E[F(S + Sk_i, t) - F(S, t)]\}$$
$$+\lambda_{N+1} E[F(S + Sk_{N+1}, t) - F(S, t)] + \lambda_{N+2} E[k_{N+2} F(S, t) - F(S, t)] = 0.$$

Defining a PIDE is important especially for the case of American options, since, as we will see below, European options may have analytic solutions. Note that for realization of the $(N + 2)^{\text{th}}$ class jump-to-ruin, the value of the contingent claim is $F = k_{N+2} F(S, t) = 0$ with $k_{N+2} = 0$. For realization of the $(N + 1)^{\text{th}}$ class jump-to-ruin, the value of contingent claim F depends on the exact nature of the claim. In the case of a call option that depends on asset S alone, $F = 0$, but in the case of a put option, F in general equals X' where most often X' equals the exercise price X unless contractually defined equal to a different value. This leads to the following two equations:

The equation for the call option is

$$\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 F}{\partial S^2} + [r - \delta^*] S \frac{\partial F}{\partial S} - \frac{\partial F}{\partial t} - (r + \lambda_{N+1} + \lambda_{N+2}) F$$
$$+ \sum_{i=1}^{N} \{\lambda_i E[F(S + Sk_i, t) - F(S, t)]\} = 0. \tag{3}$$

Effectively this is equivalent to adding the term $\lambda_{N+1} + \lambda_{N+2}$ to both r and δ. For the put option the equation is

$$\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 F}{\partial S^2} + [r - \delta^*] S \frac{\partial F}{\partial S} - \frac{\partial F}{\partial t} - (r + \lambda_{N+1} + \lambda_{N+2}) F + \sum_{i=1}^{N} \{\lambda_i E[F(S + Sk_i, t) - F(S, t)]\} + \lambda_{N+1} X = 0. \tag{3a}$$

This is again equivalent to adding the term $\lambda_{N+1} + \lambda_{N+2}$ to both r and δ, plus the additional term $\lambda_{N+1}X$. This additional term is like a continuous cash flow, and its implementation in a numerical solution framework is similar to the implementation of constant continuous interest payments in interest rate contingent claims (i.e., bonds).

The above multi-class set up is an extension of Merton (1976). As in his case, these PIDEs are difficult to solve directly in general (a rare but computationally very intensive method is that of Andersen and Andreasen, 2001). Following the approach in Merton (see also Jones, 1984), we subsequently value a European call option on asset S with time to maturity T and exercise price X, assuming independence between the different event classes and the underlying Wiener process dZ. The value of a European call option with multiple sources of jumps is given by (iterated integral):

$$\begin{aligned} & F_{call}(S, X, T, \sigma, \delta, r, \lambda_i, \gamma_i, \sigma_i) = \\ & e^{-rT} \sum_{n_1=0}^{\infty} \cdots \sum_{n_{N+2}=0}^{\infty} \{P(n_1, \ldots, n_{N+2}) \\ & \times E[(S_T - X)^+ | \ (n_1, \ldots, n_{N+2}) \ jumps]\} \\ = \ & e^{-(r+\lambda_{N+1}+\lambda_{N+2})T} \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \{P(n_1, \ldots, n_N) \\ & \times E[(S_T - X)^+ | \ (n_1, \ldots, n_N) \ jumps]\}, \end{aligned} \tag{4}$$

where $P(n_1, \ldots, n_N)$ denote the joint probabilities of any *random* realization of $n = (n_1, \ldots, n_N)$ jumps. Under the independence assumption, these joint probabilities simplify to the N-term product

$$P(n_1, \ldots, n_N) = \prod_{i=1}^{N} [e^{-\lambda_i T} (\lambda_i T)^{n_i} / n_i!].$$

The call option has positive value only when none of the jumps-to-ruin is realized, a joint event with probability equal to $P(n_{N+1} = 0, n_{N+2} = 0) = e^{-(\lambda_{N+1}+\lambda_{N+2})T}$.

In order to implement the equation for the call option we need to evaluate the risk-neutral expectation $E[(S_T - X)^+ \mid (n_1, \ldots, n_N) \ jumps]$. This

is derived along the lines of the Black-Scholes model but conditional on $n = (n_1, \ldots, n_N)$ jumps, and is:

$$E[(S_T - X)^+ \mid (n_1, \ldots, n_N)\ jumps] = S_e^{[(r-\delta^*)T + \sum_{i=1}^{N}(n_i\gamma_i)]} N(d_{1n}) - X\, N(d_{2n}), \tag{4a}$$

where

$$d_{1n} \equiv \frac{\ln(S/X) + (r-\delta^*)T + \sum_{i=1}^{N}(n_i\gamma_i) + .5\sigma^2 T + \sum_{i=1}^{N}(.5 n_i \sigma_i^2)}{[\sigma^2 T + \sum_{i=1}^{N}(n_i\sigma_i^2)]^{1/2}}$$

and

$$d_{2n} \equiv d_{1n} - [\sigma^2 T + \sum_{i=1}^{N}(n_i\sigma_i^2)]^{1/2}.$$

$N(d)$ again denotes the cumulative standard normal density evaluated at d. To operationalize the infinite sum series we truncate when two conditions are met (to a reasonable approximation level): the sum of probabilities $\sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \{P(n_1, \ldots, n_N)\}$ equals unity, and the option value does not change by adding more terms.

The value of a European put option with multiple types of jumps is similarly shown to be

$$\begin{aligned} & F_{put}(S, X, T, \sigma, \delta, r, \lambda_i, \gamma_i, \sigma_i) = e^{-rT} \sum_{n_1=0}^{\infty} \cdots \sum_{n_{N+2}=0}^{\infty} \{P(n_1, \ldots, n_{N+2}) \\ & \times E[(X - S_T)^+ \mid (n_1, \ldots, n_{N+2}) jumps]\} \\ = \ & e^{-(r+\lambda_{N+1}+\lambda_{N+2})T} \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \{P(n_1, \ldots, n_N) \\ & \times E[(X - S_T)^+ \mid (n_1, \ldots, n_N) jumps]\} + (1 - e^{-\lambda_{N+1}T}) e^{-\lambda_{N+2}T} e^{-rT} X \end{aligned} \tag{5}$$

where the last term accounts for the probability that the first catastrophic events occurs but the second does not. Again, to implement the equation for the put option we need to evaluate the risk-neutral expectation

$$\begin{aligned} & E[(X - S_T)^+ \mid (n_1, \ldots, n_N)\ jumps] \\ & = X\, N(-d_{2n}) - S_e^{[(r-\delta^*)T + \sum_{i=1}^{N}(n_i\gamma_i)]} N(-d_{1n}), \end{aligned} \tag{5a}$$

where the parameters d_{1n} and d_{2n} are again defined like in the call option. Observe that the put option differs slightly from Merton's weighted-average Black-and-Scholes-type formula by the last term. This term comes from the

possibility that jumps-to-ruin on the underlying asset may be realized before option maturity. In that case, the (present) value of the option would equal $e^{-rT}X$ times the respective probability.

The assumption of jump-to-ruin on the underlying asset was also treated in McDonald and Siegel (1986) for a perpetual option with the same result; effectively an increase in both the riskless rate and the dividend yield. This result and ours differs from the original treatment in Merton (1976) which assumed that the underlying asset process was already compensating for the knowledge on the existence of the jump by including in the drift the term $-\lambda_{N+1}\overline{k}_{N+1} = -\lambda_{N+1}(-1) = \lambda_{N+1}$. Thus, in Merton's case, for a traded asset the risk-neutral drift becomes $r - \delta - \lambda_{N+1}\overline{k}_{N+1} = r - \delta + \lambda_{N+1}$, and discounting (like in our case) is done at $r + \lambda_{N+1}$. So the final impact of this jump-to-ruin assumption for a traded asset is that only the riskless rate need be augmented by λ_{N+1}. Here we retain the assumption (similarly to McDonald and Siegel, 1986) that for a real option on an asset, the compensation term is not needed and the result is to augment both the riskless rate and the dividend yield by the intensity of the rare event.

3 Superimposing Impulse Controls with Random Outcome

We now superimpose costly impulse-type (multiplicative) controls of (random) size k, as in Martzoukos (2000), while retaining the jump-diffusion assumptions (like in Martzoukos and Trigeorgis, 2002) and the two types of jump-to-ruin. Impulse control has been used in real options literature to model change of operating scale and other similar problems (for a thorough review of such literature, see Vollert, 2003). The important difference between the traditional approach and ours is that in our case the realization of the control is random, following a specified probability distribution. This may happen with costly R&D projects, or projects of innovation adoption, like for example, in the redesign of a product with the intention to strengthen the price and/or market share, potentially with a negative customer response or reduced productivity (see discussions in Brynjolffson and Hitt, 2000). For convenience, we assume that the natural logarithm of $1+k$ is normally distributed, with mean $\gamma_c - 0.5\sigma_c^2$, variance σ_c^2, and $E[k_c] = e^{\gamma_c} - 1$. We assume that the outcome of control C is independent of the underlying asset Wiener process or any jump components, can be attained at a cost X_C, and involves a diversifiable risk. We observe that when a control is activated (and before the actual outcome is observed) it is expected to affect the value of the underlying asset and enhance its volatility. Given the volatility increase and that for a call option we would expect a positive effect (and for a put option a negative effect) on the underlying asset, it follows from the convexity of option values that the ex ante outcome is always an increase in option value. Of course, this increase must exceed the cost of this control action for control activation to be optimal.

In practice, several mutually-exclusive controls may be pursued at any time. For the European option with controls available at $t = 0$, the above distributional assumptions result in analytic solutions resembling the familiar Black-Scholes model. For more general model specifications other plausible distributional assumptions can be made, but generally one needs to resort to numerical solutions like those described in Sect. 4. Here we focus on the real investment (European call) option F to acquire underlying project value S by paying a capital cost X. A put option to sell S in order to receive X could be treated similarly. In general, the objective is to find the optimal control policy that maximizes the value of the real claim. Each control action can (only) be taken at pre-specified times $t(c)$. The decision maker has the option to activate each available control at time $t(c)$ by paying cost X_C. In this section, we derive an analytic solution for the special case of the European option, where controls are available only at time $t = 0$. More generally, the control problem can be described as follows:

$$\underset{C}{\text{Maximize}} \ \{F[t, S, X_C, t(c)]\} \text{ subject to}$$

$$\frac{dS}{S} = (r - \delta^*)dt + \sum_{i=1}^{N+1} (k_i dq_i) \text{ when controls are not activated,}$$

and

$$\frac{dS}{S} = (r - \delta^*)dt + \sum_{i=1}^{N+1} (k_i dq_i) + k_c dq_c \text{ when a control is activated.}$$

The control distributional characteristics are: $\ln(1 + k_c)$ normally distributed with mean $\gamma_c - 0.5\sigma_c^2$, variance σ_c^2 and $E[k_c] = e^{\gamma_c} - 1$.

The terminal condition at maturity T defines the particular claim as a call (or alternatively as a put) option. In (4) and (5) we use a subscript to denote if it is a call or a put option. In the expressions below, we omit the subscript when we refer to the call option. Given these assumptions, the European call option value *conditional* on activation of a random control c at time $t = 0$, $t(c) = 0$ is given by:

$$\begin{aligned} &F_{cond}[S, X, \sigma, \delta^*, \lambda_i, \gamma_i, \sigma_i, T, r, \gamma_c, \sigma_c, t(c) = 0] = \\ &e^{-(r+\lambda_{N+1}+\lambda_{N+2})T} \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \{P(n_1, \ldots, n_N) \\ &\times E[(S_T - X)^+ \mid (n_1, \ldots, n_N) \ jumps; \ t(c) = 0]\big\}. \end{aligned} \tag{6}$$

Again we need the risk-neutral expectation in order to implement the call option. This, derived along the lines of the Black-Scholes-Merton jump-diffusion model but conditional on control activation, is:

$$\begin{aligned} &E[(S_T - X)^+ \mid (n_1, \ldots, n_N) \ jumps; t(c) = 0] \\ = \ &S_e^{[(r-\delta^*)T + \sum_{i=1}^{N} (n_i\gamma_i) + \gamma_C]} N(d_{1n}) - X\, N(d_{2n}), \end{aligned} \tag{6a}$$

where

$$d_{1n} \equiv \frac{\ln(S/X) + (r - \delta^*)T + \sum_{i=1}^{N} (n_i\gamma_i) + .5\sigma^2 T + \sum_{i=1}^{N} (.5n_i\sigma_i^2) + \sigma_C^2}{[\sigma^2 T + \sum_{i=1}^{N} (n_i\sigma_i^2) + \sigma_j^2]^{1/2}}$$

and

$$d_{2n} \equiv d_{1n} - [\sigma^2 T + \sum_{i=1}^{N} (n_i\sigma_i^2) + \sigma_C^2]^{1/2},$$

with $N(d)$ denoting the cumulative standard normal density evaluated at d. Table 1 presents representative results using the analytic solutions (in parenthesis) that show the impact of the control, along with different assumptions for the intensity of ruin. The explanation of the derivation of the numerical values is in Sect. 4.

The sensitivity (the Greeks) of the European call option (conditional on control activation at $t = 0$) with respect to asset price S, the control mean γ_c, and the control volatility σ_c are as follows:

$$\begin{aligned}
\frac{\partial F}{\partial S} &= \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \{P(n_1, \ldots, n_N) \\
&\quad \times N(d_{1n}) e^{\left[-(\delta^* + \lambda_{N+1} + \lambda_{N+2})T + \sum_{i=1}^{N} (n_i\gamma_i) + \gamma_c\right]}\} > 0, \\
\frac{\partial F}{\partial \gamma_c} &= \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \{P(n_1, \ldots, n_N) \\
&\quad \times N(d_{1n}) S e^{\left[-(\delta^* + \lambda_{N+1} + \lambda_{N+2})T + \sum_{i=1}^{N} (\gamma_i) + \gamma_c\right]}\} > 0,
\end{aligned}$$

and

$$\frac{\partial F}{\partial \sigma_c} = \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \{P(n_1, \ldots, n_N) \frac{X e^{-(r + \lambda_{N+1} + \lambda_{N+2})T - \frac{d_2^2}{2}}}{\sqrt{2\pi \left[\sigma^2 T + \sum_{i=1}^{N} (n_i\sigma_i^2) + \sigma_c^2\right]}} \sigma_c\} > 0.$$

Their calculation follows the calculation of the Greeks in the Black and Scholes model (see for example, Stoll and Whaley, 1993, Chap. 11; for additional insights, see also Bergman Grundy, and Wiener, 1996). Due to the jump-diffusion, the call option is a probability-weighted average of call option values conditional on the realizations of the N jump-classes (and conditional on the impact the control action, through γ_c, and σ_c). Thus, the sensitivity to each parameter is again a probability-weighted average of the sensitivity of each term, which is calculated like in the standard Black and Scholes model.

Table 1. Real option values with rare events (jumps), control activation, and jumps-to-ruin

Ruin intensity	A. Case with jump-diffusion but without control: $\gamma_c = 0.00, \sigma_c = 0.00 (c = 0)$.		
	$S = 75$	$S = 100$	$S = 125$
$\lambda_{N+1} + \lambda_{N+2} = 0$	0.402 (0.403)	6.199 (6.192)	24.041 (24.041)
$\lambda_{N+1} + \lambda_{N+2} = 0.10$	0.364 (0.365)	5.609 (5.603)	21.753 (21.753)
$\lambda_{N+1} + \lambda_{N+2} = 0.25$	0.313 (0.314)	4.828 (4.822)	18.723 (18.723)
$\lambda_{N+1} + \lambda_{N+2} = 0.50$	0.244 (0.245)	3.760 (3.756)	14.582 (14.581)
Ruin intensity	B. Case with control (and catastrophic jumps only): $\gamma_c = 0.10$, $\sigma_c = 0.10$ $(c = 0)$.		
	$S = 75$	$S = 100$	$S = 125$
$\lambda_{N+1} + \lambda_{N+2} = 0$	0.500 (0.500)	11.416 (11.413)	34.574 (34.574)
$\lambda_{N+1} + \lambda_{N+2} = 0.10$	0.452 (0.452)	10.330 (10.327)	31.284 (31.284)
$\lambda_{N+1} + \lambda_{N+2} = 0.25$	0.389 (0.389)	8.891 (8.888)	26.926 (26.926)
$\lambda_{N+1} + \lambda_{N+2} = 0.50$	0.303 (0.303)	6.924 (6.922)	20.970 (20.870)
Ruin intensity	C. Case with jump-diffusion and control: $\gamma_c = 0.10$, $\sigma_c = 0.10$ $(c = 0)$.		
	$S = 75$	$S = 100$	$S = 125$
$\lambda_{N+1} + \lambda_{N+2} = 0$	1.618 (1.617)	13.511 (13.506)	35.666 (35.666)
$\lambda_{N+1} + \lambda_{N+2} = 0.10$	1.464 (1.463)	12.225 (12.221)	32.272 (32.272)
$\lambda_{N+1} + \lambda_{N+2} = 0.25$	1.260 (1.259)	10.522 (10.518)	27.777 (27.777)
$\lambda_{N+1} + \lambda_{N+2} = 0.50$	0.982 (0.981)	8.195 (8.192)	21.633 (21.633)

Notes: The above case involves a European (real) call option where the underlying asset (project) follows a jump-diffusion process with $N = 2$ sources of non-catastrophic rare events; in addition we consider jump-to-ruin threats on both the underlying asset (the $N + 1$ event class) and the option (the $N + 2$ event class) and managerial control activation at $t = 0$. The parameter values are: $X = 100$, $r = 0.10$, $\delta^* = \delta = 0.10$, $\sigma = 0.10$, $T = 1$, and $S = 75$, 100, 125; for the non-catastrophic rare events: $\lambda_1 = \lambda_2 = 0.50$, $\gamma_1 = 0.10$, $\gamma_2 = -0.10$, and $\sigma_1 = \sigma_2 = 0.10$; and for the control $\gamma_c = 0.10$, $\sigma_c = 0.10$, and $c = 0$; the ruin intensity for both classes of catastrophic events ranges from 0.10 to 0.50. Numerical values using a scheme with 650 refinements in the asset dimension, 80 steps in the time dimension, and a 125-nomial Markov-chain differ by no more than $\pm 0.15\%$ from the analytic results (presented in parenthesis and calculated from (4), (4a), (6), and (6a)). In the absence of both jumps and controls, the at-the-money call option value equals 3.613 (3.608).

Now we wish to verify whether activating a control maximizes option value. When there is a single control available at t(c) = 0, the optimal value for a European call option equals:

$$\begin{aligned}&\text{Max}\{F_{cond}[S, X, \sigma, \delta^*, \lambda_i, \gamma_i, \sigma_i, T, r, \gamma_c, \sigma_c, t(c) = 0]\\&-X_C, F(S, X, \sigma, \delta^*, \lambda_i, \gamma_i, \sigma_i, T, r)\}\end{aligned}$$

Similarly, in case of more than one mutually-exclusive controls (all at $t = 0$), the optimal control will be chosen among the K alternatives $\{F_{cond1}(.)$ – X_{C1} ,...,$F_{condK}(.)$ – X_{CK}, $F\}$. The problem above has analytic solutions for European call (and put) options when controls can be exercised at $t = 0$ only. More generally, one needs to resort to the numerical solutions discussed in Sect. 4.

3.1 Optimal Decision Thresholds

Suppose that the decision maker faces a situation similar to the case treated in the first part of Sect. 3. The underlying asset follows a multi-class jump-diffusion process, with two additional sources of catastrophic risk present, and a costly control that can be activated at time zero. Suppose further the control not only has an impact (with random outcome) on the underlying asset value, but it also affects (reduces) the frequency of the catastrophic events. This action has the element of pre-emption against competitive entry, etc. Empirical evidence on strategic preemption is ample in Bunch and Smiley (1992). They find that, for mature markets, advertising, branding, and product differentiation so as to fill most or all product niches, are used most frequently; for newly developed, concentrated, research intensive markets, strategic deterrence employs intensive advertising, and high R&D expenditures resulting in broad patenting practices.

For the call option we need only consider the cumulative effects ($-\lambda_C$) on the frequency of both sources of catastrophic risk. This is a typical setting characterizing any complex/sequential decision making framework before a capital intensive investment is made. We utilize the special case with analytic solutions to obtain insights on the thresholds that determine the optimal managerial decisions. In general, for increasing asset values, the following describes the alternative decision regions: $\{W, C, W, C\}$ where W stands for wait and C stands for *control activation* (early option exercise is also feasible but we avoid it here for brevity without any significant effect on the derived insights). The last region (where for large asset values control activation is optimal) arises due to the multiplicative nature of the controls, and is less interesting from an economic perspective. It is very common though for the intermediate "wait" region to vanish so we get $\{W, C\}$. As we will later see, when the W region appears twice, the first one is of more economic significance. To understand why these regions appear, one can only examine

the derivatives of the option value functions with respect to the price of the underlying asset S:

$$\frac{\partial F}{\partial S} \underset{S\to\infty}{\lim} = \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \{P(n_1,\ldots,n_N) \\ e^{\left[-(\delta^*+\lambda_{N+1}+\lambda_{N+2})T+\sum_{i=1}^{N}(n_i\gamma_i)\right]}\} > 0,$$

$$\frac{\partial F_{cond}}{\partial S} \underset{S\to\infty}{\lim} = \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \{P(n_1,\ldots,n_N) \\ e^{\left[-(\delta^*+\lambda_{N+1}+\lambda_{N+2}-\lambda_C)T+\sum_{i=1}^{N}(n_i\gamma_i)+\gamma_c\right]}\} > 0$$

and

$$\frac{dF}{dS} \underset{S\to 0}{\lim} = 0,$$
$$\frac{dF_{cond}}{dS} \underset{S\to 0}{\lim} = 0.$$

Both slopes start from a value of zero for very low S values. They also end up with the slope of the option value conditional on control activation always greater (given positive jump frequencies and positive expected impact of the control) than the slope of the option value without control activation. This implies that for very high values of S control activation is always dominant, and that for very low values of S doing nothing (waiting) is dominant given that control activation always involves a cost. Figure 1 illustrates the case where the optimal regions are just $\{W, C\}$. The threshold value S^* that separates the two regions can be obtained by equating the option value of waiting with the option value with costly control activation (and solving numerically the highly non-linear equation):

$$F_{cond}[S^*, X, \sigma, \delta^*, \lambda_i, \gamma_i, \sigma_i, T, r, \gamma_c, \sigma_c, t(c)=0] - X_C \\ = F(S^*, X, \sigma, \delta^*, \lambda_i, \gamma_i, \sigma_i, T, r).$$

Figure 1 confirms that for low values of S the optimal decision is W, but for higher S values the optimal decision is C, with the threshold level being at $S^* = 121.742$. The lower panel of the figure confirms that beyond $S = 0$ the slope of the payoff conditional on control activation is always higher than the slope of the payoff in the wait mode. Thus, when the optimal decision switches from W to C it stays there for any value of $S \geq S^* = 121.742$.

Figure 2 shows a more general case where the decision regions are $\{W, C, W, C\}$. We first observe decision W for low S values, decision C for somewhat higher S values, then decision W again, and (outside the plotted area) decision C again. In the lower panel we cannot see the slope for very high S values (it is outside the plotted area) but we know that eventually C will dominate

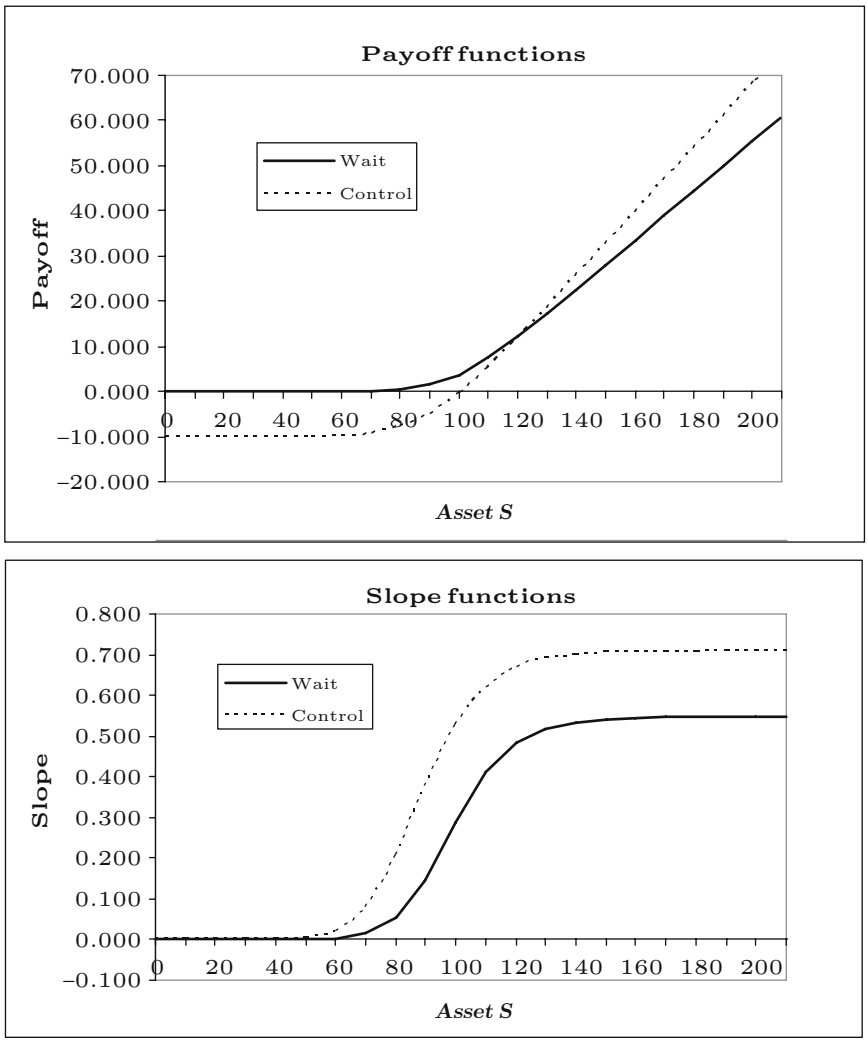

Fig. 1. Payoff and slope functions with a single decision threshold

Notes: Basic parameters are $r = \delta = 0.1$, $\sigma = 0.1$, $X = 100$, and $T = 1$. For the non-catastrophic jumps $\lambda_1 = \lambda_2 = 0.5$, $\gamma_1 = 0.1, \gamma_2 = -0.1$, $\sigma_1 = \sigma_2 = 0.1$; for catastrophic jumps-to-ruin, $\lambda_3 = \lambda_4 = 0.25$. For the control, $\gamma_C = 0.1$, $\sigma_C = 0.1$, $\lambda_C = 0.15$, and cost $X_C = 10$. The upper panel shows the payoff functions that determine the optimal decisions, with decision W (wait) dominating for low S values, and decision C (control activation) for higher S values. The lower panel shows the partial derivative of the payoff function with respect to S. The switching threshold is (numerically) estimated $S_{W \to C} = 121.742$. The slope confirms that C dominates W for high values of S

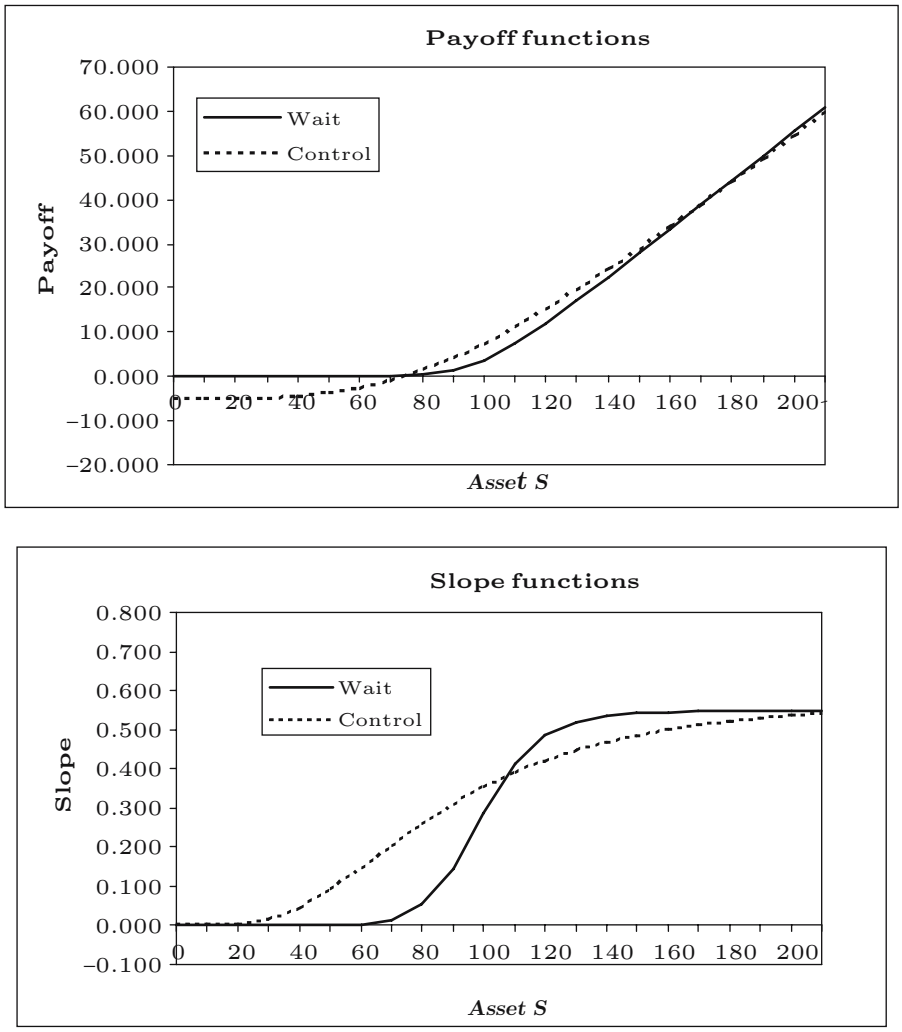

Fig. 2. Payoff and slope functions with multiple decision thresholds
Notes: Basic parameters are: $r = \delta = 0.1$, $\sigma = 0.1$, $X = 100$, and $T = 1$. For non-catastrophic jumps we have $\lambda_1 = \lambda_2 = 0.5$, $\gamma_1 = 0.1, \gamma_2 = -0.1$, $\sigma_1 = \sigma_2 = 0.1$; for catastrophic jumps-to-ruin, $\lambda_3 = \lambda_4 = 0.25$. For the control, $\gamma_C = 0.02$, $\sigma_C = 0.5$, $\lambda_C = 0$, and cost $X_C = 5$. The upper panel shows the payoff functions that determine the optimal decisions, with optimal decisions W (wait) prevailing for low S values, C (control activation) for somewhat higher S values, then W prevails again, and (outside the plotted area) C dominates again. The lower panel shows the partial derivative of the payoff function with respect to S. We know from theory that the slope (outside the plotted area) is such that C will dominate W again for very high values of S. The regions $\{W, C, W, C\}$ are separated at the (numerically estimated) thresholds $S_{W\to C} = 76.384$, $S_{C\to W} = 163.362$, and $S_{W\to C} = 445.384$

Table 2. The optimal decision thresholds

Panel A. A single decision threshold

	Threshold S^*		
	$\lambda_C = 0$	$\lambda_C = 0.15$	$\lambda_C = 0.3$
Base-case	172.496	121.742	108.815
Base-case, Cost -5	103.881	96.713	92.594
Base-case, $\sigma_C + 0.2$	165.789	103.362	94.451
Base-case, $\gamma_C + 0.2$	86.752	83.749	81.197

Note: Input is the same as in Fig. 1.

Panel B. Multiple decision thresholds

	Threshold S^*		
	$\lambda_C = 0$	$\lambda_C = 0.15$	$\lambda_C = 0.3$
Base-case	445.384	–	–
	163.362	–	–
	76.384	72.666	69.431
Base-case, Cost +2	627.859	–	–
	132.185	–	–
	86.892	81.213	76.908
Base-case, σ_C -0.1	448.427	–	–
	133.809	–	–
	85.738	81.007	77.348
Base-case, γ_C -0.01	901.932	–	–
	150.108	–	–
	77.323	73.486	70.179

Note: Input is the same as in Fig. 2.

W again. The regions $\{W, C, W, C\}$ are separated by thresholds $S_{W\to C} = 76.384$, $S_{C\to W} = 163.362$, and $S_{W\to C} = 445.384$. Table 2 provides sensitivity analysis on the optimal thresholds for these two cases, with panel A input parameters corresponding to Fig. 1, and panel B input parameters those of Fig. 2. Thus, in panel A only a single threshold appears, whereas in panel B all three thresholds may appear. Increasing the attractiveness of the control shifts the first threshold to lower S values.

Attractiveness of a control increases when its cost is lower, its mean impact is higher, its impact on (decrease of) the ruin probabilities is higher, or when its volatility is higher. Similarly in panel B we see that increasing the attractiveness of the control diminishes the second occurrence of the W region. For higher λ_C values, this region can be eliminated altogether. It would be similarly eliminated for any reason that increases the attractiveness of the control.

4 A Numerical Markov-Chain Solution Method for Valuing Claims with Controls and Multiple Sources of Jumps

For the valuation of claims with multiple types of rare events and controls in a general context, we follow a numerical approach similar to Martzoukos (2000) and Martzoukos and Trigeorgis (2002) – drawing on convergence properties of Markov-chains studied in Kushner (1977), Kushner and DiMasi (1978), and Kushner (1990). As in Amin (1993), we implement a rectangular finite-difference scheme that augments the lattice approach of Cox, Ross and Rubinstein (1979) as suggested by Jarrow and Rudd (1983). Valuation proceeds in a backward, dynamic-programming fashion. The contingent claim F is valued starting at maturity T; in the absence of rare events, (risk-neutral) valuation continues backward in the lattice until time zero, at each step using the (risk-neutral) probabilities of up or down moves of asset S and discounting expected values accordingly. The lattice expands (from time 0) in a tree-like fashion (usually binomial or trinomial). The scheme we implement is consistent with a binomial path for the underlying asset (in the absence of jumps), but is built using a rectangular "finite-difference" grid. This rectangular scheme allows implementation of the Markov-chain solution methodology because (in the joint presence of the geometric Brownian motion, the rare events, and the controls) the distribution of the value of the contingent claim F is highly skewed and cannot be approximated well with the next two (or three) points alone.

The discretization scheme is spaced in the asset dimension, $\sigma\sqrt{(\Delta \mathrm{t})}$ values apart around the logarithm of the expected asset value relative to the time-0 asset value, and it retains the logarithmic risk-neutral drift, $\alpha_{\Delta t} = [r - \delta^* - .5\sigma^2]\Delta t$. In the absence of jumps, the asset value can move up or down with equal probabilities ($p_u = p_d = 0.50$). In the presence of jumps (with random arrival times following a Poisson distribution), the value of contingent claim F depends on all possible subsequent values (with a reasonable truncation for practical purposes – see Martzoukos and Trigeorgis, 2002, for more on the implementation details of a Markov-chain finite-difference scheme).

To better understand the Markov-chain approximation scheme, we need first to understand the impact of a) the rare events, b) the Brownian motion, and c) the controls. In each time interval, the following mutually-exclusive rare events can occur (assuming that only one rare event can occur at a time):

no jump of any type with probability $P(n_i = 0 \text{ for all } i)$,
one jump of type $i = 1$ only, with probability $P(n_{i=1} = 1,\ n_{i\neq 1} = 0)$,
...
one jump of type $i = N$ only, with probability $P(n_{i=N} = 1,\ n_{i\neq N} = 0)$,
and a control (together with any of the above).

The above directly accounts for non-catastrophic jumps only, since the jumps-to-ruin are accounted for indirectly via their impact on the dividend yield and the riskless rate.

Assuming independence of these rare events, their *joint* probabilities are given from the N-term products:

$$P(n_i = 0 for all i) = \prod_{i=1}^{N} (e^{-\lambda_i T}) = e^{-T \sum_{i=1}^{N} (\lambda_i)}, \tag{7}$$

$$P(n_{i=1} = 1, n_{i \neq 1} = 0) = e^{-\lambda_1 T} \lambda_1 T \prod_{i, i \neq 1}^{N} (e^{-\lambda_i T}) = \lambda_1 T e^{-T \sum_{i=1}^{N} (\lambda_i)},$$

$$\ldots$$

$$P(n_{i=N} = 1, n_{i \neq N} = 0) = e^{-\lambda_N T} \lambda_N T \prod_{i, i \neq N}^{N-1} (e^{-\lambda_i T}) = \lambda_N T e^{-T \sum_{i=1}^{N} (\lambda_i)}.$$

In the absence of any jumps or controls, the option value at time t and state j, $F(t, j)$, is determined from the up and down values one time-step later, $F(t + \Delta t,\ j + 1)$ and $F(t + \Delta t,\ j - 1)$, using the up and down probabilities. In the presence of jumps or controls, $F(t,\ j)$ needs to be calculated from the option values for all possible states one time-step later, using their risk-neutral (Markov-chain) transition probabilities (within the finite-difference approximation scheme). We retain the assumption that the rare event is observed inside the interval Δt, and that only one rare event (of any type) can be observed within this time interval. In general, the probability $P\{.\}$ of a certain outcome (movement of the asset value S over the next period by l steps within the finite-difference grid) is approximated by

$$\begin{aligned} & P\{\ln[S(t + \Delta t)] - \ln[S(t)] = \alpha_{\Delta t} + l\sigma\sqrt{(\Delta t)}\} \\ = \ & N[(l + .5)\sigma\sqrt{(\Delta t)}] - N[(l - .5)\sigma\sqrt{(\Delta t)}], \end{aligned}$$

where $N[.]$ is the cumulative normal distribution of the logarithm of the asset value (given an occurrence of a rare event, the Brownian motion up or down move, and a control). Of course, the underlying asset S can move (by one step at a time only) up or down even in the absence of any jumps, following a geometric Brownian motion with probabilities p_u and p_d as defined earlier. The risk-neutral *transition probabilities* associated with the various jump types as well as the Brownian motion movement (in most general case with $l \neq \pm 1$) in the absence of control activation are given by:

$$\begin{aligned} & P\{\ln[S(t + \Delta t)] - \ln[S(t)] = \alpha_{\Delta t} + l\sigma\sqrt{(\Delta t)} \mid l \neq \pm 1\} = \sum_{i=1}^{N} P(n_{k=i} = 1, \\ & n_{k \neq i} = 0)\{p_u N_i[(l - 1 + .5)\sigma\sqrt{(\Delta t)}] - p_u N_i[(l - 1 - .5)\sigma\sqrt{(\Delta t)}] \\ & + p_d N_i[(l + 1 + .5)\sigma\sqrt{(\Delta t)}] - p_d N_i[(l + 1 - .5)\sigma\sqrt{(\Delta t)}]\}. \end{aligned} \tag{8}$$

Here N_i denotes the probability associated with arrival of a jump of type i. For the special case of only one up move ($l = +1$), the Markov-chain probabilities are:

$$\begin{aligned}
&P\{\ln[S(t+\Delta t)] - \ln[S(t)] = \alpha_{\Delta t} + l\sigma\sqrt{(\Delta t)} \mid l = +1\} \\
&= p_u P(n_i = 0 for all i) + \sum_{i=1}^{N} P(n_{k=i} = 1, n_{k\neq i} = 0) \\
&\times\{p_u N_i[(l-1+.5)\sigma\sqrt{(\Delta t)}] - p_u N_i[(l-1-.5)\sigma\sqrt{(\Delta t)}] \\
&+p_d N_i[(l+1+.5)\sigma\sqrt{(\Delta t)}] - p_d N_i[(l+1-.5)\sigma\sqrt{(\Delta t)}]\}, \qquad (8a)
\end{aligned}$$

where the first term is due to the Brownian motion up-movement alone. Similarly, for only one down move ($l = -1$):

$$\begin{aligned}
&P\{\ln[S(t+\Delta t)] - \ln[S(t)] = \alpha_{\Delta t} + l\sigma\sqrt{(\Delta t)} \mid l = -1\} \\
&= p_d P(n_i = 0 for all i) + \sum_{i=1}^{N} P(n_{k=i} = 1, n_{k\neq i} = 0) \\
&\times \{p_u N_i[(l-1+.5)\sigma\sqrt{(\Delta t)}] - p_u N_i[(l-1-.5)\sigma\sqrt{(\Delta t)}] \\
&+p_d N_i[(l+1+.5)\sigma\sqrt{(\Delta t)}] - p_d N_i[(l+1-.5)\sigma\sqrt{(\Delta t)}]\}. \qquad (8b)
\end{aligned}$$

In the case of control activation, (8), (8a), (8b) are replaced by the following (9). The *transition probabilities* P_c associated with control activation, various jump types, and the Brownian motion movement are generally given by:

$$\begin{aligned}
&P_c\{\ln[S(t+\Delta t)] - \ln[S(t)] = \alpha_{\Delta t} + l\sigma\sqrt{(\Delta t)}\} \\
&= P(n_i = 0 \; for \; all \; i)\{p_u N_c[(l-1+.5)\sigma\sqrt{(\Delta t)}] - p_u N_c[(l-1-.5)\sigma\sqrt{(\Delta t)}] \\
&+ p_d N_c[(l+1+.5)\sigma\sqrt{(\Delta t)}] - p_d N_c[(l+1-.5)\sigma\sqrt{(\Delta t)}]\} \\
&+ \sum_{i=1}^{N} P(n_{k=i} = 1, n_{k\neq i} = 0)\{p_u N_{i,c}[(l-1+.5)\sigma\sqrt{(\Delta t)}] \\
&- p_u N_{i,c}[(l-1-.5)\sigma\sqrt{(\Delta t)}] + p_d N_{i,c}[(l+1+.5)\sigma\sqrt{(\Delta t)}] \\
&- p_d N_{i,c}[(l+1-.5)\sigma\sqrt{(\Delta t)}]\}. \qquad (9)
\end{aligned}$$

Here $N_{i,c}$ denotes the joint probability associated with control activation and the simultaneous arrival of a jump of type i, whereas N_c denotes the probability due to control activation alone.

At each point on the rectangular grid, and in the absence of control, the value F of a European-type claim at time t and state j is obtained from:

$$\begin{aligned}
F(t,j) = e^{-(r+\lambda_{N+1}+\lambda_{N+2})\Delta T} &\sum_{l=-m}^{l=m} P\{\ln[S(t+\Delta t)] - \ln[S(t)] \\
&= \alpha_{\Delta t} + l\sigma\sqrt{(\Delta t)}\} F(t+1, j+l), \qquad (10)
\end{aligned}$$

with the summation limits m defining a suitable truncation so that the probabilities $P\{.\}$ add to unity, leading to a suitable solution in a $(2m+1)$-nomial approximation framework. Conditional on control activation, the contingent claim value $F_{cond}(t,\, j)$ is similarly obtained from

$$F_{cond}(t,j) = e^{-(r+\lambda_{N+1}+\lambda_{N+2})\Delta T} \sum_{l=-m}^{l=m} P_c\{\ln[S(t+\Delta t)] - \ln[S(t)] = \alpha_{\Delta t} + l\sigma\sqrt{(\Delta t)}\}F(t+1, j+l). \tag{11}$$

Optimal control activation involves determining F^* as the maximum of $F(t,\, j)$ and $F_{cond}(t,\, j)$ (taking into account all possible mutually-exclusive control actions). The closed-form analytic results obtained through (4–4a) and (6–6a) and provided in parenthesis in Table 1 provide a benchmark for testing the numerical accuracy of the above numerical approximation scheme. In general, our numerical scheme with 650 refinements in the asset dimension, 80 steps in the time dimension, and a 125-nomial Markov-chain approximation provides results that differ by no more than $\pm 0.15\%$. Our numerical method can readily accommodate the early exercise feature of American-type claims, as well as more complex sequential/compound options often encountered in real option applications (see Trigeorgis, 1993).

5 Conclusions

In this paper we study real (investment) options in the presence of managerial controls, and exogenous rare and catastrophic events. Managerial controls are multiplicative of the impulse-type with random outcome, they are costly, and they must be optimally activated by the firm. We assume a lognormal distribution for the effect of the control. We also incorporate rare events that arise from a multi-class Poisson process. The impact of these rare events is also multiplicative and lognormal. Two of the rare events classes are assumed to be of catastrophic nature, one affecting the underlying asset, and one affecting the contingent claim.

By studying two different types of catastrophic events we have depicted that results are the same in the case of a standard call option, but results differ significantly in the case of the put option. The assumption of lognormality for the effect of the controls allows us to use an analytic framework with a solution isomorphic to the Black and Scholes model, when a single control is optimally activated at time zero. The similar lognormality assumption for the effect of the randomly arriving rare events permits an analytic solution with both the controls and the randomly arriving jumps. We have studied the case where the control not only affects the underlying asset (by enhancing its value), but also pre-emptively affects the intensity of the catastrophic event (it reduces the intensity). We have demonstrated the optimal control activation thresholds. Increasing the attractiveness of the control widens the region

where it is optimal to activate such a control. We finally provide a numerical Markov-Chain approach for the case of sequential controls in the presence of a multi-class jump-diffusion. This framework is demonstrated for the case of lognormal effects, but it can easily be adjusted to handle other plausible distributions.

Acknowledgements

The authors are grateful for partial financial support to the HERMES European Center of Excellence on Computational Finance and Economics at the University of Cyprus. N. Koussis would like to acknowledge financial support by the Cyprus Research Promotion Foundation (ΠΕΝΕΚ ΕΝΙΣΧ/0504).

References

Amin, K. I. (1993). Jump diffusion option valuation in discrete time. *Journal of Finance*, XLVIII, 1833–1863.

Andersen, L., & Andreasen, J. (2001). Jump-diffusion processes: Volatility smile fitting and numerical methods for option pricing. *Review of Derivatives Research*, 4, 231–262.

Ball, C.A., & Torous W. N. (1985). On jumps in common stock prices and their impact on call option pricing. *Journal of Finance*, XL, 155–173.

Bardham, I., & Chao, X. (1996). On Martingale measures when asset prices have unpredictable jumps. *Stochastic Processes and their Applications*, 63, 35–54.

Bates, S. D. (1991). The crash of '87: Was it expected? The evidence from options markets. *Journal of Finance*, XLVI, 1009–1044.

Bergman, Y. Z., Grundy B. D., & Wiener Z. (1996). General properties of option prices. *Journal of Finance*, 51, 1573–1610.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81, 637–659.

Brennan, M. J. (1991). The price of convenience and the valuation of commodity contingent claims. In D. Lund, & B. Øksendal, (Eds.), *Stochastic Models and Option Values*. (pp. 33–72). Amsterdam, Netherlands: North-Holland.

Brynjolffson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives*, 14, 23–48.

Bunch, D. S., & Smiley, R. (1992). Who deters entry? Evidence on the use of strategic entry deterrents. *Review of Economics and Statistics*, 74, 509–521.

Chan, T. (1999). Pricing contingent claims on stocks driven by Levy processes. *Annals of Applied Probability*, 9, 504–528.

Constantinides, G. (1978). Market risk adjustment in project valuation. *Journal of Finance*, 33, 603–616.

Cox, J., Ross, S. A., & Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics*, 7, 229–263.

Dixit, A. K., & Pindyck, R. S. (1994). *Investment Under Uncertainty*. Princeton, New Jersey: Princeton University Press.

Henderson, V., & Hobson, D. (2003). Coupling and option price comparisons under a jump-diffusion model. *Stochastics and Stochastic Reports*, 75, 79–101.

Jarrow, R., & Rudd, A. (1983). *Option Pricing*. Homewood, Ill.: Richard D. Irwin, Inc.

Jones, E. P. (1984). Option arbitrage and strategy with large price changes. *Journal of Financial Economics*, 13, 91–113.

Kou, S. G. (2002). A jump diffusion model for option pricing. *Management Science*, 48, 1086–1101.

Kou, S. G., & Wang, H. (2004). Option pricing under a double exponential jump diffusion model. *Management Science*, 50, 1178–1192.

Kushner, H. J. (1977). *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*. New York, New York: Academic Press.

Kushner, H. J. (1990). *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*. Cambridge, MA: Birkhäuser Boston.

Kushner, H. J., & DiMasi, G. (1978). Approximations for functionals and optimal control on jump diffusion processes. *Journal of Mathematical Analysis and Applications*, 40, 772–800.

Martzoukos, S. H. (2000). Real options with random controls and the value of learning. *Annals of Operations Research*, 99, 305–323.

Martzoukos, S. H. (2003). Multivariate contingent claims on foreign assets following jump-diffusion processes. *Review of Derivatives Research*, 6, 27–46.

Martzoukos, S. H., & Trigeorgis, L. (2002). Real (investment) options with multiple types of rare events. *European Journal of Operational Research*, 136, 696–706.

McDonald, R., & Siegel, D. (1984). Option pricing when the underlying asset earns a below-equilibrium rate of return: A note. *Journal of Finance*, 39, 261–265.

McDonald, R., & Siegel, D. (1986). The value of waiting to invest. *Quarterly Journal of Economics*, 101, 707–727.

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica*, 41, 867–887.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3, 125–144.

Stoll, H. R., & Whaley, R. E. (1993). *Futures and Options: Theory and Applications*. Cincinnati, Ohio: South-Western Publishing Co.

Trigeorgis, L. (1993). The nature of option interactions and the valuation of investments with multiple real options. *Journal of Financial and Quantitative Analysis*, 28, 1–20.

Trigeorgis, L. (1996). *Real Options: Managerial Flexibility and Strategy in Resource Allocation*. Cambridge, Massachusetts: The MIT Press.

Vollert, A. (2003). *A Stochastic Control Framework for Real Options in Strategic Valuation*. Boston: Birkhäuser.

Index