Aalborg Universitet

Some statistical models for high-dimensional data

Gorst-Rasmussen, Anders

Publication date:
2011

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Gorst-Rasmussen, A. (2011). *Some statistical models for high-dimensional data*. Department of Mathematical Sciences, Aalborg University. Ph.D. Report Series No. 19

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

? Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
? You may not further distribute the material or use it for any profit-making activity or commercial gain
? You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: November 30, 2020

Preface

This thesis summarises research work carried out during my employment as a PhD student at Department of Mathematical Sciences, Aalborg University, Denmark. Part of the work was carried out while based at Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital, as an affiliate of the Nordic Centre of Excellence 'SYSDIET' funded by NordForsk.

The thesis is about statistical modelling of high-dimensional data. Behind this broad description lies an equally broad collection of seven research papers and manuscripts working in two diverse areas of application: telecommunications and medicine. The first three papers concern problems from telecommunications and were written partly under the supervision of associate professor Martin Bøgsted Hansen (previously Aalborg University) during the period 2006-2008; although some of the papers were not finished completely until much later. In 2009, practical circumstances lead to a substantial revision of my PhD study plan. The last four papers contain research carried out in the period 2009-2011 under the supervision of professor Thomas H. Scheike (University of Copenhagen) and concern high-dimensional survival regression models with medical and biotechnological applications. Each paper is self-contained, with separate section/equation numbering, separate notation, and a separate reference list.

I am grateful to my most recent supervisor Thomas Scheike for taking on the role as a long-distance principal supervisor but also to my first supervisor Martin Bøgsted Hansen who originally inspired me to work with high-dimensional problems. I am indebted to professor Kim Overvad, Aarhus University, for his practical support and willingness to share his huge knowledge about epidemiological issues. Thanks to the statisticians and others at Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital, for their support and interest. A special thanks goes to my mentor and good friend Søren Lundbye-Christensen for our many discussions on professional and non-professional matters, often taking place in or near the freezing winter waters of Limfjorden.

Thanks to Centre for Ultra-Broadband Information Networks (CUBIN), University of Melbourne, Australia, and professor Darryl Veitch for hosting me for a period in 2007-2008, during which the manuscript 'Why FARIMA Models are Brittle' was drafted, and thanks to the scholarship 'Rejselegat for Matematikere' for making this stay possible. Thanks to Department of Nutrition, Harvard School of Public Health, Harvard University, USA, and associate professor Eric B. Rimm for hosting me while drafting the paper 'Exploring Dietary Patterns by Using the Treelet Transform'.

Finally, a big thanks to my family for their support. Most of all, thanks to Dorte for her patience and encouragement, especially during the last hectic months of writing.

Aalborg, Denmark, August 2011 Anders Gorst-Rasmussen

Revised February 2012

Summary

The dimension of a mathematical entity can be loosely defined as the 'number of numbers' needed for its description. With this definition in mind, high-dimensional data is essentially just data where each observation consists of a large number of numbers. Examples could be regular measurements of an Internet data stream; or the genetic information of a human. Increasingly larger amounts of high-dimensional data are collected in medicine and technology, and the development of descriptive and inferential methods for such data is the biggest current challenge for research in statistics and probability. The research work in this thesis contributes to meeting this challenge by investigating a range of different applied statistical and probabilistic problems from telecommunications, medicine, and biotechnology.

Classically, high-dimensional data is often taken to mean data describable via a suitable stochastic process. This notion of high dimensionality is embraced in the initial three papers of the thesis, which deal with problems derived from telecommunications. Stochastic processes are convenient models for high-dimensional phenomena because of their often rich intrinsic structure. Strong use of such intrinsic structure is made in the first and third paper which rely on classical asymptotic statistical theory to investigate the sampling properties of estimators of functional parameters of an underlying stochastic process. Specifically, the first paper deals with regenerative sequences appearing in queueing theoretical models whereas the third paper concerns a certain time series model used for modelling communication systems. These two works have applications in the statistical analysis of tele-queues and in performance analysis for wireless communications, respectively. The second paper is a critical view on the routine use of a particular statistical model: fractional time series, often used as prototypical examples of long memory time series in, for example, simulation studies are shown to exhibit a rather atypical form of long memory.

In recent years, high-dimensional data has come to refer to standard regression data with the additional complication that we seek to estimate a large number of parameters compared to the number of observations. In modern genetics, for example, millions of measurements may be made on each of only a few hundred individuals. A successful approach to dealing statistically with such difficult data is to use standard 'unstructured' statistical models and impose structure at the estimation rather than at the modelling stage. This is known as regularised estimation and is one of the most active current research areas in statistics. It is also the subject of the last four papers of the thesis which contribute to both theoretical, computational, and practical aspects of regularised regression for survival data in medicine and biotechnology. In the fourth and fifth paper, we introduce a recent regularisation method, the treelet transform, to an epidemiological audience in the context of dietary pattern analysis and show how it may substantially improve over existing methods. The last two papers promote the so-called semiparametric additive hazards model for analysing survival regression data with high-dimensional explanatory variables. This flexible model is particularly well suited for regularisation purposes because of its simple analytic form and excellent computational properties. We develop in the sixth paper highly efficient coordinate descent algorithms and software for fitting the lasso regularised additive hazard model. In the seventh and final paper of the thesis, we present a method for univariate screening for survival data with high-dimensional explanatory variables, loosely based on the additive hazards model. We provide a study of the consistency properties of the method in the asymptotic regime of ultra-high dimension.

Dansk Resumé (Summary in Danish)

Dimensionen af en matematisk størrelse kan løst defineres som antallet af tal, der kræves for at beskrive størrelsen. Med afsæt i denne definition er højdimensionelle data grundlæggende blot data, hvor hver enkelt observation består af et stort antal tal. Eksempler kunne være regelmæssige målinger af en datastrøm på internettet; eller genetisk information for et menneske. Stadig større mængder højdimensionelle data indsamles i medicin og teknologi, og udvikling af deskriptions- og inferensmetoder for sådanne data er den største aktuelle udfordring for statistisk og sandsynlighedsteoretisk forskning. Forskningsarbejdet i denne afhandling bidrager til at imødegå denne udfordring ved at behandle en række anvendte statistiske og sandsynlighedsteoretiske problemstillinger fra telekommunikation samt medicin og bioteknologi.

I klassisk regi forstås ved højdimensionelle data i reglen data, som kan beskrives ved hjælp af en passende stokastisk proces. Denne beskrivelse er baggrunden for de første tre artikler i afhandlingen, som vedrører problemstillinger fra telekommunikation. Stokastiske processer er nyttige modeller for højdimensionelle fænomener som følge af deres typisk righoldige indre struktur. En sådan righoldig indre struktur spiller en central rolle i den første samt den tredje artikel, som begge benytter sig af klassisk asymptotisk statistisk teori til at undersøge fordelingsopførslen for estimatorer af funktionelle parametre i en underliggende stokastisk proces. Konkret vedrører første artikel regenerative følger fra modeller for køsystemer, mens tredje artikel behandler en bestemt tidsrækkemodel, som anvendes til at modellere kommunikationssystemer. Disse to arbejder har anvendelser inden for henholdsvis statistisk analyse af tele-køer og performanceanalyse i trådløs kommunikation. Den anden artikel er en kritisk kommentar til rutinebrugen af en konkret statistisk model: fraktionelle tidsrækker, der ofte anvendes som standardeksempler på tidsrækker med lang hukommelse i eksempelvis simulationsstudier, vises at besidde en ganske atypisk form for lang hukommelse.

I de senere år har betegnelsen højdimensionelle data typisk fundet anvendelse om standard regressionsdata med den komplikation, at man søger at estimere et stort antal parametre i forhold til antallet af observationer. I eksempelvis moderne genetik er det ikke ualmindeligt at foretage millioner af målinger på hvert enkelt af nogle få hundrede individer. En givtig statistisk tilgang til sådanne vanskelige data består i at inkorporere struktur i estimations- snarere end i modelleringsstadiet. Dette er kendt som regulariseret estimation, og er et af de for tiden mest aktive forskningsområder i statistik. Det er også emnet i de sidste fire artikler i afhandlingen, som bidrager til både teoretiske, beregningsmæssige og praktiske aspekter af regulariseret regression for overlevelsesdata i medicin og bioteknologi. I fjerde og femte artikel introduceres en nyligt opfundet regulariseringsmetode, treelet-transformen, til et epidemiologisk publikum i forbindelse med kostmønsteranalyser, og det demonstreres, hvordan metoden markant kan forbedre eksisterende analysemetoder. De sidste to artikler promoverer den såkaldte semiparametriske additive hazardmodel som et værktøj til analyse af overlevelsesdata med højdimensionelle forklarende variable. Denne fleksible model er særligt velegnet i regulariseringsøjemed på grund af dens simple analytiske form og fremragende beregningsmæssige egenskaber. I sjette artikel udvikles særdeles efficiente beregningsmetoder og software til at estimere i den lasso-regulariserede additive hazardmodel. I den syvende og sidste artikel i afhandlingen præsenteres en metode til at foretage univariat screening for overlevelsesdata med højdimensionelle forklarende variable, som er løst baseret på den additive hazardmodel. Der beskrives et studie af metodens konsistensegenskaber i en grænsesituation med ultrahøj dimension.

Contents

Preface i

Summary iii

Dansk Resumé (Summary in Danish) v

Introduction 1

1. High dimensionality in telecommunications 1
2. High dimensionality in medicine and biotechnology 5
3. Some past and future research directions 8

References. 14

1 Applications in Telecommunications 15

Paper I. Asymptotic Inference for Waiting Times and Patiences in Queues with Abandonment 17

1. Introduction 17
2. Asymptotic inference for regenerative sequences 19
3. Asymptotic inference for waiting times and patiences 21
4. Practical considerations and simulation examples 24
5. Application to real data 26

Appendix: validity of the RBB for empirical processes 28
References. 31

Paper II. Why FARIMA Models are Brittle 33

1. Introduction 33
2. Background 34
3. Fractionally differenced processes are not typical LRD processes 38
4. Closeness of the ACVF 42
5. Fractional processes are brittle 44
6. Discussion 46

Appendix: proofs 48
References. 54

Paper III. Some Statistical Properties of an Ultra-Wideband Communication Channel Model 55

1. Introduction 55
2. Model and problem statement 57
3. CLT for the infinite-length equaliser 59
4. CLT for the finite-length equaliser 64
5. Concluding remarks 66

Appendix: CLT for functionals of periodic Gaussian vector processes 67
References. 74

2 Applications in Medicine and Biotechnology 75

Paper IV. Exploring Dietary Patterns by Using the Treelet Transform 77

1. Introduction 78
2. Materials and methods 78
3. Results 82
4. Discussion 87

References. 91
Appendix 1: An invited commentary and our response 92
Appendix 2: Supplementary tables 94

Paper V. tt: Treelet Transform with Stata 99

1. Introduction 99
2. The treelet transform algorithm 100
3. The tt add-on. 103
4. A data example 105
5. Concluding remarks 111

References. 111

Paper VI. Coordinate Descent Methods for the Penalized Semiparametric Additive Hazards Model 113

1. Introduction 113
2. The semiparametric additive hazards model 114
3. Model fitting via cyclic coordinate descent 116
4. Additional details 120
5. Timings and a data example 121
6. Discussion 125

References. 128

Paper VII. Independent Screening for Single-Index Hazard Rate Models with Ultra-High Dimensional Features 129

1. Introduction 129
2. The FAST statistic and its motivation 131
3. Independent screening with the FAST statistic 133
4. Beyond simple independent screening – iterated FAST screening . . . 137
5. Simulation studies. 141
6. Application to AML data 146
7. Discussion 148

Appendix: proofs 150
References. 159

Introduction

The computational and technological advances of the last few decades have brought a veritable data revolution. In science and technology alike, ever more vast amounts of data are generated through measurements, experiments, and computer simulations. Much of this data can be described as 'high-dimensional', consisting of observations of instances of some phenomenon which lives in a high-dimensional space. Today's focus on high-dimensional data presents a range of challenges and opportunities for statistical researchers. In medicine, statistics has classically dealt with situations where a small number of carefully selected variables were measured and scrutinised for each study subject. Nowadays, genetic experiments may measure millions of biomarkers per subject and there is an acute demand for novel statistical methods for converting this data into scientific insight. In engineering sciences, the increasing reliance on computer simulations as a research tool has intensified the need for statistical models to describe and synthesise complex high-dimensional phenomena. Similar examples of the need for novel and tailored modelling strategies for high-dimensional data abound in diverse fields such as finance and economics, environmental sciences, and imaging.

This PhD thesis contributes broadly to the knowledge about statistical and stochastic modelling of high-dimensional data through its pursuit of two independent lines of research. The first line of research finds applications in telecommunications and concerns high-dimensional data that are inherently sequential and are naturally studied via stochastic process techniques. The second line of research is rooted in medicine and biotechnology and works with a more recent notion of high dimensionality in the sense of regression models with many explanatory variables. Seven different scientific papers and manuscripts are included in the thesis. These span widely not only in application areas but also in methodological scope; dealing with both mathematical theory, statistical software, and highly interdisciplinary problems. This wide scope reflects a deliberate attempt to build a contemporary research profile. Modern statistical modelling is increasingly a dynamic part of substantive research. As a research statistician, the ability to shorten the path from mathematical rigour, through software, to dialogue with substantive researchers – and back – is key to rapid and relevant progress in both the substantive research field and in statistics as a mathematical discipline.

Because of the wide application and methodological scope, the individual papers and manuscripts in this thesis necessarily target a number of different statistical audiences. The purpose of this introductory chapter is to provide an outline of each piece of research which is more broadly accessible than the concise abstract accompanying each paper. The final section of the chapter describes some future research problems that are not discussed later in the thesis.

1. High dimensionality in telecommunications

Telecommunications has been swamped by data during recent years thanks to the increasing rate of technological innovation and the focus on information exchange. Being an engineering field at heart, the utility of telecommunications is measured by the extent to which it improves technology. Accordingly, research in telecommunications is naturally product-oriented, often relying on vaguely stated mathematical models, 'proofs by computer simulation', and intuition. These are characteristics of the data analysis approach to science described by Donoho (2000), in which deep and rigorous mathematical analysis is abandoned in favour of a pragmatic, data-centric approach. In the hands of the skilled, such a pragmatic approach can be very effective. However, it falls short in situations where the domain of application becomes sufficiently complex and the data sufficiently high-dimensional. Such situations are becoming more and more common in telecommunications, creating new opportunities for researchers in statistics and probability. An important example of the success of mathematical theory in telecommunications is free probability and random matrix theory. Whilst highly theoretical in its origin, this subfield of probability theory has become an important tool for modelling multi-antenna wireless communications systems and gaining qualitative insight about system behaviour (Tulino and Verdú, 2004). Another success history is the theory of long-range dependent stochastic processes which attracted the attention of researchers in communications during the mid 90s (Leland *et al.*, 1994; Paxson and Floyd, 1995) and is now routinely used for the purpose of modelling and synthesising Internet data streams (Karagiannis *et al.*, 2004). With the ever-widening dependence on information exchange, we can only expect more of these unique opportunities for true interdisciplinary progress in telecommunications, and statistics and probability.

The work in this thesis related to telecommunications is the result of an opportunistic research programme which aimed to identify mathematical-statistical focus areas in telecommunications and conduct research work on a problem-solving basis with an emphasis on mathematical rigour. This has lead to three papers and manuscripts which work with the idea of high dimensionality from rather different perspectives. However, they all share the common trait that a stochastic process model is the fundamental model for the high-dimensional data under investigation.

Paper I. *'Asymptotic Inference for Waiting Times and Patiences in Queues with Abandonment'*

Consider a tele-queue at a call centre. Depending on the rate at which customers arrive and the rate at which they are serviced, the sequence of customer waiting times can be rather complicated and feature strong positive correlations between consecutive waiting times. However, under suitable stationarity assumptions, each time the tele-queue empties completely, the behaviour of past waiting times becomes completely irrelevant for future waiting times. This is a basic example of a regenerative sequence. The aim of Paper I is to develop methods for making statistical inferences about this type of data which are common in applications of queueing theoretical models. Since a regenerative sequence consists of independent and identically distributed (IID) blocks, so-called regenerative cycles, it is not much different from a sequence of IID observations and can in fact be treated as such for the purpose of point estimation (Leventhal, 1988; Tsai, 1998). However, inferences about standard errors, confidence intervals etc. will generally be incorrect.

A flexible tool for performing distributional inferences about IID data is the bootstrap. For example, letting $\mathbb{1}$ denote the indicator function, consider the estimator $\hat{F}(x) = n^{-1}\sum_{i=1}^{n}\mathbb{1}(X_i \leqslant x)$ of the distribution function $\mathbb{P}(X_1 \leqslant x)$ of IID observations $X_1,\ldots,X_n$. If we sample $X_1^*,\ldots,X_n^*$ independently with replacement from realisations $\{X_1,\ldots,X_n\}$ and define

$$\hat{F}^*(x) = n^{-1}\sum_{i=1}^{n}\mathbb{1}(X_i^* \leqslant x),$$

then the distributional behaviour of $\hat{F}^*$ (conditional on data), when viewed as a stochastic process indexed by the set of functions $\mathscr{F} = \{\mathbb{1}(\cdot \leq x) : x \in \mathbb{R}\}$, can be shown to be similar to that of (the unconditional) $\hat{F}$ for large n. This is the empirical process bootstrap for IID data which can be generalised far beyond simple empirical distribution functions by suitably generalising $\mathscr{F}$.

The main contribution of Paper I is to extend the empirical process bootstrap for IID observations to regenerative sequences. The natural way to do this is by resampling not among individual observations, but among the IID regenerative cycles. A detailed proof of the correctness of this regenerative blockwise bootstrap method is provided, and we demonstrate how the method can be used practically to make statistical inferences about the waiting time distribution in a queueing system (in, for example, call centres). Here we consider the added difficulty of customers which may abandon the queue while waiting. This in turn leads to a basic theory of statistical inference about regenerative survival data, including a resampling-based (two-sample) Kolmogorov-Smirnov test for equality of distribution functions derived from regenerative sequences.

Paper II. *'Why FARIMA Models are Brittle'*

Paper II derives some novel structural properties of the so-called fractional autoregressive moving average (FARIMA) time series model. This model is structurally similar to the classical ARIMA time series model but with the additional feature that it exhibits long-range dependence (LRD). Informally, LRD means that the autocorrelation function decays very slowly so that correlation between observations persists over very large time intervals. Such behaviour has been observed empirically in, for example, telecommunications, finance, and hydrology (Beran, 1994). In particular, FARIMA is commonly used for synthesising data in simulation studies related to telecommunications (for example, Taqqu and Teverovsky (1997); Abry *et al.* (2003)).

Consider a second-order stationary time series $\{X_t : t \in \mathbb{N}\}$. LRD can be characterised in terms of the rate of growth of the variance time function:

$$\omega(n) := \mathrm{Var}(X_1 + X_2 + \cdots + X_n), \quad n \in \mathbb{N}.$$

For many classical (short-range dependent) time series, the variance time function scales as if the time series consisted of (finite-variance) IID observations, whereby $n^{-1}\omega(n)$ tends to a constant when $n \to \infty$. In contrast, the variance time function of an LRD process, because of its slowly decaying autocorrelation function, grows at a faster rate. Specifically, a time series is LRD with Hurst parameter $H \in (1/2, 1)$ if and only its variance time function satisfies

$$\omega(n) = cn^{2H} + \omega_d(n), \quad n \to \infty; \tag{1}$$

where $c > 0$ and $\omega_d(n) = o(n^{2H})$ is a remainder term. The function $n \mapsto cn^{2H}$ in (1) can be identified with the variance time function of a particularly simple Gaussian long-range dependent process known as fractional Gaussian noise (fGn). Intuitively, (1) implies that an LRD process asymptotically looks like its corresponding fGn in terms of the second-order structure. The definition of LRD is a strictly asymptotic one and leaves considerable freedom. A process is 'far' from fGn if $\omega_d(n)$ converges slowly; it is 'close' to fGn if $\omega_d(n)$ converges rapidly.

In the manuscript, we use tools from harmonic analysis to show that FARIMA is extremely close to fGn in the sense that, for a constant D, the remainder ω_d satisfies

$$\omega_d(n) = D + o(1), \quad n \to \infty. \tag{2}$$

Considering that $\omega_d(n) = o(n^{2H})$ for general LRD, (2) represents a type of LRD very similar to that of fGn. We moreover argue that (2) implies an undesirable 'brittleness' of FARIMA. For example, by taking a FARIMA time series and perturbing it by adding an independent white noise sequence with variance v, the variance time function becomes $\omega_{\text{perturbed}}(n) = cn^{2H} + D + vn + o(1)$ which is much further from fGn than its non-perturbed counterpart. Hence, the LRD behaviour of FARIMA is not robust to simple white noise.

With data analysis and synthesis in mind, Paper II suggests two key limitations of FARIMA. First, FARIMA adds little beyond the simpler fGn in terms of LRD behaviour. Second, the brittleness of FARIMA indicates that it may not be ideal for modelling for real-world data where (additive) white noise is often unavoidable.

Paper III. *'Some Statistical Properties of an Ultra-Wideband Communication Channel Model'*

When a radio signal encounters an obstacle, it scatters depending on the properties of the obstacle. Accordingly, a receiver in a transmission environment with scattering obstacles typically does not see a single (noisy) copy of the transmitted signal but rather several (noisy) 'echoes', i.e. attenuated and delayed versions of the signal. This is called multipath propagation. A simple transmitter-receiver system with multipath propagation can be described formally as a complex-valued linear time-invariant system

$$Y_n = \sum_{l=0}^{L-1} H_l X_{n-l} + E_n, \quad n \in \mathbb{Z}; \tag{3}$$

where $\{E_n\}$ is white noise, $\{X_n\}$, $\{Y_n\}$ is the (stationary) transmitted/received signal and $H_0, \ldots, H_{L-1}$ are independent, mean-zero random attenuation factors for the signal 'echoes' such that $\mathbb{E}|H_0|^2, \mathbb{E}|H_1|^2, \ldots$ is an appropriately decaying sequence. The use of a discrete-time model reflects the fact that we in practice sample the continuous-time signal to roughly match the intersymbol time of the transmitter. Intuitively, we can think of the intersymbol time as the time between different pulses of information.

Paper III considers the behaviour of a specific instance of (3) in the following setting:

- The intersymbol time is small; accordingly, we sample the signal often.
- The transmission environment contains many scatterers, leading to rich multipath diversity. Thus, a large number of different 'echoes' reach the receiver.

A high sampling rate implies that we can distinguish the many 'echoes' in practice. Consequently, L is large in (3). This in turn leads to a simple statistical model for the physical behaviour in certain transmission environments of a recent technology for short-range wireless communication; so-called ultra-wideband radio. Ultra-wideband radio can be viewed as sophisticated type of Morse code which transmits information in the form of rapid pulses, potentially achieving very high data transmission rates.

Paper III analyses statistics derived from (3) in the above-described ultra-wideband limiting regime of decreasing intersymbol time/increasing L. Specifically, we prove central limit theorems for the so-called MMSE of the optimal linear estimator of X via techniques which work generally for nonlinear functionals of the discrete-time Fourier transform of $[H_0, \ldots, H_{L-1}]$ under complex Gaussianity. This solves a problem stated in Pereira *et al.* (2005) and further explored in Rubak (2007). From a general time series point of view, the problem amounts to investigating, for a specific wide-sense stationary time series, the distributional asymptotics of the statistic $\int_{-\pi}^{\pi} \phi\{I(\omega)\} \mathrm{d}\omega$ with I the sample periodogram and ϕ a nonlinear, smooth function.

2. High dimensionality in medicine and biotechnology

Some of the currently most challenging methodological issues in statistics are driven by biotechnological innovations which have rendered feasible the collection of extremely detailed biomarker data. Current mass market genetic microarrays can register in the order of millions of genetic markers per subject, and large-scale, cost-effective full genome sequencing is rapidly coming within reach (Snyder *et al.*, 2010). Modern 'omics' research areas such as proteomics, metabolomics, lipidomics etc. (Joyce and Palsson, 2006) are likewise becoming important sources of high-dimensional biomedical data.

The single most important methodological problem in this context is how to extract useful information from data when the sample size is much smaller than the number of explanatory variables, causing standard regression methods to fail. Some of the most promising approaches to this problem have come from statistical machine learning. This is a broad descriptive term for the research field on the intersection between statistics and computer science, of which an integral part is the study of the impact of computation on estimation. A severely underdetermined regression model can be regularised at the estimation and computation stage to possess a well-defined regression coefficient estimator. An important example of this is the popular lasso regression technique (Tibshirani, 1997) where an estimator is sought in a cleverly constrained space which not only ensures its well-definition but also incorporates automatic variable selection, in the sense that most coefficients will be estimated to be exactly zero. The idea of incorporating automatic variable selection is commonly referred to as sparse estimation and has evolved into one of the most active research areas in statistics (see Fan and Lv (2010) for a review). Exploring the possibilities and limitations of sparse estimation is not merely an interesting mathematical and computational problem but is also highly relevant in applications where sparsity may substantially simplify data interpretation. There has recently been an interest in utilising the fundamental computational ideas of sparse estimation to incorporate biological information such as biological pathway knowledge in estimation procedures (for example Li and Li (2008); Slawski *et al.* (2010)). Such evidence synthesis, traditionally associated with Bayesian statistics, is indicative of an increasingly opportunistic approach to high-dimensional statistical problems.

Many recent ideas in statistics and machine learning are fundamentally different from classical statistics. Where statisticians previously spoke of p-values and confidence intervals, they now speak of cross-validation and stability selection; and new and much-needed statistical methods can be hardly accessible for the uninitiated medical researcher. Conversely, the complexity and scope of modern medical research makes it equally arduous for the statistician to convert real-world problems into relevant statistical models. There is a great research potential in improving interdisciplinary dialogue. The last four papers of this thesis represent the results so far of a research programme with the long-term aim to exploit this potential. Specifically, they represent a research agenda at the intersection between statistical theory for survival data with high-dimensional explanatory variables, and applied medical and biotechnological research.

Papers IV and V. *'Exploring Dietary Patterns by Using the Treelet Transform' and 'tt: Treelet Transform with Stata'*

Epidemiological research is, to a certain extent, characterised by a conservative approach to statistics, relying on a small set of well-established core statistical models.

Conservatism is useful for the purpose of ensuring objectivity and simplifying comparisons between different studies involving standard epidemiological data. However, it is a limitation when seeking to integrate high-dimensional and complex biomarker data which are becoming increasingly cost-effective to measure even in large observational studies. The research in Papers IV-V represents an initial effort to explore and communicate the potential of statistical machine learning methods in epidemiology.

The subject of the two papers is dimension reduction via a recent statistical machine learning method, the so-called treelet transform due to Lee *et al.* (2008). Principal components analysis is a classical example of a dimension reduction method. More generally, a (linear) dimension reduction method is a rotation of the data coordinate system such that the projection of data onto the first few axes capture the 'important part' of the variation in data. These first few axes (components) are often informally viewed as 'latent variables' and subjected to interpretation; a difficult task since all entries in component vectors are nonzero and often noisy. The treelet transform, on the other hand, is a dimension reduction method which constructs an entire sequence of 'rotated' coordinate systems of interpretable sparse basis vectors/components. Starting with the canonical coordinate system, the next coordinate system in the sequence is obtained by rotating precisely two components so that they *locally* capture as much variation as possible. This leads to a multiscale decomposition akin to wavelet analysis, where the first few coordinate systems contain mostly very sparse 'detail components' and the last few contain mostly less-sparse 'sum components'. A sparse dimension reduction method results by selecting, from a single coordinate system within this sequence, a few components corresponding to large-variance projections of data.

The original motivation was to apply the treelet transform to novel adipose fatty acid measurements in a large Danish cohort study (Tjønneland *et al.*, 2007) and later incorporate genetic data in analyses. However, the dominant epidemiological application area of dimension reduction is in dietary studies which typically use principal components analysis on multivariate data sets of dietary intakes to construct and analyse components known as dietary patterns (Hu, 2002). It was natural to introduce the treelet transform in this well-known context. Paper IV is concerned with conducting and comparing two parallel standard dietary pattern analyses using the new treelet transform method and the established method of principal components analysis, respectively. The key methodological conclusion is that the treelet transform seems to offer results largely comparable to those obtained from a principal components analysis but with a simpler interpretation due to the sparsity of treelet components.

The statistical software program Stata (StataCorp, 2009) is widely used by medical researchers and a comprehensive add-on for Stata was developed as a part of the work on Paper IV in order to encourage experimentation with the treelet transform among epidemiologists. A short and non-technical introduction to the features and usage of this add-on has been published in Paper V.

Paper VI. *'Coordinate Descent Methods for the Penalized Semiparametric Additive Hazards Model'*

A flexible but not widely known regression model for survival data is the semiparametric additive hazards model. This model asserts that the hazard function for the survival time conditionally on some p-dimensional explanatory variable Z is

$$h(t|Z) = h_0(t) + Z^\top \beta^0; \tag{4}$$

with h_0 an unspecified baseline hazard and $Z^\top\beta^0$ a linear regression function. It turns out that the natural estimating equations in this statistical model are of the simple linear form $D\beta = d$ where the $p \times p$ matrix D is symmetric. Equivalently, β^0 can be estimated by minimising the loss function $\beta \mapsto \beta^\top D\beta - 2\beta^\top d$. The estimation problem is notably similar to that of the simple linear regression model. The least-squares form not only makes the additive hazards model computationally suitable for dealing with very high-dimensional data; it also enables many of the machine learning methods developed for the linear regression model to be adapted to a survival setting. Demonstrations of this were given by, for example, Martinussen and Scheike (2009) and Martinussen and Scheike (2010). The principal aim of the methodological research leading to the last two papers of this thesis was to pursue more such adaption opportunities.

Paper VI grew from a need for efficient computational methods for the additive hazards model in connection with this principal research aim. It evolved into a highly optimised algorithm and software for solving the constrained optimisation problem:

$$\hat\beta(\gamma) = \operatorname{argmin}_\beta\big(\beta^\top D\beta - 2\beta^\top d\big), \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq \gamma. \tag{5}$$

The quantity $\hat\beta(\gamma) = [\hat\beta_1(\gamma),\ldots,\hat\beta_p(\gamma)]^\top$ is the lasso penalised estimator in the additive hazards model. By choosing γ small enough, the geometry of the constraint in (5) implies that only a few $\hat\beta_j(\gamma)$ will be nonzero, while the rest will be exactly zero (Tibshirani, 1997). This is the variable selection property of the lasso. The property is key to computing $\hat\beta(\gamma)$ efficiently since, in principle, we need only pay attention to nonzero entries of $\hat\beta(\gamma)$. To convert principles to practice, we use the method of cyclic coordinate descent which calculates $\hat\beta(\gamma)$ by cycling through all coordinate-wise optimisation problems. It turns out that this method can operate using essentially only the part of the matrix D and the vector d that pertains to nonzero $\hat\beta_j(\gamma)$s. This and a few other tricks enable our algorithm to handle extremely large problems (p in the order of hundreds of thousands).

Essential to promoting a less well known statistical method is the availability of high-quality software. The most substantial part of the work related to Paper VI has been the implementation of the developed algorithms in the package **ahaz** (Gorst-Rasmussen, 2011) for the statistical software R (R Development Core Team, 2011). The package was also designed to be useful in future work on machine learning methods in relation to the additive hazards model. For example, it includes general and efficient methods for calculating the statistics D and d.

Paper VII. *'Independent Screening for Single-Index Hazard Rate Models with Ultra-High Dimensional Features'*

A popular approach to dealing with high dimensionality in real-life applications of regression models is to initially ignore the additional information offered by the multidimensional structure of explanatory variables. Instead, all univariate regression models are fitted and a small number of 'relevant' variables are retained for further analysis based on, for example, their p-values in these univariate models. It is obvious that this crude approach to model selection generally leads to a loss of information compared to an approach that respects the multivariate structure of data. The question is when such independent screening leads to a sensible result. Consider the linear regression model $y = Z\beta^0 + \varepsilon$ for some $n \times p$ design matrix Z and an n-vector ε of IID (Gaussian) errors. The normal equations are $Z^\top y = (Z^\top Z)\beta$. Consider the limit when

$n \to \infty$ and p is fixed. Root-n consistency implies that we can consistently infer the nonzero entries in β^0 by truncating the component-wise absolute feature-response correlations $n^{-1}|Z^\top y|$ with some sequence γ_n converging to zero slower than $n^{-1/2}$, provided that the covariance matrix $n^{-1}\mathbb{E}(Z^\top Z)$ is simple (diagonal, say). Fan and Lv (2008) went a step further and showed that this can also be done for a suitable choice of γ_n when p grows exponentially fast with n. This is the sure screening property of independent screening in ultra-high dimension. It is a result of practical interest: it provides (partial) justification for the otherwise *ad hoc* use of independent screening to reduce an extremely high-dimensional feature space to a moderate dimension where more sophisticated modelling techniques can be applied.

Independent screening has been extended to more general regression models (Fan and Song, 2010; Fan *et al.*, 2010) but lacks theoretical justification for the case of survival regression models. Paper VII proposes to perform independent screening for survival data by using the Feature Aberration at Survival Times (FAST) statistic, defined as:

$$d := n^{-1} \int_0^\infty \sum_{i=1}^n \left\{ Z_i - \frac{\sum_{j=1}^n Y_j(t) Z_j}{\sum_{j=1}^n Y_j(t)} \right\} \mathrm{d}N_i(t).$$

Here $N_i(t)$ is the counting process counting the number of events for individual i up to time t and $Y_i(t)$ is the at-risk-indicator which is 1 if individual i is at risk at time t and 0 otherwise. This simple statistic turns out to work more or less like feature-response correlations in a linear regression model. For right-censored survival times, when the censoring mechanism and covariance structure of explanatory variables is sufficiently simple, the FAST statistic leads to a sure screening property in ultra-high dimension when survival times are from a general class of single-index hazard rate models, in which the conditional hazard depends on Z_i only through some linear functional $Z_i^\top \beta^0$.

Despite being essentially model-free, the FAST statistic is closely related to the semiparametric additive hazards model. We exploit this connection to introduce multi-step procedures (iterated independent screening, Fan *et al.* (2009)) which combine FAST screening with penalised regression in order to deal with situations where the covariance assumptions of plain FAST screening fail. We also present simulation studies which indicate that our procedures are very competitive with existing methods in terms of computational speed and empirical model selection properties.

3. Some past and future research directions

We conclude this introductory chapter by briefly outlining some research problems that will *not* be revisited in more detail later in the thesis. These are problems that either have been partially explored, will be explored, or would be interesting to explore in future studies. All relate to the ongoing research programme with medical and biotechnological applications described in Section 2. Since the problems will not be revisited later, the presentation in this section is technical at times, assuming some familiarity with the terminology of the relevant research papers to follow.

3.1. *The treelet transform*

An important property of the treelet transform is that it obtains components only by utilising information about the internal relationships between explanatory variables. This is a limitation when seeking to use components for prediction purposes, in which case it makes more sense to use a supervised method that also utilises information in the response variable and produces components that lead to strong predictors. For example, a popular supervised ‘analogue’ of principal components analysis is partial least squares (Rosipal and Krämer, 2006). It is of interest to develop a similar supervised analogue of the treelet transform. The challenge of incorporating supervision into the treelet transform was brought up by discussants of the paper by Lee *et al.* (2008). For example, Meinshausen and Bühlmann (2008) discussed a semi-supervised approach, combining the treelet transform with the use of a supervised, nonuniform choice of cut-level to maximise the predictive potential of components. Another semi-supervised option is to disregard the tree structure entirely and use variable selection methods such as lasso to select predictors from the union of all coordinate systems over all cut-levels (Bickel and Ritov, 2008).

In a suitably abstract form, the treelet transform consists of two rules: a rule for choosing which two variables to merge, and a rule for choosing an orthogonal linear transformation of two variables for performing the actual merge. It would be desirable if supervision could be incorporated in the treelet transform without destroying this fundamental structure. While it is not difficult to devise supervised variants of each rule separately, it is far from obvious how to devise supervised variants which work well together. Further research will be needed to explore this issue.

3.2. *The semiparametric additive hazards model*

Consider a collection of n independently right-censored survival-times, each represented by a counting process $N_i(t)$ which counts events for subject i up to time t and an at-risk-process $Y_i(t)$. Let Z_i be the associated vector of explanatory variables. Following Lin and Ying (1994), the natural estimating equations for the semiparametric additive hazards model (4) take the form $D\beta = d$ where

$$D := \int_0^\tau \sum_{i=1}^n \{Z_i - \bar{Z}(t)\}\{Z_i - \bar{Z}(t)\}^\top Y_i(t)\mathrm{d}t$$

$$d := \int_0^\tau \sum_{i=1}^n \{Z_i - \bar{Z}(t)\}\mathrm{d}N_i(t);$$

where $[0,\tau]$ denotes the observation time window and $\bar{Z}(t) := \sum_{i=1}^n Z_i Y_i(t) / \sum_{i=1}^n Y_i(t)$ is the at-risk-average of the Z_is.

Bootstrapping the additive hazards lasso. Lasso regression works well for point estimation whereas inference about standard errors, confidence intervals etc. is difficult. For example, the sandwich variance estimator suggested by Tibshirani (1997) and Fan and Li (2001) only works for nonzero coefficients. Developing more flexible tools for distributional inference in the lasso penalised additive hazards model seems worthwhile. One of the first efforts in the research programme described in Section 2 was actually to investigate bootstrap methods for the lasso penalised additive hazards model. As will be explained, these efforts were discontinued because of the apparently limited applicability of the results.

It holds generally that $d - D\beta^0 = \sum_{i=1}^n \int_0^\tau \{Z_i - \bar{Z}(t)\}\mathrm{d}M_i(t)$ where the M_is are martingales (Lin and Ying, 1994). Asymptotically, this defines an IID decomposition of $d - D\beta^0$ which we can estimate by substituting the estimator $\mathrm{d}\hat{M}_i(t;\hat{\beta}) = \mathrm{d}N_i(t) - Y_i(t)\mathrm{d}\hat{\Lambda}_0(t;\hat{\beta}) - Y_i(t)Z_i^\top\hat{\beta}$ for $\hat{\Lambda}_0(t;\beta) = \sum_{i=1}^n \{\mathrm{d}N_i(s) - Y_i(s)Z_i^\top\beta\mathrm{d}s\}/\{\sum_{i=1}^n Y_i(s)\}$

the Breslow estimator of the baseline cumulative hazard. Let $G_1,\dots,G_n$ be IID mean-zero, finite-variance random variables and introduce

$$d^*(\hat\beta^{\mathrm{p}}) := D\hat\beta^{\mathrm{p}} + \int_0^\tau \sum_{i=1}^n G_i\{Z_i - \bar Z(t)\}\mathrm{d}\hat M_i\big(t;\hat\beta^{\mathrm{p}}\big)$$

where $\hat\beta^{\mathrm{p}}$ is a root-n consistent pilot estimator of β^0. Consider the lasso penalised estimator and a (weighted) bootstrapped analogue similar to the residual bootstrap for the linear regression model:

$$\begin{aligned}\hat\beta &:= \mathrm{argmin}_\beta\big\{\beta^\top D\beta - 2\beta^\top d + \lambda_n\|\beta\|_1\big\};\\ \hat\beta^* &:= \mathrm{argmin}_\beta\big\{\beta^\top D\beta - 2\beta^\top d^*(\hat\beta^{\mathrm{p}}) + \lambda_n\|\beta\|_1\big\},\end{aligned}$$

where $\|\cdot\|_1$ denotes the ℓ^1-norm. When $\lambda_n = O(n^{1/2})$ as $n\to\infty$, it can be shown (Martinussen and Scheike, 2009) that $\hat\beta$ is root-n-consistent. In view of the results of Scheike (2001) for the non-penalised problem ($\lambda_n \equiv 0$), we might expect the bootstrapped lasso to be consistent as well in the sense that the conditional law of $n^{1/2}(\hat\beta^* - \hat\beta^{\mathrm{p}})$, with $\hat\beta^{\mathrm{p}} := \hat\beta$, converges to the law of the random variable $n^{1/2}(\hat\beta - \beta^0)$ when $n\to\infty$. This is *not* the case. In fact, it can be shown that the law of $n^{1/2}(\hat\beta^* - \hat\beta)$ converges to a random measure, by arguing as in Chatterjee and Lahiri (2010). This happens essentially because the lasso is not model selection consistent when $\lambda_n = O(n^{1/2})$, in the sense that $\hat\beta^p$ does not asymptotically capture the correct sign of variables j for which $\beta^0_j = 0$ (Zou, 2006). However, by following the approach of Chatterjee and Lahiri (2009), it can be shown that one can enforce consistency of the bootstrapped lasso by making the additional assumption of (weak) selection consistency of $\hat\beta^{\mathrm{p}}$ in the sense that

$$\mathbb{P}\{\mathrm{sign}(\hat\beta^{\mathrm{p}}_j) = \mathrm{sign}(\beta^0_j)\} \to 1, \quad n\to\infty;$$

for $j = 1,\dots,p$. A pilot estimator $\hat\beta^{\mathrm{p}}$ satisfying this can be obtained by using an oracle estimator such as the adaptive lasso estimator (Martinussen and Scheike, 2009). Simpler yet, we can follow Chatterjee and Lahiri (2009) and take

$$\hat\beta^{\mathrm{p}}_j := \hat\beta_j \mathbb{1}(|\hat\beta_j| \geqslant a_n); \tag{6}$$

where $\hat\beta$ is the lasso penalised estimator and the pre-defined sequence a_n converges to zero at some rate slower than $n^{-1/2}$.

The efforts described above indicate that it is theoretically possible to enforce consistency of the bootstrapped lasso. Unfortunately, our limited experiments suggests that its usefulness in practice may be very limited because its finite-sample properties are completely dominated by the choice of pilot estimator. This is also the reason why the work on the this method was discontinued. The fact that the choice of pilot estimator plays a crucial role may not be so surprising. For example, the pilot estimator (6) can be identified with the superefficient Hodges' estimator which is notorious for its erratic finite-sample behaviour. See Leeb and Pötscher (2008) for related reservations about the lasso.

Sparse partial least squares. Partial least squares (PLS) is a class of regularised estimation methods for regression models with strong collinearity or a large number of explanatory variables. Classically, PLS is a type of shrinkage estimator, see Rosipal and Krämer (2006). In the case of the additive hazards model, it was introduced by Martinussen and Scheike (2009) who showed that a natural PLS estimator can be defined explicitly as

$$\hat{\beta}^{\mathrm{PLS}} = R(R^\top DR)^{-1}R^\top d; \tag{7}$$

where the columns of the matrix R are given by the Krylov sequence $d,\dots,D^{K-1}d$.

Chun and Keleş (2010) recently argued that PLS is not well suited for very high-dimensional data and developed a sparse variant of PLS for the standard linear regression model (possibly with multivariate responses). Their algorithm is straightforward to adapt to the additive hazards model. Specifically, given a regularisation parameter λ, the K-component sparse PLS solution is obtained as follows. Set $w := d/|d|$, $A = \emptyset$, and $\hat{\beta}^{\mathrm{PLS}} := 0$. For $k = 1,2,\dots,K$ do:

1. Calculate the sparse 'direction vector' $\tilde{w}$ with entries $\tilde{w}_j := (|w_j| - \lambda)_+\mathrm{sign}(w_j)$.
2. Update the active set as $A = \{j : \tilde{w}_j \neq 0\} \cup \{j : \hat{\beta}_j^{\mathrm{PLS}} \neq 0\}$
3. Calculate the corresponding k-component PLS estimator $\hat{\beta}^{\mathrm{PLS}}$ based on explanatory variables in A only, setting $\hat{\beta}_j^{\mathrm{PLS}} := 0$ for $j \notin A$.
4. Update $w \leftarrow w - D\hat{\beta}^{\mathrm{PLS}}$ and normalise to unit length.

In view of (7), for the case of $K = 1$, the above simply corresponds to setting $\hat{\beta}^{\mathrm{PLS}}$ equal to a soft-thresholded, scaled version of d. This type of sparse PLS shares similarities with the (iterated) independent screening method of Paper VII. It is an interesting future research problem to explore this connection in more detail, and to assess the performance of additive hazards sparse PLS in practice.

Interpretable hazard regression. James *et al.* (2009) presented a framework for functional regression problems which uses penalised regression to estimate a regression function with 'sparse derivatives', i.e. a regression function whose derivative of a certain order is identically equal to zero on most of its domain. This effectively regularises the functional form of the regression function and can greatly simplify interpretation. Their framework extends to the semiparametric additive hazards model. Assume for simplicity that explanatory variables are univariate and that the conditional hazard function takes the form

$$h(t|Z) = h_0(t) + \beta(t)Z.$$

Let $B(t) := [b_1(t), b_2(t),\dots,b_p(t)]^\top$ be some collection of basis function (indicator functions, splines, wavelets etc.) and suppose that $\beta(t) = B(t)^\top\eta$ for some $\eta \in \mathbb{R}^p$. Then we may estimate the regression function $\beta(t)$ as $D^{-1}dB(t)$, taking

$$D = \sum_{i=1}^n \int_0^\tau \left[\{Z_i - \bar{Z}(t)\}B(t)^\top\right]^{\otimes 2} Y_i(t)\mathrm{d}t, \quad d = \sum_{i=1}^n \int_0^\tau \{Z_i - \bar{Z}(t)\}B(t)^\top \mathrm{d}N_i(t).$$

This is the basis function approach to incorporating time-varying regression coefficients. Following James *et al.* (2009), consider now a grid $t_1,\dots,t_p$ of evenly spaced points in the observation time window $[0,\tau]$ and let A be the $p \times p$ matrix $A := [D^kB(t_1),\dots,D^kB(t_p)]$ where D^k is the kth difference operator. Setting $\gamma := A\eta$, it follows that γ_j is an approximation to $\beta^{(k)}(t_j)$. Suppose that A is invertible. Parameterising the additive hazards model in terms of γ, we obtain

$$D_\gamma = \sum_{i=1}^{n} \int_0^\tau \left[\{Z_i - \bar{Z}(t)\}B(t)^\top A^{-1}\right]^{\otimes 2} Y_i(t)\mathrm{d}t, \quad d_\gamma = \sum_{i=1}^{n} \int_0^\tau \{Z_i(t) - \bar{Z}(t)\}B(t)^\top A^{-1}\mathrm{d}N_i(t).$$

By using penalised regression such as the lasso with D_γ and d_γ, we effectively estimate the regression function $t \mapsto \beta(t)$ while regularising its functional form. For example, taking $k = 1$, we can use the above construction to estimate a piecewise constant regression function. The methods presented in James *et al.* (2009) also enable us to consider combinations of derivatives of different orders, as well as non-invertible A.

We have experimented informally with this technique, taking $b_1, \ldots, b_p$ to be a collection of indicator functions of a partition of $[0, \tau]$ into intervals. For survival data, although computationally convenient, this particular basis is not the best choice since the nonuniform distribution of survival times leads to unstable estimates. Future efforts will consider more complex choices of bases.

The recent work by Fan and James (2011), which uses the group lasso to do variable selection with functional predictors and simultaneously estimate their functional form, could be similarly generalised to the survival setting; leading to a novel variable selection method for the general (nonparametric) Aalen model.

References

Abry, P., Flandrin, P., Taqqu, M. S. and Veitch, D. (2003) Self-similarity and long-range dependence through the wavelet lens. In *Theory and Applications of Long-Range Dependence* (eds. P. Doukhan, G. Oppenheim and M. S. Taqqu), 527–556. Birkhäuser.

Beran, J. (1994) *Statistics for Long-Memory Processes*. Chapman and Hall, New York.

Bickel, P. and Ritov, Y. (2008) Discussion of: Treelets – an adaptive multi-scale basis for sparse unordered data. *Annals of Applied Statistics*, **2**, 474–477.

Chatterjee, A. and Lahiri, S. N. (2009) Bootstrapping lasso estimators. Tech. rep., Department of Statistics, Texas A&M University.

Chatterjee, A. and Lahiri, S. N. (2010) Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, **138**, 4497–4509.

Chun, H. and Keleş, S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society Series B*, **72**, 3–25.

Donoho, D. L. (2000) High-dimensional data analysis: The curses and blessings of dimensionality. Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century.

Fan, J., Feng, Y. and Wu, Y. (2010) *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*, chap. High-dimensional variable selection for Cox's proportional hazards model. Institute of Mathematical Statistics.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan, J. and Lv, J. (2008) Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society Series B*, **70**, 849–911.

Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20**, 101–148.

Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research*, **10**, 2013–2038.

Fan, J. and Song, R. (2010) Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics*, **38**, 3567–3604.

Fan, Y. and James, G. M. (2011) Functional additive regression. Tech. rep., Marshall School of Business, University of Southern California.

Gorst-Rasmussen, A. (2011) ***ahaz**: Regularization for semiparametric additive hazards regression*. URL `http://cran.r-project.org/package=ahaz`. R package.

Hu, F. B. (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. *Current Opinion in Lipidology*, **13**, 3–9.

James, G. M., Wang, J. and Zhu, J. (2009) Functional linear regression that's interpretable. *Annals of Statistics*, **37**, 2083–2108.

Joyce, A. R. and Palsson, B. O. (2006) The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, **7**, 198–210.

Karagiannis, T., Molle, M. and Faloutso, M. (2004) Long-range dependence – ten years of Internet traffic modeling. *IEEE Internet Computing*, **8**, 57–64.

Lee, A. B., Nadler, B. and Wasserman, L. (2008) Treelets – an adaptive multi-scale basis for sparse unordered data. *Annals of Applied Statistics*, **2**, 435–471.

Leeb, H. and Pötscher, B. M. (2008) Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, **142**, 201–211.

Leland, W. E., Taqqu, M. S., Willinger, W. and Wilson, D. V. (1994) On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, **2**, 1–15.

Leventhal, S. (1988) Uniform limit theorems for Harris recurrent Markov chains. *Probability Theory and Related Fields*, **80**, 101–118.

Li, C. and Li, H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Lin, D. Y. and Ying, Z. (1994) Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61–71.

Martinussen, T. and Scheike, T. H. (2009) Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics*, **36**, 602–619.

Martinussen, T. and Scheike, T. H. (2010) The additive hazards model with high-dimensional regressors. *Lifetime Data Analysis*, **15**, 330–342.

Meinshausen, N. and Bühlmann, P. (2008) Discussion of: Treelets – an adaptive multi-scale basis for sparse unordered data. *Annals of Applied Statistics*, **2**, 478–481.

Paxson, V. and Floyd, S. (1995) Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, **3**, 226–244.

Pereira, S., Kyritsi, P., Papanicolaou, G. and Paulraj, A. (2005) Asymptotic properties of richly scattering ultrawideband channels. Stanford University. Unpublished manuscript.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Rosipal, R. and Krämer, N. (2006) Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection Techniques* (eds. C. Saunders, M. Grobelnik, S. Gunn and J. Shawe-Taylor), 34–51. New York: Springer.

Rubak, E. (2007) *Central limit theorems for weakly dependent stochastic processes*. Master's thesis, Aalborg University.

Scheike, T. H. (2001) The additive nonparametric and semiparametric Aalen model as the rate function for a counting process. *Lifetime Data Analysis*, **8**, 247–262.

Slawski, M., Castell, W. and Tutz, G. (2010) Feature selection guided by structural information. *Annals of Applied Statistics*, **4**, 1056–1080.

Snyder, M., Du, J. and Gerstein, M. (2010) Personal genome sequencing: current approaches and challenges. *Genes and Development*, **24**, 423–431.

StataCorp (2009) *Stata Statistical Software: Release 11.* StataCorp LP, College Station, TX.

Taqqu, M. S. and Teverovsky, V. (1997) Robustness of Whittle-type estimators for time series with long-range dependence. *Stochastic Models*, **13**, 323–357.

Tibshirani, R. (1997) The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385–395.

Tjønneland, A., Olsen, A., Boll, K. *et al.* (2007) Study design, exposure variables, and socioeconomic determinants of participation in Diet, Cancer and Health: a population-based prospective cohort study of 57,053 men and women in Denmark. *Scand J Public Health.*, **35**, 432–441.

Tsai, T. H. (1998) *The Uniform CLT and LIL for Markov Chains*. Ph.D. thesis, University of Wisconsin.

Tulino, A. and Verdú, S. (2004) *Random Matrix Theory and Wireless Communications*. Hannover, MA: Now Publishers.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

Part 1

Applications *in* Telecommunications

Paper I

Asymptotic Inference for Waiting Times and Patiences in Queues with Abandonment

Author list

Anders Gorst-Rasmussen
Aalborg University, Denmark

Martin B. Hansen
Aalborg University, Denmark

Summary

Motivated by applications in call centre management, we propose a framework based on empirical process techniques for inference about the waiting time and patience distribution in multiserver queues with abandonment. The framework rigorises heuristics based on survival analysis of independent and identically distributed observations by allowing correlated successive waiting times. Assuming a regenerative structure of the sequence of offered waiting times, we establish asymptotic properties of estimators of limiting distribution functions and derived functionals. We discuss construction of bootstrap confidence intervals and statistical tests, including a simple bootstrap two-sample test for comparing patience distributions. The methods are exemplified in a small simulation study, and a real data example is given involving comparison of patience distributions for two customer classes in a call centre.

Supplementary info

This paper extends results from the MSc thesis of Anders Gorst-Rasmussen (*Empirical Processes for Regenerative Sequences*, Aalborg University, 2006). The paper first appeared as:

> Gorst-Rasmussen A, Hansen MB (2007). Asymptotic Inference for Waiting Times and Patiences in Queues with Abandonment. *Technical report R-2007-14*. Department of Mathematical Sciences, Aalborg University

It was later published by Taylor & Francis as:

> Gorst-Rasmussen A, Hansen MB (2009). Asymptotic Inference for Waiting Times and Patiences in Queues with Abandonment. *Communications in Statistics: Simulation and Computation*; **38**(2):318-334

The version here is the journal version with minor typographical edits.

1. Introduction

In a queueing system with abandonment, customers may abandon the waiting line before being serviced. This leads to right-censored waiting times where offered waiting times in the queue without abandonment are censored by random customer patiences. Models for queues with abandonment are of practical interest when designing and analysing call centres where abandonment may considerably affect performance

(Garnett *et al.*, 2002). There has recently been a surge of interest in empirical applications of queueing models with abandonment to running call centres for which detailed call-by-call data are available. Statistical analyses of such data can provide both quantitative measures of performance and quality of service, as well as offer valuable insight into the qualitative nature of customer abandonment. This was demonstrated by Brown *et al.* (2005), who applied methods from classical survival analysis to estimate cumulative distribution functions (CDFs) of waiting times and patiences, hazard rates, and related functionals. However, positive correlation of successive waiting times generally invalidates the asymptotic theory classically used to derive interval estimates and statistical tests. As pointed out by Gans *et al.* (2003), there is a need to develop survival analytic methods which are capable of providing confidence intervals and statistical tests for call-by-call data from queues with abandonment.

Nonparametric survival techniques for dependent observations have previously been studied in the literature under mixing assumptions, and include Kaplan-Meier estimation (Cai, 2001), quantile estimation (Cai and Kim, 2003), and hazard rate estimation (Cai, 1998). The techniques rely on mixing assumptions for the observation sequence, and computation of confidence intervals and statistical testing is often difficult and case-specific. In the present paper, we assume that the sequence of offered waiting times is regenerative. Informally, this means that the waiting time sequence splits into IID random blocks of random lengths. The assumption of regenerative offered waiting times is satisfied by the widely used *GI*/*G*/*m* queueing model under weak assumptions (Asmussen, 2003, Theorem XII.2.2), with blocks defined by system-wide busy periods. Regenerativity of the offered waiting times extends to independently right-censored waiting times:

$$(1) \qquad \widetilde{W}_n := \min\{W_n, P_n\}, \quad n \in \mathbb{N},$$

with $\{W_n\}$ the individual customer offered waiting times and $\{P_n\}$ the individual IID customer patiences, which we assume independent of $\{W_n\}$. Regenerativity of the offered waiting times is not a special property of the *GI*/*G*/*m* queueing model. It remains a valid model whenever the arrival and service time sequences are stationary, and the waiting time sequence splits into independent blocks. The latter happens, for example, if the queueing system restarts at fixed time points, as is often the case in call centres.

In the present paper, we show how the assumption of regenerativity, when combined with techniques from the theory of empirical processes, can be used to rigorise methods for analysing waiting times and patiences in queues. From a practical perspective, regenerativity justifies the use of various resampling methods to obtain confidence intervals and statistical tests for parameters. Emphasis will be placed on a simple blockwise bootstrap resampling technique. Besides from contributing tools for practical inference, the paper contributes to the limited literature on nonparametric inference for queueing systems using empirical processes; see for example Bingham and Pitts (1999a); Bingham and Pitts (1999b) – or Hansen and Pitts (2006) for statistical inference involving empirical processes of regenerative observations. We remark that while this paper deals specifically with inference about waiting times and patiences, the empirical process techniques discussed here apply also to estimators for other types of regenerative sequences.

The paper is organised as follows. In Section 2, we review basic empirical process techniques for regenerative observations and state a new result concerning the validity of a functional blockwise bootstrap. Section 3 describes estimation of CDFs, nonparametric two-sample testing for the patience CDF, and estimation of various functionals of the waiting time and patience CDF of interest in call centre management. Section 4 presents a discussion of the practical use of the framework together with a simulation study. Finally, Section 5 illustrates a selection of the procedures applied to real-world data.

2. Asymptotic inference for regenerative sequences

Consider a sequence $\{C_n : n \in \mathbb{N}_0\}$ of random cycles taking values in $\bigcup_{m \geqslant 0} \mathbb{R}^m$, with $C_1, C_2, \dots$ independent and identically distributed (IID) and independent of C_0. Thus each C_i is a block of random variables of random length. Defining X_n to be the nth real-valued observation in $\{C_n : n \in \mathbb{N}_0\}$, the sequence of random variables $X = \{X_n : n \in \mathbb{N}\}$ is called a regenerative sequence. The first cycle C_0 is known as the delay of the regenerative sequence. We denote by ℓ_n the length of C_n, define the renewal sequence $T_{n+1} := \ell_n + T_n$ (letting $T_0 := 0$), and let $\tau_n := \inf\{m \geqslant 1 : T_m > n\} - 1$ be the number of complete, observed cycles at time n. We assume ℓ_1 to be non-lattice with finite expectation. Then X admits a limiting distribution $\mathbb{P}$ (Asmussen (2003), Corollary VI.1.5), in the sense that $X_n \to \mathbb{P}$ in total variation where

$$\mathbb{P}(\cdot) = \frac{1}{\mathbb{E}(\ell_1)} \mathbb{E}\left\{ \sum_{i=T_1+1}^{T_2} \mathbb{1}(X_i \in \cdot) \right\}. \tag{2}$$

Nonparametric statistical methods for regenerative sequences use regenerative analogues of the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT) to establish consistency and asymptotic distributional properties of estimators. Adequately general forms of these limit results come from the theory of empirical processes which concerns the asymptotic behaviour of functional estimators of the form

$$\mathrm{P}_n(f) = n^{-1} \sum_{i=1}^{n} \{f(X_i) - \mathbb{P}f\}, \quad f \in \mathscr{F}, \tag{3}$$

uniformly over a set of measurable real-valued functions $\mathscr{F}$. The sequence of stochastic processes $\{\mathrm{P}_n(f) : f \in \mathscr{F}\}$ is called an empirical measure. A detailed review of limit results for empirical processes of IID observations can be found in van der Vaart and Wellner (1996). Limit results for empirical processes of regenerative observations have received limited attention in the literature; see Leventhal (1988) and Tsai (1998). In this paper, we restrict ourselves to discussing the use of empirical process theory for estimating the limiting CDF of a regenerative sequence, $F(\cdot) := \mathbb{P}(-\infty, \cdot]$. This is not contrived: as we shall explain, a 'good' estimator of F can be used to define 'good' estimators of a range of functionals of the form $\phi(F)$.

From observations $X_1, \dots, X_n$ of a regenerative sequence, we may estimate F using the empirical CDF defined for $x \in \mathbb{R}$ by $F_n(x) := n^{-1} \sum_{i=1}^{n} \mathbb{1}(X_i \leqslant x)$. The sequence $\{F_n\}$ is the empirical measure of $\mathscr{F} = \{\mathbb{1}(\cdot \leqslant x) : x \in \mathbb{R}\}$ and defines a sequence in the space $D(\mathbb{R})$ of real cadlag functions equipped with the supremum norm $\|\cdot\|_\infty$. A Vapnik-Cervonenkis argument (Pollard, 1984, p. 16) and the limit theorems of Leventhal (1988) immediately lead to regenerative analogues of the classical Glivenko-Cantelli (uniform LLN) and Donsker theorems (uniform CLT).

THEOREM 1 (Regenerative Glivenko-Cantelli/Donsker). *Let X be a regenerative process satisfying $\mathbb{E}(\ell_1) < \infty$, and denote by F the CDF of the limiting distribution of X. Then*

$$\|F_n - F\|_\infty \to 0, \quad \textit{in probability.}$$

If also $\mathbb{E}(\ell_1^2) < \infty$ *then there exists a centered tight Gaussian process* H_F *on* $\mathbb{R}$ *such that*

$$n^{1/2}(F_n - F) \xrightarrow{d} \mathrm{H}_F,$$

where $\xrightarrow{d}$ *denotes weak convergence in* $D(\mathbb{R})$.

The precise meaning of weak convergence in $D(\mathbb{R})$ is that $\mathbb{E}^*\{\varphi(F_n)\} \to \mathbb{E}\{\varphi(\mathrm{H}_F)\}$ for bounded, continuous, real-valued functions φ where $\mathbb{E}^*$ denotes outer expectation. This general form of weak convergence is required since F_n is generally non-measurable when $D(\mathbb{R})$ is equipped with the supremum norm and the Borel σ-field.

Theorem 1 in theory allows for approximating the sampling distribution of functionals of $F_n - F$ from the limiting Gaussian process H_F. However, this result is of little practical use since the covariance function of H_F depends on X in a nontrivial manner, precluding construction of distribution-free statistics in general. Instead, resampling methods can be used, i.e. methods which utilise (random) subsets of data to approximate sampling distributions. The strong mixing property of regenerative sequences (Thorrison, 2000, Theorem 3.3) in principle enables application of the method of functional subsampling (Wolf *et al.*, 1999) and, under additional mixing assumptions, the moving blocks bootstrap (Naik-Nimbalkar and Rajarshi, 1994). However, the performance of either method relies on complex preliminary calibrations which again depend on the statistic under investigation. We suggest a simpler alternative which utilises the intrinsic structure of regenerative sequences. Here resampling is performed by sampling with replacement among regenerative cycles rather than individual observations, extending the naive bootstrap idea of sampling with replacement from IID observations (Efron, 1979) to regenerative sequences. This regenerative block bootstrap (RBB) has previously been studied for the case of inference for the mean (Athreya and Fuh, 1989; Datta and McCormick, 1993; Bertail and Clémençon, 2006) and is described algorithmically below.

ALGORITHM 1 (Regenerative blockwise bootstrap).
Given observations $\{X_i : i \leqslant n\}$ of X, let $\theta_n := \theta_n(X_1,\ldots,X_n)$ denote a statistic.

1. Divide $\{X_i : i \leqslant n\}$ into regenerative cycles $C_1,\ldots,C_{\tau_n}$.
2. Conditionally on $\{X_i : i \leqslant n\}$ and τ_n, sample $C_1^*,\ldots,C_{\tau_n}^*$ with replacement from $\{C_1,\ldots,C_{\tau_n}\}$.
3. Define the bootstrapped sample $\{X_i^* : i = 1,2,\ldots,n_*\}$ where X_i^* is the ith real-valued observation of $\{C_1^*,\ldots,C_{\tau_n}^*\}$, $T_{i+1}^* := T_i^* + l_i^*$ (taking $T_1^* := 0$ and l_i^* to be the length of C_i^*), and $n_* := T_{\tau_n+1}^*$.
4. Compute $\theta_n^* := \theta_n(X_1^*,\ldots,X_{n_*}^*)$.

Approximate the law of θ_n by the conditional law of θ_n^* given $\{X_i : i \leqslant n\}$.

In the present paper, we need validity of an empirical process version of the RBB where $\theta_n := F_n$ is the empirical CDF and $\theta_n^* =: F_n^*$ its bootstrapped counterpart. Validity of the RBB in this setting may be defined in terms of a distance d metrising weak convergence on $D(\mathbb{R})$ by requiring

$$d\{n_*^{1/2}(F_n^* - F_n), \mathrm{H}_F\} \to 0, \quad \text{in probability;} \tag{4}$$

where the ‘in probability’ statement is relative to the law governing the observations. This in turn implies that the RBB estimator $n_*^{1/2}(F_n^* - F_n)$ is a consistent estimator of

$n^{1/2}(F_n - F)$ in the sense that their d-distance tends to zero in probability as $n \to \infty$. Typically, d will be the dual bounded Lipschitz distance on $D(\mathbb{R})$ (van der Vaart and Wellner, 1996, p. 73). Validity of the empirical process RBB has been investigated by Radulović (2004) for a class of empirical processes with observations from a discrete atomic Markov chain. In the appendix, we give a short proof of validity in the sense of (4) of the RBB for general empirical processes under the assumptions of the uniform CLT for regenerative observations of Tsai (1998). For the case of the RBB for the empirical CDF, the validity result reads as follows.

THEOREM 2 (Bootstrap validity). *Let X be a regenerative sequence with $\mathbb{E}(\ell_1^2) < \infty$. Denote by F the CDF of the limiting distribution of X and let F_n^* be the CDF obtained from the RBB. Then* (4) *holds.*

Estimation of the sampling distribution of F_n alone is of limited interest in applications, and it is desirable to extend the asymptotic results above to general functionals of F_n (plugin estimators). The continuous mapping theorem ensures that the RBB works for continuous real-valued functions of F_n. Another versatile tool not restricted to real-valued statistics is a functional analogue of the finite-dimensional delta-method. With the notation of Algorithm 1, let θ_n be a statistic of regenerative observations $X_1,\dots,X_n$, taking values in a normed space V, and denote by θ_n^* the bootstrapped statistic obtained using the RBB. Suppose that $\phi\colon V \to W$ for some normed space W is a mapping for which there is a bounded linear operator $\mathrm{d}\phi_\theta : V \to W$ satisfying $\sup_{h\in K} \|t^{-1}\{\phi(\theta + th) - \phi(\theta)\} - \mathrm{d}\phi_\theta(h)\| \to 0$ when $t \to 0$ for every compact set $K \subseteq V$. Then ϕ is called Hadamard differentiable at θ. The next result follows from Theorem 3.9.4 and Theorem 3.9.11 of van der Vaart and Wellner (1996).

THEOREM 3 (Functional delta-method). *Assume that there exists $\theta \in V$ and $r_n \uparrow \infty$ such that $r_n(\theta_n - \theta) \to^d T$ for a tight random element T, and that the RBB estimator $r_n(\theta_n^* - \theta)$ is a consistent estimator of T. If ϕ is Hadamard differentiable at θ with derivative* $\mathrm{d}\phi_\theta$ *then $r_n\{\phi(\theta_n) - \phi(\theta)\} \to^d$* $\mathrm{d}\phi_\theta(T)$*, and the RBB estimator $r_n\{\phi(\theta_n^*) - \phi(\theta_n)\}$ is a consistent estimator of $r_n\{\phi(\theta_n) - \phi(\theta)\}$.*

If T is tight Gaussian, linearity of $\mathrm{d}\phi_\theta$ implies that $\mathrm{d}\phi_\theta(T)$ is also tight Gaussian. One reason why the functional delta-method is so useful is the chain rule of Hadamard differentiation (van der Vaart and Wellner, 1996, Lemma 3.9.3). This allows one to establish the asymptotics of a complicated statistic by representing it as a composition of simpler Hadamard differentiable maps applied to the empirical CDF.

RBB-based confidence intervals can be constructed using Efron's percentile method (Efron, 1979). Namely if θ_n is an estimator of a real-valued parameter θ, and θ_n^* is obtained from the RBB using Algorithm 1, an approximate $(1 - \alpha - \beta) \times 100\%$ confidence interval for θ is given by $[\theta_n - \xi_{n,\beta}^*, \theta_n - \xi_{n,1-\alpha}^*]$ where $\xi_{n,\gamma}^*$ is the upper γth percentile of the bootstrap distribution of $\theta_n^* - \theta_n$, that is, the largest value x which satisfies $\mathbb{P}^*(\theta_n^* - \theta_n \geqslant x) \geqslant 1 - \gamma$. The RBB confidence interval asymptotically has level $1 - \alpha - \beta$, whenever the statistic θ_n is a continuous or Hadamard differentiable function of the empirical CDF.

3. Asymptotic inference for waiting times and patiences

Let $\widetilde{W}_1,\dots,\widetilde{W}_n$ be right-censored waiting times from a queueing system, defined as in (1) so that the underlying offered waiting times are assumed to form a regenerative sequence and the patiences are assumed to be IID random variables. Observations take the form

$$(\widetilde{W}_1,\delta_1),\dots,(\widetilde{W}_n,\delta_n), \tag{5}$$

where δ_i is the non-censoring indicator of $\widetilde{W}_i$. If we seek features of the waiting time distribution, censoring occurs when the customer abandons the queue and vice versa for the patience distribution. Inferential procedures for such observations can be investigated with the empirical process methods of the previous section. This leads to a qualitative description of estimator asymptotics which, when combined with resampling techniques, can be used quantitatively to construct confidence intervals and statistical tests. We shall consider resampling using the RBB, but other resampling methods (see the discussion preceding Algorithm 1) may also be used to infer sampling distributions of the estimators of this section.

Denote by F the limiting CDF of uncensored observations from (5). A basic problem is how to estimate F from the censored observations. We suggest to use the product-limit (or Kaplan-Meier) estimator,

$$F_n(t) := 1 - \prod_{i:\widetilde{W}_{(i)}\leqslant t}\left(1-\frac{n-i}{n-i+1}\right)^{\delta_{(i)}},$$

where $\widetilde{W}_{(i)}$ is the ith order statistic of $\widetilde{W}_1,\dots,\widetilde{W}_n$ and $\delta_{(i)}$ the corresponding indicator of non-censoring. The asymptotic properties of F_n can be established using Theorems 1 and 3. Denote by $H^{\mathrm{uc}}(t) := \mathbb{P}(\widetilde{W}\leqslant t,\delta=1)$ the limiting subdistribution function of the uncensored observations and by $\overline{H}(t) := \mathbb{P}(\widetilde{W}\geqslant t)$ the limiting tail function of observations. A classical result from survival analysis (Gill and Johansen, 1990) states that F can be obtained from $(\overline{H},H^{\mathrm{uc}})$ via the mappings

$$(\overline{H},H^{\mathrm{uc}}) \xmapsto{\alpha} \int_{[0,\cdot]}\overline{H}(s)^{-1}\mathrm{d}H^{\mathrm{uc}}(s) =: \Lambda \xmapsto{\beta} \prod_{s\in(0,\cdot]}\{1-\mathrm{d}\Lambda(s)\} = 1-F.$$

Here Λ is the cumulative hazard rate, and $\prod_{s\in(0,t]}$ denotes the product integral over $(0,t]$. Then F_n is in fact the plugin estimator $\beta\{\alpha(\overline{H}_n,H_n^{\mathrm{uc}})\}$ where

$$H_n^{\mathrm{uc}}(t) = n^{-1}\sum_{i=1}^{n}\delta_i\mathbb{1}\big(\widetilde{W}_i\leqslant t\big), \quad \overline{H}_n(t) = n^{-1}\sum_{i=1}^{n}\mathbb{1}\big(\widetilde{W}_i\geqslant t\big).$$

It can be shown (Gill and Johansen, 1990) that each of α, β, then $\beta\circ\alpha$ are Hadamard differentiable at $(H^{\mathrm{uc}},\overline{H})$ when the latter is viewed as an element of $D[0,\tau]\times D[0,\tau]$ for some τ with $\overline{H}(\tau)>0$. Combining this with Theorem 1-3, we conclude that the product-limit estimator based on regenerative observations is consistent, asymptotically Gaussian and can be bootstrapped. So we can use the RBB to construct both pointwise confidence bands for F (by estimating the distribution of $F(t)$ for each t) and uniform confidence bands (by estimating the distribution of $\sup_{t\in[0,\tau]}|F(t)|$). Examples will follow in the next section. By similar arguments, one obtains consistency, asymptotic Gaussianity, and bootstrap validity for the Nelson-Aalen-type estimator $\Lambda_n := \alpha(\overline{H}_n,H_n^{\mathrm{uc}})$ of the cumulative hazard rate. Estimates of functions relating to the (cumulative) hazard rate have previously been used to explore abandonment behaviour of customers in a call centre (Brown *et al.*, 2005). Note that empirical process theory, although a powerful framework, essentially deals with inference using step functions (empirical measures) and does not lend itself towards methods for smooth estimation of, for example, densities or hazard rates. Smooth estimation procedures for censored sequences under mixing assumptions are discussed by Cai (1998).

One may ask whether estimators of expectations or quantiles of F based on plugging in the product-limit estimator F_n in the formulas $\mathbb{E}\{\xi(X)\} = \int_0^\infty \xi(x)F(\mathrm{d}x)$ and $F^{-1}(p) := \inf\{x : F(x) \geqslant p\}$ inherit the nice asymptotic properties. Such statistics may arise as key performance indicators in call centre managing, where one seeks summary statistics such as expected waiting times and patiences; or median waiting times and patiences (Nederlof and Anton, 2002). If the largest observation is censored, the product-limit estimator is not a CDF and plugging it in the definition of the expectation will produce infinite values. Instead, one can estimate the truncated expectation from $\int_0^\tau \xi(x)F_n(\mathrm{d}x)$ where τ satisfies $\mathbb{P}(\tilde{W} \leqslant \tau) < 1$. Consistency, asymptotic Gaussianity, and bootstrap validity of this estimator follows from Lemma 3.9.17 of van der Vaart and Wellner (1996) and Theorem 3. Note that this truncated expectation is a negatively biased estimator of $\mathbb{E}\{\xi(X)\}$ and should be interpreted with care. Similarly for quantiles of F, Lemma 3.9.20 of van der Vaart and Wellner (1996) implies Hadamard differentiability of the mapping taking F to its pth percentile, whenever F has a strictly positive derivative at $F^{-1}(p)$. Theorem 3 again implies consistency, asymptotic Gaussianity, and bootstrap validity for the estimator of the pth percentile based on F_n.

We next consider the issue of how to formally test equality of two limiting patience CDFs from right-censored regenerative patiences. This problem has to the best of our knowledge not been considered previously, but is of relevance when comparing abandonment behaviour of two customer classes in a call centre. Assume that we have available two independent samples of the form (5) (with censoring when the customer is serviced) of sizes n and m, such that the limiting CDFs of uncensored observations are F and G, respectively, and the limiting CDFs of the censored observations are H and I. Denote by F_n and G_n the product-limit estimators of the CDFs, and let τ be such that $H(\tau) < 1$ and $I(\tau) < 1$. We seek to test the null hypothesis

$$H_0 : F(t) = G(t), \quad \forall\, t \in [0,\tau] \tag{6}$$

against the two-sided alternative $F \neq G$. Denote by W the common tight Gaussian limit of $n^{1/2}(F_n - F)$ and $m^{1/2}(G_m - G)$ under the null hypothesis. Define the test statistic

$$D_{n,m} := \sqrt{(nm)/(n+m)}\|(F_n - F) - (G_m - G)\|_\infty,$$

where $\|\cdot\|_\infty$ denotes supremum over the interval $[0,\tau]$, and suppose $nm/(n+m) \to \lambda \in (0,1)$ when $n,m \to \infty$. Then, under the null hypothesis, the continuous mapping theorem implies $D_{n,m} \to^d \|W\|_\infty$. The distribution of the supremum $\|W\|_\infty$ is intractable and must be approximated by resampling techniques. To this end, define the bootstrapped counterpart of $D_{n,m}$ by

$$D^*_{n,m} = \sqrt{(n_* m_*)/(n_* + m_*)}\|(F^*_n - F_n) - (G^*_m - G_m)\|_\infty.$$

Here the quantities n_*, F^*_n and m_*, G^*_m are obtained by applying the RBB to each censored sample separately. The map $(A,B) \mapsto A - B$ is Hadamard differentiable on $(D[0,\tau])^2$. Theorem 3, Slutsky's lemma for the bootstrap (Radulović, 2004, Lemma 3.1), and Theorem 1-2 together with the continuous mapping theorem implies consistency of $D^*_{n,m}$ as an estimator of $D_{n,m}$ as $n,m \to \infty$. So the conditional distribution of the bootstrapped test statistic $D^*_{n,m}$ may be used to define critical levels for the null hypothesis (6): if $\xi^*_{n,m,\alpha}$ is the upper α percentile of the RBB distribution $\mathrm{P}^*(D^*_{n,m} \leqslant \cdot)$, then H_0 is rejected at approximate level α if $\sqrt{mn/(m+n)}\|F_n - G_m\|_\infty > \xi^*_{n,m,\alpha}$. This essentially corresponds to constructing an $(1-\alpha)\times 100\%$ uniform confidence band for $F-G$ and rejecting H_0 at level α if the band does not contain the zero function. Analogous procedures with potentially better power properties are easily defined for other smooth 'discrepancy functionals' $(F,G)\mapsto \phi(F,G)$ than the difference: for example the odds ratio or the cumulative hazard ratio of two limiting CDFs – or weighted versions hereof.

The above approach to hypothesis testing (constructing confidence intervals by resampling and checking whether zero is contained in the interval) applies generally to simple hypotheses $H_0 : \theta_1 = \theta_2$ whenever consistent estimators $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$ of θ_1 and θ_2 exist which are asymptotically Gaussian and can be bootstrapped. This in turn yields a method for rigorous empirical comparison of for example medians, probabilities, and expectations. Note that, in the case of inference for expectations with respect to the limiting distribution, more efficient RBB-methods based on the percentile t-method (Hall, 1992) exist (Bertail and Clémençon, 2007).

4. Practical considerations and simulation examples

In the previous section, we discussed methods for qualitatively and quantitatively investigating properties of estimators from right-censored waiting times. The key was the asserted regenerative structure of the offered waiting times which enabled regenerative empirical process techniques to be applied. The assumption of regenerativity is often a reasonable and parsimonious model. It holds in the general $GI/G/m$-queuing model with regeneration occurring when all servers are idle (Asmussen, 2003, Theorem XII.2.2), allowing regenerative cycles to be constructed whenever all such regeneration points have been identified in an observation sequence. In call centers, with many servers and high load, there may be few or no system wide idle periods during a typical day of operation. On the other hand, if a regenerative model is adopted, forced regeneration occurs at the end of every day when the call center closes. This suggests that (a subset of) the waiting time sequence for each separate day of operation can be used to define regenerative cycles. This idea is not restricted to $GI/G/m$-type queuing systems, but applies to any queuing system for which independent and identically distributed cycles of waiting times can be defined. Stationarity of the cycle sequence can be checked empirically by investigating stationarity of a sequence of real-valued statistics calculated from the cycles (averages, variances etc.), for example using time series plots. A sufficient condition for cycle stationarity is stationarity of the underlying observation and cycle length sequence.

We performed a small simulation study to illustrate coverages of RBB confidence intervals for selected statistics of waiting times and patiences, as well as level and power of the two-sample RBB test for patience CDFs. In all experiments, we considered an $M/M/15$ queuing system with an arrival rate of 13.5 customers per minute and a service rate of 1 customer per minute, corresponding to a system load of 90%. Waiting times were right-censored with IID patiences from various distributions. Each regenerative block used in the RBB was simulated independently and comprised 15 minutes of observations following a 15 minute start-up period, corresponding to blocks of approximately 200 successive observations in the stationary regime. A start-up period was used solely for computational reasons: the inferential methods also apply in the transient regime, but are not easily compared with theoretical results.

A typical sequence of right-censored waiting times from an $M/M/15$ queuing system with exponential patiences is shown in Figure 1 (left). Observe the positive correlation between successive observations which precludes the use of standard statistical methods for IID data. In Figure 1 (right), an example of the estimated patience CDF (superimposed on the true patience CDF) is shown.

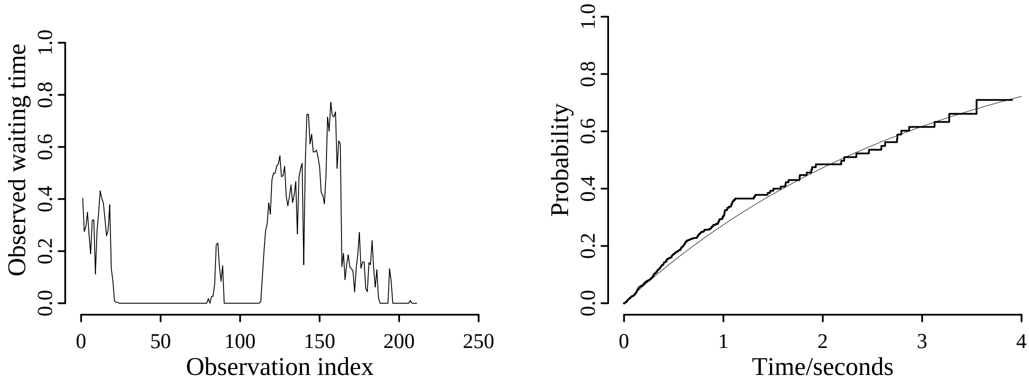

Figure 1. Left: Example of a waiting time sequence in an $M/M/15$ queue with an arrival rate of 13.5 customers per minute, a service rate of 1 customer per minute, and exponential patiences. Right: An estimate of the patience CDF (thick line) in the same queuing system system, superimposed on the true patience CDF (thin line).

Table 1 shows estimated coverages of RBB-confidence intervals for a selection of statistics of the right-censored waiting times. Observe that coverages are subject to sampling variation which can be quantified using standard methods for binomial proportions. All confidence intervals have been calculated using the percentile method. The estimated coverages in Table 1 are generally close to their nominal values, although the confidence intervals appear slightly anticonservative. We found that decreasing the rate of abandonment did not markedly impact coverage for estimates from the patience distribution, although quantile estimation becomes difficult when the rate of abandonment is small. This is due to the product-limit estimator having an atom at infinity if the largest observation is censored, frequently leading to infinite quantile estimates in the case of heavy censoring. The uniform confidence intervals and the corresponding coverages are calculated for the respective CDFs over the fixed interval $[0,1.5]$ for all simulations. The estimated coverages for the uniform confidence intervals were sensitive to the choice of interval – too large intervals lead to poor coverages. In applications, one would typically use the interval ranging from zero to the largest uncensored observation of the sample.

The estimated level and power of the RBB two-sample test for two different types of patience distributions (exponential and lognormal with fixed logarithmic variance 1) is shown in Table 2. Test statistics were calculated over the fixed interval $[0,1.5]$ for all simulations. The parameter of each patience distribution was adjusted to provide rates of abandonment of 20%, 10%, and 5%, respectively. The level of the test was estimated for each rate of abandonment. We also estimated the power to detect a supremum distance deviation of 0.05, 0.1, and 0.2 from these reference patience distributions, letting each comparison distribution be stochastically larger than its reference counterpart. The test exhibits reasonable power properties, considering the small rate of abandonment: more detailed power assessments are difficult due to the lack of reference methods. The estimated levels suggest that the test is slightly conservative. As was the case for uniform RBB confidence intervals for CDFs, the test was sensitive to the choice of interval over which the test statistic was calculated.

5. Application to real data

As an application of the methods of this paper, we considered inference from real data given by call logs from a call center of a small Israeli bank. See Brown *et al.* (2005) for a detailed description and statistical analysis of the data. We extracted right-censored waiting times for all customers of the call center arriving during the period 2 p.m.–3 p.m. on ordinary Israeli weekdays (Sunday-Thursday) in November and December. This is representative of customer waiting experience during peak hours and would be of particular interest to a call center manager. We obtained 36 observation sequences of average length 139. Observation sequences of separate days were assumed independent. The assumption of stationarity of blocks was assessed by checking the sufficient condition of stationarity of the waiting time sequence, using time series plots and visual inspection of estimates of waiting time and patience distribution CDFs for different weekdays. We did not find evidence against the stationarity assumption.

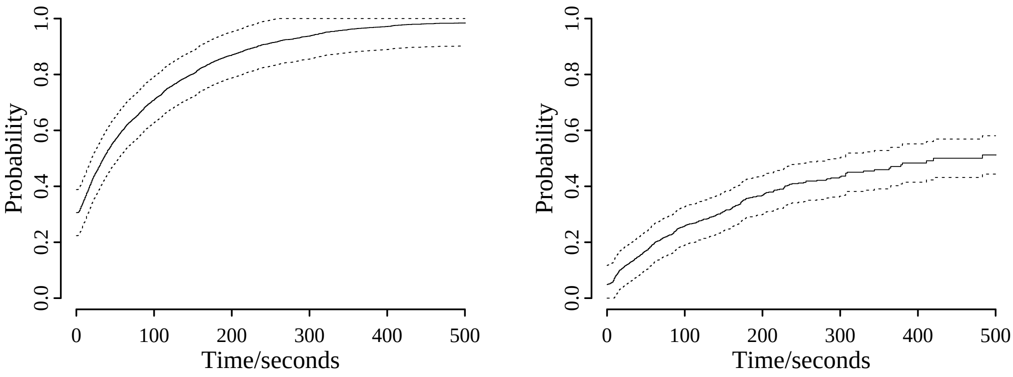

Figure 2. Left: Estimated waiting time CDF (solid line) with RBB-based 95% uniform confidence bands (dotted lines). Right: Estimated patience CDF (solid line) with RBB-based 95% uniform confidence bands (dotted lines). Observations used are for customers arriving between 2 p.m. and 3 p.m. on ordinary weekdays (Sunday-Thursday).

In the following, estimates are presented as estimate (95% confidence interval). All interval estimates were constructed using the percentile method, using 4,000 replications using the RBB on the 36 blocks. The product-limit estimates with uniform 95% confidence bands for the waiting time and patience CDFs are shown in Figure 2. The median waiting time was 37 seconds (21-53), while the probability of waiting more than 3 minutes was 0.15 (0.11-0.20). The tail of the waiting time distribution is reasonably well estimated (Figure 2, left), so in this case it is meaningful to estimate the expected waiting time using the tail formula (truncating the product-limit estimate at the largest observation). The value was 81 seconds (63-98). The 20th upper percentile of the patience distribution was 52 (47-86), while the probability of having a patience greater than 3 minutes was 0.64 (0.61-0.68).

To illustrate the application of the RBB two-sample test, we considered comparison

Table 1. Observed coverage of RBB confidence intervals for functionals of the patience CDF F in an $M/M/15$ queue with an arrival rate of 13.5 customers per minute, a service rate of 1 customer per minute, and exponential patiences. The parameter of each patience distribution was adjusted to provide the given rate of abandonment. Each figure is based on 500 independent simulations of a sequence of 25 IID blocks of average length 200. For each simulation, 4000 bootstrap replications were used. The statistic $\|F\|_\infty$ was calculated over the fixed interval $[0,1.5]$.

		Coverage					
		Waiting times			Patiences		
Abandonment	$1-\alpha$	$F(1)$	$F^{-1}(0.5)$	$\|F\|_\infty$	$F(1)$	$F^{-1}(0.2)$	$\|F\|_\infty$
20%	0.90	0.84	0.97	0.85	0.87	0.90	0.93
	0.95	0.91	0.92	0.91	0.93	0.94	0.96
10%	0.90	0.87	0.88	0.85	0.88	0.86	0.90
	0.95	0.91	0.94	0.92	0.94	0.90	0.96
5%	0.90	0.84	0.86	0.82	0.85	0.33	0.90
	0.95	0.89	0.93	0.90	0.91	0.26	0.95

Table 2. Observed level and power of the RBB two-sample test for detecting a difference of Δ between patience CDFs in an $M/M/15$ queue with an arrival rate of 13.5 customers per minute, a service rate of 1 customer per minute, and exponential/lognormal patience distributions. Parameters of the three reference patience distributions were adjusted to provide the given rates of abandonments (logarithmic variance of lognormal distribution fixed to 1). Comparison distributions were chosen stochastically larger than their reference distributions. Each figure is based on 500 independent simulations of a sequence of 25 IID blocks of expected length 200. For each simulation, 4000 bootstrap replications were used. The two-sample test statistic was calculated over the interval $[0,1.5]$.

			Exponential patience				Lognormal patience			
Abandonment	$1-\alpha$		Level	Power to detect Δ			Level	Power to detect Δ		
		Δ		0.05	0.10	0.20		0.05	0.10	0.20
20%	0.90		0.93	0.53	0.91	1.00	0.94	0.55	0.89	1.00
	0.95		0.98	0.31	0.79	0.99	0.99	0.31	0.80	0.99
10%	0.90		0.92	0.37	0.81	0.98	0.95	0.47	0.89	1.00
	0.95		0.96	0.22	0.65	0.95	0.98	0.38	0.79	0.98
5%	0.90		0.93	0.38	0.62	0.95	0.93	0.59	0.93	1.00
	0.95		0.97	0.11	0.46	0.87	0.97	0.41	0.85	0.99

of patience CDFs of two different priority groups of stock market customers. We used censored waiting times collected on ordinary weekdays (Sunday-Thursday) in the period 8 a.m.–8 p.m. The large time interval was used to obtain a reasonable number of observed patiences, although waiting times are unlikely to be stationary over such an interval. For the framework of this paper, however, nonstationarity is not a theoretical issue: we only require blocks to be stationary (and independent), corresponding to the heuristic assumption that the different days of operation are 'stochastically similar'. We obtained 36 blocks of average length 170. Product-limit estimates of the CDFs are shown in Figure 3, left. Using 4,000 replications in the RBB, we accepted the hypothesis of equal patience distributions, with a p-value of 0.07. To further explore the nature of the (nonsignificant) difference between the patience distribution, their absolute difference was plotted alongside a uniform 95% confidence band (Figure 3, right). There appears to be a borderline significant discrepancy around 500 seconds, indicating that patience distributions for the two customer classes may differ in the tails.

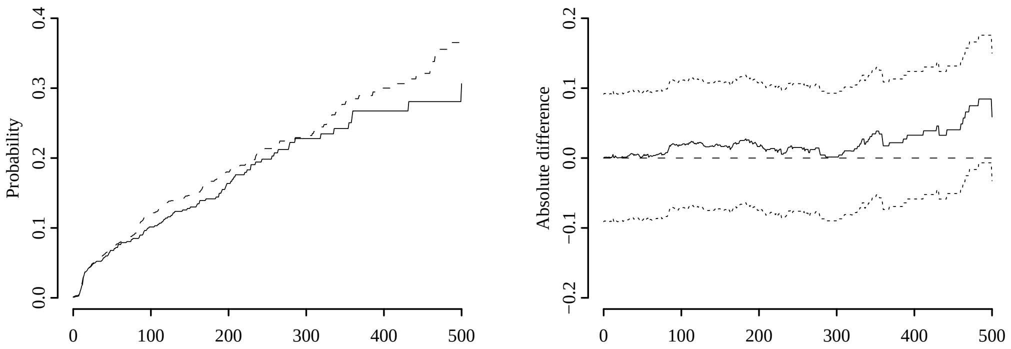

Figure 3. Left: Estimated patience CDF for regular stock market customers (thick line) and priority stock market customers (thin line) arriving between 8 am and 8 pm on ordinary weekdays (Sunday-Thursday). Right: Estimated absolute distance between the two priority groups' CDFs (solid line) with uniform 95% confidence band (dotted lines).

Appendix: validity of the RBB for empirical processes

For definiteness, we assume the regenerative sequence X to be defined canonically in terms of the cycles $\{C_n : n \in \mathbb{N}_0\}$ which are given by the coordinate sequence on an infinite product space $(\Omega,\mathscr{B},\mathbb{Q}) := (\tilde{\Omega},\tilde{\mathscr{G}},\mathbb{Q}') \otimes \prod_{n\geqslant 1}(\tilde{\Omega},\tilde{\mathscr{G}},\mathbb{Q}^*)$ where $\tilde{\Omega} = \bigcup_{m\geqslant 0}\mathbb{R}^m$ and $\tilde{\mathscr{G}}$ is the natural σ-algebra generated by $\bigcup_{n\geqslant 1}\mathscr{B}^n$ for the Borel σ-algebra $\mathscr{B}$ on $\mathbb{R}$. Recall that the total variation limit $\mathbb{P}$ is defined in (2).

The empirical process corresponding to the empirical measure (3) for a class of real-valued measurable functions $\mathscr{F}$ with values in $\mathbb{R}$ is the $\mathscr{F}$-indexed stochastic process $\{\mathrm{G}_n(f) : f \in \mathscr{F}\}$ where $\mathrm{G}_n(f) := n^{1/2}\mathrm{P}_n(f)$. The corresponding bootstrapped empirical process $\{\mathrm{G}_n^*(f) : f \in \mathscr{F}\}$ is given by

$$\mathrm{G}_n^*(f) := n_*^{1/2}\left(n_*^{-1}\sum_{i=1}^{n_*} f(X_i^*) - \underline{n}^{-1}\sum_{i=1}^{\underline{n}} f(X_i)\right), \qquad n \in \mathbb{N},\ f \in \mathscr{F};$$

with n_* and $\{X_i^* : i = 1,\ldots,n_*\}$ obtained from Algorithm 1, and $\underline{n} := T_{\tau_n+1}$. Each of G_n and

G_n^* are viewed as functions with values in the metric space $\ell^\infty(\mathscr{F})$ of uniformly bounded real-valued functions on $\mathscr{F}$ equipped with the uniform norm $\|K\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} |K(f)|$.

The theorem below is the bootstrap variant of the uniform CLT by Leventhal (1988) and Tsai (1998). It bears some similarities to the bootstrap uniform CLT by Radulović (2004) for a class of empirical processes with observations from a discrete atomic Markov chain. However, our method of proof is distinct from his in that we avoid assuming mixing properties for the regenerative sequence and imposing bracketing conditions on the function class $\mathscr{F}$. Also, our approach uses Poissonization, so that we can use the strategy of Giné and Zinn (1990) to give a concise proof based on multiplier inequalities.

For a measure γ on $(\mathbb{R},\mathscr{B})$, the $\mathscr{L}^p(\gamma)$ ε-covering number $N_p(\mathscr{F},\varepsilon,\gamma)$ of $\mathscr{F}$ for some $\varepsilon > 0$ is the smallest number of $\mathscr{L}^p(\gamma)$ ε-balls needed to cover $\mathscr{F}$. The following combinatorial entropy is due to Pollard (1982)

$$N_p(\varepsilon,\mathscr{F}) := \sup_\gamma N_p(\mathscr{F},\varepsilon,\gamma),$$

where the supremum runs over finitely supported measures γ on $(\mathbb{R},\mathscr{B})$. Recall that an envelope function F for $\mathscr{F}$ is any (measurable) real-valued function on Λ satisfying $f(\lambda) \leq F(\lambda)$ for all λ and f. To simplify our derivation, we assume in the following that $\mathscr{F}$ is sufficiently regular to ensure measurability of suprema of processes. Following Leventhal (1988) (see also Pollard (1984), Appendix C), we require that $\mathscr{F}$ is permissible, i.e. that $\mathscr{F}$ can be indexed by an analytic subset T of a compact metric space equipped with the Borel σ-field such that the evaluation map $(t,x) \mapsto f_t(x)$, $t \in T$, $x \in \mathbb{R}$, is jointly measurable.

THEOREM A1. *Suppose that* $\mathbb{E}(\ell_1^2) < \infty$. *Let* $\mathscr{F}$ *be a class of measurable real-valued functions on* $\mathbb{R}$ *with envelope function* F *such that*

$$\int_0^\infty \sqrt{\log N_2(\varepsilon,\mathscr{F})}\,\mathrm{d}\varepsilon < \infty, \qquad \mathbb{E}\Big\{\sum_{i=T_1+1}^{T_2} F(X_i)\Big\}^2 < \infty.$$

Under further measurability assumptions on $\mathscr{F}$, *there exists a tight, centered Gaussian process* $\mathrm{H}_\mathbb{P}$ *on* $\ell^\infty(\mathscr{F})$ *such that* $\mathrm{G}_n \to^d \mathrm{H}_\mathbb{P}$ *where* $\to^d$ *denotes weak convergence in* $\ell^\infty(\mathscr{F})$, *and the RBB is valid for the empirical process of* G_n *in the sense that, for d dual bounded Lipschitz distance on* $\ell^\infty(\mathscr{F})$ *(van der Vaart and Wellner, 1996, p. 73), it holds that*

$$d(\mathrm{G}_n^*,\mathrm{H}_\mathbb{P}) \to 0, \quad \textit{in probability } (\mathbb{Q}). \tag{A1}$$

Proof. By Theorem 4.3 of Tsai (1998), the hypotheses imply that G_n converges weakly in $\ell^\infty(\mathscr{F})$ to a tight, centered Gaussian process $\mathrm{H}_\mathbb{P}$. Following Giné and Zinn (1990), bootstrap validity holds if we can show the analogue of (A1) for the finite-dimensional distributions of G_n^* and stochastic asymptotic equicontinuity in probability with respect to a totally bounded semimetric ρ on $\mathscr{F}$. The latter means that

$$\lim_{\delta\downarrow 0}\lim_n \|\mathrm{G}_n^*\|_{\mathscr{F}_\delta} = 0, \quad \text{in probability } (\mathbb{Q}),$$

where $\|K\|_{\mathscr{F}_\delta} := \sup\{|K(f) - K(g)| : \rho(f,g) < \delta\}$ for $K \in \ell^\infty(\mathscr{F})$. Additionally, it must hold that ρ makes $\mathrm{H}_\mathbb{P}$ uniformly equicontinuous. As shown in Tsai (1998), our assumptions imply that $\mathscr{F}$ is totally bounded in $\mathscr{L}^2(\mathbb{P})$ and G_n asymptotically $\mathscr{L}^2(\mathbb{P})$-equicontinuous. By basic properties of $\mathscr{L}^p$-seminorms, both properties also hold for $\mathscr{L}^1(\mathbb{P})$-seminorm. Theorem 1.5.7 of van der Vaart and Wellner (1996) then implies that $\mathrm{H}_\mathbb{P}$ is uniformly $\mathscr{L}^1(\mathbb{P})$-equicontinuous. So we can use $\mathscr{L}^1(\mathbb{P})$-seminorm in the definition of $\|\cdot\|_{\mathscr{F}_\delta}$.

The result (A1) for finite-dimensional distributions follows from the Cramér-Wold device (Billingsley, 1995, Theorem 29.4) and Theorem 2.1 in Radulović (2004). The latter concerns convergence of finite-dimensional distributions for observations from a discrete Markov chain; using basic asymptotics of renewal/regenerative processes (Asmussen, 2003, Section V.6 and VI.3), the proof also works for regenerative sequences.

We proceed to show stochastic $\mathscr{L}^1(\mathbb{P})$-equicontinuity of G_n^*. Define for $j = 1,\dots,\tau_n$ stochastic processes

$$Z_j(f) := \sum_{i=T_j+1}^{T_{j+1}} f(X_i), \quad Z_j^*(f) := \sum_{i=T_j^*+1}^{T_{j+1}^*} f(X_i^*), \quad f \in \mathscr{F}.$$

Denote by γ the distribution of the bootstrapped observations obtained from Algorithm 1 and by $\mathbb{E}_\gamma$ expectation with respect to γ and take $\mu := \mathrm{E}(\ell_1)$. Define $a_n := [n/\mu]$. Then

$$\begin{aligned}\|(n_*/a_n)^{1/2}\mathrm{G}_n^*\|_{\mathscr{F}_\delta} \quad &\leqslant \quad \left\|a_n^{-1/2}\sum_{i=1}^{\tau_n}(Z_i^* - Z_i)\right\|_{\mathscr{F}_\delta} + (\tau_n/a_n)^{1/2}\left\|Y_n a_n^{-1}\sum_{i=1}^{\tau_n}(Z_i - \mu\mathbb{P})\right\|_{\mathscr{F}_\delta} \\ &\quad\quad + (\tau_n/a_n)^{3/2}\|Y_n\mu\mathbb{P}\|_{\mathscr{F}_\delta} + (n_*/\underline{n})\underline{a}_n^{-1/2}\sum_{i=T_0+1}^{T_1}|F(X_i)| \\ &=: \quad A(n,\delta) + (\tau_n/a_n)^{1/2}B(n,\delta) + (\tau_n/a_n)^{3/2}C(n,\delta) + (n_*/\underline{n})D(n),\end{aligned}$$

where $Y_n := (a_n/\underline{n}) \times \tau_n^{-1/2}(\underline{n} - n_*)$. By Slutsky's lemma for the bootstrap (Radulović, 2004, Lemma 3.1), it is enough to show convergence in probability as $n \to \infty$, $\delta \downarrow 0$ of $A(n,\delta)$, $B(n,\delta)$, $C(n,\delta)$, and $D(n)$ separately.

It is immediate that $D(n) \to 0$ almost surely. Concerning $C(n,\delta)$, define $\bar{\ell}_{\tau_n} = \tau_n^{-1}\sum_{i=1}^{\tau_n}\ell_i$. Then $n_* - \underline{n} = \sum_{i=1}^{\tau_n}(\ell_i^* - \bar{\ell}_{\tau_n})$ is of order $O_{\mathbb{Q}}(\sqrt{n})$ as $n \to \infty$. This follows since the ℓ_i^*s are conditionally IID, so that by Markov's inequality

$$\gamma\left(\sum_{i=1}^{\tau_n}\ell_i^* > \sqrt{n}M\right) \leqslant \tau_n\gamma\big(\ell_1^* > \sqrt{n}M\big) \leqslant M^{-2}n^{-1}\sum_{i=1}^{\tau_n}\ell_i^2 \to 0, \qquad n, M \to \infty;$$

almost surely, by the Law of Large Numbers. Slutsky's lemma for bootstrapped processes (Radulović, 2004, Lemma 3.1) then implies $Y_n = O_{\mathbb{Q}}(1)$. Recalling our choice of semimetric in the definition of $\|\cdot\|_{\mathscr{F}_\delta}$, we obtain $C(n,\delta) \leqslant |Y_n|\mu\delta$ which converges to zero in probability as $n \to \infty$, $\delta \downarrow 0$.

Convergence of $B(n,\delta)$ to zero in probability follows from Slutsky's lemma and arguments as in the proof of Lemma 4.6 of Tsai (1998). Since $Y_n = O_{\mathbb{Q}}(1)$, we have $\lim_{\delta\downarrow 0}\lim_n \|B(n,\delta)\|_{\mathscr{F}_\delta} = 0$ in probability.

Finally, regarding $A(n,\delta)$, fix $\varepsilon > 0$, $\delta > 0$. By Markov's inequality

$$\begin{aligned}\gamma\left(\left\|a_n^{-1/2}\sum_{i=1}^{\tau_n}(Z_i^* - Z_i)\right\|_{\mathscr{F}_\delta} > \varepsilon\right) &\leqslant \gamma\left(\left\{\left\|a_n^{-1/2}\sum_{i=1}^{\tau_n}(Z_i^* - Z_i)\right\|_{\mathscr{F}_\delta} > \varepsilon\right\} \cap \{|\tau_n - a_n| \leqslant a_n\}\right) \\ &\quad + \mathbb{1}(|\tau_n - a_n| > a_n) \\ &\leqslant \varepsilon^{-1}\mathbb{E}_\gamma\left(\left\|a_n^{-1/2}\sum_{i=1}^{\tau_n}(Z_i^* - Z_i)\right\|_{\mathscr{F}_\delta}\mathbb{1}(|\tau_n - a_n| \leqslant a_n)\right) + o_{\mathbb{Q}}(1).\end{aligned}$$

To bound the last expectation, we use Poissonization. Let $\{N_n\}$ be a sequence of IID symmetrised Poisson random variables with parameter 1/2 independent of X, T, defined on the same probability space. To simplify notation, we implicitly assume all of the calculations in the following to be conditionally on $|\tau_n - a_n| \leqslant a_n$. By Lemma 3.6.6 of van der Vaart and Wellner (1996),

$$\mathbb{E}_\gamma \left\| a_n^{-1/2} \sum_{i=1}^{\tau_n} (Z_i^* - Z_i) \right\|_{\mathscr{F}_\delta} \leqslant 4\mathbb{E}_N \left\| a_n^{-1/2} \sum_{i=1}^{\tau_n} N_i Z_i \right\|_{\mathscr{F}_\delta}.$$

Since $\mathbb{E}\|W_1\|_{\mathscr{F}_\delta} \leqslant \mathbb{E}\|W_1 + W_2\|_{\mathscr{F}_\delta}$ for centered, independent processes W_1, W_2 by Jensen's inequality,

$$\begin{aligned} \mathbb{E}_N \left\| a_n^{-1/2} \sum_{i=1}^{\tau_n} N_i Z_i \right\|_{\mathscr{F}_\delta} &\leqslant \mathbb{E}_N \left\| a_n^{-1/2} \sum_{i=1}^{a_n} N_i Z_i \right\|_{\mathscr{F}_\delta} + \mathbb{E}_N \left\| a_n^{-1/2} \sum_{i=a_n+1}^{\tau_n} N_i Z_i \right\|_{\mathscr{F}_\delta} \\ &\leqslant 2\mathbb{E}_N \left\| a_n^{-1/2} \sum_{i=1}^{a_n} N_i Z_i \right\|_{\mathscr{F}_\delta}. \end{aligned}$$

Taking expectations $\mathbb{E}_X$ with respect to X, T everywhere, conclude that for some universal constant C

$$\mathbb{E}_X \gamma \left(\left\| a_n^{-1/2} \sum_{i=1}^{\tau_n} (Z_i^* - Z_i) \right\|_{\mathscr{F}_\delta} > \varepsilon \right) \leqslant C\varepsilon^{-1} \mathbb{E} \left\| a_n^{-1/2} \sum_{i=1}^{a_n} N_i Z_i \right\|_{\mathscr{F}_\delta}.$$

The multiplier inequality argument in the proof of Theorem 3.6.3 of van der Vaart and Wellner (1996) implies convergence to zero of the right hand side of the display as $n \to \infty, \delta \downarrow 0$. This proves stochastic equicontinuity in probability of $A(n, \delta)$ and so $\mathbb{G}_n^*$ is stochastically equicontinuous in probability ($\mathbb{Q}$). Combining this with convergence of finite-dimensional distributions, we obtain the desired result. ■

References

Asmussen, S. (2003) *Applied Probability and Queues*. Springer, second edn.

Athreya, K. B. and Fuh, C. D. (1989) Bootstrapping Markov chains: countable case. Tech. Rep. 89-7, Institute of Statistical Science, Academia Sinica, Taiwan.

Bertail, P. and Clémençon, S. (2006) Regenerative block-bootstrap for Markov chains. *Bernoulli*, **12**, 689–712.

Bertail, P. and Clémençon, S. (2007) Second order properties of regeneration-based bootstrap for Markov chains. *Test*, **16**, 109–122.

Billingsley, P. (1995) *Probability and Measure*. Wiley, third edn.

Bingham, N. H. and Pitts, S. M. (1999a) Nonparametric estimation for the $M/G/\infty$ queue. *Ann. Inst. Statist. Math.*, **51**, 71–97.

Bingham, N. H. and Pitts, S. M. (1999b) Nonparametric inference from the $M/G/1$ busy periods. *Commun. Statist. Stoch. Models*, **15**, 247–272.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005) Statistical analysis of a telephone call center: A queuing science perspective. *J. Amer. Statist. Assoc.*, **100**, 36–50.

Cai, J. and Kim, J. (2003) Nonparametric quantile estimation with correlated failure time data. *Lifetime Data Anal.*, **9**, 357–371.

Cai, Z. (1998) Kernel density estimation and hazard rate estimation for censored dependent data. *J. Multivariate Anal.*, **67**, 23–34.

Cai, Z. (2001) Estimating a distribution function for censored time series data. *J. Multivariate Anal.*, **78**, 299–318.

Datta, S. and McCormick, W. P. (1993) Regeneration-based bootstrap for Markov chains. *Canad. J. Statist.*, **21**, 181–193.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

Gans, N., Koole, G. and Mandelbaum, A. (2003) Commissioned paper: Telephone call centers: Tutorial, review and research prospects. *Manuf. Serv. Op.*, **5**, 79–141.

Garnett, O., Mandelbaum, E. and Reiman, M. (2002) Designing a call center with impatient customers. *Manuf. Serv. Op.*, **4**, 208–228.

Gill, R. and Johansen, S. (1990) A survey of product-integration with a view towards application in survival analysis. *Ann. Statist.*, **18**, 1501–1555.

Giné, E. and Zinn, J. (1990) Bootstrapping general empirical measures. *Ann. Probab.*, **18**, 851–869.

Hall, P. (1992) *The Bootstrap and Edgeworth Expansion*. Springer.

Hansen, M. B. and Pitts, S. M. (2006) Nonparametric inference from the $M/G/1$ workload. *Bernoulli*, **12**, 737–759.

Leventhal, S. (1988) Uniform limit theorems for Harris recurrent Markov chains. *Probab. Theory Rel. Fields*, **80**, 101–118.

Naik-Nimbalkar, U. V. and Rajarshi, M. B. (1994) Validity of blockwise bootstrap for empirical processes with stationary observations. *Ann. Statist.*, **22**, 980–994.

Nederlof, A. and Anton, J. (2002) *Customer Obsession: Your Roadmap to Profitable CRM*. Anton Press.

Pollard, D. (1982) A central limit theorem for empirical processes. *J. Austr. Math. Soc. (Ser. A)*, **33**, 235–248.

Pollard, D. (1984) *Convergence of Stochastic Processes*. Springer.

Radulović, D. (2004) Renewal type bootstrap for Markov chains. *Test*, **13**, 147–192.

Thorrison, H. (2000) *Coupling, Stationarity and Regeneration*. Springer.

Tsai, T. H. (1998) *The Uniform CLT and LIL for Markov Chains*. Ph.D. thesis, University of Wisconsin.

van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. Springer.

Wolf, M., Politis, D. N. and Romano, J. P. (1999) Weak convergence of dependent empirical measures with application to subsampling in function spaces. *J. Statist. Plann. Infer.*, **79**, 179–191.

Paper II

Why FARIMA Models are Brittle

Author list

Anders Gorst-Rasmussen
Aalborg University, Denmark

Darryl Veitch
University of Melbourne, Australia

András Gefferth
Morgan Stanley Analytics Hungary, Hungary

Summary

The FARIMA models, which have long-range-dependence (LRD), are widely used in many areas. Through deriving a precise characterisation of the spectrum, autocovariance function, and variance time function, we show that this family is atypical among LRD processes, being extremely close to the fractional Gaussian noise in a precise sense. Furthermore, we show that this closeness property is not robust to additive noise. We argue that the use of FARIMA, and more generally fractionally differenced time series, should be reassessed in some contexts, in particular when convergence rate under rescaling is important and noise is expected.

Supplementary info

The work on this manuscript began in 2007-2008 during a visit to Centre for Ultra-Broadband Information Networks (CUBIN), University of Melbourne, Australia, but was not finished completely until early 2011.

We plan to submit the manuscript to *Fractals*.

1. Introduction

For a wide variety of purposes including data modelling, synthetic data generation, and the testing of statistical estimators, tractable and flexible time series models are indispensable. The well known autoregressive moving average (ARMA) family, for example, allows for a wide variety of short range correlation structures, and has been used in many contexts. Long-range dependence (LRD), or long memory, in stationary time series is a phenomenon of great importance (Taqqu, 2002). The fractional autoregressive integrated moving average (FARIMA) models (Hosking, 1981; Granger and Joyeux, 1980) are very widely used as a class which inherits the advantages of ARMA, while exhibiting LRD with tunable Hurst parameter, the scaling parameter of LRD. They have in particular been widely used to parsimoniously model data sets exhibiting LRD (Ilow, 2000), and more importantly for our purposes here, they have also been employed to make quantitative assessments of the behaviour of stochastic systems in the face of LRD (for example Barbe and McCormick (2010)). A good example is in relation to estimators of the Hurst parameter H. FARIMA models have been used in order to evaluate the performance of H estimators under circumstances more challenging than that of the canonical fractional Gaussian Noise (fGn), in particular to assess small sample size performance using Monte Carlo simulation (for example

Taqqu *et al.* (1995); Taqqu and Teverovsky (1997); Abry *et al.* (2003)). Although explicit claims of the generality of the FARIMA family are not made, implicitly it is taken to be a typical class of LRD time series in some sense, and so results obtained using it are taken to be representative for LRD inputs in general.

In fact, no parametric model can be truly typical. However, for a model class to be useful, it should be representative for the purposes to which it is commonly put. In this paper, we show that FARIMA time series, and more generally time series whose LRD scaling derives directly from fractional differencing such as the FEXP models (Robinson, 1994), are far from typical when it comes to their LRD character, the very quality for which they were first introduced. In a sense we make precise, out of all possible LRD time series, their LRD behaviour is in fact 'as close as possible' to that of fGn. A key technical consequence is ultra-rapid convergence to fGn under the rescaling operation of aggregation. The implications for the role of the family is strong, namely that, in regards to LRD behaviour, *FARIMA offers no meaningful diversity beyond fGn*. A second key consequence is that the addition of additive noise (of almost any kind) pushes a FARIMA process out of the immediate neighbourhood of fGn, changing the convergence rate. In other words, FARIMA is structurally unstable in this sense, or brittle, and is therefore unsuited for use as a class of LRD time series representing real-world signals.

This work arose out of our prior study of (second-order) self-similarity of stationary time series (Gefferth *et al.*, 2003), which highlighted the benefits of the variance time function (VTF) formulation of the autocovariance structure, over the more commonly used autocovariance function (ACVF) formulation. Using the VTF, questions of process convergence under rescaling to exactly (second-order) self-similar limits can often be more simply stated and studied. The paper is structured as follows. After Section 2 on background material, Section 3 establishes the main results. It begins by characterising a link between a fractionally differenced process and fGn in the spectral domain. Using it, we prove that related Fourier coefficients in the time domain decay extremely quickly, and then show that as a result the VTFs of the fractionally differenced process and fGn are extremely close. We then explain why this behaviour is so atypical, and how it results in fast convergence to fGn. Finally, we go on to provide distinct direct proofs of closely related results for the ACVF and spectral formulation which are of independent interest. In particular, they lead to additional closeness results for the spectrum. In Section 5, we explain why fractional processes are not robust to the addition of additive noise, even noise of particularly non-intrusive character. We also provide numerical illustrations of this brittleness, and of the fast convergence to fGn of FARIMA processes. We discuss possible implications of our findings in Section 6.

Very early versions of this work appear in the papers Gefferth *et al.* (2002a,b).

2. Background

Let $\{X(t) : t \in \mathbb{Z}\}$ be a discrete-time second-order stationary stochastic process. The mean μ and variance $\mathcal{V} > 0$ of such a process are independent of t, the autocovariance function (ACVF), $\gamma(k) := \mathbb{E}[\{X(t)-\mu\}\{X(t+k)-\mu\}]$, depends only on the lag $k \in \mathbb{Z}$, and $\gamma(k) = \gamma(-k)$.

A description of the autovariance structure which is entirely equivalent to γ is the variance time function, defined as $\omega(n) = (\mathbf{I}\gamma)(n) := \sum_{k=0}^{n-1}\sum_{i=-k}^{k}\gamma(i)$, $n = 1, 2, 3, \ldots$, where $\mathbf{I}$ denotes the double integration operator acting on sequences. Its normalised form, the correlation time function (CTF), is just $\phi(n) = \omega(n)/\omega(1) = \omega(n)/\mathcal{V}$. In terms of the original process, $\omega(n)$ is just the variance of the sum $\sum_{t=1}^{n} X(t)$. It is convenient to symmetrically extend ω and ϕ to $\mathbb{Z}$ by setting $\omega(n) := \omega(-n)$ for $n < 0$ and $\omega(0) = 0$.

2.1. *LRD, second-order self-similarity, and comparing to fGn*

There are several definitions of long-range dependence, all of which encapsulate the idea of slow decay of correlations over time. Common definitions include power-law tail decay of the ACVF, $\gamma(n) \overset{n\to\infty}{\sim} c_\gamma n^{2H-2}$, or power-law divergence of the spectral density at the origin, $f(x) \overset{x\to 0}{\sim} c_f |x|^{-(2H-1)}$ for related constants c_γ and c_f (Taqqu (2002), Section 4).

The well known fractional Gaussian noise (fGn) family, parameterised by the Hurst parameter $H \in [0,1]$ and variance $\mathcal{V} > 0$, has $\omega(m) = w^*_{H,\mathcal{V}}(m) := \mathcal{V} m^{2H}$ (to lighten notation we usually write ω^*_H or simply ω^*). It has long memory iff $H \in (1/2, 1]$. In this paper we compare against fGn with $H \in (1/2, 1]$ as it plays a special role among among LRD processes; that of being a family of second-order self-similar time series[1]. To understand how this comparison can be made, we must define self-similarity and related notions.

Self-similarity relates to invariance with respect to a rescaling operation. In the present context, the time rescaling is provided by what is commonly called aggregation. For a fixed $m \in \mathbb{N}$, the aggregation of level m of the original process X is the process $X^{(m)}$ defined as

$$X^{(m)}(t) := \frac{1}{m} \sum_{j=m(t-1)+1}^{mt} X(j).$$

The functions γ, ω, ϕ and the variance of the m-aggregated process will be written $\gamma^{(m)}$, $\omega^{(m)}$, $\phi^{(m)}$ and $\mathcal{V}^{(m)}$ respectively. It is not difficult to show (Gefferth *et al.*, 2003) that

$$\omega^{(m)}(n) = m^{-2}\omega(mn), \qquad \mathcal{V}^{(m)} = m^{-2}\omega(m).$$

To seek invariance, the time rescaling must be accompanied by a compensating amplitude rescaling. This is performed naturally by dividing by $\mathcal{V}^{(m)}$, which amounts to examining the effect of aggregation on the correlation structure. Combining the time and amplitude rescalings yields the correlation renormalisation

$$\phi^{(m)}(n) = \frac{\phi(mn)}{\phi(m)} = \frac{\omega(mn)}{\omega(m)}. \tag{1}$$

We can now define second-order self-similarity as the fixed points of this operator.

DEFINITION 1. A process is second-order self-similar iff $\phi^{(m)} = \phi$, for all $m \in \mathbb{N}$.

Clearly fGn, with $\phi(m) = \phi^*_H(m) := m^{2H}$, satisfies this definition for all $H \in [0,1]$.

Given a fixed point ϕ^*_H, we define its domain of attraction (DoA) to be those time series which converge to it pointwise under the action of (1). This definition is very general, in particular it includes processes whose VTFs have divergent slowly varying prefactors, as these cancel following normalisation (see Section 3.3). It provides a natural way to define LRD which subsumes and generalises most other definitions including those above (Gefferth *et al.*, 2003): a time series is long-range dependent if and only if it is in the domain of attraction of ϕ^*_H for some $H \in (0.5, 1]$.

With the above definitions, the DoA are revealed as the natural way to partition the space of all LRD processes, namely into sets of processes each corresponding to the same unique normalised fGn fixed point. Since all processes within a DoA converge to the same fixed point, their asymptotic structure can be meaningfully compared both against each other and to the fixed point itself. Alternatively, if two processes were in different DoAs then they cannot be close asymptotically as they would converge to different processes. Section 3.2 provides a precise characterisation of the closeness of a fractionally differenced process to its corresponding fixed point, and its associated fast convergence under renormalisation.

---

[1] Until recently, fGn was considered to be the only such family. A second (and final) family was discovered recently (Gefferth *et al.*, 2004).

Within a given DoA, one can further partition processes according to some measure of distance from the common fixed point. Section 3.3 establishes such a notion, enabling a comparison of this closeness to that of other members of the DoA to be made.

2.2. *Fractionally differenced processes and FARIMA*

Let B denote the backshift operator and Γ the gamma function. The fractional differencing operator of order $d > -1$ is given by

$$(1-B)^d := \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} B^j.$$

Let $\{Y(t) : t \in \mathbb{Z}\}$ be a second-order stationary stochastic process. If $H \in (0,1)$ then the process

$$X := (1-B)^{-(H-1/2)} Y$$

is called a fractionally differenced process with differencing parameter $H - 1/2$ driven by Y. If h is the spectral density of Y then X has spectral density (Brockwell and Davis (1991), Thm. 4.10.1)

$$f_H(x) = h(x)|1-\mathrm{e}^{2\pi i x}|^{-(2H-1)} = h(x)|2\sin \pi x|^{-(2H-1)}, \quad x \in [-1/2, 1/2]. \tag{2}$$

In this paper, we assume that Y is short-range dependent, and that h satisfies:

- $h(x) > 0$ and is continuous for all $x \in [-1/2, 1/2]$ (and is therefore bounded);
- h is three times continuously differentiable on $(-1/2, 1/2)$ (and is therefore in C^3).

Under such conditions, the ACVF of X exists and satisfies $\gamma_H(n) \sim c_\gamma n^{2H-2}$ for some constant c_γ (Brockwell and Davis (1991), Thm. 13.2.2). Hence, when $H \in (1/2, 1)$ the process X is LRD with Hurst parameter H.

An important example of a fractionally differenced process is the FARIMA class (Hosking, 1981) where h is the spectral density of a causal invertible ARMA model. This family includes the ARMA family as the special case $H = 1/2$. Another class is the class of FEXP models (for example Bloomfield (1973); Robinson (1994); Beran (1993)) which comes from taking the logarithm of h to be a trigonometric polynomial, i.e. $\log h(x) = \theta_1 \cos x + \theta_2 \cos(2x) + \cdots + \theta_{q-1} \cos\{(q-1)x\}$ for real coefficients $\theta_1, \ldots, \theta_{q-1}$. FARIMA and FEXP models are widely used in statistical applications since, in addition to exhibiting LRD, they both enable modelling of arbitrary short-range correlation structures.

2.3. *Normalising a fractionally differenced process to its fGn limit*

To identify the fGn fixed point of a fractionally differenced time series, only the value of H need be determined. When aggregating an unnormalised fractionally differenced time series, however, to identify the corresponding limiting fGn time series we must in addition know the correct variance $\mathcal{V}$. The purpose of this section is to define notation to make this simple and along the way to provide useful expressions for the spectra of these processes.

The ACVF, VTF, and spectral density corresponding to the fixed point are denoted γ_H^*, ω_H^*, and f_H^*, respectively. The latter is given by (Samorodnitsky and Taqqu, 1994)

$$f_H^*(x) = c_f^* \pi^{-2}(2\pi)^{2H+1}\sin^2(\pi x)\sum_{j=-\infty}^{\infty}|2\pi j + 2\pi x|^{-(2H+1)} \overset{x\to 0}{\sim} c_f^*|x|^{-(2H-1)}, \quad x \in [-1/2, 1/2]; \tag{3}$$

where $c_f^* = \mathcal{V}(2\pi)^{2-2H}C(H) > 0$ is the prefactor of the power-law at the origin, and $C(H) = \pi^{-1}H\Gamma(2H)\sin(H\pi)$ (see Samorodnitsky and Taqqu (1994), pp. 333-4, but note that the change to normalised frequency multiplies f_H^* by 2π, and c_f^* by $(2\pi)^{2-2H}$).

We denote by γ_H, ω_H and f_H the ACVF, VTF, and spectral density of a fractional process with Hurst parameter $H \in (1/2, 1)$. In view of (2), the latter is given by

$$f_H(x) = h(x)|2\sin\pi x|^{-(2H-1)} = c_f(2\pi)^{2H-1}\frac{h(x)}{h(0)}|2\sin\pi x|^{-(2H-1)} \overset{x\to 0}{\sim} c_f|x|^{-(2H-1)}, \quad x \in [-1/2, 1/2], \tag{4}$$

where $c_f = (2\pi)^{1-2H}h(0) > 0$. In the case of a pure fractionally differenced process, denoted FARIMA$(0, d, 0)$, it holds that $h(x) = h(0) = 2\pi$, and $c_f = (2\pi)^{2-2H}$ (note again the changes related to normalised frequency, in particular the factor of 2π is built into $h(0)$).

To conclude, the particular fGn to which the fractionally differenced process will converge under renormalisation is the one such that $c_f^* = c_f$. From this, the value of $\mathcal{V}$ can be obtained using the expressions for c_f^* and c_f above, if needed.

2.4. *Regularity and other notations*

Denote for $\alpha \geqslant 0$ by Λ_α the normed space of uniformly α-Hölder continuous functions defined on $[-1/2, 1/2]$,

$$\Lambda_\alpha := \{\varphi\colon [-1/2, 1/2] \to \mathbb{R} : \|\varphi\|_{\Lambda_\alpha} < \infty\};$$

where $\|\cdot\|_{\Lambda_\alpha}$ is the α-Hölder norm

$$\|\varphi\|_{\Lambda_\alpha} := \sup_{x,y\in[-1/2,1/2]} |\varphi(x) - \varphi(y)||x-y|^{-\alpha}.$$

Clearly $\Lambda_\alpha \supseteq \Lambda_\beta$ whenever $\alpha \leqslant \beta$. The space Λ_α is closed under pointwise multiplication, addition, and composition with functions in Λ_1. In particular, the subset of Λ_α whose members are bounded away from zero is closed under reciprocation (i.e. if $g \in \Lambda_\alpha$, and g is bounded away from zero, then so is $1/g$). Observe that $\varphi \in \Lambda_1$ whenever φ' exists and is bounded. Functions in Λ_α are absolutely continuous.

The linear space V of functions of bounded variation on $[-1/2, 1/2]$ is defined by

$$\mathrm{V} := \{\varphi\colon [-1/2, 1/2] \to \mathbb{R} : \|\varphi\|_{\mathrm{V}} < \infty\},$$

where $\|\cdot\|_{\mathrm{V}}$ is the total variation norm:

$$\|\varphi\|_{\mathrm{V}} := \sup\left\{\sum_{i=1}^{n}|\varphi(x_i) - \varphi(x_{i-1})| : -1/2 \leqslant x_0 < x_1 < \cdots < x_n \leqslant 1/2, n \in \mathbb{N}\right\}.$$

The space V is also closed under pointwise multiplication and addition (Apostol (1974), Thm. 6.9), and reciprocation of those functions in V bounded away from zero (Apostol

(1974), Thm. 6.10). Any differentiable function with bounded derivative on $(-1/2, 1/2)$ is of bounded variation on $[-1/2, 1/2]$ (Apostol (1974), Thm. 6.6).

We shall use the notation $\star$ for convolution of sequences. For sequences a and b

$$(a \star b)_n = \sum_{j=-\infty}^{\infty} a_j b_{n-j}, \quad n \in \mathbb{Z}.$$

The convolution is said to exist if the infinite sum converges for all n. When needed for clarity, we also use $(a \star b)(n)$ to denote $(a \star b)_n$.

Throughout, by a smooth function, we mean one in C^∞.

3. Fractionally differenced processes are not typical LRD processes

The goal of this section is to establish our main results, rigorous characterisations of the closeness of the asymptotic covariance structure of a fractionally differenced process to that of fGn. Our approach is simple and can be described as follows. We begin in the spectral domain where the relationship between the processes can be simply stated through a function g by defining

$$f_H(x) = f_H^*(x) g(x). \tag{5}$$

The simple closed form of the spectra (3) and (4) allow g to be explicitly written. We study the properties of g, obtaining a characterisation of the closeness of the processes in the spectral domain (Theorem 1). This leads to a convolution formulation $\gamma_H = \gamma_H^* \star G$ in the time domain, where G is the sequence of Fourier coefficients of g, and thereby to a similar relationship for the VTFs, where the fast decay of the Fourier coefficients can be used to characterise the closeness (Theorem 2). The VTF result then allows the closeness within the DoA and the convergence speed to be easily established (Theorem 3). Finally we also provide direct closeness results for the ACVF (Theorem 4).

3.1. *Closeness of the spectrum*

We are ultimately interested in characterising the closeness of the covariance structure of a fractionally differenced process to that of its fGn fixed point at large lags. The rate of decay of the sequence of Fourier coefficients of a function is well known to be closely connected to its smoothness properties. It is, therefore, unsurprising that a notion of closeness in the spectral domain can take the form of statements about smoothness of the function g in (5).

The following spectral closeness result is the crucial basis for both the VTF and ACVF results to come.

THEOREM 1. *Assume that* $H \in [1/2, 1)$. *Define* $g(x) := f_H(x)/f_H^*(x)$ *for* $x \neq 0$ *and* $g(0) := \lim_{x\to 0} g(x)$. *Then it holds that* $g(0) = c_f/c_f^* = h(0)/\{2\pi \mathcal{V} C(H)\}$ *and* g *satisfies the following on* $[-1/2, 1/2]$*:*

(i). *g is even, continuous, positive, bounded, and L^p, $p > 0$;*
(ii). *g is twice differentiable, and smooth away from $x = 0$;*
(iii). $g'' \in \Lambda_{2H-1} \cap \mathrm{V}$, *but* $g'' \notin \Lambda_{\beta'}$ *for* $\beta' > 2H - 1$*;*
(iv). *g admits a Fourier series with coefficients $\{G_j\}$ such that $\sum_{j=-\infty}^{\infty} j^2 |G_j| < \infty$ and $G_n = O(n^{-3})$. In particular, $\sum_{j=-\infty}^{\infty} |G_j| < \infty$ and $\sum_{j=-\infty}^{\infty} j^\alpha |G_j| < \infty$ for $1 < \alpha < 2$.*

Proof. Unless otherwise specified, we consider the domain $x \in [-1/2, 1/2]$. First, since $f_H(x) \overset{x \to 0}{\sim} c_f^* |x|^{-(2H-1)}$ and $f_H^*(x) \overset{x \to 0}{\sim} c_f |x|^{-(2H-1)}$, $g(0) := \lim_{x \to 0} g(x) = c_f / c_f^*$.

The proof of (i) is straightforward; details are provided in the appendix. To prove the smoothness properties (ii) and (iii), we first establish those of $\tilde{g}$ defined as

$$\begin{aligned} (6) \qquad \tilde{g}(x) &:= \frac{c_f \pi^{2H+1}}{c_f^* h(0)} \cdot \frac{h(x)}{g(x)} \\ (7) \qquad &= \left| \frac{\sin(\pi x)}{\pi x} \right|^{2H+1} + |\sin(\pi x)|^{2H+1} \sum_{\substack{j=-\infty \\ j \neq 0}}^{\infty} |\pi j + \pi x|^{-(2H+1)} \\ (8) \qquad &:= |a(x)|^{2H+1} + |b(x)|^{2H+1} c(x). \end{aligned}$$

It is not difficult to show (see the appendix) that $\tilde{g}$ is smooth everywhere except at the origin where its smoothness is controlled by that of $|b|^{2H+1}$, which we now study.

Let $\beta = 2H - 1$. Since b is smooth and $\beta \in (0,1)$, $|b|^{\beta+2}$ is twice differentiable at the origin. The smoothness of its second derivative is controlled by $(b')^2 |b|^\beta$, which, since $b \in \Lambda_1$ and $x \mapsto |x|^\beta$ is in Λ_β, is also in Λ_β by the multiplicative and compositional closure properties of Λ_β. It follows that $\tilde{g}''$ exists and is in Λ_β. Since however $x \mapsto |x|^\beta$ is not in $\Lambda_{\beta'}$ for any $\beta' > \beta$, and moreover $b(x) \overset{x \to 0}{\sim} \pi x$ and $b'(0) \neq 0$, $\tilde{g}''$ is not in $\Lambda_{\beta'}$ for any $\beta' > \beta$.

Since smooth functions are in V, by similar arguments using the closure properties of V, we have $\tilde{g}'' \in \mathrm{V}$ if $|b|^\beta \in \mathrm{V}$. The latter holds since it is easy to see that $|b|^\beta$ is monotone (with total variation 2).

We have shown that $\tilde{g}''$ exists and is in $\Lambda_{2H-1} \cap \mathrm{V}$, but not in $\Lambda_{\beta'}$ for any $\beta' > 2H - 1$. We now prove the same for g using (6). It suffices to consider $1/\tilde{g}$ since h''' exists. Since $\tilde{g}$ is bounded away from zero, (ii) follows since $(1/\tilde{g})'' = 2(\tilde{g}')^2/\tilde{g}^3 - \tilde{g}''/\tilde{g}^2$ clearly exists, and is smooth away from the origin. Now consider (iii). It follows from the last expression and the fact that $\tilde{g} > 0$ that $(1/\tilde{g})''$ and hence g'' are in V and Λ_{2H-1} by applying the respective closure properties. Finally, since $1/\tilde{g}^2(0) \neq 0$, the smoothness of $(1/\tilde{g})''$ is controlled by that of $\tilde{g}''$ and so $(1/\tilde{g})'' \notin \Lambda_{\beta'}$ for any $\beta' > 2H - 1$. This completes the proof of (iii).

We now prove (iv). Since each of g, g', and g'' are continuous and bounded, the Fourier series for each exists and are related by term-by-term differentiation (Champeney (1990), Thm. 15.19). In particular, $g(x) = \sum_{j=-\infty}^{\infty} G_j e^{2\pi i j x}$, and we can write $g''(x) = -4\pi^2 \sum_{j=-\infty}^{\infty} j^2 G_j e^{2\pi i j x}$. Now Zygmund (2002), Thm. VI.3.6, states that the Fourier series of a function in $\Lambda_\beta \cap \mathrm{V}$ for some $\beta > 0$ converges absolutely. This applies to g'' and hence $\sum_{j=-\infty}^{\infty} j^2 |G_j| < \infty$ as claimed. Finally, since $g'' \in \mathrm{V}$, the magnitude of its Fourier coefficients decay as $O(|j|^{-1})$ (Zygmund (2002), Thm. II.4.12), so that $G_j = O(j^{-3})$. ■

The result suggests that fractionally differenced processes are not typical; for a general LRD process, only boundedness of g at the origin would be automatic. In contrast, the present g is a very well behaved function. A plot of g is provided in Figure 1 which shows its flatness at the origin (it also suggests that g is monotone increasing over $[0, 1/2]$, though this plays no role in what follows). Here we have set $c_f = c_f^*$, so that its value at the origin is just 1. It is interesting to note that since g is positive, even, and square integrable, it is the spectral density of some second-order stationary time series.

3.2. *Closeness of the VTF*

The first step in elucidating the relationship between ω_H and ω_H^* is to confirm that the relationship $f_H(x) = f_H^*(x) g(x)$ between the spectral densities translates to the expected convolution relationship $\gamma_H = \gamma_H^* \star G$ between the ACVFs. It is straightforward to confirm that, thanks to the nice behaviour of g and G detailed in Theorem 1, this is indeed the case.

LEMMA 1. *The ACVFs γ_H and γ_H^* are related through the convolution $\gamma_H = \gamma_H^* \star G$.*

For completeness, a proof is given in the appendix.

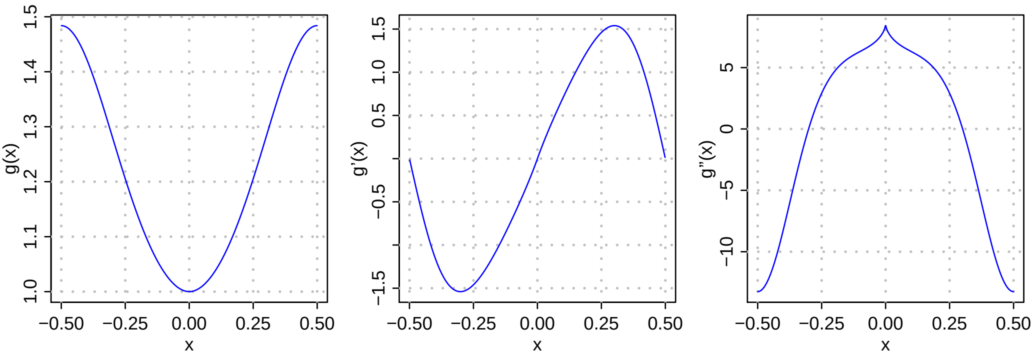

Figure 1. Graphs of the function $g(x) = f_H(x)/f_H^*(x)$ and its first two derivatives in the canonical case of a pure fractionally differenced process (FARIMA$(0,d,0)$) with $H = 0.8$ and $c_f = c_f^*$.

Since $\omega_H = \mathbf{I}\gamma_H$, it is tempting to seek a relationship of the form $\omega_H = G \star \omega_H^*$ through taking the 'double integral' of $\gamma_H = G \star \gamma_H^*$. However, since $\omega_H^*(m) = \mathcal{V} m^{2H}$ diverges with m, this operation is not necessarily well defined. The following lemma provides a sufficient condition for the existence of such a convolution, as well as some of its important properties which will be crucial in what follows.

LEMMA 2. *Assume $1 < \alpha < 2$ and let $a := \{|n|^\alpha : n \in \mathbb{Z}\}$. Let b be a symmetric sequence satisfying $\sum_{j=1}^\infty j^\alpha |b_j| < \infty$. Then $S_b := \sum_{j=-\infty}^\infty b_j$ and the symmetric sequence $c := a \star b$ exist, and $(c_n - S_b a_n) \overset{n\to\infty}{\to} 0$.*

For a proof, see the appendix. The proof of the last part is based on the monotonicity of a function which generalises γ_H^* to two parameters (Lemma A2 in the appendix).

COROLLARY 1. *The convolution $G \star \omega_H^*$ exists for $H \in (1/2, 1)$.*

Proof. Set $b = G$ in Lemma 2. The condition on b holds since $\sum_{j=1}^\infty j^\alpha |G_j| < \sum_{j=1}^\infty j^2 |G_j|$ which is finite, from Theorem 1. The result then following immediately by identifying α with $2H$ and a with ω_H^*. ■

The following lemma shows that, if existence is granted, taking the 'double integral' of a convolution is straightforward, provided that a double counting issue at the origin is allowed for.

LEMMA 3. *Let a, b be symmetric sequences and assume that $c := a \star b$ exists. Then $\mathbf{I}c$ exists, and if $(\mathbf{I}a) \star b$ exists, then $\mathbf{I}c = (\mathbf{I}a) \star b - \{(\mathbf{I}a) \star b\}_0$.*

The proof of this result is based on a careful rearrangement of terms justified by the repeated use of the existence of $(\mathbf{I}a) \star b$. It is given in the appendix.

We are now able to prove our main result on the VTF.

THEOREM 2. *Let* ω_H *denote the VTF of a fractionally differenced process for which* $H \in (1/2, 1)$ *and with* c_f *chosen equal to* c_f^*. *Then*

$$\omega_H(n) = \omega_H^*(n) + D + o(1);$$

where $D = -2\sum_{j=1}^{\infty} j^{\alpha}|G_j| < 0$ *is a constant.*

Proof. Since each of $\gamma_H^* \star G$ and $\omega_H^* \star G$ exist, Lemma 3 applies upon identifying $a = \gamma_H^*$, $b = G$ and $c = \gamma_H$ and states that $\omega_H = \omega_H^* \star G - \{\omega_H \star G\}(0)$. From Lemma 2 with $b = G$, $S_G = \sum_{j=-\infty}^{\infty} G_j < \infty$ exists. By introducing the term $S_G\omega_H^*$ we obtain

$$\omega_H = S_G\omega_H^* + \left(\omega_H^* \star G - S_G\omega_H^*\right) - (\omega_H \star G)(0) = S_G\omega_H^* + o(1) - 2\sum_{j=1}^{\infty} j^{\alpha}|G_j|,$$

by the final part of Lemma 2. Since $S_G = g(0) = c_f/c_f^* = 1$, the result follows. ■

The key property underlying this result is $(\omega_H^* \star G - S_G\omega_H^*) \overset{n\to\infty}{\to} 0$, which shows that G is 'compact' enough to act as an aggregate multiplier S_G asymptotically. This is analogous to the role the covariance sum $S_\gamma := \sum_{k=-\infty}^{\infty} \gamma(k)$ plays in the asymptotic variance of aggregated short-range dependence processes (Gefferth *et al.*, 2003).

3.3. *Atypicality and speed of convergence*

Theorem 2 showed that the VTF of a fractionally differenced process is asymptotically equal to the VTF of its fGn fixed point up to an additive constant. This makes fractionally differenced processes atypical among LRD processes. We show this first for the VTF itself, and then for the speed of convergence of the CTF to the fixed point.

Without loss of generality, the VTF of any time series in the domain of attraction of a given fGn can be expressed as

$$\omega_H(n) = \omega_H^*(n) + \omega_d(n); \tag{9}$$

where ω_d represents the distance of the VTF from its limiting fGn counterpart. By definition, $\omega_d(n) = o(n^{2H})$, but otherwise the growth rate of ω_d is not constrained, implying that there is considerable variety within the domain of attraction.

One way of characterising the size of the difference $\omega_d(n)$ is to use regular variation (Bingham *et al.*, 1987; Gefferth *et al.*, 2003). A regularly varying function $f(n)$ of index β and integer argument $n \in \mathbb{N}$ satisfies $\lim_{k\to\infty} f(kn)/f(k) = n^{\beta}$, $\beta \in \mathbb{R}$. Assume without loss of generality that ω_d is upper bounded by a regularly varying function of index $\beta \in [0, 2H]$, that is

$$\omega_d(n) = O\{s(n)n^{\beta}\}, \tag{10}$$

where s is a slowly varying function (that is regularly varying with index 0), and β is the infimum of indices for which (10) holds. A notion of closeness of the process to the limiting fGn can then be defined in terms of β, where the smaller the index, the closer the process.

According to this scheme, Theorem 2 states that fractionally differenced processes belong in the closest layer of the hierarchy, corresponding to $\beta = 0$. Furthermore, the theorem shows that $s(n)$ (which could in general diverge, for example $s(n) \overset{n\to\infty}{\sim} \log(n)$) tends to a constant. Thus, the VTF of a fractionally differenced process lies in a very tight neighbourhood indeed of the VTF of its limiting fixed point. Far from being typical LRD processes, fractionally differenced processes deviate only in very subtle ways from fGn in terms of their large lag behaviour.

From (1), there is a direct relationship between closeness in the above sense and speed of convergence of the CTF to its fixed point under aggregation.

THEOREM 3. *Let* ϕ_H *denote the CTF of a fractionally differenced process in the domain of attraction of* ϕ^*_H *with* $H \in (1/2, 1)$. *Then*

$$\phi_H^{(m)}(n) = \phi^*_H(n) + D\left(1 - n^{2H}\right)m^{-2H} + o\left(m^{-2H}\right) = \phi^*_H(n) + O\left(m^{-2H}\right),$$

where D is the constant from Theorem 2.

Proof. The result follows from substituting $\omega_H(n) = \omega^*_H(n) + D + o(1)$ from Theorem 2 in (1) and using that $(1+x)^{-1} = 1 - x + O(x^2)$. ■

Beginning from (9), it holds generally for LRD processes in the DoA of ϕ^*_H that $\phi_H^{(m)}(n) = \phi^*_H(n) + O\{s(m)m^{-2H+\beta}\}$. It follows that fractionally differenced processes for which $\beta = 0$ and $s(m)$ is identically equal to a constant converge faster to the fixed point compared to all other processes in the DoA. Examples are provided in Section 5.

4. Closeness of the ACVF

Recall that $\omega = \mathbf{I}\gamma$. Because the double sum operator $\mathbf{I}$ smooths out local variations, Theorem 2 cannot be used to derive an explicit characterisation of the closeness in terms of the ACVF. We therefore set out to provide a closeness result for the ACVF here. Not only is this of interest in its own right, it also provides an alternative way of demonstrating the closeness to fGn, and leads to an additional result on the spectral closeness to fGn in an additive sense.

The following lemma is the analogue of Lemma 2 used for the ACVF. A proof is given in the appendix.

LEMMA 4. *Assume* $-1 \leqslant \alpha < 0$ *and let a be the symmetric positive sequence* $a_n := |n|^\alpha$, $n \neq 0$ *and* $a_0 > 0$. *Let b be a symmetric sequence with* $|b_0| < \infty$ *for which there exists* $\beta \in [0,2]$ *such that* $\sum_{j=1}^\infty j^\beta |b_j| < \infty$ *and* $|b_n| = O(n^{-(\beta+1)})$. *Then* $S_b := \sum_{j=-\infty}^\infty b_j$ *and the symmetric sequence* $c := a \star b$ *exist, and* $c_n - S_b a_n = O(n^{\alpha-\beta})$ *as* $n \to \infty$.

We can now prove the ACVF closeness result

THEOREM 4. *Let* γ_H *denote the ACVF of a fractionally differenced process for which* $H \in (1/2, 1)$ *and with* c_f *chosen equal to* c^*_f. *Then* $\gamma_H(n) = \gamma^*_H(n) + O(n^{2H-4})$.

Proof. The exact ACVF of a unit variance fGn with Hurst parameter H is given by

$$\gamma^*_H(n) = \frac{1}{2}\left\{(n+1)^{2H} + (n-1)^{2H} - 2n^{2H}\right\},$$

for $n \geqslant 0$ and $\gamma^*_H(n) = \gamma^*_H(-n)$ for $n < 0$. Then $\gamma^*_H(0) = 1$, and for $n \neq 0$ $\gamma^*_H(n) = (1/2)|n|^{2H} k(|n|^{-1})$ where $k(x) := (1+x)^{2H} + (1-x)^{2H} - 2$. Expanding k in a Taylor series around the origin, we obtain the following series representation:

$$\gamma^*_H(n) = \sum_{j=1}^\infty c_j f_j(n), \quad c_j := \frac{\prod_{i=0}^{2j-1}(2H-i)}{(2j)!}, \quad f_j(n) := \begin{cases} |n|^{2H-2j} & n \neq 0 \\ \mathbb{1}(j=1)/c_1 & n = 0 \end{cases}$$

which is uniformly absolutely convergent since $\{c_j\}$ is absolutely convergent by the ratio test.

Now $\gamma_H(n) = (\gamma_H^* \star G)(n) = \sum_{k=-\infty}^{\infty} G(k) \sum_{j=1}^{\infty} c_j f_j(n-k) = \sum_{j=1}^{\infty} c_j (f_j \star G)(n)$ where the existence of $\gamma_H^* \star G$ and γ_H^* as absolutely convergent series justifies the interchange of summations (Apostol (1974), Thm. 8.43). We can now compare γ_H and γ_H^* as

$$
\begin{aligned}
&(11) \qquad |\gamma_H(n) - \gamma_H^*(n)| &= \; & \sum_{j=1}^{\infty} |c_j| \big|(f_j \star G)(n) - f_j(n)\big| \\
&(12) & \leqslant \; & |c_1| \big|(f_1 \star G)(n) - f_1(n)\big| + \sum_{j=2}^{\infty} |c_j| \big|(f_j \star G)(n)\big| + O(n^{2H-4}).
\end{aligned}
$$

We shall show that each of the terms on the right hand side are of order $O(n^{2H-4})$.

The result for the first term follows immediately from Lemma 4 upon identifying f_1 with a, $2H-2$ with α, G with b, setting $\beta = 2$ (justified by Theorem 1iii), and noting that $\sum_{j=-\infty}^{\infty} G_j = 1$ by the assumption $c_f = c_f^*$.

Now consider the second term. Recall from Theorem 1 that $G_n = O(n^{-3})$, i.e. there exists $K > 0$ such that $G_n \leq K|n|^{-3}$ for n sufficiently large. Thus, when $j \geqslant 2$ and for $n > 0$ large enough

$$
\begin{aligned}
|(f_j \star G)(n)| &= \sum_{k=-\infty}^{\infty} |f_j(k)||G_{n-k}| = \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} |k|^{2H-2j} |G_{n-k}| \leq \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} |k|^{2H-4} |G_{n-k}| \\
&= \sum_{k=1}^{\infty} |k|^{2H-4} |G_{n+k}| + \sum_{k=1}^{\lfloor n/2 \rfloor} |k|^{2H-4} |G_{n-k}| + \sum_{k=\lfloor n/2 \rfloor + 1}^{\infty} |k|^{2H-4} |G_{n-k}| \\
&\leq K \sum_{k=1}^{\infty} |k|^{2H-4} (n+k)^{-3} + K \sum_{k=1}^{\lfloor n/2 \rfloor} |k|^{2H-4} (n-k)^{-3} + \left| \frac{n}{2} \right|^{2H-4} \sum_{k=\lfloor n/2 \rfloor + 1}^{\infty} |G_{n-k}| \\
&\leq K \sum_{k=1}^{\infty} (kn + k^2)^{2H-4} + K \sum_{k=1}^{\lfloor n/2 \rfloor} (kn - k^2)^{2H-4} + \left| \frac{n}{2} \right|^{2H-4} \sum_{k=-\infty}^{\infty} |G_{n-k}| \\
&< K \sum_{k=1}^{\infty} (kn)^{2H-4} + K \sum_{k=1}^{\lfloor n/2 \rfloor} (kn/2)^{2H-4} + \left| \frac{n}{2} \right|^{2H-4} \sum_{k=-\infty}^{\infty} |G_k| \\
&= O(n^{2H-4});
\end{aligned}
$$

using that $2H - 4 \geq -3$, that $kn + k^2 \geqslant kn$ for all k, $kn - k^2 \geq nk/2$ for $1 \leqslant k \leqslant n/2$, the absolute summability of G, and the fact that $\sum_{k=1}^{\infty} |k|^{2H-4} < \infty$. Hence the right hand side of (11) is of order $O(n^{2H-4})$. ■

In Section 3.1, we derived a result which may best be described as 'multiplicative closeness' for the spectrum of a fractionally differenced process. This form of closeness was natural for providing a subsequent link to the time domain. However, when calculations in the frequency domain are of specific interest, an additive closeness result for the spectrum is useful. Such a result can be derived from the above theorem.

COROLLARY 2. *It holds that* $f_H(x) = f_H^*(x) + \varphi(x)$ *where* φ *is differentiable, and satisfies* $\varphi' \in \Lambda_\alpha$ *if* $\alpha < 2 - 2H$, *and* $\varphi(0) = 0$. *Moreover,* $\varphi(x) = O(x^{-2H+3})$ *as* $x \to 0$.

Proof. Let $\varphi := f_H - f_H^*$. The Fourier series of φ exists and equals φ, and its coefficients are given by $d_n = \gamma_H(n) - \gamma_H^*(n)$, which by Theorem 4 is $O(|n|^{2H-4})$. Since $2H - 4 < -2$ the first absolute moment of the coefficients exists, so Thm. 7.19 in Kufner and Kadlec (1971) applies and shows that φ' exists and $\varphi' \in \Lambda_{2-2H}$. By the definition of $g(0)$, $\varphi(0) = \lim_{x\to 0}\{f_H(x) - f_H(x)/f_H^*(x) f_H^*(x)\} = 0$. The last claim follows by straightforward Taylor expansion of $f_H(x) - f_H^*(x)$ around $x = 0$. Details are given in the appendix. ■

The additive closeness of the spectrum is a highly non-trivial result: from the usual spectrum definition of LRD (Section 2.1), LRD with Hurst parameter H implies only that the ratio between f_H/f_H^* is bounded at the origin whereas the difference $f_H - f_H^*$ generally diverges. That the difference is not only a bounded function but tends to zero, and is also differentiable, emphasises in yet another way how unusual fractionally differenced processes are among LRD processes. To explore this in more detail, observe that the statement of Corollary 2 can be written

$$f_H + \varphi^- = f_H^* + \varphi^+$$

with $\varphi^- \geq 0$ and $\varphi^+ \geq 0$. Both φ^+ and φ^- define spectral densities with $\varphi^+(0) = \varphi^-(0) = 0$. We then (Brockwell and Davis (1991), Cor. 4.3.1) obtain a probabilistic variant of the closeness result: a fractionally differenced process is equal in the distributional sense to its limiting fGn up to additive independent processes with spectra φ^+, φ^-, both of which have the property of having a vanishing covariance sum $S_\gamma = \sum_{j=-\infty}^{\infty} \gamma_j$. Such processes (called constrained short-range dependent (CSRD) in Gefferth *et al.* (2003)), lie in the DoA of an fGn with Hurst parameter $H' \in [0, 1/2)$. In contrast, for short-range dependent (SRD) processes (those in the DoA of a fGn with $H' = 1/2$), S_γ is finite but positive. A graph of a particular φ and its first derivative is shown in Figure 2. The plot suggests that $\varphi^- \equiv 0$; whereby FARIMA would be equal in distribution to fGn plus an independent CSRD process.

To conclude our treatment of the ACVF, observe that a slightly weaker form of the closeness result of Theorem 2 can be derived from Theorem 4. Indeed, the identity $\omega_H(n) - \omega_H^*(n) = (\mathbf{I}d)_n$ implies

$$\left|\omega_H(n) - \omega_H^*(n)\right| = \left|\sum_{k=0}^{n-1}\left(\sum_{j=-\infty}^{\infty} d_j - \sum_{j=-k}^{k} d_j\right)\right| \leqslant 2\sum_{k=0}^{n-1}\sum_{j=k+1}^{\infty} |d_j| \leqslant O(1)\sum_{k=0}^{n-1} k^{2H-3} = O(1),$$

using that $d_n = O(|n|^{2H-4})$ implies $\sum_{j=k+1}^{\infty} |d_j| = O(1)\sum_{j=k+1}^{\infty} j^{2H-4} = O(k^{2H-3})$. The $O(1)$ remainder term simply corresponds to a bounded function; this is clearly weaker than the asymptotically constant remainder term appearing in Theorem 2.

We recently became aware of Lieberman and Phillips (2008) who provide an asymptotic expansion for a class of fractionally differenced processes similar to (2), though h is required to be smooth rather than C^3. Using the first two terms of this expansion and comparing with an expansion for $\gamma_H^*(m)$, one can recover the $O(n^{2H-4})$ term of Theorem 4. The work of Lieberman and Phillips (2008) focuses on numerical approximation through infinite-order asymptotic expansions and does not compare against fGn or draw conclusions on convergence speed or brittleness as we do here.

5. Fractional processes are brittle

As pointed out in Section 3.3, fractionally differenced processes converge 'almost immediately' to their fGn fixed point compared to other processes in the domain of attraction, and this is true in terms of each of the VTF, ACVF, and spectrum. In this section, we point out and illustrate a key consequence of this fact, namely the brittleness of fractionally differenced models.

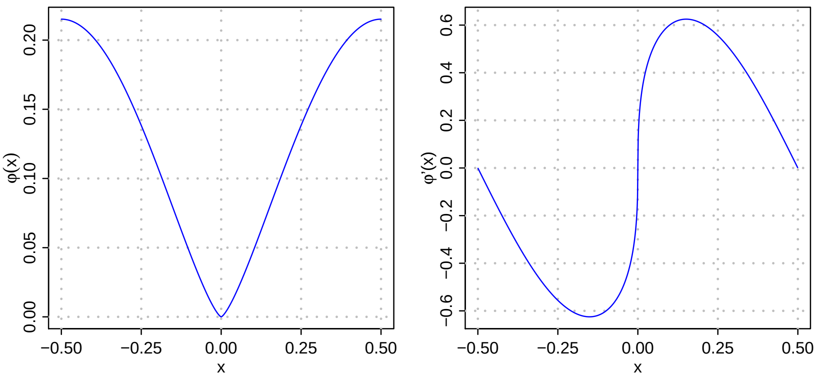

Figure 2. Graphs of the function $\varphi(x) = f_H(x) - f_H^*(x)$ and its first derivative in the canonical case of a pure fractionally differenced process (FARIMA(0, d, 0)) with $H = 0.8$ and $c_f = c_f^*$.

5.1. *Brittleness*

Experimental data, especially data measured on a continuous scale, is very rarely clean. Imperfections in physical measurement are often treated through the concept of observation noise, modelled as a random process which perturbs the underlying observables. A very common choice is that of additive independent Gaussian noise, either white or coloured. In the present context, this corresponds to adding to the original VTF (or ACVF, or spectrum) the VTF (respectively ACVF, spectrum) of an SRD noise process.

As argued at the end of the previous section, we can essentially think of a fractionally differenced process as an fGn to which a CSRD process has been added. Adding an SRD noise to this will change the asymptotic behaviour, because the SRD asymptotics (with $S_\gamma > 0$) is ‘stronger’ than CSRD asymptotics (with $S_\gamma = 0$). In terms of the hierarchy within the DoA described by the index β from (10), whereas the original process lies very close to the centre with $\beta = 0$, the SRD-perturbed process will lie considerably further out, with $\beta = 1$. A similar observation can be made if we instead add a noise with LRD with $H' < H$ (resulting in $\beta \in (1, 2H)$), or even another CSRD process with $H' > 0$ (resulting in $\beta \in (0, 1)$). This last result follows from the fact that Theorem 2 implies that the ‘error’ processes are so special that they are not only CSRD, but correspond to the extreme case of $H' = 0$, resulting in $\beta = 0$.

Since the addition of even trace amounts of noise of diverse kinds will change the asymptotics, pushing the process further from its fGn limit and therefore slowing its convergence rate to it under aggregation, fractional differencing models are ‘brittle’ or non-robust in this sense. Properties of systems driven by such processes may therefore differ qualitatively from properties of the same system once noise is added. The precise impact of the noise is beyond the scope of this paper (see the discussion). It will depend on both the application and the class of noise and must be determined case by case.

5.2. *Numerical illustrations*

In this section we illustrate the brittle nature of fractionally differenced process through high accuracy numerical evaluation of the VTF of FARIMA time series, both with and without additive noise.

Three examples will be considered, two with SRD-noise and one with LRD-noise. Specifically, the perturbed processes are $Z_i(t) = X_i(t) + \sqrt{0.1}Y_i(t)$ for $i = 1,2,3$, with

1) X_1: unit variance FARIMA$(0,0.3,0)$;
   Y_1: unit variance Gaussian white noise,
2) X_2: unit variance FARIMA$(1,0.3,1)$ with ARMA parameters $(\phi_1,\theta_1) = (0.3,0.7)$;
   Y_2: unit variance ARMA$(1,1)$ also with ARMA parameters $(\phi_1,\theta_1) = (0.3,0.7)$,
3) X_3: unit variance FARIMA$(0,0.3,0)$;
   Y_3: unit variance FARIMA$(0,0.2,0)$.

See Brockwell and Davis (1991), Def. 13.2.2, for the general definition of FARIMA(p,d,q). In each case, the original process X_i and the perturbed process Z_i share a common fGn fixed point, but have unequal variances. It may seem unfair to compare results for processes with different variances, however, the opposite is true. In fact, if the variances of Z_i and X_i were chosen equal, this would mean that $c_f \neq c_f^*$, and so their fGn limits would be different, rendering meaningful comparison impossible. To see this more directly, from the definitions in Section 2.1 it is clear that adding a perturbation corresponding to a smaller H value does not alter the fixed point. On the other hand the variance must increase when an independent noise is added.

For each example $i = 1,2,3$, we calculate the VTF of Z_i and X_i and normalise them by dividing by their common fGn limit ω_H^*. Closeness to fGn can then be evaluated by investigating how the normalised VTF deviates from 1 for each lag. Maple 13 (Maplesoft, 2009) was used to numerically evaluate the variance time functions to a high degree of precision.

Figure 3 displays the normalised VTFs for lags 1-10 for aggregation levels $m =$ 1, 10, and 100, with one example per column. The graphs clearly demonstrate that even a small departure from FARIMA takes the process much further away from its corresponding fGn. Indeed, after an aggregation of level 100, in each case the VTF of the original process is visually indistinguishable from its fGn limits compared to their perturbed versions.

Note that both the second and third columns in the figure give examples where before aggregation ($m = 1$) the perturbed process was in fact *closer* to the fixed point over the first few lags, where most of the obvious autocovariance lies. Under aggregation however, this quickly reverses as the different asymptotic behaviours of the original and perturbed processes manifest and become dominant at all lags.

6. Discussion

We have shown that fractionally differenced processes have an asymptotic autocovariance structure which is extremely close to that of the fGn, more specifically, to that of the fGn fixed point to which the given process will tend under aggregation based renormalisation. We have shown this independently for each of three equivalent views of the autocovariance structure, namely behaviour of the spectral density at the origin, and each of the ACVF and the variance time function in the large lag limit.

We showed that the natural class of processes against which this behaviour should be compared are those in the domain of attraction of the fGn fixed point limit. Using regular variation to provide a measure of distance from this fixed point within the DoA, we were able to precisely quantify the nature of this 'closeness', and to confirm that the fractionally differenced class are indeed exceptionally unusual in this regard,

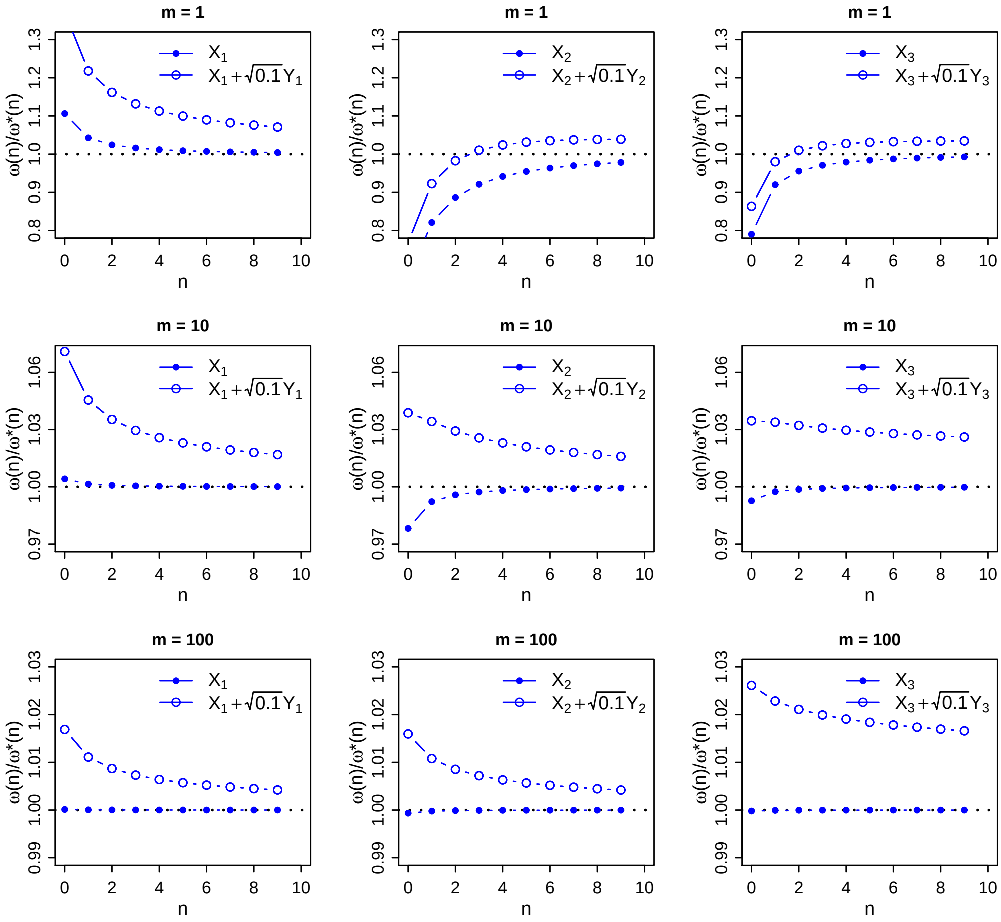

Figure 3. Ratios of VTFs of original FARIMA and perturbed processes to their fGn limit, both originally and under aggregation, one column per example. The solid circles denote unperturbed FARIMA; the hollow circles the perturbed ones. It is seen that the VTFs for unperturbed FARIMA converge much faster than their perturbed counterparts.

resulting in very fast convergence to fGn under renormalisation. We then used this fact to point out that the fractionally differenced process class is brittle, that is, non-robust to the presence of noise. In particular we showed that the addition of arbitrarily small amounts of independent noise, not only Gaussian white noise but also noises which are much gentler in a precise sense, changes the asymptotic covariance structure qualitatively. This fact has not been appreciated in the literature where such models, for example the FARIMA class, are widely used in time series modelling, synthetic data generation, and to drive more complex stochastic systems such as queueing systems, without regard to robustness with respect to the model in this sense.

The assessment of the impact of the brittleness of fractionally differenced models is beyond the scope of this work, as it will depend intimately on each particular application as well as the nature of the noise in question. However, we argue that conclusions based on the perception that FARIMA and related models represent ‘typical’ LRD behaviour need to be reassessed, in particular in contexts where noise is important to consider. To give an example of a possible impact in the noiseless case, we conclude by expanding upon the comments given in the introduction on statistical estimation.

The closeness of a process to its fGn fixed point in functional terms is directly related to the speed of convergence of that process to the fixed point under aggregation. One application where this fact carries direct implications is the performance of statistical estimators for the Hurst parameter H. Fundamentally, semi-parametric estimators of scaling parameters such as H are based on underlying estimates made at a set of 'aggregations' at different levels, that is, at multiple scales (Robinson, 1994; Beran, 1994; Abry *et al.*, 1998; Taqqu *et al.*, 1995). The sophistication of particular estimators notwithstanding, this is true regardless of whether they are based in the spectral, time, or wavelet domains, though the technical details vary considerably. In the time domain using time domain aggregation the link is of course direct, and reduces to looking at the asymptotically power-law nature of $\mathcal{V}^{(m)} = \omega^{(m)}(0)$ as a function of m in some form. This is precisely where fractionally differenced processes are at a real advantage, as this quantity converges extremely quickly to that of the fGn fixed point, whose ideal power-law behaviour $\mathcal{V}^{(m)} = \mathcal{V} m^{2H}$ allows H to be easily recovered. As a result, estimator performance evaluated through the use of fractionally differenced models would be superior to that for LRD processes more generally. Note that we are not recommending that H-estimation be performed directly in the time domain by regressing $\hat{\mathcal{V}}^{(m)}$ on m, indeed we have argued the opposite (Abry *et al.*, 1998). Our point is that the extreme closeness of such models to fGn must ultimately manifest in simpler asymptotic behaviour which will, in general, translate to improved estimation. Indeed, in the spectral domain, the importance of the degree of smoothness at the origin for the ultimate limits on estimator performance has already been noted (Giraitis *et al.*, 1997). Note that the above observations in no way put into question findings of prior work on estimation of fractional processes in noise.

Appendix: proofs

The appendix is split according to results relating to spectral closeness (Section 3.1), closeness of the VTF (Section 3.2), and of the ACVF (Section 4). For convenience, the statement of results proved here are generally repeated.

Spectrum

Details of the proof of Theorem 1. **(i).** It is well known, and can be verified by examining (3) and (4), that each of $f_H^*(x)$ and $f_H(x)$ diverge to infinity at $x = 0$ but are otherwise even, positive and continuous. Since $g(0) > 0$ is finite, g is positive and continuous on a compact domain and hence bounded, and even. Since g is continuous, it is integrable (Champeney (1990), p. 9), and by similar arguments, g is in L^p.

(ii). Since a is strictly positive, $\tilde{g}$ is bounded away from zero. Each of a, b, and c are smooth. The latter follows from the fact that for each $j \neq 0$ the term $|\pi j + \pi x|^{-(2H+1)}$ is infinitely differentiable for $x \in [-1/2, 1/2]$. By comparing against $\sum_{j=1}^{\infty}(j - 1/2)^{-(2H+1)} < \infty$ the Weierstrass M-test shows that the defining sum for c, and the sum of the term by term first derivatives, each converge uniformly. A classical result on the differentiability of infinite series (Apostol (1974), Thm. 9.14) then shows that c' is given by the latter sum. Using exactly the same M-test, this can be repeated for derivatives of all orders, proving that c is smooth.

Since a is smooth and bounded, $|a|^{2H+1}$ is smooth on $[-1/2, 1/2]$, and the same is true for $|b|^{2H+1}$ away from the origin. It follows that $\tilde{g}$ is smooth everywhere except at the origin where its smoothness is controlled by that of $|b|^{2H+1}$.

Variance time function

LEMMA A1. *The ACVFs γ_H and γ_H^* are related through the convolution $\gamma_H = G \star \gamma_H^*$.*

Proof. The right-hand side of $\gamma_H = G \star \gamma_H^*$ exists since

$$\sum_{j=-\infty}^{\infty} G_j \gamma_N^*(n-j) \le \sum_{j=-\infty}^{\infty} |G_j||\gamma_N^*(n-j)| \le \gamma_N^*(0) \sum_{j=-\infty}^{\infty} |G_j| < \infty,$$

from Theorem 1. For the left-hand side, we can write

$$\begin{aligned} \gamma_H(n) &= \int_{-1/2}^{1/2} f_H(x)\mathrm{e}^{2\pi i x n}\mathrm{d}x = \int_{-1/2}^{1/2} g(x) f_H^*(x)\mathrm{e}^{2\pi i x n}\mathrm{d}x &\text{(A1)} \\ &= \int_{-1/2}^{1/2} \left(\sum_{j=-\infty}^{\infty} G_j \mathrm{e}^{-2\pi i x j} \right) f_H^*(x)\mathrm{e}^{2\pi i x n}\,\mathrm{d}x &\text{(A2)} \end{aligned}$$

since the Fourier series for g converges absolutely everywhere (Theorem 1). Now

$$\int_{-1/2}^{1/2} \left(\sum_{j=-\infty}^{\infty} |G_j \mathrm{e}^{-2\pi i x j}| \right) f_H^*(x)\mathrm{e}^{2\pi i x n}\,\mathrm{d}x = \sum_{j=-\infty}^{\infty} |G_j| \int_{-1/2}^{1/2} f_H^*(x)\mathrm{e}^{2\pi i x n}\,\mathrm{d}x = \gamma_H^*(n) \sum_{j=-\infty}^{\infty} |G_j| \; < \infty.$$

This justifies the use of Fubini's Theorem (Taylor (1973), Theorem 6.5) on the iterated integral (A2) to reverse the order of integration and summation. Using the evenness of G and f_H^*, this yields

$$\begin{aligned} \gamma_H(n) &= \sum_{j=-\infty}^{\infty} G_j \int_{-1/2}^{1/2} f_H^*(x) \cos(2\pi x j) \cos(2\pi x n)\,\mathrm{d}x \\ &= \sum_{j=-\infty}^{\infty} G_j \int_{-1/2}^{1/2} f_H^*(x) \frac{1}{2} \Big[\cos\{2\pi x(j-n)\} + \cos\{2\pi x(j+n\} \Big]\,\mathrm{d}x \\ &= \frac{1}{2} \sum_{j=-\infty}^{\infty} G_j \{\gamma_H^*(j-n) + \gamma_H^*(j+n)\} = \frac{1}{2} \left(\sum_{j=-\infty}^{\infty} G_j \gamma_H^*(n-j) + \sum_{j=-\infty}^{\infty} G_j \gamma_H^*(-j-n) \right) \\ &= \frac{1}{2} \{ (G \star \gamma_H^*)(n) + (G \star \gamma_H^*)(-n) \} = (G \star \gamma_H^*)(n); \end{aligned}$$

using the evenness of γ_H^* and $G \star \gamma_H^*$, and the existence of $G \star \gamma_H^*$ to justify the splitting of the sum. ■

LEMMA A2. *Assume* $1 < \alpha < 2$ *and define* $f_\alpha(x,y) := |x-y|^\alpha + (x+y)^\alpha - 2x^\alpha$ *for* $x \ge 0$, $y > 0$. *For each* y, $f_\alpha(\cdot, y)$ *is positive, strictly decreasing, and* $\lim_{x \to \infty} f_\alpha(x,y) = 0$.

Proof. Fix $y > 0$. We split the domain of $f_\alpha(\cdot, y)$ and consider two cases.

Suppose $x \geqslant y$. It follows that $f_\alpha'(\cdot, y) = \alpha f_{\alpha-1}(\cdot, y)$. Define $g(x) = x^\alpha$. Since $g'(x) = \alpha x^{\alpha-1}$ is strictly concave, $(x-y)^{\alpha-1} + (x+y)^{\alpha-1} < 2x^{\alpha-1}$ and so $f_\alpha'(\cdot, y) < 0$ and $f_\alpha(\cdot, y)$ is strictly decreasing. To prove $\lim_{x\to\infty} f_\alpha(x,y) = 0$, we apply the mean value theorem twice to g, and then once to g' to obtain:

$$f_\alpha(x,y) = \big\{(x+y)^\alpha - x^\alpha\big) - \big(x^\alpha - (x-y)^\alpha\big\} < \alpha y\big\{(x+y)^{\alpha-1} - (x-y)^{\alpha-1}\big\} < 2\alpha(\alpha-1)y^2(x-y)^{\alpha-2}$$

(since g' is strictly increasing and g'' strictly decreasing), which tends to zero as $x \to \infty$.

Assume $x < y$. In this case, the derivative with respect to x is

$$\begin{aligned} f'_\alpha(x,y) &= \alpha\big\{(x+y)^{\alpha-1} - (y-x)^{\alpha-1} - 2x^{\alpha-1}\big\} \\ &< \alpha\big\{(x+y)^{\alpha-1} - (y-x)^{\alpha-1} - (2x)^{\alpha-1}\big\} \\ &= \alpha\big\{h_x(y) - h_x(x)\big\} \end{aligned}$$

where $h_x(y) = (x+y)^{\alpha-1} - (y-x)^{\alpha-1}$. Since the derivative of h_x is negative for $x > 0$, h_x is strictly decreasing. Hence $f'_\alpha(\cdot,y) < 0$ and so $f_\alpha(\cdot,y)$ is likewise strictly decreasing.

Finally, since $f_\alpha(\cdot,y)$ is strictly decreasing for all $y > 0$ and tends to zero, it is positive. ■

LEMMA A3. *Assume* $1 < \alpha < 2$ *and let* $a := \{|n|^\alpha : n \in \mathbb{Z}\}$. *Let* b *be a symmetric sequence satisfying* $\sum_{j=1}^\infty j^\alpha |b_j| < \infty$. *Then* $S_b := \sum_{j=-\infty}^\infty b_j$ *and the symmetric sequence* $c := a \star b$ *exist, and* $(c_n - S_b a_n) \overset{n\to\infty}{\to} 0$.

Proof. Since $\alpha > 1$, $\sum_{j=-\infty}^\infty |b_j| \le |b_0| + 2\sum_{j=1}^\infty j^\alpha |b_j| < \infty$, so b is absolutely summable and hence summable. Now consider c. Clearly $c_0 = \sum_{j=-\infty}^\infty |-j|^\alpha b_j$ exists by the assumptions on b, and for $n > 0$

$$\begin{aligned} |c_n| = |(a \star b)_n| &\leqslant \sum_{j=-\infty}^{-n} |n-j|^\alpha |b_j| + \sum_{j=-n+1}^{n-1} |n-j|^\alpha |b_j| + \sum_{j=n}^{\infty} |n-j|^\alpha |b_j| \\ &\leqslant \sum_{j=n}^{\infty} (2j)^\alpha |b_j| + \sum_{j=-n+1}^{n-1} |n-j|^\alpha |b_j| + \sum_{j=n}^{\infty} j^\alpha |b_j| < \infty. \end{aligned}$$

Since both a and b are symmetric, c_n also exists for $n < 0$, and so c exists and is symmetric.

For the last part, since $c_n - S_b a_n$ is symmetric in n, assume $n \ge 0$ and rewrite as

$$\sum_{j=-\infty}^{\infty} |n-j|^\alpha b_j - n^\alpha \sum_{j=-\infty}^{\infty} b_j = n^\alpha b_0 + \sum_{j=1}^{\infty} (n+j)^\alpha b_j + \sum_{j=1}^{\infty} |n-j|^\alpha b_j - n^\alpha \sum_{j=-\infty}^{\infty} b_j = \sum_{j=1}^{\infty} T_n^j b_j$$

where $T_n^j := |n-j|^\alpha + (n+j)^\alpha - 2n^\alpha$, $n \ge 0$, $j > 0$. Noticing that $T_n^j = f_\alpha(n,j)$ from Lemma A2, we have that $T_n^j < T_0^j = 2j^\alpha$ for each fixed j, and so

$$|c_n - S_b a_n| \leqslant \sum_{j=1}^{N} |T_n^j||b_j| + \sum_{j=N+1}^{\infty} |T_n^j||b_j| < \sum_{j=1}^{N} |T_n^j||b_j| + 2\sum_{j=N+1}^{\infty} j^\alpha |b_j|.$$

Now, given any $\varepsilon > 0$, a $N(\varepsilon) > 1$ can be found such that $\sum_{j=N+1}^\infty j^\alpha |b_j| < \varepsilon/4$. Next, since $T_n^j \overset{n\to\infty}{\to} 0$ for any fixed j (Lemma A2), there exists an $n_0(N)$ such that $\sum_{j=1}^N |T_n^j||b_j| < \varepsilon/2$ when $n \geqslant n_0$. Hence $|c_n - S_b a_n| < \varepsilon$ for $n \geqslant n_0$ and so $(c_n - S_b a_n) \overset{n\to\infty}{\to} 0$. ■

LEMMA A4. *Let* a, b *be symmetric sequences and assume that* $c := a \star b$ *exists. Then* $\mathbf{I}c$ *exists, and if* $(\mathbf{I}a) \star b$ *exists, then* $\mathbf{I}c = (\mathbf{I}a) \star b - ((\mathbf{I}a) \star b)_0$.

Proof. Since the quantity $(\mathbf{I}c)_n$ is a finite sum of elements of c, it exists for each n. Now, the expression

$$(\mathbf{I}c)_n = \sum_{k=0}^{n-1} \sum_{i=-k}^{k} \sum_{j=-\infty}^{\infty} a_j b_{i-j}$$

can be rewritten as $(\mathbf{I}c)_n = \sum_{j=-\infty}^{\infty} a_j H_n(j)$ where $H_n(j) := \sum_{k=0}^{n-1} \sum_{i=-k}^{k} b_{i-j}$, since a finite sum of convergent series is convergent. Since $(\mathbf{I}a)_{j-1} - 2(\mathbf{I}a)_j + (\mathbf{I}a)_{j+1} = a_{-j} + a_j = 2a_j$, we have

$$(\mathbf{I}c)_n = \sum_{j=-\infty}^{\infty} a_j H_n(j) \tag{A3}$$

$$= \frac{1}{2} \sum_{j=-\infty}^{\infty} \{(\mathbf{I}a)_{j-1} - 2(\mathbf{I}a)_j + (\mathbf{I}a)_{j+1}\} H_n(j) \tag{A4}$$

$$= \frac{1}{2} \left\{ \sum_{j=-\infty}^{\infty} (\mathbf{I}a)_{j-1} H_n(j) - 2 \sum_{j=-\infty}^{\infty} (\mathbf{I}a)_j H_n(j) + \sum_{j=-\infty}^{\infty} (\mathbf{I}a)_{j+1} H_n(j) \right\} \tag{A5}$$

$$= \frac{1}{2} \sum_{j=-\infty}^{\infty} (\mathbf{I}a)_j \{H_n(j+1) - 2H_n(j) + H_n(j-1)\}. \tag{A6}$$

Here the rewrite (A5) is justified since each of the sums is convergent, because each can be written as a finite sum of series of the form $\sum_{j=-\infty}^{\infty} (\mathbf{I}a)_j b_{m-j}$ for some m. But this is simply $\{(\mathbf{I}a) \star b\}_m$ which exists by assumption. Now,

$$\begin{aligned}
H_n(j+1) - 2H_n(j) + H_n(j-1) &= \{H_n(j-1) - H_n(j)\big) - \big(H_n(j) - H_n(j+1)\} \\
&= \sum_{k=0}^{n-1} \left\{ \left(\sum_{i=-k-j+1}^{k-j+1} b_i - \sum_{i=-k-j}^{k-j} b_i \right) - \left(\sum_{i=-k-j}^{k-j} b_i - \sum_{i=-k-j-1}^{k-j-1} b_i \right) \right\} \\
&= \sum_{k=0}^{n-1} \left\{ \big(b_{k-j+1} - b_{-k-j}\big) - \big(b_{k-j} - b_{-k-j-1}\big) \right\} \\
&= \sum_{k=0}^{n-1} \big(b_{k-j+1} - b_{k-j}\big) - \sum_{k=0}^{n-1} \big(b_{-k-j} - b_{-k-j-1}\big) \\
&= (b_{n-j} - b_{-j}) - (b_{-j} - b_{-n-j}) = b_{n-j} + b_{-n-j} - 2b_{-j}.
\end{aligned}$$

The result then follows by substitution into (A6), using the existence of $(\mathbf{I}a) \star b$ to justify splitting the sum, and finally by the symmetry of $\mathbf{I}a$ and b. ■

Autocovariance function

LEMMA A5. *Assume $\alpha < 0$ and define $f_\alpha(x,y) := |x-y|^\alpha + (x+y)^\alpha - 2x^\alpha$ for $x > y > 0$. Then f_α satisfies $f_\alpha(x,y) < 2\alpha(\alpha-1)y^2(x-y)^{\alpha-2}$.*

Proof. Since $x > y$ it follows that $f'_\alpha(\cdot,y) = \alpha f_{\alpha-1}(\cdot,y)$. Define $g(x) := x^\alpha$. Since $g'(x) = \alpha x^{\alpha-1}$ is strictly concave, $\alpha(x-y)^{\alpha-1} + \alpha(x+y)^{\alpha-1} < 2\alpha x^{\alpha-1}$ and so $f'_\alpha(\cdot,y) < 0$ whereby $f_\alpha(\cdot,y)$ is strictly decreasing. Now apply the mean value theorem twice to g, and then once to g', to obtain:

$$f_\alpha(x,y) = \{(x+y)^\alpha - x^\alpha\} - \{x^\alpha - (x-y)^\alpha\} < \alpha y\{(x-y)^{\alpha-1} - (x+y)^{\alpha-1}\} < 2\alpha(\alpha-1)y^2(x-y)^{\alpha-2};$$

since g' is strictly increasing and g'' strictly decreasing. ■

LEMMA A6. *Assume* $-1 \leqslant \alpha < 0$ *and let* a *be the symmetric positive sequence* $a_n := |n|^\alpha$, $n \neq 0$ *and* $a_0 > 0$. *Let* b *be a symmetric sequence with* $|b_0| < \infty$ *for which there exists* $\beta \in [0,2]$ *such that* $\sum_{j=1}^{\infty} j^\beta |b_j| < \infty$ *and* $|b_n| = O(n^{-(\beta+1)})$. *Then* $S_b := \sum_{j=-\infty}^{\infty} b_j$ *and the symmetric sequence* $c := a \star b$ *exist, and satisfies* $c_n - S_b a_n = O(n^{\alpha-\beta})$ *as* $n \to \infty$.

Proof. We have $\sum_{j=-\infty}^{\infty} |b_j| \leqslant b_0 + 2\sum_{j=1}^{\infty} j^\beta |b_j| < \infty$ so b is absolutely summable and therefore summable. Then S_b exists. Moreover, $|c_n| \leqslant \sum_{j=-\infty}^{\infty} |b_j||a_{n-j}| \le |b_n|a_0 + \sum_{j=-\infty}^{\infty} |b_j| < \infty$. We conclude that c_n exists for each $n \in \mathbb{Z}$, and that c is symmetric, by symmetry of a and b. Define $T_n^j := a_{|n-j|} + a_{n+j} - 2a_n$, and using symmetry of a and b, rewrite $c_n - S_b a_n$ as

$$(a \star b)_n - S_b a_n = a_n b_0 + \sum_{j=1}^{\infty} a_{|n-j|} b_j + \sum_{j=1}^{\infty} a_{n+j} b_j - S_b a_n = \sum_{j=1}^{\infty} T_n^j b_j.$$

To prove the last part of the lemma it suffices to consider $n \ge 0$, since c is symmetric, and as we are interested in large-n asymptotics, we restrict to $n > 2$. The sum for $|c_n|$ can be decomposed as

$$|c_n - S_b a_n| \leqslant \sum_{j=1}^{\lfloor n/2 \rfloor} |T_n^j||b_j| + \sum_{j=\lfloor n/2 \rfloor + 1}^{2n} |T_n^j||b_j| + \sum_{j=2n+1}^{\infty} |T_n^j||b_j| =: A_n + B_n + C_n.$$

We shall show that each of A_n, B_n, and C_n are of order $O(n^{\alpha-\beta})$.

The definition of A_n implies $n > j > 0$, so Lemma A5 applies to $T_n^j = f_\alpha(n,j)$, and implies the existence of a constant $K > 0$ such that $|T_n^j| \le K j^2 (n-j)^{\alpha-2} < K j^2 (n/2)^{\alpha-2}$ when $j \le n/2$. Thus

$$K^{-1} 2^{\alpha-2} A_n \leqslant n^{\alpha-2} \sum_{j=1}^{\lfloor n/2 \rfloor} j^{2-\beta} j^\beta |b_j| \leqslant n^{\alpha-2} n^{2-\beta} \sum_{j=1}^{\lfloor n/2 \rfloor} j^\beta |b_j| \leqslant n^{\alpha-\beta} \sum_{j=1}^{\infty} j^\beta |b_j| = O\big(n^{\alpha-\beta}\big).$$

For B_n, where $n \neq j$ and $n, j > 0$, we have $|T_n^j| < 2|n-j|^\alpha$, while $|T_n^n| = a_0 + (2^\alpha - 2)a_n = O(1)$. Then, for sufficiently large n, by assumption there exists a $K > 0$ such that

$$\begin{aligned}
B_n &\leqslant 2 \sum_{j=\lfloor n/2 \rfloor + 1}^{n-1} (n-j)^\alpha |b_j| + 2 \sum_{j=n+1}^{2n} (j-n)^\alpha |b_j| + |T_n^n||b_n| \\
&< 2K(n/2)^{-(\beta+1)} \left\{ \sum_{j=\lfloor n/2 \rfloor + 1}^{n-1} (n-j)^\alpha + \sum_{j=n+1}^{2n} (j-n)^\alpha + |T_n^n|/2 \right\} \\
&< 2^{\beta+3} K n^{-(\beta+1)} \sum_{j=1}^{n} j^\alpha + O(n^{-(\beta+1)}) \\
&< 2^{\beta+3} K n^{-(\beta+1)} \left(1 + \int_1^n x^\alpha \mathrm{d}x\right) + O\big(n^{-(\beta+1)}\big) = O\big(n^{\alpha-\beta}\big).
\end{aligned}$$

For C_n, where $j \geqslant 2n$, we have $T_n^j < 2n^\alpha$. Since

$$\sum_{j=2n+1}^{\infty} |b_j| \leqslant (2n)^{-\beta} \sum_{j=2n+1}^{\infty} j^\beta |b_j| = o(n^{-\beta}),$$

we get

$$C_n \leqslant \sum_{j=2n+1}^{\infty} 2n^\alpha |b_j| \leqslant 2n^\alpha \sum_{j=2n+1}^{\infty} |b_j| = o\big(n^{\alpha-\beta}\big).$$

Conclude that $|c_n - S_b a_n| = O(n^{\alpha-\beta})$ as $n \to \infty$. ∎

Details of the proof of Corollary 2. We explain here why $\varphi(x) = f_H(x) - f_H^*(x) = O(x^{-2H+3})$ as $x \to 0$. Calculate the first few derivatives of the analytic function $x \mapsto (\sin(x)/x)^{-2H+1}$ (set to 1 at $x = 0$) and expand in a Taylor series around the origin to find that $(\sin(x)/x)^{-2H+1} = 1 + O(x^2)$. It follows that $\sin(x)^{-2H+1} = x^{-2H+1} + O(x^{-2H+3})$ for $x \neq 0$. The function h is assumed three times (continuously) differentiable. Symmetry implies $h'(0) = 0$ so that by Taylors theorem, $h(x) = h(0) + O(x^2)$. Thus

$$h(x)\sin(\pi x)^{-2H+1} = h(0)\pi^{-2H+1}x^{-2H+1} + O\big(x^{-2H+3}\big), \quad x \neq 0,$$

while it can be shown that

$$x^{-2H-1}\sin^2(\pi x)/2 = \{1-\cos(2\pi x)\}x^{-2H-1} = 2\pi^2 x^{-2H+1} + O\big(x^{-2H+3}\big), \quad x \neq 0,$$

and

$$(1/2)\sin^2(\pi x)\sum_{j\neq 0}|\pi j + \pi x|^{-2H-1} = \{1-\cos(2\pi x)\}\sum_{j\neq 0}|\pi j + \pi x|^{-2H-1} = O(x^2).$$

Then, as $x \to 0$,

$$\begin{aligned}
f_H(x) - f_H^*(x) &= \big[h(x)\{2\sin(\pi x)\}^{-2H+1}\big] - \big[h(0)2^{-2H}\pi^{-2H-1}\{1-\cos(2\pi x)\}x^{-2H-1}\big] + O(x^2) \\
&= \big\{h(0)\pi^{-2H+1}2^{-2H+1}x^{-2H+1} + O(x^{-2H+3})\big\} \\
&\qquad - \big\{h(0)2^{-2H+1}\pi^{-2H+1}x^{-2H+1} + O\big(x^{-2H+3}\big)\big\} + O\big(x^2\big) \\
&= O\big(x^{-2H+3}\big).
\end{aligned}$$

References

Abry, P., Flandrin, P., Taqqu, M. S. and Veitch, D. (2003) Self-similarity and long-range dependence through the wavelet lens. In *Theory and Applications of Long-Range Dependence* (eds. P. Doukhan, G. Oppenheim and M. Taqqu), 527–556. Birkhäuser.

Abry, P., Veitch, D. and Flandrin, P. (1998) Long-range dependence: revisiting aggregation with wavelets. *Journal of Time Series Analysis (Bernoulli Society)*, **19**, 253–266.

Apostol, T. M. (1974) *Mathematical Analysis*. Addison Wesley, second edn.

Barbe, P. and McCormick, W. (2010) An extension of a logarithmic form of Cramér's ruin theorem to some FARIMA and related processes. *Stochastic Processes and their Applications*, **120**, 801–828.

Beran, J. (1993) Fitting long memory processes by generalized regression. *Biometrika*, **80**, 817–822.

Beran, J. (1994) *Statistics for Long-Memory Processes*. Chapman and Hall, New York.

Bingham, N., Goldie, C. and Teugels, J. (1987) *Regular Variation*. Cambridge University Press, Cambridge England.

Bloomfield, P. (1973) An exponential model for the spectrum of a scalar time series. *Biometrika*, **60**, 217–226.

Brockwell, P. and Davis, R. (1991) *Time Series: Theory and Methods*. Springer, second edn.

Champeney, D. (1990) *A Handbook of Fourier Theorems*. Cambridge University Press.

Gefferth, A., Veitch, D., Maricza, I. and Molnár, S. (2002a) Farima models for long-range dependent traffic. In *International Workshop on High Speed Networking*. Budapest, Hungary.

Gefferth, A., Veitch, D., Maricza, I., Molnár, S. and Ruzsa, I. (2003) The Nature of Discrete Second-Order Self-Similarity. *Advances in Applied Probability*, **35**, 395–416.

Gefferth, A., Veitch, D. and Molnár, S. (2002b) Convergence speed of asymptotically self-similar fARIMA processes. In *International Workshop on High Speed Networking*, 67–71. Budapest, Hungary.

Gefferth, A., Veitch, D., Ruzsa, I., Maricza, I. and Molnár, S. (2004) A New Class of Second Order Self-Similar Processes. *Stochastic Models*, **20**, 381–389.

Giraitis, L., Robinson, P. and Samarov, A. (1997) Rate optimal semiparametric estimation of the memory parameter of the Gaussian time series with long range dependence. *Journal of Time Series Analysis*, **18**, 49–61.

Granger, C. and Joyeux, R. (1980) An introduction to long-memory times series models and fractional differencing. *Journal of Time Series Analysis*, **1**, 15–29.

Hosking, J. (1981) Fractional differencing. *Biometrika*, **68**, 165–176.

Ilow, J. (2000) Forecasting network traffic using FARIMA models with heavy tailed innovations. In *ICASSP '00: Proceedings of the Acoustics, Speech, and Signal Processing, 2000. on IEEE International Conference*, 3814–3817. Washington, DC, USA: IEEE Computer Society.

Kufner, A. and Kadlec, J. (1971) *Fourier Series*. Iliffe Books.

Lieberman, O. and Phillips, P. C. (2008) A complete asymptotic series for the autocovariance function of a long memory process. *Journal of Econometrics*, **147**, 99–103.

Maplesoft (2009) *Maple 13*. Waterloo Maple Inc.

Robinson (1994) Semiparametric analysis of long-memory time series. *Annals of Statistics*, **22**, 515–539.

Samorodnitsky, G. and Taqqu, M. (1994) *Stable Non-Gaussian Random Processes*. Chapman and Hall.

Taqqu, M. (2002) Fractional Brownian Motion and Long-Range Dependence. In *Theory and Applications of Long-Range Dependence* (eds. P. Doukhan, G. Oppenheim and M. S. Taqqu), 6–38. Birkhäuser.

Taqqu, M. and Teverovsky, V. (1997) Robustness of Whittle-type estimators for time series with long-range dependence. *Stochastic Models*, **13**, 323–357.

Taqqu, M., Teverovsky, V. and Willinger, W. (1995) Estimators for long-range dependence: an empirical study. *Fractals*, **3**, 785–798. Reprinted in *Fractal Geometry and Analysis*, C.J.G. Evertsz, H-O Peitgen and R.F. Voss, editors. World Scientific Publishing Co., Singapore, 1996.

Taylor, S. (1973) *Introduction to Measure and Integration*. Cambridge University Press.

Zygmund, A. (2002) *Trigonometric Series*. Cambridge University Press, third edn.

Paper III

Some Statistical Properties of an Ultra-Wideband Communication Channel Model

Author list

Anders Gorst-Rasmussen
Aalborg University, Denmark

Summary

An ultra-wideband (UWB) radio transmits information in the form of a series of closely spaced, ultra-short pulses. In contrast to conventional radio signals, a UWB signal is well localised in the time domain but poorly localised in the frequency domain. The temporal localisation enables very high data transmission rates over short ranges and an ability to distinguish closely spaced multipath components of a received signal. This makes UWB suited for use in rich multipath environments such as office buildings or industrial environments. In the present paper, we consider a basic narrowband wireless channel model with Rayleigh fading in a UWB regime where the spectral bandwidth is very large and there is infinitely rich multipath diversity. Within this limiting regime, we establish a central limit theorem for the minimum mean squared error of both the infinite-length and finite-length optimal linear equaliser. Our approach relies on general central limit results for nonlinear functionals of continuous-time Gaussian vector processes and works generally for nonlinear statistics of the channel frequency response under Rayleigh fading.

Supplementary info

This manuscript is an *advanced draft* on the strictly mathematical aspects of the problem described in the summary. Journal publication has not been pursued yet; a natural target would be an engineering/signal processing journal, in which case a more detailed assessment of the practical impact and utility of the mathematical results is needed.

1. Introduction

A conventional radio signal is an electromagnetic wave which has all its power concentrated in a relatively narrow band of frequencies within the radio spectrum. By strictly regulating spectrum bandwidth usage, different radio technologies can coexist without risking harmful interference. However, since the amount of information carriable by a radio signal is proportional to its bandwidth, the data transmission rate of conventional narrowband radio is limited. Narrowband radio is also not ideal for deployment in complex transmission environments with many obstacles such as office buildings or industrial environments due to poor penetration properties and the risk of multipath fading. The latter is a descriptive term for the situation where multiple 'echoes' of the radio signal, arising due to scattering off objects in the transmission environment, interfere and potentially decrease signal strength.

Ultra-wideband (UWB) radio has recently attracted much attention for its potential to alleviate some of the limitations of conventional radio. A UWB signal is comparable to Morse code, consisting of a series of closely spaced pulses of extremely short duration. Because a UWB signal is well localised in the time domain, it is poorly localised in the frequency domain in accordance with the Fourier uncertainty principle, distributing its power over several GHz of spectral bandwidth. See Figure 1. Co-existence with conventional radio technology is possible because the signal power at any given frequency is required to be small. This limits the range of UWB but also makes it a promising short-range wireless technology for low-powered devices. It is intuitively clear that UWB can support data transmission rates much higher than those of narrowband radio. Because UWB signals span a large range of frequencies, they are less susceptible to multipath fading and have excellent penetration properties (Win and Scholtz, 1998).

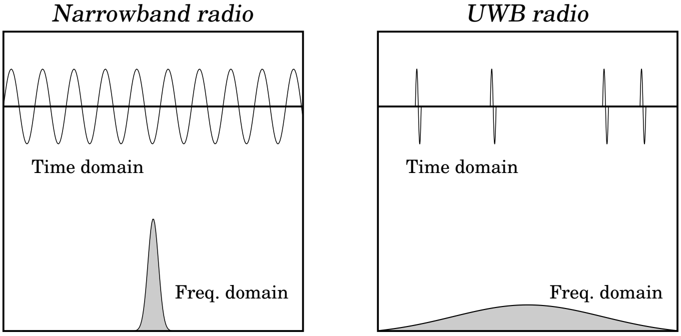

Figure 1. Conceptual difference between narrowband and UWB. In contrast to narrowband signals, UWB signals are well localised in the time domain and accordingly poorly localised in the frequency domain.

Broadly described, the present paper deals with theoretical assessments of performance limits of UWB systems. One way to approach such assessments is to consider a conventional transmitter-receiver model for a band-limited radio signal and study it in the limiting regime where the bandwidth grows large. A basic such wireless channel model for use in stationary environments asserts that the received signal y is the output of a linear time-invariant system of the form:

$$y(t) = \int_0^\infty h(s)x(t-s)\mathrm{d}s + e(t), \qquad t \in \mathbb{R}; \tag{1}$$

where x is the transmitted signal, e is noise, and h depends on the transmission environment (the wireless channel); all quantities being wide-sense stationary, complex-valued stochastic processes. The so-called channel filter h models the multipath propagation where x reaches the receiver in the form of multiple 'echoes' due to scattering off objects surrounding transmitter and receiver. A physically inspired model for h sets $h(t) = \sum_{l=0}^{L-1} h_l \delta(t-\tau_l)$ for δ the Dirac delta-function so that y is the sum of L noisy copies of x, attenuated by random factors $h_0, \ldots, h_{L-1}$ and delayed by random times $\tau_0, \ldots, \tau_{L-1}$. However, one need not restrict h in this manner *a priori*. In fact, suppose that x is band-limited in the frequency domain with bandwidth B. The sampling theorem (for example, Jerri (1977)) then implies that (1) is (approximately) equivalent to a discrete-time, sampled model

$$Y_n = \sum_{l=0}^{L-1} H_l X_{n-l} + E_n, \qquad n \in \mathbb{Z}, \tag{2}$$

for a suitably large L. Here $Y_n := y(nT)$, $X_n := x(nT)$, and $E_n := e(nT)$ with $T \approx 1/B$ the so-called intersymbol time which intuitively is the time between different pulses of information. It is convenient to think of H_l as the samples $h(lT)$; more accurately, it is a sequence of samples from a lowpass-filtered version of h (Tse and Visnawath (2005), Section 2.2.3). The complex random variables $H_0,\dots,H_{L-1}$ are called channel taps.

In a narrowband system, L is usually assumed to take on a small value to reflect rapid convergence to zero of the sequence $\{\mathbb{E}|h(nT)|^2\}$; either because T is relatively large (small B) or/and $t \mapsto \mathbb{E}|h(t)|^2$ is rapidly decaying because the signal only travels along a small number of different paths (low multipath diversity). For UWB radios in challenging (rich multipath diversity) transmission environments, different considerations apply:

1. As the intersymbol time T grow smaller (B grows larger), the effect of the channel filter h at essentially all delays (i.e. all paths) can be resolved by the receiver.
2. As multipath diversity increases, we effectively think of h as a continuous process, i.e. there is no natural upper limit on the number of taps L (although the variance of each tap will tend to zero with increasing l to reflect conservation of energy).

This points to a theoretical analysis of the model (2) for UWB radio in a regime where L and B grow large while the sequence $\mathbb{E}|H_0|^2, \mathbb{E}|H_1|^2, \dots$ tends to zero at a suitable rate.

This paper concerns central limit theorems (CLTs) for statistics derived from a sampled model of the form (2) in the 'large B, large L' (large bandwidth, rich multipath) regime. We focus on a classical statistic, the minimum mean squared error (MMSE) of the optimal linear estimator of the signal X (the Wiener filter estimator). The problem of establishing a CLT for the MMSE of this estimator was originally proposed and explored in an unpublished manuscript by Pereira *et al.* (2005) who attempted to use a CLT for mixing sequences stated in Ibragimov and Linnik (1971). The problem was revisited in the MSc thesis by Rubak (2007) who formalised the problem and explored the approach based on CLTs for mixing sequences in more depth. However, an actual proof of the asserted CLT was not found. In this paper, we take a different approach and provide a solution to the problem based on an extension of the general CLT for nonlinear functionals of Gaussian vector processes due to Bardet and Surgailis (2011).

2. Model and problem statement

2.1. *Channel model*

Recall that a complex-valued random variable Z is called complex Gaussian with mean μ and variance $\sigma^2 > 0$ iff the vector $[\mathrm{Re}\,Z, \mathrm{Im}\,Z]^\top$ is bivariate Gaussian with mean μ and covariance matrix $I\sigma^2/2$ for I the identity matrix.

We consider the sampled model (2) under the following detailed assumptions:

(i). $X := \{X_n : n \in \mathbb{Z}\}$ is a sequence of independent and identically distributed (IID) complex random variables.

(ii). $E := \{E_n : n \in \mathbb{Z}\}$ is a sequence of IID mean-zero complex Gaussian random variables with variance σ_E^2.

(iii). $H^L := \{H_l : l = 0,1,\dots,L-1\}$ is a sequence of independent mean-zero complex Gaussian random variables with variance

$$\mathbb{E}|H_l|^2 = \int_{l/B}^{(l+1)/B} p(t)\mathrm{d}t, \quad l = 0,\dots,L-1;$$

with $p : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ a probability density function satisfying $\int_0^{L/B} p(t)\mathrm{d}t \to 1$, $L,B \to \infty$.

By the mean value theorem, (iii) implies that each $\mathbb{E}|H_l|^2$ is proportional to $1/B \approx T$ with a constant of proportionality that varies according to the smoothness properties of p. The last part of assumption (iii) has the physical interpretation that we account for all energy in the signal X when taking into account all possible signal paths ($L, B \to \infty$). Slightly more generally, p could be the density of some finite measure on $\mathbb{R}_0^+$.

The assumption (iii) of independent Gaussian H_ls is known as a Rayleigh fading model to reflect the fact that $|H_0|, \dots, |H_{L-1}|$ are independently Rayleigh distributed. This is a classical wireless channel model (Tse and Visnawath (2005), Section 2.4.2) based on the physical assumption that each channel tap 'aggregates' propagation along a large number of independent paths. A limiting argument based on the CLT then implies approximately Gaussian taps. It has been argued (Molisch (2005) and references herein) that neither approximate Gaussianity nor independence of taps are universally tenable assumptions for UWB radio because of its ability to resolve closely spaced, not necessarily independent paths; such as different paths due to scattering off the same obstacle. On the other hand, empirical support for a standard Rayleigh fading model for UWB signals was provided by Schuster and Bölcskei (2007).

2.2. *Problem statement*

Consider the so-called equalisation problem of estimating the signal sequence X from Y under the channel model of the preceding section, assuming knowledge of channel taps H^L (which one obtains in practice by probing the channel with a test signal known by the receiver). A particularly simple type of estimator is a linear deconvolution estimator of the form $\hat{X}_n = \sum_{k=-\infty}^{\infty} \hat{W}_k Y_{n-k}$ for a fixed sequence $\hat{W}$. If we choose $\hat{W}$ such that the mean squared error (MSE) $\mathbb{E}(|X_n - \hat{X}_n|^2 | H^L)$ is minimised, we obtain the classical (non-causal) Wiener filter estimator of X. Note that the MSE is independent of n, by stationarity.

It can be shown that the minimum MSE (MMSE) is given by

$$M_{L,B} =: \min_{\hat{W} \in \mathbb{C}^\infty} \mathbb{E}\left(\left| X_n - \sum_k \hat{W}_k Y_{n-k} \right|^2 \middle| H^L \right) = B^{-1} \int_0^B \frac{\sigma_E^2}{|\mathscr{H}_{L,B}(\omega)|^2 + \rho^{-1}} \, \mathrm{d}\omega; \tag{3}$$

where $\rho := \sigma_X^2 / \sigma_E^2$ is the signal-to-noise ratio and $\mathscr{H}_{L,B}$ is the discrete-time Fourier transform of H^L, more commonly known as the channel frequency response,

$$\mathscr{H}_{L,B}(\omega) := \sum_{l=0}^{L-1} H_l \mathrm{e}^{-i2\pi\omega l/B}, \qquad \omega \in \mathbb{R}. \tag{4}$$

The quantity (3) will be referred to as the MMSE of the infinite-length MMSE equaliser.

In practice, one can only use a finite number of observations N for estimation. Defining for each i the vectors $\mathbf{Y}_i^N := [Y_i, \dots, Y_{i-N+1}]^\top$, $\mathbf{X}_i^N := [X_i, \dots, X_{i-N+1}]^\top$, $\mathbf{E}_i^N := [E_i, \dots, E_{i-N+1}]^\top$, and $\mathbf{H}_L := [H_0, \dots, H_{L-1}]$, the finite submodel derived from (2) takes the form:

$$\mathbf{Y}_i^N = H\mathbf{X}_i^N + \mathbf{E}_i^N, \quad \text{where } H := \begin{bmatrix} \mathbf{H}_L & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{H}_L & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{H}_L & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & & \mathbf{H}_L \end{bmatrix} \tag{5}$$

We then estimate $X_{i-\Delta}$ by an inner product $\hat{\mathbf{W}}^*\mathbf{Y}_i^N$ where the introduction of a delay $0 \leqslant \Delta \leqslant N-1$ ensures causality of the filter in practice. The finite-dimensional analogue of (3) becomes

$$M_{L,B}^N := \inf_{\hat{\mathbf{W}}\in\mathbb{C}^N} \min_{0\leqslant\Delta\leqslant N-1} \mathbb{E}\big(X_{i-\Delta} - \hat{\mathbf{W}}^*\mathbf{Y}_i^N \,|\, H^L\big)^2 = \sigma_E^2 \min \operatorname{diag}(H^*H + \rho^{-1}I)^{-1}; \tag{6}$$

with I the $(N+L-1)\times(N+L-1)$ identity matrix. This is the MMSE for the finite-length MMSE equaliser.

The formal derivation of (3) and (6) is quite lengthy and involves stating and solving the Wiener-Hopf equations for the respective models (2) and (5). Details can be found in, for example, Kurzweil (2000), Chapter 10; or Cioffi (2003), Chapter 3. These references also describe how the actual filter coefficients $\hat{W}$ and $\hat{\mathbf{W}}$ look.

The problem in this paper is simple: we seek conditions under which it holds that

$$\begin{aligned} B^{1/2}\big(M_{L,B} - \mu_1\big) &\xrightarrow{D} \mathrm{N}(0,\sigma_1^2), \qquad L,B \to \infty, \\ B^{1/2}\big(M_{L,B}^N - \mu_2\big) &\xrightarrow{D} \mathrm{N}(0,\sigma_2^2), \qquad L,B,N \to \infty; \end{aligned}$$

for suitable μ_i, σ_i^2, where $\to^D$ denotes convergence in distribution and $\mathrm{N}(\mu,\sigma^2)$ is the normal distribution with mean μ and variance σ^2.

3. CLT for the infinite-length equaliser

We start by investigating the behaviour of the channel frequency response $\mathscr{H}_{L,B}$ from (4) in the limit when $L,B \to \infty$. It is clear that $\mathscr{H}_{L,B}$ is a Gaussian stochastic process. Because the channel taps H^L are assumed to be independent, $\mathscr{H}_{L,B}$ is wide-sense stationary with auto-covariance function (ACVF) given by

$$r_{L,B}(\omega) := \mathbb{E}\big\{\mathscr{H}_{L,B}(\omega)\overline{\mathscr{H}_{L,B}(0)}\big\} = \sum_{l=0}^{L-1} \mathrm{e}^{-i2\pi\omega l/B} \int_{l/B}^{(l+1)/B} p(t)\mathrm{d}t, \quad \omega \in \mathbb{R}. \tag{7}$$

Denote henceforth

$$\begin{aligned} P(t) &:= \int_0^t p(s)\mathrm{d}s, \quad t \geqslant 0, \\ P^{-1}(x) &:= \inf\{t \in \mathbb{R} : x \leqslant P(t)\}, \quad 0 < x < 1; \end{aligned}$$

and define a complex-valued Gaussian process $\mathscr{H}_\infty$ by the stochastic integral

$$\mathscr{H}_\infty(\omega) := \int_0^1 \mathrm{e}^{-i2\pi\omega P^{-1}(x)}\mathrm{d}W(x), \qquad \omega \in \mathbb{R},$$

with W standard complex Brownian motion on the unit interval, i.e. $W = (W_1 + iW_2)/\sqrt{2}$ with W_1, W_2 independent standard Brownian motions on $[0,1]$. By the Itô isometry, $\mathscr{H}_\infty$ is wide-sense stationary with ACVF

$$r_\infty(\omega) := \mathbb{E}\big\{\mathscr{H}_\infty(\omega)\overline{\mathscr{H}_\infty(0)}\big\} = \int_0^1 \mathrm{e}^{-i2\pi\omega P^{-1}(x)}\mathrm{d}x = \int_0^\infty \mathrm{e}^{-i2\pi\omega x} p(x)\mathrm{d}x =: \hat{p}(\omega), \quad \omega \in \mathbb{R}. \tag{8}$$

This is simply the Fourier transform of p.

It holds that $\mathscr{H}_\infty$ is the mean square limit of $\mathscr{H}_{L,B}$ when $L,B \to \infty$. To see this, note that we may without loss of generality assume the sequence of channel taps H^L to be given by the increments of W,

$$H_l = W[P\{(l+1)/B\}] - W[P(l/B)], \quad l = 0,\dots,L-1,$$

Defining intervals $I_l := [P(l/B), P\{(l+1)/B\}]$ for $l = 0,\dots,L-1$, we may then write

$$\mathscr{H}_{L,B}(\omega) = \int_0^1 \sum_{l=0}^{L-1} \mathbb{1}(x \in I_l) \mathrm{e}^{-i2\pi\omega l/B} \, \mathrm{d}W(x),$$

where $\mathbb{1}(\cdot \in A)$ is the indicator function of a set A. The following Lipschitz property holds generally for the complex exponential:

$$|\mathrm{e}^{ia\nu} - \mathrm{e}^{ia\omega}| \leqslant \sqrt{2}a|\nu - \omega|, \quad a, \nu, \omega \in \mathbb{R}. \tag{9}$$

Using this, the Itô isometry, and Jensen's inequality, we obtain

$$\begin{aligned}
\mathbb{E}|\mathscr{H}_\infty(\omega) - \mathscr{H}_{L,B}(\omega)|^2 &= \int_0^1 \Bigg| \sum_{l=0}^{L-1} \mathbb{1}(x \in I_l)\big(\mathrm{e}^{-i2\pi\omega P^{-1}(x)} - \mathrm{e}^{-i2\pi\omega l/B}\big) && (10)\\
&\qquad + \mathbb{1}\{x \in (P(L/B), 1]\} \mathrm{e}^{-i2\pi\omega P^{-1}(x)} \Bigg|^2 \mathrm{d}x && (11)\\
&\leqslant 2 \int_0^1 \sum_{l=0}^{L-1} \mathbb{1}(x \in I_l) \big|\mathrm{e}^{-i2\pi\omega P^{-1}(x)} - \mathrm{e}^{-i2\pi\omega l/B}\big|^2 \mathrm{d}x + 2\{1 - P(L/B)\} && (12)\\
&\leqslant 16\pi^2\omega^2 \sum_{l=0}^{L-1} \int_{l/B}^{(l+1)/B} (x - l/B)^2 p(x) \mathrm{d}x + 2\{1 - P(L/B)\} && (13)\\
&\leqslant 16\pi^2\omega^2 P(L/B) B^{-2} + 2\{1 - P(L/B)\}; && (14)
\end{aligned}$$

which converges to zero when $L,B \to \infty$, by the assumption $P(L/B) \to 1$ from Section 2.1.

Define the following limiting variant of the infinite-length equaliser MMSE:

$$M_B^\infty := B^{-1} \int_0^B \frac{\sigma_E^2}{|\mathscr{H}_\infty(\omega)|^2 + \rho^{-1}} \, \mathrm{d}\omega. \tag{15}$$

This quantity exists (as a Riemann integral) almost surely because almost all sample paths of $\mathscr{H}_\infty$ are continuous almost everywhere. Our first CLT will relate to M_B^∞ which is simpler to deal with than $M_{L,B}$ since the integrand does not depend on L,B. As we will show, $M_{L,B}$ and M_B^∞ converge in distribution to the same limit when $L,B \to \infty$.

The CLT for M_B^∞ requires a few preliminary lemmas. The first summarises essential regularity properties of the function acting on $\mathscr{H}_\infty$ in the integrand in (15). Refer to the appendix for an explanation of the notion of Hermite rank of a function.

LEMMA 1. *Let Z_1, Z_2 be independent standard Gaussian random variables and define the function $\tilde{\varphi}: \mathbb{R}^2 \to \mathbb{R}^+$ by*

$$\tilde{\varphi}(x,y) := \frac{\sigma_E^2}{x^2 + y^2 + \rho^{-1}} - \mathbb{E}\left(\frac{\sigma_E^2}{Z_1^2 + Z_2^2 + \rho^{-1}}\right). \tag{16}$$

Then $\tilde{\varphi}$ has Hermite rank 2. Also, $\tilde{\varphi}$ is uniformly bounded and Lipschitz continuous.

Proof. The first few nontrivial Hermite polynomials are given by $H_1(x) = x$ and $H_2(x) = x^2 - 1$. Since $\tilde{\varphi}(Z_1, Z_2)Z_i$ is symmetrically distributed, it follows that $\mathbb{E}\{\tilde{\varphi}(Z_1,Z_2)Z_i\} = 0$ for $i = 1,2$. On the other hand, the random variable $\tilde{\varphi}(Z)Z_1^2$ is strictly positive, implying $\mathbb{E}\{\tilde{\varphi}(Z_1,Z_2)Z_1^2\} > 0$. Hence $\tilde{\varphi}$ has Hermite rank 2.

Clearly, $\tilde{\varphi}$ is uniformly bounded (by $\sigma_E^2\rho$). Lipschitz continuity follows from the mean value theorem since the partial derivatives of $\tilde{\varphi}$ are also uniformly bounded. ∎

LEMMA 2. *Consider the wide-sense stationary complex Gaussian process* $\theta(\omega) := \int_0^1 g(x)\mathrm{e}^{i\omega f(x)}\mathrm{d}W(x)$ *where* W *is a standard complex Brownian motion on the unit interval and* f, g *are suitably regular real-valued functions on* $[0,1]$. *Denote* $r_\theta(\omega) := \mathbb{E}\{\theta(\omega)\overline{\theta(0)}\}$. *Set* $\theta_1 := \mathrm{Re}\,\theta$, $\theta_2 := \mathrm{Im}\,\theta$, *and* $r^{(ij)}(\omega) := \mathbb{E}\{\theta_i(\omega)\theta_j(0)\}$, $i,j = 1,2$. *Then* $\max_{i,j}|r^{(ij)}|^2 \leqslant 2|r_\theta|^2$.

Proof. The result follows from the identities $|r^{(11)}| + |r^{(22)}| = |\mathrm{Re}\,r_\theta|$ and $|r^{(12)}| + |r^{(21)}| = |\mathrm{Im}\,r_\theta|$ which are easily verified by direct calculations using that

$$\sqrt{2}\theta(\omega) = \int_0^1 \left[g(x)\cos\{\omega f(x)\}\mathrm{d}W_1(x) - g(x)\sin\{\omega f(x)\}\mathrm{d}W_2(x)\right]$$
$$+ i\int_0^1 \left[g(x)\sin\{\omega f(x)\}\mathrm{d}W_1(x) + g(x)\cos\{\omega f(x)\}\mathrm{d}W_2(x)\right];$$

for W_1, W_2 independent standard Brownian motions on $[0,1]$. Combining these identities with Jensen's inequality, we find that $\max_{i,j}|r^{(ij)}|^2 \leqslant (\sum_{i,j}|r^{(ij)}|)^2 \leqslant 2|r_\theta|^2$ as desired. ∎

We now return to the the limiting variant of the infinite-length equaliser MMSE. Provided that p is square-integrable, the variance of M_B^∞ decays at a rate B^{-1}, as recorded in the following proposition.

PROPOSITION 1. *Denote* $\mu := \mathbb{E}(M_B^\infty)$ *and* $\sigma^2 := \lim_{B\to\infty}\mathrm{Var}(B^{1/2}M_B^\infty)$. *Suppose that* $p \in \mathscr{L}^2(\mathbb{R}_0^+)$. *Then* μ *and* σ^2 *are well-defined and given by*

$$\mu = \int_0^\infty \frac{\sigma_E^2}{\omega + \rho^{-1}}\mathrm{e}^{-\omega}\mathrm{d}\omega, \quad \textit{and } \sigma^2 = 2\int_0^\infty r_\infty^\varphi(\omega)\mathrm{d}\omega; \tag{17}$$

where $r_\infty^\varphi(\omega) := \mathbb{E}[\varphi\{\mathscr{H}_\infty(\omega)\}\varphi\{\mathscr{H}_\infty(0)\}]$ *and*

$$\varphi(z) := \frac{\sigma_E^2}{|z|^2 + \rho^{-1}}, \quad z \in \mathbb{C}. \tag{18}$$

Proof. The expression for $\mu = \mathbb{E}(M_B^\infty)$ follows immediately by noting that the modulus squared of a standard complex Gaussian random variable is exponentially distributed with rate parameter 1.

Concerning σ^2, it holds that

$$\mathrm{Var}(B^{1/2}M_B^\infty) = B^{-1}\mathbb{E}\left[\int_0^B \varphi\{\mathscr{H}_\infty(\omega)\}\mathrm{d}\omega\right]^2 = 2B^{-1}\int_0^B (B-\omega)r_\infty^\varphi(\omega)\mathrm{d}\omega; \tag{19}$$

using symmetry of r_∞^φ around 0 and the following basic identity for integrable g,

$$\int_0^a\int_0^a g(v-\omega)\mathrm{d}v\mathrm{d}\omega = \int_{-a}^a (a - |v|)g(v)\mathrm{d}v. \tag{20}$$

Combining Lemma A2 in the appendix with Lemma 1-2, there exists $C > 0$ such that $r_\infty^\varphi(\omega) \leqslant C|r_\infty(\omega)|^2 = C|\hat{p}(\omega)|^2$. Since $p \in \mathscr{L}^2(\mathbb{R}_0^+)$, the right-hand side of (19) converges to σ^2 in (17) when $B \to \infty$, by dominated convergence. ∎

We can now prove the CLT for M_B^∞ (with μ, σ^2 defined in Proposition 1).

THEOREM 1. *Assume that* $p \in \mathscr{L}^2(\mathbb{R}_0^+)$. *Then* $B^{1/2}(M_B^\infty - \mu) \xrightarrow{D} \mathrm{N}(0,\sigma^2)$ *when* $B \to \infty$.

Proof. We will apply Theorem A1 to the bivariate Gaussian process defined by $(X_1,X_2) := \sqrt{2}(\mathrm{Re}\,\mathscr{H}_\infty, \mathrm{Im}\,\mathscr{H}_\infty)$ and the function $\tilde{\varphi}(\cdot/\sqrt{2})$ with $\tilde{\varphi}$ defined in (18). By Lemma 1, $\tilde{\varphi}(\cdot/\sqrt{2})$ satisfies the relevant assumptions of the theorem with Hermite rank $\tau = 2$. We proceed to check assumptions A and B of Theorem A1.

For assumption A, Lemma 2 implies

$$\sup_{\omega\in[0,B]} \int_0^B |r^{(ij)}(\nu-\omega)|^2 \mathrm{d}\nu \leqslant 2 \sup_{\omega\in[0,B]} \int_0^B |r_\infty(\nu-\omega)|^2 \mathrm{d}\nu \leqslant \int_{-B}^B |\hat{p}(\nu)|^2 \mathrm{d}\nu;$$

using that $-B \leqslant \nu - \omega \leqslant B$ for $0 \leqslant \nu, \omega \leqslant B$. The right-hand side of the display is finite, by the assumptions and Plancherel's theorem. Hence assumption A holds. Assumption B also holds since, given m and taking $S_{B,m} := \{(\nu,\omega) \in [0,B]^2 : m \leqslant |\nu-\omega| \leqslant B-m\}$, the identity (20) implies

$$B^{-1}\int_{S_{B,m}} |r^{(ij)}(\nu-\omega)|^2 \mathrm{d}\nu\,\mathrm{d}\omega \leqslant 2B^{-1}\int_m^{B-m} (B-\omega)|r_\infty(\omega)|^2 \mathrm{d}\omega \leqslant 2\int_m^\infty |\hat{p}(\omega)|^2 \mathrm{d}\omega; \tag{21}$$

which can be made arbitrarily small by choice of m, by Plancherel's theorem. ■

We next seek to extend Theorem 1 to $M_{L,B}$. This will again be done by applying Theorem A1. The main nuisance is to ensure a sufficiently rapid convergence to zero of the ACVF $r_{L,B}$ of $\mathscr{H}_{L,B}$. Since $r_{L,B}$ is formally a (partial) Fourier series, its convergence properties are determined by the regularity properties of the sequence of 'coefficients' $\int_{l/B}^{(l+1)/B} p(t)\mathrm{d}t$, which in turn depend on regularity properties of p.

Denote by $\mathrm{V}_a^b(f)$ the total variation of a real function f over an interval $[a,b] \subseteq \mathbb{R}$, i.e.

$$\mathrm{V}_a^b(f) := \sup\left\{\sum_{i=1}^n |f(x_i) - f(x_{i-1})| : a \leqslant x_0 < \cdots < x_n \leqslant b,\ n \in \mathbb{N}\right\}.$$

We can then bound $r_{L,B}$ explicitly as follows.

LEMMA 3. *Suppose that p is continuous. For* $0 < |\omega| < B$, *it holds that* $|r_{L,B}(\omega)| \leqslant B^{-1}\{p(0) + \mathrm{V}_0^{L/B}(p)\}|\sin(\pi\omega B^{-1})|^{-1}$.

Proof. It is a well known result for the Dirichlet kernel that $|\sum_{k=0}^n \mathrm{e}^{ikx}| \leqslant |\sin(x/2)|^{-1}$ for $n \in \mathbb{N}$ and $0 < |x| < 2\pi$. Hence

$$\left|\sum_{l=0}^k \mathrm{e}^{-i2\pi\omega l/B}\right| \leqslant |\sin(\pi\omega B^{-1})|^{-1}, \qquad k \in \mathbb{N}, \quad 0 < |\omega| < B.$$

Denote $a_l := \mathrm{e}^{-i2\pi\omega l/B}$, $A_k := \sum_{l=0}^k a_l$, and $b_l := \int_{l/B}^{(l+1)/B} p(t)\mathrm{d}t$. Summing by parts,

$$|r_{L,B}(\omega)| = \left|A_{L-1}b_{L-1} + \sum_{l=0}^{L-2} A_l(b_{l+1} - b_l)\right| \leqslant |\sin(\pi\omega B^{-1})|^{-1} \max\left\{|b_{L-1}|, \sum_{l=0}^{L-2} |b_{l+1} - b_l|\right\}. \tag{22}$$

By the mean value theorem, there exists $(l-1)/B \leqslant \nu_l \leqslant l/B$ and $l/B \leqslant \omega_l \leqslant (l+1)/B$ for $l = 0,\ldots,L-1$ such that

$$\sum_{l=0}^{L-2} |b_{l+1} - b_l| = B^{-1}\sum_{l=0}^{L-2} |p(\nu_l) - p(\omega_l)| \leqslant B^{-1}\mathrm{V}_0^{L/B}(p).$$

Moreover, for $x \in [0,L/B]$, it holds that $p(x) \leqslant p(0) + |p(x) - p(0)| \leqslant p(0) + \mathrm{V}_0^{L/B}(p)$. Then $b_{L-1} \leqslant B^{-1}\{p(0) + \mathrm{V}_0^{L/B}(p)\}$, by the mean value theorem. Comparing with (22) yields the statement of the lemma. ■

We have the following CLT for $M_{L,B}$. Again, μ and σ^2 are defined in Proposition 1.

THEOREM 2. *Suppose that p is continuous, $p \in \mathscr{L}^2(\mathbb{R}_0^+)$, and $\mathrm{V}_0^{L/B}(p) = O(1)$, $L,B \to \infty$. Then*

$$B^{1/2}(M_{L,B} - \mu) \xrightarrow{D} \mathrm{N}(0,\sigma^2), \qquad L,B \to \infty. \tag{23}$$

Proof. It holds that $\mathbb{E}(M_{L,B}) \to \mu$ by dominated convergence, since $\mathscr{H}_{L,B}$ is complex Gaussian with $\mathbb{E}|\mathscr{H}_{L,B}(\omega)|^2 = \int_0^{L/B} p(t)\mathrm{d}t \to 1$ when $L,B \to \infty$. We moreover claim that

$$\lim_{L,B\to\infty} \mathrm{Var}(B^{1/2}M_{L,B}) = \sigma^2. \tag{24}$$

To see this, set $r_{L,B}^{\varphi}(\omega) := \mathbb{E}[\varphi\{\mathscr{H}_{L,B}(\omega)\}\varphi\{\mathscr{H}_{L,B}(0)\}]$ with φ defined in (18). As in (19), it holds that

$$\mathrm{Var}(B^{1/2}M_{L,B}) = 2B^{-1}\int_0^B (B-\omega)r_{L,B}^{\varphi}(\omega)\mathrm{d}\omega. \tag{25}$$

The pointwise mean square convergence of $\mathscr{H}_{L,B}$ to $\mathscr{H}_\infty$ established in (10)-(14) implies weak convergence of $\mathscr{H}_{L,B}$ to $\mathscr{H}_\infty$, in the sense of finite-dimensional distributions. Since φ is continuous and bounded (Lemma 1), we have pointwise convergence for $\omega \in \mathbb{R}$,

$$r_{L,B}^{\varphi}(\omega) \to r_\infty^{\varphi}(\omega), \qquad L,B \to \infty. \tag{26}$$

Observe that boundedness of φ is not essential for this result; we could instead rely on uniform integrability arguments as in Billingsley (1995), pp. 338-339.

Lemma A2 and Lemma 2 implies the existence of $C > 0$ such that $|r_{L,B}^{\varphi}(\omega)| \leqslant C|r_{L,B}(\omega)|^2$. Set $A_{m,B} := [m, B-m]$, taking $A_{m,B} = \emptyset$ if $m \geqslant B/2$. Since $x < \tan x$ for $0 < x < \pi/2$, symmetry of $r_{L,B}$ around $\omega = B/2$ together with Lemma 3 implies

$$B^{-1}\left|\int_{A_{m,B}} (B-\omega)r_{L,B}^{\varphi}(\omega)\mathrm{d}\omega\right| \leqslant 2C\int_m^{B/2} |r_{L,B}(\omega)|^2\mathrm{d}\omega \tag{27}$$

$$\leqslant 2CB^{-1}\{p(0) + \mathrm{V}_0^{L/B}(p)\}^2|\tan(\pi B^{-1}m)|^{-1} \tag{28}$$

$$\leqslant 2C\{p(0) + \mathrm{V}_0^{L/B}(p)\}^2\pi^{-1}m^{-1}, \tag{29}$$

which, uniformly in L,B, can be made arbitrarily small by choice of m. Hence the sequence of integrands on the right hand side of (25) is tight (and obviously uniformly integrable, being bounded). Combining this with (26), Vitali's convergence theorem (Folland (1999), p. 187) then implies (24).

Having established the asymptotics of the mean and variance of $M_{L,B}$, we proceed to prove the CLT. Set $\tilde{\sigma}^2 := \mathrm{Var}\{\mathrm{Re}\,\mathscr{H}_{L,B}(0)\} = \mathrm{Var}\{\mathrm{Im}\,\mathscr{H}_{L,B}(0)\}$. As in the proof of Theorem 1, we will apply Theorem A1 to $(X_1,X_2) := \tilde{\sigma}^{-1}(\mathrm{Re}\,\mathscr{H}_{L,B}, \mathrm{Im}\,\mathscr{H}_{L,B})$ and the function $\varphi_{\tilde{\sigma}}(\cdot) := \tilde{\varphi}(\,\cdot\,\tilde{\sigma})$ which satisfies the relevant assumptions with Hermite rank $\tau = 2$. Moreover,

$$\tilde{\sigma}^2 = \frac{1}{2}\int_0^{L/B} p(t)\,\mathrm{d}t \to \frac{1}{2}, \qquad L,B \to \infty,$$

so dominated convergence implies $\varphi_{\tilde{\sigma}} \to \tilde{\varphi}(\,\cdot\,/\sqrt{2})$, $L,B \to \infty$, in (Gaussian) mean square.

We may replace the integrands $|r^{(ij)}(\nu-\omega)|^{\tau}$ in assumptions A and B of Theorem A1 with $2|r_{L,B}(\nu-\omega)|^2$ (Lemma 2). Regarding assumption A, let $a_l := \int_{l/B}^{(l+1)/B} p(t)\mathrm{d}t$. The collection of functions $\{\nu \mapsto \exp(i2\pi B^{-1}n\nu) : n \in \mathbb{Z}\}$ is $\mathscr{L}^2[0,B]$-orthogonal. Then, by Jensen's inequality,

$$\sup_{\omega\in[0,B]} \int_0^B |r_{L,B}(\nu-\omega)|^2 \mathrm{d}\nu = B\sum_{l=0}^{L-1} a_l^2 + \sup_{\omega\in[0,B]} \int_0^B \sum_{l\neq m} a_l a_m \mathrm{e}^{-i2\pi(\nu-\omega)l/B} \mathrm{e}^{i2\pi(\nu-\omega)m/B} \mathrm{d}\nu \tag{30}$$

$$\leqslant \sum_{l=0}^{L-1} \int_{l/B}^{(l+1)/B} p(t)^2 \mathrm{d}t, \tag{31}$$

which is finite since $p \in \mathscr{L}^2(\mathbb{R}_0^+)$. Turning to assumption B, the identity (20) and calculations analogous to those in (27)-(29) imply

$$B^{-1}\int_{S_{B,m}} |r_{L,B}(\nu-\omega)|^2 \mathrm{d}\nu\mathrm{d}\omega \leqslant 2\int_m^{B-m} |r_{L,B}(\omega)|^2 \mathrm{d}\omega \leqslant 4\{p(0)+\mathrm{V}_0^{L/B}(p)\}^2 \pi^{-1} m^{-1}; \tag{32}$$

which converges to zero when $m \to \infty$. This proves (38). ■

The convergence result (24) does not depend crucially on Lemma 3 (and hence on continuity and bounded variation of p); by the convergence theorem in Pratt (1960), it suffices to assume mean square convergence in the sense $\int_0^B |r_{L,B}(\omega)|^2 \mathrm{d}\omega \to \int_0^\infty |\hat{p}(\omega)|^2 \mathrm{d}\omega$ when $L,B \to \infty$. On the other hand, it is not obvious how to show convergence of the left-hand side of (32) without making use of the explicit bound in Lemma 3.

4. CLT for the finite-length equaliser

In this section, we derive a CLT for the MMSE of the finite-length equaliser. With μ from Proposition 1, we can write

$$B^{1/2}\big(M_{L,B}^N - \mu\big) = B^{1/2}\big(M_{L,B}^N - M_{L,B}\big) + B^{1/2}(M_{L,B} - \mu). \tag{33}$$

Consequently, if the CLT holds for $M_{L,B}$, Slutsky's lemma implies that we need only prove $B^{1/2}\mathbb{E}|M_{L,B}^N - M_{L,B}| \to 0$ when $N,L,B \to \infty$. The convergence in mean is a consequence of the following sandwich inequality for $M_{L,B}^N$ due to Pereira *et al.* (2005).

LEMMA 4. *Denote $K := N+L-1$ and let φ be defined as in Lemma 1. Then*

$$M_{L,B} \leqslant M_{L,B}^N \leqslant K^{-1}\sum_{j=1}^{K} \varphi\{\mathscr{H}_{L,B}(jB/K)\} + \sigma_X^2 L K^{-1}. \tag{34}$$

Proof. The model associated with the finite-length equaliser MMSE is a submodel of the model associated with the infinite-length equaliser MMSE, hence $M_{L,B} \leqslant M_{L,B}^N$.

Establishing the upper bound for $M_{L,B}^N$ requires more work. Denote in the following by λ_i^A the ith eigenvalue of a Hermitian square matrix A, arranged in order of decreasing absolute value. Consider the matrix H defined in (5). By the spectral theorem, we can write $H^*H = U^*\Lambda U$ for a unitary matrix U and a diagonal matrix Λ consisting of the K nonnegative eigenvalues of H^*H. Then $(H^*H+\rho^{-1}I)^{-1} = U^*(\Lambda+\rho^{-1}I)^{-1}U$ so that the (strictly positive) eigenvalues of $(H^*H+\rho^{-1}I)^{-1}$ are $(\lambda_j^{H^*H}+\rho^{-1})^{-1}$ for $j=1,\dots,K$. Consequently,

$$M^N_{L,B} \leqslant K^{-1}\sigma_E^2 \operatorname{tr}\{(H^*H+\rho^{-1}I)^{-1}\} = K^{-1}\sigma_E^2 \sum_{j=1}^{K} \left(\lambda_j^{H^*H}+\rho^{-1}\right)^{-1}. \tag{35}$$

We next approximate H^*H with a circulant matrix for which we can evaluate eigenvalues explicitly. Specifically, append to the $N\times K$ matrix H an $(L-1)\times K$ matrix E such that the $K\times K$ matrix $\tilde{H} := [H\,E]^\top$ is circulant. By matrix block multiplication, $\tilde{H}^*\tilde{H} = H^*H + E^*E$. Viewing E^*E as a perturbation of H^*H, Hermiticity alongside Weyl's inequalities (Bhatia (1997), Section III.2) imply that

$$\lambda_{i+j-1}^{\tilde{H}^*\tilde{H}} \leqslant \lambda_i^{H^*H} + \lambda_j^{E^*E}, \quad i+j-1 \leqslant K.$$

Since E has $L-1$ rows, E^*E can have rank at most $L-1$, implying $\lambda_j^{E^*E}=0$ for $j\geqslant L$. In particular, the above display implies $\lambda_{i+L-1}^{\tilde{H}^*\tilde{H}} \leqslant \lambda_i^{H^*H}$ for $i\leqslant N$. Then

$$\sum_{j=1}^{K}\left(\lambda_j^{H^*H}+\rho^{-1}\right)^{-1} \leqslant \sum_{j=1}^{N}\left(\lambda_{j+L-1}^{\tilde{H}^*\tilde{H}}+\rho^{-1}\right)^{-1} + \sum_{j=N+1}^{N+L-1}\left(\lambda_j^{H^*H}+\rho^{-1}\right)^{-1} \tag{36}$$

$$\leqslant \sum_{j=1}^{N}\left(\lambda_j^{\tilde{H}^*\tilde{H}}+\rho^{-1}\right)^{-1} + L\rho. \tag{37}$$

From the standard formula for the eigenvalues of a circulant matrix,

$$\lambda_j^{\tilde{H}} = \sum_{l=0}^{L-1} H_l \mathrm{e}^{-i2\pi jl/K} = \mathscr{H}_{L,B}(jB/K);$$

whereby $\lambda_j^{\tilde{H}^*\tilde{H}} = |\mathscr{H}_{L,B}(jB/K)|^2$. Combining this with (35)-(37), the asserted upper bound for $M^N_{L,B}$ in (34) follows. ■

We then have the following CLT for $M^N_{L,B}$ (with μ, σ^2 defined in Proposition 1).

THEOREM 3. *Suppose that $LB^{1/2}=o(N)$ when $N,L,B\to\infty$. Assume that the CLT* (23) *for $M_{L,B}$ holds. Then*

$$B^{1/2}\left(M^N_{L,B}-\mu\right) \to \mathrm{N}(0,\sigma^2), \quad L,B,N\to\infty. \tag{38}$$

Proof. From the decomposition (33), we need only show $B^{1/2}\mathbb{E}|M^N_{L,B}-M_{L,B}|\to 0$, $L,B\to\infty$.

We will use Lemma 4. Denote in the following $K:=N+L-1$ and take $v_j := jB/K$, $j=0,\dots,K$. The upper bound in (34) can be written as the integral of a step function approximation to the function φ defined in (18),

$$K^{-1}\sum_{j=1}^{K}\varphi\{\mathscr{H}_{L,B}(jBK^{-1})\} = B^{-1}\int_0^B \sum_{j=1}^{K} \mathbb{1}\{\omega\in(v_{j-1},v_j]\}\varphi\{\mathscr{H}_{L,B}(v_j)\}\mathrm{d}\omega \tag{39}$$

$$=: B^{-1}\int_0^B \chi_{L,B}(\omega)\mathrm{d}\omega. \tag{40}$$

Denote $\delta := B/K$. By Lipschitz continuity of φ (Lemma 1), there exists a universal constant $C>0$ such that

$$\left|\varphi\{\mathscr{H}_{L,B}(\omega)\} - \chi_{L,B}(\omega)\right| \leqslant C\sup\{|\mathscr{H}_{L,B}(v)-\mathscr{H}_{L,B}(\omega)| : v,\omega\in[0,B],\ |v-\omega|\leqslant\delta\} = Cw_\delta(\mathscr{H}_{L,B});$$

where the right-hand side is a random variable, by a separability argument. Combining Lemma 4 with (40) and applying Hölder's inequality,

$$B^{1/2}\mathbb{E}\big|M^N_{L,B} - M_{L,B}\big| \leqslant CB^{1/2}\big(\mathbb{E}|w_\delta(\mathscr{H}_{L,B})|^2\big)^{1/2} + O(B^{1/2}L/K). \tag{41}$$

Suppose $|\omega - \nu| \leqslant \delta$. Invoking Lipschitz continuity (9) of the complex exponential and independence of channel taps, it holds that

$$\begin{aligned}\mathbb{E}|\mathscr{H}_{L,B}(\nu) - \mathscr{H}_{L,B}(\omega)|^2 &\leqslant \sum_{0\leqslant l,m<L} |\mathbb{E}(H_l\bar{H}_m)|\big|\mathrm{e}^{-i2\pi\nu l/B} - \mathrm{e}^{-i2\pi\omega l/B}\big|\big|\mathrm{e}^{i2\pi\nu m/B} - \mathrm{e}^{i2\pi\omega m/B}\big| \\ &\leqslant 8\pi^2\delta^2 \sum_{0\leqslant l,m<L} |\mathbb{E}(H_l\bar{H}_m)| lm/B^2 \\ &\leqslant 8\pi^2\delta^2 B^{-2}\sum_{l=0}^{L-1} l^2\mathbb{E}|H_l|^2.\end{aligned}$$

Since $\sum_{l=0}^{L-1} l^2\mathbb{E}|H_l|^2 \leqslant L^2$ and $\delta = B/K$, we obtain from (41)

$$B^{1/2}\mathbb{E}|M^N_{L,B} - M_{L,B}| \leqslant B^{1/2}\big[O\{\delta^2L^2B^{-2}\}\big]^{1/2} + O(B^{1/2}L/N) = O(B^{1/2}L/K) + O(B^{1/2}L/N);$$

which converges to zero when $L,B,N \to \infty$, by the assumptions. This implies (38). ■

5. Concluding remarks

The main implications of the results in this paper can be stated concisely as follows:

1. Assuming rich multipath diversity, the MMSE tends, with increasing bandwidth B, to its mean at a rate $B^{1/2}$; and the mean depends only on the variance on noise variance and the signal-to-noise ratio.
2. Under the same assumptions, the (scaled) MMSE is asymptotically Gaussian.

The first implication can be viewed as a general statement about the performance limits of UWB radio and how rapidly we can achieve the asymptotic equalisation performance when increasing the bandwidth. The second statement is interesting from a practical point of view since it in principle enables an experimenter to make approximate probabilistic statements about the MMSE using only knowledge about the 'average' behaviour of a given random wireless channel (in the form of the probability density p). In practice, however, the asymptotic variance in Proposition 1 is difficult to compute for a given p.

We have focused strictly on establishing central limit results for the MMSE. However, our approach will work generally for statistics of the form

$$B^{-1}\int_0^B \xi\{\mathscr{H}_{L,B}(\omega)\}\mathrm{d}\omega; \tag{42}$$

provided that $\xi\colon \mathbb{C} \to [0,\infty)$, when identified with a real-valued function on $\mathbb{R}^2$, is Lipschitz continuous and has Hermite rank at least 2. An important example of a function satisfying these criteria is $\xi(z) = \log(1+\rho|z|^2)$. Using this ξ in (42) leads to the so-called capacity, for which Barriac and Madhow (2004) derived a CLT based on results from Serfling (1968) and heuristic Riemann sum approximation arguments. Our approach is a different and more rigorous way of establishing CLTs for the capacity.

There are several relevant extensions of the results in this paper, one of which would be to allow for correlated channel taps. Nonzero correlation between taps will have substantial implications for the analysis since it entails nonstationarity of $\mathscr{H}_{L,B}$. We have relied heavily on stationarity throughout but the CLT of Bardet and Surgailis (2011) can in fact be modified to avoid the stationarity assumption in Theorem A1. However, it is a nontrivial problem to devise a suitable weak correlation structure on the channel tap sequence which is analytically tractable.

Another relevant extension would be to allow for non-Gaussian channel taps. This is an even more challenging extension since Gaussianity is essential for the CLT in Theorem A1; no general similar CLT results exist in the non-Gaussian case. On the other hand, it may still be possible to establish useful moment bounds and convergence rates even if asymptotic normality fails. The most promising starting point for investigating such extensions comes from the field of time series analysis. Specifically, one may note that $|\mathscr{H}_{L,B}|^2$ is actually the periodogram of the sequence $H_0,\dots,H_{L-1}$. Previous works have investigated the asymptotics of nonlinear functions of the periodogram for general second-order stationary time series (for example, Faÿ *et al.* (2002); Faÿ (2010)). Unfortunately, most of this work has focused on the periodogram asymptotics at Fourier frequencies only. The problem of investigating the asymptotics of nonlinear functions of the periodogram as a continuous-time process is surprisingly difficult, as discussed by Deo and Chen (2000).

Appendix: CLT for functionals of periodic Gaussian vector processes

In this appendix, we describe in detail how to extend the discrete-time CLT in Bardet and Surgailis (2011) so that it applies to a class of continuous-time, possibly periodic Gaussian vector processes. For simplicity, we consider only the case of (wide-sense) stationary processes.

We start out by recalling some properties of multivariate Hermite polynomials (Arcones, 1994): for $x = (x_1,\dots,x_d) \in \mathbb{R}^d$ and a multi-index $\mathbf{k} = (k^{(1)},\dots,k^{(d)}) \in \mathbb{N}_0^d$, the product Hermite polynomial associated with $\mathbf{k}$ is defined as

$$H_{\mathbf{k}}(x) := \prod_{i=1}^{d} H_{k^{(i)}}(x_i),$$

where H_n denotes the nth real Hermite polynomial:

$$H_n(x) := (-1)^n \mathrm{e}^{x^2/2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \mathrm{e}^{-x^2/2}, \quad n \in \mathbb{N}_0.$$

The collection $\{H_{\mathbf{k}} : \mathbf{k} \in \mathbb{N}_0^d\}$ of d-dimensional product Hermite polynomials forms an orthogonal basis for $\mathscr{L}^2(\mathrm{Q})$ with Q the standard Gaussian measure on $\mathbb{R}^d$. In particular, if $\varphi \in \mathscr{L}^2(\mathrm{Q})$ then the following series expansion holds in the $\mathscr{L}^2(\mathrm{Q})$-sense:

$$\varphi(x) = \sum_{|\mathbf{k}| \geqslant \tau} \frac{J_\varphi(\mathbf{k})}{\mathbf{k}!} H_{\mathbf{k}}(x), \qquad x \in \mathbb{R}^d,$$

with $\mathbf{k}! := k^{(1)}! \cdots k^{(d)}!$, $|\mathbf{k}| := k^{(1)} + \cdots + k^{(d)}$, $J_\varphi(\mathbf{k}) := \int H_{\mathbf{k}}(x)\varphi(x)\mathrm{Q}(\mathrm{d}x)$, and τ the Hermite rank of φ defined as

$$\tau := \min\{|\mathbf{k}| : \mathbf{k} \in \mathbb{N}_0^d, J_\varphi(\mathbf{k}) = 0\}.$$

THEOREM A1. *Let* $\{X_n(t) : 0 \leqslant t \leqslant n\} = \{[X_n^1(t),\dots,X_n^d(t)]^\top : 0 \leqslant t \leqslant n\} \in \mathbb{R}^d$, $n \in \mathbb{N}$, *be a triangular array of continuous-time Gaussian vector processes with* X_n *wide-sense stationary for each* n. *Assume that* $\mathbb{E}\{X_n(t)\} = 0$, $\mathbb{E}\{X_n^i(t)X_n^j(t)\} = \delta_{ij}$, *and denote*

$$r_n^{(ij)}(t) := \mathbb{E}\{X_n^i(|t|)X_n^j(0)\}, \quad t \in \mathbb{R}.$$

Let $\{\varphi_n : n \in \mathbb{N}\} \subseteq \mathscr{L}^2(\mathrm{Q})$ *be a collection of functions with Hermite rank at least* τ *satisfying* $\mathrm{Q}\varphi_n = 0$, *and* $\varphi_n \to \varphi \in \mathscr{L}^2(\mathrm{Q})$ *in* Q*-mean square. Assume that for* $1 \leqslant i,j \leqslant d$

$$\text{(A).} \quad \sup_{n \geqslant 1} \sup_{t \in [0,n]} \int_0^n |r_n^{(ij)}(s-t)|^\tau \mathrm{d}s < \infty,$$

$$\text{(B).} \quad \lim_{m\to\infty} \sup_{n \geqslant 1} n^{-1} \int_{S_{n,m}} |r_n^{(ij)}(s-t)|^\tau \mathrm{d}s\mathrm{d}t = 0;$$

with $S_{n,m} := \{(s,t) \in [0,n]^2 : m \leqslant |s-t| \leqslant n-m\}$. *Then there exists* $\sigma^2 < \infty$ *such that*

$$n^{-1/2} \int_0^n \varphi\{X_n(t)\}\mathrm{d}t \xrightarrow{D} \mathrm{N}(0,\sigma^2), \quad n \to \infty.$$

We first state two auxiliary result used to prove Theorem A1. The first is well known; see Bardet and Surgailis (2011) for a proof based on characteristic functions.

LEMMA A1. *Suppose that* $\mathbb{E}(Z_n) = 0$, $\mathbb{E}(Z_n^2) < \infty$ *and* $\lim_{n\to\infty}\mathbb{E}(Z_n^2) = \sigma^2 < \infty$. *Assume that for each* $\varepsilon > 0$ *there exists* $Z_{n,\varepsilon}$ *such that* $\mathbb{E}|Z_n - Z_{n,\varepsilon}|^2 < \varepsilon$ *whenever* $n \geqslant n_0$ *for some* $n_0(\varepsilon)$ *and that* $Z_{n,\varepsilon} \to^D \mathrm{N}(0,\sigma_\varepsilon^2)$. *Then* $Z_n \to^D \mathrm{N}(0,\sigma^2)$.

Denote in the following by $\|\cdot\|$ the mean square norm with respect to Q. The following Gaussian moment bound is a crucial ingredient in the proof of Theorem A1.

LEMMA A2 (Arcones' inequality). *Assume that the real-valued measurable functions* φ_1,φ_2 *on* $\mathbb{R}^d$ *have Hermite rank* τ *and satisfy* $\mathrm{Q}\varphi_1 = \mathrm{Q}\varphi_2 = 0$ *and* $\varphi_1,\varphi_2 \in \mathscr{L}^2(\mathrm{Q})$. *Let* $X = [X_1,\dots,X_d]^\top$ *and* $Y = [Y_1,\dots,Y_d]^\top$ *be standard* d*-dimensional Gaussian vectors with* $\mathrm{Cov}(X_i,Y_j) = r^{(ij)}$, $1 \leqslant i,j \leqslant d$. *Define* $r := \max_{i,j}|r^{(ij)}|$. *Then*

$$|\mathbb{E}\{\varphi_1(X)\varphi_2(Y)\}| \leqslant \|\varphi_1\|\|\varphi_2\|(dr)^\tau.$$

Proof. When $r \leqslant d^{-1}$, the result is Lemma 1 of Arcones (1994) (see also Soullier (2001)). When $r > d^{-1}$, Cauchy-Schwarz's inequality implies

$$|\mathbb{E}\{\varphi_1(X)\varphi_2(Y)\}| \leqslant \|\varphi_1\|\|\varphi_2\| \leqslant \|\varphi_1\|\|\varphi_2\|d^\tau r^\tau.$$ ■

The lemma below implies that in the proof of Theorem A1, we may restrict the analysis to functions φ which are finite Hermite polynomials.

LEMMA A3. *Under the assumptions of Theorem A1, suppose that for each* $M \geqslant \tau$ *there exists* σ_M^2 *such that*

$$\text{(A1)} \qquad n^{-1/2} \int_0^n \sum_{\tau \leqslant |\mathbf{k}| \leqslant M} \frac{J_\varphi(\mathbf{k})}{\mathbf{k}!} H_{\mathbf{k}}\{X_n(t)\}\mathrm{d}t \xrightarrow{D} \mathrm{N}\big(0,\sigma_M^2\big), \quad n \to \infty.$$

Then $n^{-1/2}\int_0^n \varphi_n\{X_n(t)\}\mathrm{d}t \to^D \mathrm{N}(0,\sigma^2)$ *when* $n \to \infty$ *where* $\sigma^2 = \lim_{M\to\infty}\sigma_M^2$.

Proof. With the notation in the statement of Theorem A1, denote

$$C := \sup_{n\geqslant 1} \sup_{t\in[0,n]} \max_{i,j} \int_0^n |r_n^{(ij)}(s-t)|^\tau \mathrm{d}s.$$

From Lemma A2 and assumption A in Theorem A1, we get that

$$\text{(A2)} \qquad \mathbb{E}\left(n^{-1/2}\int_0^n \left[\varphi_n\{X_n(t)\} - \varphi\{X_n(t)\}\right]\mathrm{d}t\right)^2 \leqslant \|\varphi_n - \varphi\|^2 d^\tau C \to 0, \quad n\to\infty.$$

Then Slutsky's lemma implies that we need only prove asymptotic normality of $n^{-1/2}\int_0^n \varphi\{X_n(t)\}\mathrm{d}t$. By similar arguments, it also holds that

$$\text{(A3)} \qquad \sigma^2 = \lim_{n\to\infty} \mathbb{E}\left[n^{-1/2}\int_0^n \varphi\{X_n(t)\}\mathrm{d}t\right]^2 \leqslant \|\varphi\|^2 d^\tau C < \infty.$$

Suppose that $\sigma^2 > 0$; otherwise the result is trivial. Denote for $n\in\mathbb{N}$ and $M\geqslant\tau$

$$Z_{n,M} := n^{-1/2}\int_0^n \sum_{\tau\leqslant|\mathbf{k}|\leqslant M} \frac{J_\varphi(\mathbf{k})}{\mathbf{k}!} H_{\mathbf{k}}\{X_n(t)\}\mathrm{d}t, \quad \text{and } Z_n := n^{-1/2}\int_0^n \varphi\{X_n(t)\}\mathrm{d}t.$$

Since the partial Hermite series of φ is convergent in Q-mean square, arguments as in (A2) imply that for each $\varepsilon > 0$ we can choose $M(\varepsilon)$ large enough so that

$$\text{(A4)} \qquad \|Z_n - Z_{n,M(\varepsilon)}\| \leqslant \varepsilon;$$

uniformly in n. Lemma A2 then allows us to conclude that $Z_n \to^D \mathrm{N}(0,\sigma^2)$ when $n\to\infty$. From the reverse triangle inequality, (A4) also implies $\sigma_M^2 \to \sigma^2$ when $M\to\infty$. ■

We will rely on cumulants (Brillinger (1975), Section 2.3) to prove (A1). Recall that the joint cumulant of random variables $X_1,\dots,X_n$ is defined as

$$\mathrm{cum}(X_1,\dots,X_n) := (-i)^n \frac{\partial^n \log\Phi(z_1,\dots,z_n)}{\partial z_1\cdots\partial z_n}\bigg|_{z_1=\cdots=z_n=0}$$

where Φ is the joint characteristic function of $X_1,\dots,X_n$. The cumulant of order $p\in\mathbb{N}$ of a random variable X, denoted $\mathrm{cum}^{(p)}(X)$, is defined as

$$\mathrm{cum}^{(p)}(X) := \mathrm{cum}(\underbrace{X,\dots,X}_{p \text{ times}}).$$

The joint cumulant is multi-linear, i.e. if $X_1,\dots,X_m$ are random variables and $Y_i = \sum_{j=1}^m c_{ij}X_j$ for real numbers c_{ij}, $i=1,\dots,n$, then

$$\mathrm{cum}(Y_1,\dots,Y_n) = \sum_{j_1,\dots,j_n=1}^m c_{1j_1}\cdots c_{nj_m}\mathrm{cum}(X_{j_1},\dots,X_{j_m}).$$

It is well known that cumulants of order $p>2$ of the normal distribution are zero. By the method of moments, it follows that if a sequence of random variables satisfies $\lim_{n\to\infty}\mathrm{cum}^{(p)}(X_n) = 0$ for all $p>2$, then $X_n \to^D X$ for some Gaussian X.

The main ingredient in the proof of Theorem A1 is the existence of explicit formulas for cross-moments of product Hermite polynomials applied to Gaussian vectors. These formulas are typically established within the so-called diagram formalism. Consider the p-row, not necessarily rectangular array

$$T(\mathbf{k}_1,\dots,\mathbf{k}_p) := \begin{bmatrix} (1,1) & (1,2) & \cdots & (1,k_1) \\ (2,1) & (2,2) & \cdots & (2,k_2) \\ \cdots & & & \\ (p,1) & (p,2) & \cdots & (p,k_p) \end{bmatrix}$$

where $k_j =: |\mathbf{k}_j| = k_j^{(1)} + \cdots + k_j^{(d)}$, $\mathbf{k}_1,\dots,\mathbf{k}_p \in \mathbb{N}_0^d$. We refer to $T(\mathbf{k}_1,\dots,\mathbf{k}_p)$ as a table. A diagram is a table $T(\mathbf{k}_1,\dots,\mathbf{k}_p)$ alongside a partition γ into pairs of entries of T so that entries comprising each pair belong to different rows. Write $\Gamma(\mathbf{k}_1,\dots,\mathbf{k}_p)$ for the collection of all diagrams associated with the table $T(\mathbf{k}_1,\dots,\mathbf{k}_p)$. An element of the partition into pairs γ is called an edge. The number of edges between rows u and v in γ is denoted $\ell_{uv}(\gamma)$ or simply ℓ_{uv}.

A subtable of $T(\mathbf{k}_1,\dots,\mathbf{k}_p)$ is simply a table composed of a subset of rows from T. We refer to a diagram as connected if it cannot be written as the union of disjoint subtables T_1, T_2 such that no edge passes between T_1 and T_2. We write $\Gamma_{\text{con}}(\mathbf{k}_1,\dots,\mathbf{k}_p)$ for the set of all connected diagrams over $T(\mathbf{k}_1,\dots,\mathbf{k}_p)$.

For more details on the following bound for cross-moments of Hermite polynomials of dependent Gaussian variables, see Surgailis (2000) or Surgailis (2003).

THEOREM A2 (Diagram moment bound). *Let* $X_i := [X_i^1,\dots,X_i^d]^\top$ *for* $i = 1,\dots,p$ *be* d*-dimensional standard Gaussian random vectors. Denote* $r_{ij} := \max_{l,k} |\mathbb{E}(X_i^l X_j^k)|$, $i \neq j$. *With* $\mathbf{k}_1,\dots,\mathbf{k}_p \in \mathbb{N}_0^d$, *it holds that*

$$\left|\operatorname{cum}\{H_{\mathbf{k}_1}(X_1),\dots,H_{\mathbf{k}_p}(X_p)\}\right| \leqslant \sum_{\gamma\in\Gamma_{\text{con}}(\mathbf{k}_1,\dots,\mathbf{k}_p)} \prod_{1\leqslant i<j\leqslant p} r_{ij}^{\ell_{ij}(\gamma)}.$$

We can now prove Theorem A1. The strategy of the proof follows closely that of Bardet and Surgailis (2011), except that the relevant processes are now continuous and the various regions of integration are modified to allow for periodicity of the stochastic processes involved.

Proof of Theorem A1. Denote in the sequel $r_n(t) := \max_{1\leqslant i,j\leqslant d} |r^{(ij)}(t)|$ and set

$$C := \sup_{n\geqslant 1} \sup_{t\in[0,n]} \int_0^n |r_n(s-t)|^\tau \mathrm{d}s.$$

By Lemma A3, it suffices to show that

$$n^{-1/2} \int_0^n \sum_{\tau\leqslant|\mathbf{k}|\leqslant M} \frac{J_\varphi(\mathbf{k})}{\mathbf{k}!} H_{\mathbf{k}}\{X_n(t)\}\mathrm{d}t \xrightarrow{D} \mathrm{N}(0,\sigma_M^2),$$

for each $M > \tau$. By the discussion of cumulants on the preceding page, asymptotic normality of the left-hand side of the display will follow if, for each $p > 2$ and $\tau \leqslant |\mathbf{k}_i| \leqslant M$, we have

$$\text{(A5)} \qquad \operatorname{cum}\left[\int_0^n H_{\mathbf{k}_1}\{X_n(t_1)\}\mathrm{d}t_1,\dots,\int_0^n H_{\mathbf{k}_p}\{X_n(t_p)\}\mathrm{d}t_p\right] = o(n^{p/2}), \quad n\to\infty.$$

Fix $M > \tau$, $p > 2$ and denote in the following

$$\operatorname{cum}(t_1,\dots,t_p) := \operatorname{cum}\left[H_{\mathbf{k}_1}\{X_n(t_1)\},\dots,H_{\mathbf{k}_p}\{X_n(t_p)\}\right],$$

where the index n is implicit in the quantity on the left-hand side to simplify notation. Let Π be the set of all partitions of $\{1,\ldots,p\}$. From the Leonov-Shirayev formula (Leonov and Shiryaev, 1959) for joint cumulants and Fubini's theorem, we get

$$\begin{aligned}\operatorname{cum}\left[\int_0^n H_{\mathbf{k}_1}\{X_n(t_1)\}\mathrm{d}t_1,\ldots,\int_0^n H_{\mathbf{k}_p}\{X_n(t_p)\}\mathrm{d}t_p\right]&\\ =\left|\sum_{\pi\in\Pi}(|\pi|-1)!(-1)^{|\pi|-1}\prod_{B\in\pi}\mathbb{E}\left[\prod_{i\in B}\int_0^n H_{\mathbf{k}_i}\{X_n(t_i)\}\mathrm{d}t_i\right]\right|&\\ \leqslant\int_{[0,n]^p}|\operatorname{cum}(t_1,\ldots,t_p)|\mathrm{d}t_1\cdots\mathrm{d}t_p.&\end{aligned}$$

Assuming that $0\leqslant m\leqslant n/2$, take

$$\begin{aligned}A_{n,m}&:=\left\{(t_1,\ldots,t_p)\in[0,n]^p\,:\,|t_i-t_j|\leqslant m\vee|t_i-t_j|>n-m,\,\forall i,j\right\},\\ B_{n,m}^{(\alpha\beta)}&:=\left\{(t_1,\ldots,t_p)\in[0,n]^p\,:\,m\leqslant|t_\alpha-t_\beta|\leqslant n-m\}\right\},\quad 1\leqslant\alpha,\beta\leqslant p;\end{aligned}$$

and let

$$\begin{aligned}\Sigma_n(m)&:=\int_{A_{n,m}}|\operatorname{cum}(t_1,\ldots,t_p)|\,\mathrm{d}t_1\cdots\mathrm{d}t_p,\\ \Sigma_{n,\alpha\beta}(m)&:=\int_{B_{n,m}^{(\alpha\beta)}}|\operatorname{cum}(t_1,\ldots,t_p)|\,\mathrm{d}t_1\cdots\mathrm{d}t_p.\end{aligned}$$

Then

$$\int_{[0,n]^p}|\operatorname{cum}(t_1,\ldots,t_p)|\,\mathrm{d}t_1\cdots\mathrm{d}t_p\leqslant\Sigma_n(m)+\sum_{\substack{1\leqslant\alpha,\beta\leqslant p\\ \alpha\neq\beta}}\Sigma_{n,\alpha\beta}(m).$$

We will show that each term on the right-hand side is of order $o(n^{p/2})$ when $n\to\infty$.

First,

$$\Sigma_n(m)=O(nm^{p-1})=o(n^{p/2}),\quad p>2,$$

since $(t_1,\ldots,t_p)\mapsto\operatorname{cum}(t_1,\ldots,t_p)$ is bounded on $[0,n]^d$ and $\int_{A_{n,m}}\mathrm{d}t_1\cdots\mathrm{d}t_p\leqslant B_n(m)+C_n(m)$ where

$$B_n(m)=\int_{[0,n]^p\cap\{|t_i-t_j|\leqslant m,\forall i,j\}}\mathrm{d}t_1\cdots\mathrm{d}t_p\leqslant\int_0^n\mathrm{d}t_1\int_{t_1-m}^{t_1+m}\mathrm{d}t_2\cdots\int_{t_{p-1}-m}^{t_{p-1}+m}\mathrm{d}t_p=n(2m)^{p-1}$$

and

$$C_n(m)\leqslant n\left(\int_{[0,n]^2\cap\{|s-t|>n-m\}}\mathrm{d}s\mathrm{d}t\right)^{(p-1)/2}\leqslant nm^{p-1};$$

since (for q even), $\{(t_1,\ldots,t_q)\in[0,n]^q\,:\,|t_i-t_j|>n-m,\,\forall i,j\}\subseteq\prod_{(i_1,i_2)\in\pi}\{(t_{i_1},t_{i_2})\in[0,n]^2\,:\,|t_{i_1}-t_{i_2}|>n-m\}$ where π is any partition by pairs of $\{1,\ldots,q\}$.

We proceed to show that for each pair α,β with $\alpha\neq\beta$, there exists a sequence $\delta(m)\to 0$, $m\to\infty$, such that

$$\Sigma_{n,\alpha\beta}(m)\leqslant\delta(m)n^{p/2}.$$

Denote $r_n(t):=\max_{i,j}|r_n^{(ij)}(t)|$. Theorem A2 implies

$$|\operatorname{cum}(t_1,\ldots,t_p)|\leqslant\sum_{\gamma\in\Gamma_{\mathrm{con}}(\mathbf{k}_1,\ldots,\mathbf{k}_p)}\prod_{1\leqslant i<j\leqslant p}r_n(t_i-t_j)^{\ell_{ij}(\gamma)}.$$

Observe that $\ell_{ij} = \ell_{ji}$ and that $\sum_{j=1}^{p} \ell_{ij}/k_i = 1$ for $i = 1,\dots,p$. Consider the form of Hölder's inequality which states that, for functions $h_1,\dots,h_k$,

$$\int h_1 h_2 \cdots h_k \leqslant \prod_{j=1}^{k} \left(\int |h_j|^{\beta_j} \right)^{1/\beta_j}, \quad \text{where } \sum_{j=1}^{k} 1/\beta_j = 1.$$

Repeated application of this result, as in Giraitis and Surgailis (1985), yields

$$\begin{aligned}
&\int_{[0,n]^p} \prod_{1 \leqslant u < v \leqslant p} r(t_u - t_v)^{\ell_{uv}} \mathrm{d}t_1 \cdots \mathrm{d}t_p \\
&\quad = \int_{[0,n]^{p-1}} \left\{ \int_0^n \prod_{j=2}^{p} r_n(t_1 - t_j)^{\ell_{1j}} \mathrm{d}t_1 \right\} \prod_{2 \leqslant u < v \leqslant p} r(t_u - t_v)^{\ell_{uv}} \mathrm{d}t_2 \cdots \mathrm{d}t_p \\
&\quad \leqslant \int_{[0,n]^{p-1}} \left[\prod_{j=2}^{p} \left\{ \int_0^n r_n(t_1 - t_j)^{k_1} \mathrm{d}t_1 \right\}^{\ell_{1j}/k_1} \right] \prod_{2 \leqslant u < v \leqslant p} r_n(t_u - t_v)^{\ell_{uv}} \mathrm{d}t_2 \cdots \mathrm{d}t_p \\
&\quad \leqslant \int_{[0,n]^{p-2}} \left[\int_0^n \left\{ \int_0^n r_n(t_1 - t_2)^{k_1} \mathrm{d}t_1 \right\}^{k_2/k_1} \mathrm{d}t_2 \right]^{\ell_{12}/k_2} \prod_{j=3}^{p} \left[\left\{ \int_0^n r_n(t_1 - t_j)^{k_1} \mathrm{d}t_1 \right\}^{\ell_{1j}/k_1} \times \right. \\
&\qquad \left. \left\{ \int_0^n r_n(t_2 - t_j)^{k_2} \mathrm{d}t_2 \right\}^{\ell_{2j}/k_2} \right] \prod_{3 \leqslant u < v \leqslant p} r_n(t_u - t_v)^{\ell_{uv}} \mathrm{d}t_3 \cdots \mathrm{d}t_p \\
&\quad \leqslant \cdots \\
&\quad \leqslant \prod_{1 \leqslant i < j \leqslant p} \left[\int_0^n \left\{ \int_0^n r_n(s-t)^{k_i} \mathrm{d}s \right\}^{k_j/k_i} \mathrm{d}t \right]^{\ell_{ij}/k_j}.
\end{aligned}$$

By symmetry of the above manipulations in u, v, the definition of $\Sigma_{n,\alpha\beta}(m)$ implies

$$\Sigma_{n,\alpha\beta}(m) \leqslant \prod_{1 \leqslant i < j \leqslant p} R_{ij} \wedge \prod_{1 \leqslant i < j \leqslant p} R_{ji}, \tag{A6}$$

where, recalling that $S_{n,m} := \{(s,t) \in [0,n]^2 : m \leqslant |s-t| \leqslant n-m\}$,

$$R_{ij} := \begin{cases} \left[\int_0^n \left\{ \int_0^n r_n(s-t)^{k_i} \mathrm{d}s \right\}^{k_j/k_i} \mathrm{d}t \right]^{\ell_{ij}/k_j} & (i,j) \neq (\alpha,\beta),(\beta,\alpha); \\ \left[\int_0^n \left\{ \int_0^n r_n(s-t)^{k_\alpha} \mathbb{1}\{(s,t) \in S_{n,m}\} \mathrm{d}s \right\}^{k_\beta/k_\alpha} \mathrm{d}t \right]^{\ell_{\alpha\beta}/k_\beta} & (i,j) = (\alpha,\beta); \\ \left[\int_0^n \left\{ \int_0^n r_n(s-t)^{k_\beta} \mathbb{1}\{(s,t) \in S_{n,m}\} \mathrm{d}s \right\}^{k_\alpha/k_\beta} \mathrm{d}t \right]^{\ell_{\alpha\beta}/k_\alpha}, & (i,j) = (\beta,\alpha). \end{cases}$$

By assumption A of the theorem, we have

$$C := \sup_{n \geqslant 1} \sup_{t \in [0,n]} \int_0^n r_n(s-t)^{\tau} \mathrm{d}s < \infty. \tag{A7}$$

Since $k_i \geqslant \tau$ for $i = 1,\dots,p$ by the definition of Hermite rank, it follows immediately that

$$\left[\int_0^n \left\{ \int_0^n r_n(s-t)^{k_i} \mathrm{d}s \right\}^{k_j/k_i} \mathrm{d}t \right]^{\ell_{ij}/k_j} \leqslant C^{\ell_{ij}/k_i} n^{\ell_{ij}/k_j}.$$

Suppose that $k_\beta \leqslant k_\alpha$. Then by Jensen's inequality,

$$\int_0^n \mathrm{d}t \left\{ \int_0^n r_n(s-t)^{k_\alpha} \mathbb{1}\{(s,t) \in S_{n,m}\} \mathrm{d}s \right\}^{k_\beta/k_\alpha} \leqslant n \left\{ n^{-1} \int_{[0,n]^2} r_n(s-t)^{k_\alpha} \mathbb{1}\{(s,t) \in S_{n,m}\} \mathrm{d}s \mathrm{d}t \right\}^{k_\beta/k_\alpha}.$$

Also, by the monotonicity property of $\mathscr{L}^q$-norms, there exists $D > 0$ such that

$$\int_0^n \mathrm{d}t \left\{ \int_0^n r_n(s-t)^{k_\beta} \mathbb{1}\{(s,t) \in S_{n,m}\} \mathrm{d}s \right\}^{(1/k_\beta)k_\alpha} \leqslant D \int_{[0,n]^2} r_n(s-t)^{k_\alpha} \mathbb{1}\{(s,t) \in S_{n,m}\} \mathrm{d}s\mathrm{d}t.$$

Thus, by assumption B of the theorem, there exists a sequence $\tilde{\delta}(m) \to 0$ when $m \to \infty$ such that

$$R_{ij} \leqslant \begin{cases} \tilde{C} n^{\ell_{ij}/k_j} & (i,j) \neq (\alpha,\beta),(\beta,\alpha); \\ \tilde{\delta}(m) n^{\ell_{\alpha\beta}/k_\beta} & (i,j) = (\alpha,\beta); \\ \tilde{\delta}(m) n^{\ell_{\alpha\beta}/k_\alpha} & (i,j) = (\beta,\alpha). \end{cases}$$

Comparing with (A6), we will have $\Sigma_{n,\alpha\beta}(m) = \delta(m) n^{p/2}$ for some $\delta(m) \to 0$, $m \to \infty$, if

$$\left\{ \sum_{1 \leqslant i < j \leqslant p} \ell_{ij}/k_i \right\} \wedge \left\{ \sum_{1 \leqslant i < j \leqslant p} \ell_{ij}/k_j \right\} \leqslant p/2.$$

But this follows since $a + b \leqslant c$ implies $a \wedge b \leqslant c/2$ for $a, b, c \in \mathbb{N}$, and we have

$$\sum_{1 \leqslant i < j \leqslant p} \ell_{ij}/k_i + \sum_{1 \leqslant i < j \leqslant p} \ell_{ij}/k_j = \sum_{i=1}^{p} \sum_{\substack{j=1 \\ j \neq i}}^{p} \ell_{ij}/k_i = p.$$

This proves (A5) from which the statement of the theorem follows. ■

References

Arcones, M. A. (1994) Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors. *Annals of Probability*, **22**, 2242–2274.

Bardet, J. and Surgailis, D. (2011) A general moment bound for a product of Gaussian vector's functionals and a central limit theorem for subordinated Gaussian triangular arrays. Tech. rep., Université Paris. `arXiv:1104.4732v1`.

Barriac, G. and Madhow, U. (2004) Characterizing outage rates for space-time communication over wideband channels. *IEEE Transactions on Communications*, **52**, 2198–2208.

Bhatia, R. (1997) *Matrix Analysis*. Springer.

Billingsley, P. (1995) *Probability and Measure*. Wiley, third edn.

Brillinger, D. R. (1975) *Time Series: Data Analysis and Theory*. Holt, Rinehart and Winston.

Cioffi, J. (2003) Digital Communications. Stanford University – EE379 Course Reader.

Deo, R. S. and Chen, W. W. (2000) On the integral of the squared periodogram. *Stochastic Processes and their Applications*, **85**, 159–176.

Faÿ, G. (2010) Moment bounds for non-linear functionals of the periodogram. *Stochastic Processes and their Applications*, **120**, 983–10009.

Faÿ, G., Moulines, E. and Soulier, P. (2002) Nonlinear functionals of the periodogram. *Journal of Time Series Analysis*, **23**.

Folland, J. B. (1999) *Real Analysis*. Wiley Interscience, second edn.

Giraitis, L. and Surgailis, D. (1985) CLT and other limit theorems for functionals of Gaussian processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **70**, 191–212.

Ibragimov, I. and Linnik, Y. V. (1971) *Independent and Stationary Sequences of Random Variables*. Groningen: Wolters-Noordhoff.

Jerri, A. J. (1977) The Shannon sampling theorem – its various extensions and applications: a tutorial review. *Proceedings of the IEEE*, **65**.

Kurzweil, J. (2000) *An Introduction to Digital Communications*. Wiley.

Leonov, V. P. and Shiryaev, A. N. (1959) On a method of calculation of semi-invariants. *Theory of Probability and its Applications*, **4**, 319–329.

Molisch, A. F. (2005) Ultrawideband propagation channels – theory, measurement, and modeling. *IEEE Transaction on Vehicular Technology*, **54**, 1528–1545.

Pereira, S., Kyritsi, P., Papanicolaou, G. and Paulraj, A. (2005) Asymptotic properties of richly scattering ultrawideband channels. Stanford University. Unpublished manuscript.

Pratt, J. W. (1960) On interchanging limits and integrals. *Annals of Mathematical Statistics*, **31**, 74–77.

Rubak, E. (2007) *Central limit theorems for weakly dependent stochastic processes*. Master's thesis, Aalborg University.

Schuster, U. G. and Bölcskei, H. (2007) Ultrawideband channel modeling on the basis of information-theoretic criteria. *IEEE Transactions on Wireless Communications*, **6**, 2464–2475.

Serfling, R. J. (1968) Contributions to central limit theory for dependent variables. *Annals of Mathematical Statistics*, **39**, 1158–1175.

Soullier, P. (2001) Moment bounds and central limit theorem for functions of Gaussian vectors. *Statistics and Probability Letters*, **54**, 193–203.

Surgailis, D. (2000) Long-range dependence and Appell rank. *Annals of Probability*, **28**, 478–497.

Surgailis, D. (2003) CLTs for polynomials of linear sequences: Diagram formula with illustrations. In *Theory of and Applications of Long-Range Dependence* (eds. P. Doukhan, G. Oppenheim and M. S. Taqqu). Birkhäuser.

Tse, D. and Visnawath, P. (2005) *Fundamentals of Wireless Communications*. Cambridge University Press.

Win, M. Z. and Scholtz, R. A. (1998) On the robustness of ultra-wide bandwidth signals in dense multipath environments. *IEEE Communications Letters*, **2**, 51–53.

Part 2

Applications *in* Medicine and Biotechnology

Paper IV

Exploring Dietary Patterns by Using the Treelet Transform

Author list

Anders Gorst-Rasmussen
Aalborg University, Denmark

Christina C. Dahm
Aarhus University, Denmark

Claus Dethlefsen
Aarhus University, Denmark

Thomas Scheike
University of Copenhagen, Denmark

Kim Overvad
Aarhus University, Denmark

Summary

Principal component analysis (PCA) has been used extensively in nutritional epidemiology to derive patterns summarizing food and nutrient intake but its interpretation can be difficult. The authors propose the use of a new statistical technique, the treelet transform (TT), as an alternative to PCA. TT combines the quantitative pattern extraction capabilities of PCA with the interpretational advantages of cluster analysis and produces patterns involving only naturally grouped subsets of the original variables. The authors compared patterns derived using TT with those derived using PCA in a study of dietary patterns and risk of myocardial infarction (MI) among 26,155 male participants in a prospective Danish cohort. Over a median of 11.9 years of follow-up, 1,523 incident cases of MI were ascertained. The 7 patterns derived with TT described almost as much variation as the first 7 patterns derived with PCA, for which interpretation was less clear. Using multivariate Cox regression models to estimate relative risk of MI, the two methods lead to comparable significant risk factors. The study shows that TT may be a useful alternative to PCA in epidemiological studies, leading to patterns which possess comparable explanatory power and are simple to interpret.

Supplementary info

Part of this paper was prepared while based at Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital, as an affiliate of the Nordic Centre of Excellence ‘SYSDIET’ funded by NordForsk. The paper has been published by Oxford University Press (OUP) as:

> Gorst-Rasmussen A, Dahm CC, Dethlefsen C, Scheike TH, Overvad K (2011). Exploring dietary patterns by using the treelet transform. *American Journal of Epidemiology*; **173**(10):1097–1104.

The version here is based on the journal version, in accordance with OUP’s Publication Rights Policies. For convenience, the Web appendix accompanying the published paper is included in Appendix 2.

The published paper was commented upon by Drs. Imamura and Jacques in the same issue of *American Journal of Epidemiology*. We summarise their comment and list our full response in Appendix 1.

1. Introduction

In epidemiological studies involving many variables with complex interrelationships, analysis of each variable in an independent fashion provides an incomplete picture. The standard example is studies of dietary intake, where foods are consumed in combination and analysis on a per-food basis can be misleading (Willet (1998), p. 22). Alternative approaches to analysis include prior construction of scores with a biological rationale (Kennedy *et al.*, 1995; Trichopoulou *et al.*, 1995); or an exploratory approach, using statistical dimension reduction methods on the data at hand to extract the essential information in the original variables (Michels and Schulze, 2005). Principal component analysis (PCA) is by far the most popular dimension reduction method in studies of dietary patterns (Hu, 2002; Newby and Tucker, 2004). PCA works by compressing data into weighted averages of a small number of 'latent' patterns, called components or factors, obtained by analyzing the covariance or correlation matrix of the original variables. PCA can be efficient in producing factors which are associated with risk of disease (Slattery *et al.*, 1998; Hu *et al.*, 2000; DiBello *et al.*, 2008). However, since each factor involves all of the original variables, qualitative interpretation of PCA results is challenging and may require detailed prior knowledge about plausible groupings among variables, alongside considerable subjective judgment about which variables dominate a factor. When knowledge about plausible groupings is lacking, interpretation of factors becomes particularly difficult. Cluster analysis would seem a useful tool for discovering hidden groupings among variables in such cases, but does not offer generic techniques for constructing numeric summary variables. Indeed, applications of cluster analysis to studies of dietary patterns have used clustering among individuals rather than variables (Newby and Tucker, 2004), addressing a somewhat different question than PCA (Moeller *et al.*, 2007). A dimension reduction method which enables simple construction of numeric summary variables and offers more easily interpretable factors is desired.

The treelet transform (TT) is a dimension reduction method developed by Lee *et al.* (2008) which combines the strengths of PCA and cluster analysis. TT works on a covariance or correlation matrix to produce a collection of factors in the same manner as PCA. However, in contrast to PCA, each TT factor involves only a smaller number of naturally grouped variables, with no remaining variable contributing to the factor. Additionally, TT improves interpretation by producing a hierarchical grouping structure among variables, visualizable as a cluster tree.

The aim of the present methodological study was to illustrate the use of TT as an exploratory technique in an epidemiological context. Using data from a large Danish prospective cohort study, we compared PCA and TT as exploratory methods in a study of dietary patterns and the risk of myocardial infarction (MI) among middle-aged men. Associations between empirically derived dietary patterns and cardiovascular disease have been investigated extensively in the literature (Schulze and Hoffmann, 2006), making this an ideal application for critically evaluating a new exploratory technique.

2. Materials and methods

2.1. *Study population*

The study population included participants in the Danish cohort study Diet, Cancer and Health (Tjønneland *et al.*, 2007). This prospective cohort was initiated in 1993-1997 and included 57,053 Danish born residents aged 50-64 years and free of cancer at the time of registration. Analysis of the entire cohort would require stratification by sex throughout and is beyond the illustrative scope of this paper. Due to a greater incidence of MI (with correspondingly more stable association estimates), analysis was restricted to men only ($n = 27{,}178$). At enrolment, participants underwent clinical examination and a lifestyle survey. The latter included a self-administered 192 item food-frequency questionnaire, details of which have been described elsewhere (Overvad *et al.*, 1991; Tjønneland *et al.*, 1991). Briefly, participants were asked to report their average intake of different food and beverage items over the past 12 months within 12 categories ranging from never to more than 8 times per day. Daily intakes of specific foods were calculated for each participant using the software program Food Calc (Lauritsen, 1998) using specially developed standardized recipes and portion sizes (Møller and Saxholt, 1996). For the present study, the 192 foods included in the food-frequency questionnaire were aggregated into 42 groups.

2.2. *Exclusions and follow-up*

Participants with incomplete questionnaires were excluded ($n = 59$), as were participants with a cancer diagnosis that was not, at the time of invitation, registered in the Danish Cancer Registry due to processing delay ($n = 233$). Participants registered with a prior diagnosis of MI or cardiac arrest were also excluded ($n = 731$). The final study sample included 26,155 male participants.

Participants were followed up from date of enrolment until the end of April 2008 or the occurrence of fatal/non-fatal MI, emigration, or death. Follow-up was done by linkage with central Danish registries via the unique identification number assigned to all Danish citizens (Pedersen *et al.*, 2006). We identified participants registered with a first-time discharge diagnosis of MI or cardiac arrest (International Classification of Diseases, Eighth Revision, codes 410-410.99 and 427.27; International Classification of Diseases, Tenth Revision, codes I21.0-I21.9 and I46.0-I46.9) in the Danish National Patient Registry (Andersen *et al.*, 1999) from the date of enrolment until 31 December 2003. Medical records were subsequently reviewed and MI cases identified (Joensen *et al.*, 2009) using the criteria of Luepker *et al.* (2003). From 1 January 2004 onwards, we used register information and restricted the case definition to patients with MI discharged from wards and patients with a diagnosis of cardiac arrest and verified MI. Additionally, participants registered with MI or cardiac arrest as cause of death in the Causes of Death Register (Juel and Helweg-Larsen, 1999) were included as cases.

2.3. *Statistical methods*

PCA and TT are so-called linear dimension reduction methods which work on a covariance or correlation matrix to produce sample-wide orthogonal vectors (factors) across variables such that the original multi-dimensional data can be approximated as weighted averages of factors within individuals. The factors are pragmatically conceptualized as latent variables, revealing the ‘intrinsic structure’ of data. The numeric size of a variable within a factor is called the variable loading. Often, variables are standardized prior to analysis whereby a negative loading corresponds to a smaller-than-average value of the variable; a zero loading corresponds to an average value; and a positive loading to a larger-than-average value. The individual-level weights associated with each factor are called the factor scores, the variances of which are referred to as the factor variances.

PCA enjoys the optimality property that each successive term in the weighted average of factors accounts for the most variance possible for any linear dimension reduction method. Each factor involves all the original variables, i.e. all loadings are non-zero. In contrast, TT balances the ability to explain variation with factor simplicity. This is accomplished by introducing sparsity among factor loadings, i.e. making many loadings exactly zero. Informally, TT can be viewed as an amalgamation of PCA and hierarchical clustering methods. The output of TT applied to a collection of interrelated variables consists of two parts:

1. A cluster tree where branches indicate related groups of variables
2. at each level of the cluster tree, a collection of orthogonal factors where non-zero loadings reflect the grouping structure conveyed by the cluster tree at that particular level.

Technically, TT works by way of local PCA. Starting with all the original variables, the algorithm locates the two variables with the largest correlation and performs PCA on them. A merge is indicated in the cluster tree, and the two variables are replaced with a sum factor representing their maximal-variance weighted average, and an orthogonal residual factor. This scheme is repeated until all variables have joined the cluster tree. By keeping track of factors, a coordinate system for the data becomes available at each level of the cluster tree. It is comprised by the sum factors at that level, residual factors for tree nodes at or below that level, and 'single-variable factors' for variables which have not joined the cluster tree yet. This is known as a multi-resolution decomposition: at each level of the cluster tree, the most recent sum factors encodes low-resolution information about variables included so far, while residual factors encode information at an increasingly greater resolution. Consequently, the factors produced by TT convey information on both global and local relationships among variables.

Unlike PCA, TT does not automatically provide high-variance factors. To find such factors, Lee *et al.* (2008) suggest to first cut the cluster tree at a given level; second, to extract factors at this level based on their variances. The cut-level influences the sparsity of factors. When the cluster tree is cut near its root, more variation can be explained at the cost of factor sparsity. However, the increase in explained variation may be modest compared to the increase in factor complexity for a range of levels close to the root. When the number of retained factors is a fixed number, say k, the cut-level can be chosen in an informed manner by 10-fold cross-validation as follows (Lee *et al.*, 2008). First, the data is split randomly into 10 roughly equal-sized subsets. Second, for each cut-level and using data from 9 out of 10 subsets, the k highest-variance factors are calculated, and the sum of variances of scores based on these factors are calculated using the omitted subset. This is repeated 10 times, each time leaving out a different subset. Third, the cross-validation score at a particular cut-level is calculated by averaging the resulting 10 sums of variances. Fourth, an optimal cut-level is found by locating a 'knee' on the graph of cross-validation scores against cut-level, i.e. a point where increasing the cut-level does not substantially increase the cross-validation score.

We applied PCA and TT to the covariance matrix of standardized dietary intake data (viz. the correlation matrix) to prevent undue influences of food groups with large variances. Preliminary transformations of data towards normality were investigated but deemed unnecessary. Factor variances were used to guide the decision on how many factors to retain for further analysis. In contrast to PCA, TT scores have non-zero correlation across factors, so we used the method of Gervini and Rousson (2004) to assess factor variances. The cut-level for TT was selected using cross-validation as described above, alongside the following heuristic: retain scores within 5% of the cross-validation score for the maximal cut-level, and locate a 'knee' as the point of maximum curvature on the graph of these scores. To further assess sensitivity to the choice of cut-level, we repeated TT analyses at ±3 levels of the optimal level. PCA factors are commonly rotated to simplify interpretation but selecting a suitable rotation can be difficult (Martinez *et al.*, 1998). To assess objectively the extent to which factor rotation might improve interpretation of PCA, we used Procrustes rotation (Gower, 1995) to calculate the orthogonal rotation which brought the retained PCA factors closest to their TT counterparts.

Stability of TT factors was investigated by subsampling. We first calculated the sign pattern among loadings for each of the k retained factors, so that e.g. a factor with loadings $(1,-0.5,0,1,1)$ corresponded to the sign pattern $(+,-,0,+,+)$. We then performed TT on a random sample of 80% of the original data and determined sign patterns among the k new highest-variance factors. This procedure was repeated 500 times. The frequencies of each of the original k sign patterns among the 500 groups of subsampled sign patterns were used as measures of stability. Stability of PCA factors was assessed by a split-sample technique in the spirit of Lau *et al.* (2008) in each of two random split samples, PCA factor scores were obtained and additional scores calculated based on PCA factors from the other split sample. For each factor, the average absolute correlation between the resulting two sets of scores was used as a stability measure. The results were further averaged over 100 independent repetitions to reduce sampling error.

Figure 1 provides a schematic view of our proposed strategy for applying TT.

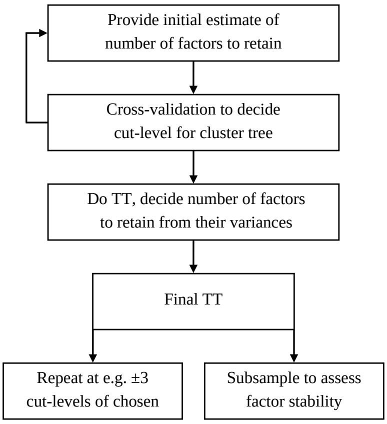
Provide initial estimate of number of factors to retain
Cross-validation to decide cut-level for cluster tree
Do TT, decide number of factors to retain from their variances
Final TT
Repeat at e.g. ±3 cut-levels of chosen
Subsample to assess factor stability

Figure 1. The proposed approach to applying the treelet transform (TT).

To investigate the association between factors and risk of MI, we divided factor scores into population quintiles and used Cox proportional hazards regression with age as time axis and delayed entry to calculate hazard ratios (HRs) and 95% confidence intervals, with the lowest quintile as reference. We adjusted for the following potential confounders: total energy intake (continuous variable), body mass index (<25, 25-29, and ⩾ 30 kg/m^2), education (<8, 8-10, and >10 years), smoking status (never, former, and currently smoking 1-14, 15-24, or ⩾25 g tobacco/day), leisure-time physical activity (<3.5 and ⩾3.5 hours/week), and history of hypertension (yes, no, and do not know). Two-sided tests for trend were calculated by entering quintile levels of exposure as a continuous ordinal variable in the Cox regression model. A P-value less than 0.05 was considered statistically significant. Akaike's information criterion was used to compare the relative fit of non-nested regression models.

R version 2.10 was used for all analyses (R Development Core team, 2009). For TT, we used the 'treelet' package for R. An add-on for Stata 10 with similar functionality has been developed (Gorst-Rasmussen, 2011).

3. Results

PCA and TT were applied to describe variation in the 42 different food groups (baseline characteristics and food groups are reported in Supplementary Tables 1-3 in the appendix). Results of the analyses, in the form of plots of factor loadings, are shown in Figure 2. For PCA, 12 factors had a variance greater than 1. To simplify reporting, we used the criteria (Slattery *et al.*, 1998) of a factor variance ⩾1.25. This yielded 7 factors to be retained for further analysis.

For TT, a plot of the cross-validation scores indicated a 'knee' in the graph at level 29 when the number of factors retained was in the range 4-9 (Figure 3). We cut the cluster tree at this level. To increase comparability with PCA and simplify reporting, the 7 highest-variance factors were retained (8 factors had a variance ⩾1; 5 factors had a variance ⩾1.25). The TT cluster tree is shown in Figure 4, with the 7 highest-variance factors indicated by numbered nodes: leaves descending from these nodes indicate non-zero loadings in the given factor.

Percentage factor variances are presented in Table 1. The first 7 PCA factors accounted for 36.9% of the variance versus 31.0% for the first 7 TT factors.

Table 1. Factor Variances for Principal component Analysis and Treelet Transform Applied to Dietary Data from 26,155 Men in a Prospective Cohort Study of Dietary Patterns and Risk of Myocardial Infarction, Denmark, 1993-2008.

	Factor							
	1	2	3	4	5	6	7	Total
Principal components analysis	10.7	7.1	4.8	4.2	3.7	3.3	3.1	36.9
Treelet transform[a]	9.9	4.6	3.0	3.6	3.6	2.9	2.6	31.0

[a] Adjusted for correlation between factor scores using the method of Gervini and Rousson (2004).

From Figure 2, the structure of factor 1 was similar between PCA and TT (a claim supported by a correlation between scores of 0.98), with both factors characterized by a high intake of red meat alongside items generally considered healthy (fish, poultry, fruit and vegetables; excluding potatoes). TT factor 2 was characterized by a high intake of eggs and refined foods (mayonnaises, processed meat, margarines, sugar/honey, butter, refined cereals). PCA factor 2 was more composite: it appeared to be a contrast of the intake of tea, wine, and selected vegetables (fruity vegetables, leafy vegetables, cabbages, legumes, and other root vegetables) seen in TT factor 3 with the intake of soya, red meat, and the foods seen in the TT factor 2. The correlation between scores from TT factor 2 and PCA factor 2 was 0.77. It was less obvious how to characterize PCA factors 3-7 whereas sparsity of TT factor loadings facilitated characterization, either as averages of normalized dietary intakes or as contrasts between two groups of normalized dietary intake.

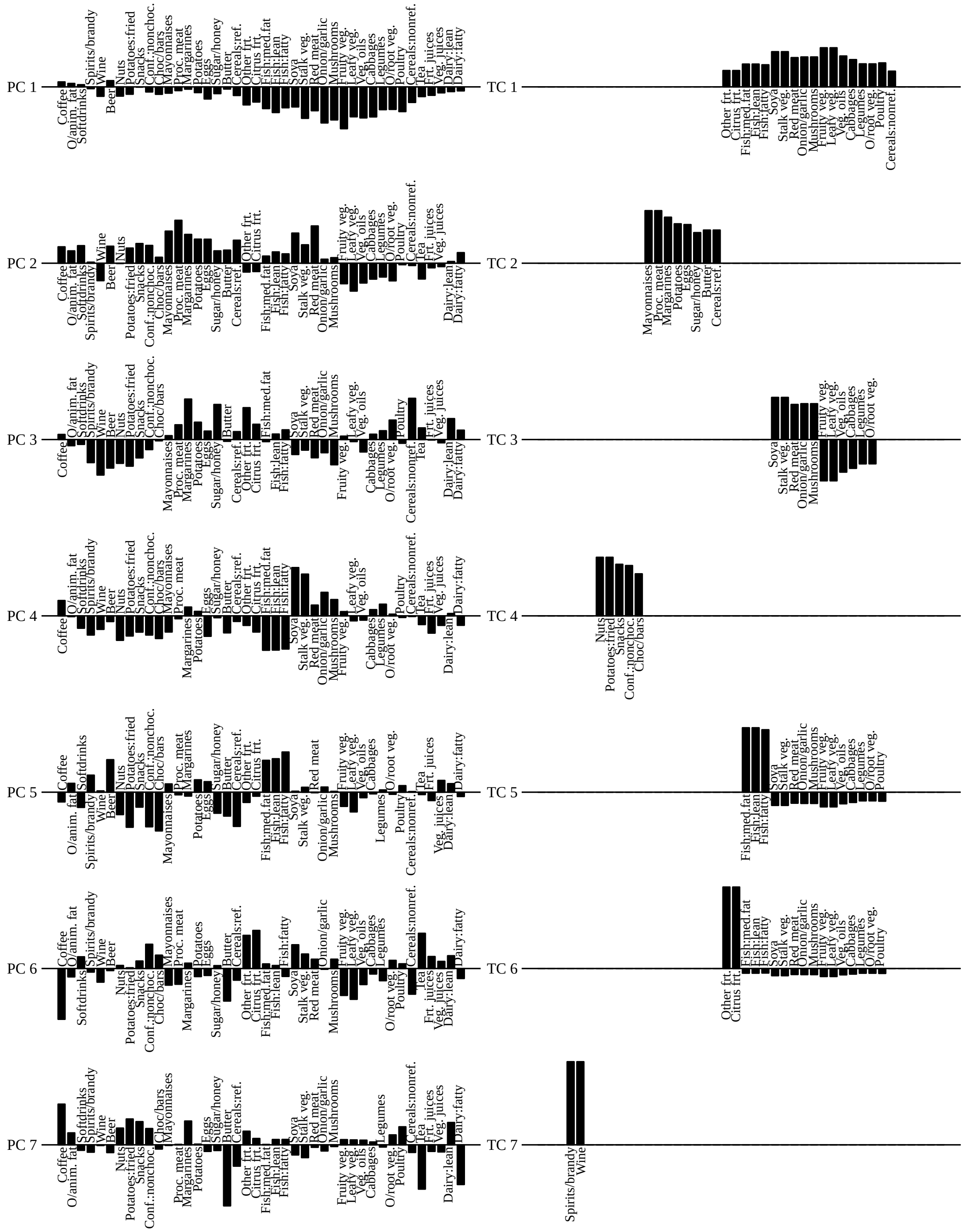

Figure 2. Loading plots for the first 7 factors from principal component analysis (PCA) and treelet transform (TT), for 26,155 men in a prospective cohort study of dietary patterns and risk of myocardial infarction, Denmark, 1993-2008. The axis is oriented in the reading direction so that loadings above the line are positive, and those below are negative; e.g. all loadings in TT factor 1 are positive.

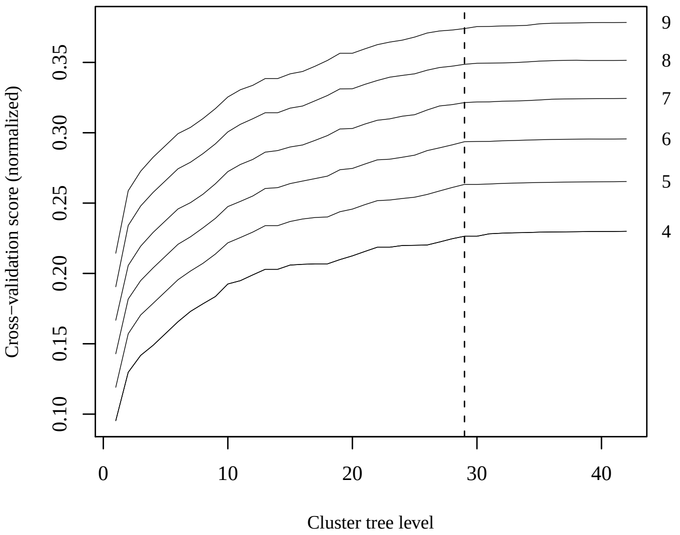

Figure 3. Cross-validation scores for the treelet transform applied to dietary data from 26,155 men in a prospective cohort study of dietary patterns and risk of myocardial infarction, Denmark, 1993-2008. The number of highest-variance factors retained for each curve of cross-validation scores are indicated in the right margin. The dashed line indicates the selected cut-level (level 29) in the cluster tree produced by the treelet transform.

Correlations between scores of the Procrustean-rotated 7 PCA factors and the original TT scores were 0.99, 0.91, 0.92 0.90, 0.85, 0.64 and 0.48 (factor loadings shown in Figure 5).

Stability analyses are reported in Table 2. For PCA, the numerically large correlations for factors 1-5 indicated stability whereas factors 6 and 7 seemed less stable. Similarly, the TT solution was stable for factors 1-6, which appeared in over 90% of the subsampling repetitions. TT factor 7 appeared in only 55% of the subsampling repetitions; competing primarily with a pattern contrasting refined cereals and butter with the remaining elements of factor 2 and appearing in 40% of subsampling repetitions.

TT analyses at cut-levels 26 and 32 (29±3) produced similar factors to those discussed, although with slightly different ordering. TT factor 7 was replaced with a factor loading solely on refined cereals and butter, confirming the instability of this factor discovered through the stability analyses.

Table 2. Stability Analyses for Principal component Analysis and Treelet Transform Applied to Dietary Data from 26,155 Men in a Prospective Cohort Study of Dietary Patterns and Risk of Myocardial Infarction, Denmark, 1993-2008.

	Factor						
	1	2	3	4	5	6	7
Principal components[a]	1.00	1.00	0.99	0.97	0.96	0.82	0.80
Treelet transform[b]	100	95	100	99	96	99	55

[a] Correlations between factor scores, evaluated using a split-sample technique.
[b] Frequencies of factor sign patterns among subsampled factor sign patterns.

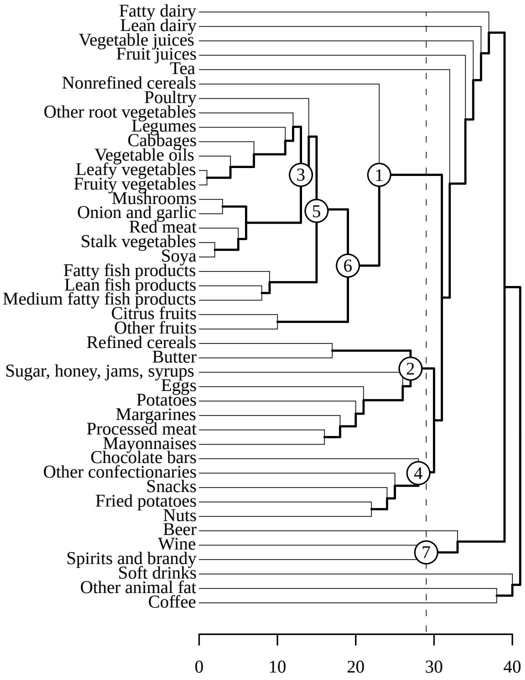
Fatty dairy
Lean dairy
Vegetable juices
Fruit juices
Tea
Nonrefined cereals
Poultry
Other root vegetables
Legumes
Cabbages
Vegetable oils
Leafy vegetables
Fruity vegetables
Mushrooms
Onion and garlic
Red meat
Stalk vegetables
Soya
Fatty fish products
Lean fish products
Medium fatty fish products
Citrus fruits
Other fruits
Refined cereals
Butter
Sugar, honey, jams, syrups
Eggs
Potatoes
Margarines
Processed meat
Mayonnaises
Chocolate bars
Other confectionaries
Snacks
Fried potatoes
Nuts
Beer
Wine
Spirits and brandy
Soft drinks
Other animal fat
Coffee
①
②
③
④
⑤
⑥
⑦
0
10
20
30
40

Figure 4. Cluster tree produced by the treelet transform applied to dietary data from 26,155 men in a prospective cohort study of dietary patterns and risk of myocardial infarction, Denmark, 1993-2008. The dashed line indicates the selected cut-level (level 29) for the cluster tree. Numbered circles indicate highest-variance factors at this cut-level; leaves descending from these nodes (food groups, left) correspond to non-zero loadings in the given factor.

3.1. *Associations between dietary patterns and risk of myocardial infarction*

During a median follow-up time of 11.9 years, we identified 1,523 incident cases of MI. Hazard ratios of MI by quintiles of factor scores, adjusted for confounders, are presented in Table 3. Unless otherwise mentioned, results are from these adjusted analyses.

Correlations between TT factor scores were modest (<0.15), except for the scores of factors 2 and 3 (correlation = 0.27). This justified univariate regression analysis on TT factor scores, as is conventionally done for the uncorrelated PCA factor scores.

Both PCA and TT factor 2 were positively associated with risk of MI. Similarly, PCA factor 3 was associated with risk of MI. We hypothesized that the association for PCA factor 3 was attributable to the large negative loadings on alcohol (beer, wine, spirits/brandy). Indeed, when loadings for alcoholic beverages were set to zero for this factor, the association vanished. This suggested that PCA factor 3 conveyed some of the same information as TT factor 7. Lastly, PCA factor 5, with large loadings on e.g. fish and beer, was associated with risk of MI. This factor was not rediscovered among the TT factors.

In crude analyses (Supplementary Table 4 in the appendix), PCA factor 1, TT factor 3, and TT factor 1 had a strong positive respectively, negative association with risk of MI. The weak adjusted association between PCA factor 5 and risk of MI was not apparent in crude analyses.

We constructed the two Cox regression models including confounding variables and

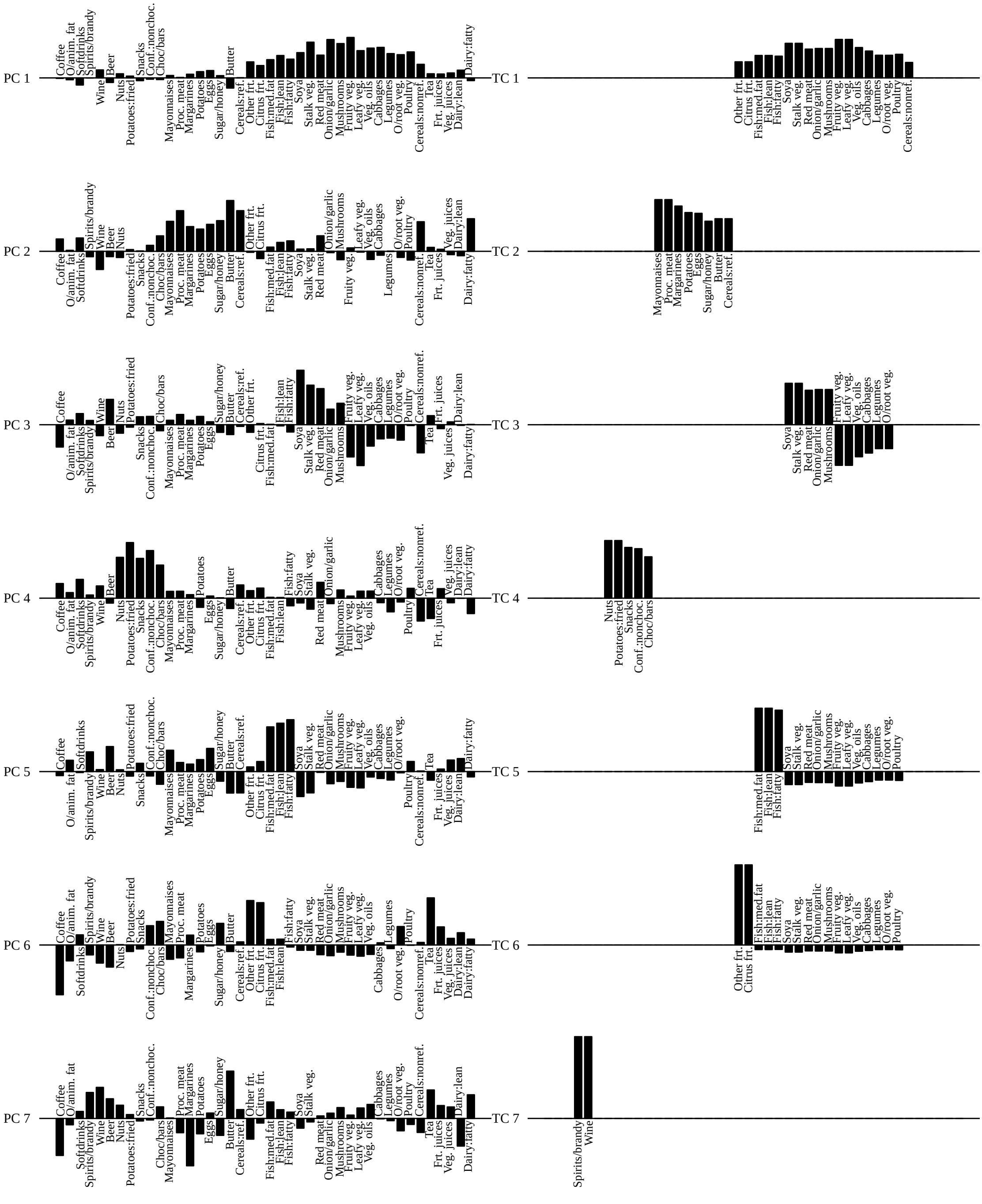

Figure 5. Loading plots of the 7 factors from a Procrustes-rotated principal component analysis (PCA) and the original treelet transform (TT), for 26,155 men in a prospective cohort study of dietary patterns and risk of myocardial infarction, Denmark, 1993-2008. The axis is oriented in the reading direction so that loadings above the line are positive, and those below are negative; e.g. all loadings in TT factor 1 are positive.

Table 3. Hazard Ratios of Risk of Myocardial Infarction According to Quintiles of Factor Scores for 26,155 Men in a Prospective Cohort Study of Dietary Patterns and Risk of Myocardial Infarction, Denmark, 1993-2008.

	Quintile 1	Quintile 2		Quintile 3		Quintile 4		Quintile 5		*P* for
	HR^a	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI	trend
PCA										
Factor 1	1.0	0.91	0.76,1.08	1.00	0.84,1.19	1.05	0.88,1.25	0.98	0.81,1.19	0.6
Factor 2	1.0	1.06	0.90,1.26	1.01	0.85,1.21	1.17	0.98,1.40	1.33	1.09,1.62	0.004
Factor 3	1.0	1.07	0.91,1.26	0.99	0.84,1.17	1.21	1.03,1.42	1.25	1.06,1.48	0.003
Factor 4	1.0	1.10	0.93,1.30	1.11	0.94,1.32	1.14	0.96,1.35	1.12	0.95,1.33	0.19
Factor 5	1.0	1.01	0.85,1.18	0.90	0.76,1.06	0.90	0.77,1.07	0.83	0.70,0.98	0.009
Factor 6	1.0	0.98	0.83,1.15	0.90	0.76,1.07	1.05	0.89,1.24	1.10	0.94,1.30	0.14
Factor 7	1.0	1.06	0.90,1.24	0.94	0.79,1.11	1.09	0.93,1.29	1.16	0.99,1.36	0.06
TT										
Factor 1	1.0	1.08	0.93,1.26	0.89	0.76,1.05	0.93	0.78,1.10	0.99	0.83,1.18	0.3
Factor 2	1.0	1.12	0.94,1.33	1.28	1.07,1.53	1.33	1.11,1.60	1.53	1.24,1.88	<0.001
Factor 3	1.0	0.86	0.72,1.02	1.01	0.86,1.19	1.03	0.87,1.21	1.08	0.91,1.27	0.09
Factor 4	1.0	1.01	0.86,1.17	0.93	0.79,1.09	0.97	0.82,1.14	1.12	0.95,1.33	0.4
Factor 5	1.0	1.02	0.86,1.20	1.06	0.90,1.24	0.98	0.83,1.15	0.91	0.77,1.07	0.2
Factor 6	1.0	1.06	0.90,1.25	1.03	0.87,1.21	1.04	0.88,1.22	0.98	0.83,1.15	0.7
Factor 7	1.0	0.84	0.73,0.98	0.82	0.70,0.96	0.71	0.61,0.84	0.80	0.69,0.94	<0.001

Abbreviations:
PCA, principal component analysis; TT, treelet transform; CI, confidence interval; HR, hazard ratio.

[a] Hazard ratios were adjusted for total energy intake (continuous variable), body mass index (weight in kilograms divided by height in meters squared) (<25, 25-29, and ⩾ 30), educational level (<8, 8-10, and >10 years), smokingstatus (never, former, and current smoker of 1-14, 15-24, or ⩾25 g tobacco/day), leisure-time physical activity (<3.5 and ⩾3.5 hours/week), and history of hypertension (yes, no, and do not know).

score quintiles of all 7 factors from PCA and TT, respectively. The model for PCA gave an Akaike's information criterion value of 478 versus 486 in the model for TT. Hence, there was negligible difference in overall goodness-of-fit between PCA and TT factors as predictors of risk of MI.

4. Discussion

The use of PCA and related dimension reduction methods in nutritional epidemiology remains controversial. Critics point to the questionable biological relevance of mathematical factors, the poor generalizability of exploratory techniques, and specific technical issues (Martinez *et al.*, 1998; Jacques and Tucker, 2001). Key points in the latter category pertain to the challenge of interpreting factors; the subjectivity inherent in deciding which factor loadings to report; and the use of arbitrary post-hoc factor rotations.

In this methodological paper, TT has been proposed to address these technical shortcomings of PCA. TT combines ideas from cluster analysis with those of PCA. It endows the collection of variables under study with both a hierarchical grouping structure and a collection of factors with loading sparsity patterns reflecting the grouping structure. The hierarchical grouping of variables distinguishes TT from another recent statistical technique, sparse PCA (Zou *et al.*, 2006), which yields sparse loadings but in a more black-box manner. TT is closer akin to techniques proposed to study gene expressions (Hastie *et al.*, 2001), where ‘factors’ result by averaging variables on a pre-constructed cluster tree.

In a study of dietary patterns and risk of MI among middle-aged men in a large Danish cohort, we demonstrated that TT may offer several advantages over PCA. TT identified a similar number of factors responsible for the main variation in dietary intake, explaining almost as much variation as the factors derived from PCA. A key property of TT is its multi-scale nature, which leads to sparse loadings and enables detection of localized sources of variation in the data. In the present study, TT identified a factor loading solely on refined foods and eggs, factor 2, and a factor loading on alcoholic beverages, factor 7. Both factors were associated with risk of MI. Refined foods and alcoholic beverages were also identified by PCA as positive, respectively negative risk factors for MI (factors 2 and 3) but their interpretation was complicated by the many non-zero loadings. PCA identified an additional risk factor, factor 5, which was not recognizable among TT factors. However, its complex loading pattern rendered interpretation challenging.

The associations found in this study were comparable to what has been observed elsewhere. Except for the large loading on red meat, both PCA and TT factor 1 resembled a typical ‘prudent pattern’, which has been shown to be associated with lower cardiovascular disease risk in some other studies (Hu, 2002; DiBello *et al.*, 2008; Osler *et al.*, 2001; Iqbal *et al.*, 2008) but not all (Martínez-Ortiz *et al.*, 2006; Shimazu *et al.*, 2007). In the present study, these factors were, however, not linked to the risk of MI after confounder adjustment. The TT factor 2 bore similarities to a typical ‘Western pattern’ which has been linked to increased risk of cardiovascular disease (Hu, 2002; Osler *et al.*, 2001; Iqbal *et al.*, 2008; Shimazu *et al.*, 2007).

As a contribution to the literature on dietary patterns and disease risk, strengths of the present study included the large sample size, the prospective design, the use of validated questionnaires, and the high quality of follow-up. Differential recall and selection bias are thus unlikely to have had a major effect on our findings. The main limitation of the study was the possibility of residual confounding, particularly in relation to physical activity and socioeconomic status. Confounding from other MI risk factors not taken into account remains a possible explanation for the observed associations.

Factor rotation is commonly used to simplify interpretation of PCA-based dietary patterns. In the present study, large correlations between scores for the first 5 Procrustes-rotated PCA factors and their TT counterparts suggested that a post-hoc rotation of PCA might be able approximate the more easily interpretable TT factors. However, PCA factor rotation is a controversial procedure with no theoretical support: it requires several arbitrary decisions with a potentially large impact on the final solution, in addition to destroying key properties of PCA (Jolliffe and Morgan, 1992; Jolliffe, 1989, 1995). In contrast, TT is able to automatically untangle the data complexity, essentially by providing localization to the global information conveyed by PCA factors. In fact, the agreement in the Procrustes analysis suggests that one may informally interpret TT as a ‘de-noised’ version of PCA.

The sparsity of TT factors may seem an unnatural feature of a dietary pattern. It is important to emphasize that dimension reduction methods, with their focus on variables, lead to statements about the variance structure among intake variables; not statements about adherences to real-world dietary patterns. Sparse dietary patterns simply convey the data reduction assumption of a distinct block correlation structure (high intra-block, low inter-block correlation), an assumption which is clarified graphically by the treelet cluster tree.

While TT addresses some of the criticisms levelled at PCA, it also has several limitations. The most important limitation is the necessity of deciding a cut-level for the cluster tree before factors can be extracted. This represents a model selection problem similar to the problem of selecting the number of clusters in a cluster analysis (Michels and Schulze, 2005). The cut-level may influence both sparsity and composition of the factors; it can be selected using cross-validation. There will typically be a range of different cut-levels which lead to similar fits, but with slightly different factor compositions. While discouraging, this Rashomon effect is probably a more truthful account of our actual state of knowledge compared to PCA where model selection is done post-hoc and less explicitly. The inconclusivity may be handled proactively by performing TT at several cut-levels. Likewise, factor composition may be sensitive to perturbations in the data, reflecting instabilities in the cluster tree. We assessed stability of factors and the cluster tree simultaneously by subsampling loading sign patterns, an approach which resembles techniques from cluster analysis (Ben-Hur *et al.*, 2002). We found an unusually high degree of stability of TT; in our experience, it is more common to see patterns differing on one or two variables compete in stability analyses, as was also seen for TT factor 7. Lastly, it must be stressed that TT factor scores have non-zero correlation, an issue to be aware of when fitting regression models to factor scores. Also, TT may not always be appropriate: it is designed to perform well primarily for collections of variables exhibiting a distinct block structure and with a simple inter-block correlation structure (Lee *et al.*, 2008).

In conclusion, we believe TT to be useful in future studies of dietary patterns as well as epidemiological studies of more complex and novel data such as dietary fatty acids, patterns among interrelated biomarkers, and 'omics' data. TT is one example from the growing array of sparse estimation methods which are promising for epidemiological research in general; providing generic frameworks for model selection in a multi-dimensional setting and leading to more interpretable models with the potential for greater insight into disease etiology. Supervised sparse estimation methods, which take into account an outcome variable when deriving patterns, would seem a particularly promising addition to the epidemiologist's toolbox. Their scope and proper use within epidemiology is an important future research topic.

References

Andersen, T. F., Madsen, M., Jørgensen, J. *et al.* (1999) The Danish National Hospital Register. a valuable source of data for modern health sciences. *Dan Med Bull.*, **46**, 263–268.

Ben-Hur, A., Elisseeff, A. and Guyon, I. (2002) A stability based method for discovering structure in clustered data. *Pac Symp Biocomput.*, 6–17.

DiBello, J. R., Kraft, P., McGarvery, S. T. *et al.* (2008) Comparison of 3 methods for identifying dietary patterns associated with risk of disease. *Am J Epidemiol.*, **168**, 1433–1443.

Gervini, D. and Rousson, V. (2004) Criteria for evaluating dimension-reducing components for multivariate data. *Am Stat.*, **58**, 72–76.

Gorst-Rasmussen, A. (2011) tt: Treelet transform with Stata. Tech. Rep. R-2011-09, Aalborg University.

Gower, J. C. (1995) Orthogonal and projection Procrustes analysis. In *Recent advances in descriptive multivariate analysis* (ed. W. J. Krzanowski), 113–134. Oxford University Press.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) Supervised harvesting of expression trees. *Genome Biol.*, **2**, research0003.

Hu, F. B. (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol.*, **13**, 3–9.

Hu, F. B., Rimm, E. B., Stampfer, M. J. *et al.* (2000) Prospective study of major dietary patterns and risk of coronary heart disease in men. *Am J Clin Nutr.*, **72**, 912–921.

Iqbal, R., Anand, S., Ounpuu, S. *et al.* (2008) Dietary patterns and the risk of acute myocardial infarction in 52 countries: results of the INTERHEART study. *Circulation*, **118**, 1929–1937.

Jacques, P. F. and Tucker, K. K. (2001) Are dietary patterns useful for understanding the role of diet in chronic disease? *Am J Clin Nutr.*, **73**, 1–2.

Joensen, A. M., Jensen, M. K., Overvad, K. *et al.* (2009) Predictive values of acute coronary syndrome discharge diagnoses differed in the Danish National Patient Registry. *J Clin Epidemiol.*, **62**, 188–194.

Jolliffe, I. T. (1989) Rotation of ill-defined principal components. *Appl Statist.*, **38**, 139–147.

Jolliffe, I. T. (1995) Rotation of principal components: choice of normalization constraints. *J Appl Statist.*, **22**, 29–35.

Jolliffe, I. T. and Morgan, B. J. (1992) Principal component analysis and exploratory factor analysis. *Stat Methods Med Res.*, **1**, 69–95.

Juel, K. and Helweg-Larsen, K. (1999) The Danish registers of causes of death. *Dan Med Bull.*, **46**, 354–357.

Kennedy, E. T., Ohls, J., Carlson, S. *et al.* (1995) The healthy eating index: design and applications. *J Am Diet Assoc.*, **95**, 1103–1108.

Lau, C., Glümer, C., Toft, U. *et al.* (2008) Identification and reproducibility of dietary patterns in a Danish cohort: the Inter99 study. *Br J Nutr.*, **99**, 1089–1098.

Lauritsen, J. (1998) *FoodCalc 1.3*. URL: http://www.ibt.ku.dk/jesper/FoodCalc/ (Accessed 1 September, 2010).

Lee, A. B., Nadler, B. and Wasserman, L. (2008) Treelets – an adaptive multi-scale basis for sparse unordered data. *Ann Appl Stat.*, **2**, 435–471.

Luepker, R. V., Apple, F. S., Christenson, R. H. *et al.* (2003) Case definitions for acute coronary heart disease in epidemiology and clinical research studies: a statement from the AHA Council on Epidemiology and Prevention; AHA Statistics Committee; World Heart Federation Council on Epidemiology and Prevention; the European Society of Cardiology working group on epidemiology and prevention; Centers for Disease Control and Prevention; and the National Heart, Lung, and Blood Institute. *Circulation*, **108**, 2543–2549.

Martinez, M. E., Marshall, J. R. and Lee, S. (1998) Invited commentary: Factor analysis and the search for objectivity. *Am J Epidemiol.*, **148**, 17–19.

Martínez-Ortiz, J. A., Fung, T. T., Baylin, A. *et al.* (2006) Dietary patterns and risk of nonfatal acute myocardial infarction in Costa Rican adults. *Eur J Clin Nutr.*, **60**, 770–777.

Michels, K. B. and Schulze, M. B. (2005) Can dietary patterns help us detect diet-disease associations? *Nutr Res Rev.*, **18**, 241–248.

Moeller, S. M., Reedy, J., Millen, A. E. *et al.* (2007) Dietary patterns: challenges and opportunities in dietary patterns research an experimental biology workshop, April 1, 2006. *J Am Diet Assoc.*, **107**, 1233–1239.

Møller, A. and Saxholt, E. (1996) *Food composition tables 1996 (In Danish).* National Food Agency.

Newby, P. K. and Tucker, K. L. (2004) Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr Rev.*, **62**, 177–203.

Osler, M., Heitmann, B. L., Gerdes, L. U. *et al.* (2001) Dietary patterns and mortality in Danish men and women: a prospective observational study. *Br J Nutr.*, **85**, 219–225.

Overvad, K., Tjønneland, A., Haraldsdóttir, J. *et al.* (1991) Development of a semiquantitative food frequency questionnaire to assess food, energy and nutrient intake in Denmark. *Int J Epidemiol.*, **20**, 900–905.

Pedersen, C. B., Gotzsche, H., Møller, J. O. *et al.* (2006) The Danish Civil Registration System. a cohort of eight million persons. *Dan Med Bull.*, **53**, 441–449.

R Development Core team (2009) *R: A Language and Environment for Statistical Computing*.

Schulze, M. B. and Hoffmann, K. (2006) Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. *Br J Nutr.*, **95**, 860–869.

Shimazu, T., Kuriyama, S., Hozawa, A. *et al.* (2007) Dietary patterns and cardiovascular disease mortality in Japan: a prospective cohort study. *Int J Epidemiol.*, **36**, 600–609.

Slattery, M. L., Boucher, K. M., Caan, B. J. *et al.* (1998) Eating patterns and risk of colon cancer. *Am J Epidemiol.*, **148**, 4–16.

Tjønneland, A., Olsen, A., Boll, K. *et al.* (2007) Study design, exposure variables, and socioeconomic determinants of participation in Diet, Cancer and Health: a population-based prospective cohort study of 57,053 men and women in Denmark. *Scand J Public Health.*, **35**, 432–441.

Tjønneland, A., Overvad, K., Haraldsdóttir, J. *et al.* (1991) Validation of a semiquantitative food frequency questionnaire developed in Denmark. *Int J Epidemiol.*, **20**, 906–912.

Trichopoulou, A., Kouris-Blazos, A., Wahlqvist, M. L. *et al.* (1995) Diet and overall survival in the elderly. *BMJ*, **311**, 1457–1460.

Willet, W. (1998) *Nutritional Epidemiology*. Oxford University Press.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *J Comput Graph Statist.*, **15**, 265–286.

Appendix 1: An invited commentary and our response

Summary of 'Imamura F and Jacques PF (2011). Invited Commentary: Dietary Pattern Analysis. American Journal of Epidemiology; ***173****(10):1105–1108'*

In their commentary, Imamura and Jacques discuss the distinguishing aspects of dietary pattern analysis and argue that the sparsity of TT factors may be inappropriate in dietary pattern analysis since 'the cumulative role of all foods is important for biologic influence of diet and public health messages'. They proceed to comment on the overlaps and differences of pattern analysis in dietary epidemiology and genetics:

1. For use in connection with disease prediction. Here they point out the usefulness of being able to determine a small set of variables responsible for exposure-disease associations in both genetic and dietary epidemiological studies alike.

2. For use in connection with adjustment for pattern confounding. Here they argue that there is typically no need for the key features of sparsity and clearer interpretation offered by TT; except in a qualitative and exploratory setting.

Imamura and Jacques conclude by discussing the issue of validity and suggest that there are two forms of validity for dietary patterns: validity in the sense of patterns capturing true dietary patterns – and validity with respect to disease prediction. They argue that, internally, the first type of validity can be assessed indirectly by checking if similar patterns result from different analytic approaches, among which TT would seem a promising option.

Gorst-Rasmussen A, Dahm CC, Dethlefsen C, Scheike T, Overvad K (2011). Response to invited commentary: Gorst-Rasmussen et al. respond to "Dietary Pattern Analysis". American Journal of Epidemiology; ***173****(10):1109–1110*

We thank Imamura and Jacques [1] for their insightful commentary on our article [2], in which they go beyond the treelet transform (TT) to critically discuss the relevance of sparsity in dietary pattern analysis. We limit this response to challenging a fundamental premise in their discussion, which is that sparsity is not a natural property of a dietary pattern because a dietary pattern should reflect the cumulative effect of all foods. We acknowledge the intuitive appeal of directly connecting the concept of a diet with dietary patterns, but diets remain individual-specific constructs, whereas dietary patterns are population-based and usually observational. Attempts to provide a universal, isolated understanding of the concept of a dietary pattern will lead to subjective and ambiguous definitions at best. It would be akin to Wittgenstein's famous beetle-in-a-box analogy [3]: Suppose that everyone has a beetle in a box and that no one can see anyone else's beetle. The actual content of our private boxes would thus be completely irrelevant for our public discussion of beetles. How, then, can we ever hope to discuss beetles scientifically? To avoid such issues, we consider an ostensive definition of dietary patterns more appropriate: A dietary pattern is a pattern produced by a dietary pattern analysis. More operationally, it is a means of data reduction [4]. Principal component analysis produces patterns that are eigenvectors of a correlation matrix of foods; cluster analysis produces patterns that show food averages within clusters; and TT produces patterns by aggregating foods according to correlation.

Different methods might or might not [5, 6] reflect similar aspects of data; some may even produce patterns that can somehow be translated to an actual diet. However, no one method can claim more validity per se than any other, be it sparse or not. This does not make dietary pattern analysis a vacuous exercise, but simply implies that it must be judged strictly externally, in terms of its usefulness: for predicting disease, for generating hypotheses, and for communicating public health messages. Within this view, we agree with Imamura and Jacques that there are situations in which sparsity is less useful. Confounding by dietary patterns [7] is one such example. Conversely, sparsity appears useful in confirmative factor analytic studies, as observed by Imamura and Jacques. In addition, as we argued in our original article, sparsity certainly seems useful in the majority of factor-analytic dietary pattern analyses, in which pattern sparsity is currently approximated by intricate exercises of factor rotation and loading truncation [6].

TT seems a promising technique for dietary pattern analysis because it could offer directly what researchers seek from a dietary pattern: a simplified interpretation without sacrifice of predictive properties [2]. However, TT is no silver bullet. As with any statistical method, its usefulness must stand the test of time and be subjected to the usual vigilance regarding underlying assumptions when applied in practice. TT will sometimes fail, but so will any method of dietary pattern analysis. Imamura and Jacques [1] mention a scenario in which the sum of systolic and diastolic blood pressure appears as a factor, although the difference is more relevant for disease prediction [8]. Any method of pattern analysis that disregards the outcome would fail in this example, which only serves to emphasize the relevance of supervision in pattern analyses.

Do we consider sparsity essential to dietary pattern analysis? No. Sparsity simply represents one promising way of enriching methodology for pattern analysis with additional structure so that it may support the scientific process rather than developing into a series of mysteries to be untangled ad hoc and case by case. The endeavor to ensure methodological transparency is essential, both in nutritional epidemiology and elsewhere.

References

1. Imamura F, Jacques PF (2011). Invited commentary: dietary pattern analysis. *Am J Epidemiol*; **173**:1105–1108.
2. Gorst-Rasmussen A, Dahm CC, Dethlefsen C, et al. (2011). Exploring dietary patterns by using the treelet transform. *Am J Epidemiol*; **173**:1097–1104.
3. Wittgenstein L (1967). *Philosophical Investigations*. 3rd ed. Anscombe GEM, trans. Oxford, UK: Basil Blackwell and Mott.
4. Slattery ML (2008). Defining dietary consumption: is the sum greater than its parts? *Am J Clin Nutr*; **88**:14–15.
5. Reedy J, Wirfält E, Flood A, et al. (2010) Comparing 3 dietary pattern methods – cluster analysis, factor analysis, and index analysis – with colorectal cancer risk: the NIH-AARP Diet and Health Study. *Am J Epidemiol*; **171**:479–487.
6. Newby PK, Tucker KL (2009). Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr Rev.*; **62**:177–203.
7. Imamura F, Lichtenstein AH, Dallal GE, et al. Confounding by dietary patterns of the inverse association between alcohol consumption and type 2 diabetes risk. *Am J Epidemiol*;**170**:37–45.
8. Lawlor DA, Ebrahim S, May M, et al (2004). (Mis)use of factor analysis in the study of insulin resistance syndrome. *Am J Epidemiol*; **159**:1013–1018.

Appendix 2: Supplementary tables

Supplementary Table 1. Baseline Characteristics for 26,155 Men in a Prospective Cohort Study of Dietary Patterns and Risk of Myocardial Infarction, Denmark, 1993-2008[a].

Variable	Value
Energy intake (kJ/day)	10930 (2760)
BMI (kg/m^2)	26.6 (3.6)
Education	
<8 yrs (%)	34
8-10 yrs (%)	42
>10 yrs (%)	24
Physical activity ⩾3.5 hrs/week (%)	65
Hypertension at baseline (%)	69
Smoking (%)	
Never	26
Former	34
Current 1-14 g tobacco/day	11
Current 15-24 g tobacco/day	17
Current ⩾25 g tobacco/day	12

[a] Data are presented as mean values (with standard deviation in parentheses) or percentages.

Supplementary Table 2. Dietary Intake Data for 26,155 Men in a Prospective Cohort Study of Dietary Patterns and Risk of Myocardial Infarction, Denmark, 1993-2008[a].

Variable	Abbreviation	Intake (g/day)
Fatty dairy products	Dairy:fatty	64.1 (64.1, 518.6)
Lean dairy products	Dairy:lean	159.0 (159.0, 922.0)
Vegetable juices	Veg. juices	0.0 (0.0, 16.4)
Fruit juices	Frt. juices	8.4 (8.4, 100.5)
Tea		28.6 (28.6, 900.0)
Nonrefined cereals	Cereals:nonref.	140.0 (140.0, 280.6)
Poultry		19.7 (19.7, 61.3)
Other root vegetables	O/root veg.	16.2 (16.2, 90.8)
Legumes		0.3 (0.3, 2.7)
Cabbages		14.4 (14.4, 43.8)
Vegetable oils	Veg. oils	2.1 (2.1, 17.4)
Leafy vegetables	Leafy veg.	7.2 (7.2, 37.7)
Fruity vegetables	Fruity veg.	54.7 (54.7, 137.0)
Mushrooms		9.6 (9.6, 33.0)
Onion and garlic	Onion/garlic	18.2 (18.2, 50.1)
Red meat		100.0 (100.0, 190.3)
Stalk vegetables	Stalk veg.	7.4 (7.4, 21.9)
Soya		0.1 (0.1, 0.5)
Fatty fish products	Fish:fatty	13.4 (13.4, 43.5)
Lean fish products	Fish:lean	19.0 (19.0, 49.8)
Medium fat fish products	Fish:med.fat	6.1 (6.1, 19.9)
Citrus fruits	Citrus frt.	10.7 (10.7, 101.2)
Other fruits	Other frt.	89.9 (89.9, 362.6)
Refined cereals	Cereals:ref.	54.8 (54.8, 140.7)
Butter		12.8 (12.8, 40.3)
Sugar, honey, jams, syrup	Sugar/honey	24.8 (24.8, 128.3)
Eggs		23.8 (23.8, 70.9)
Potatoes		146.1 (146.1, 344.1)
Margarines		13.4 (13.4, 48.0)
Processed meat	Proc. meat	35.0 (35.0, 89.8)
Mayonnaises		2.6 (2.6, 18.0)
Chocolate bars		4.2 (4.2, 22.0)
Other confectionaries	Conf.:nonchoc.	9.0 (9.0, 53.1)
Snacks		0.8 (0.8, 4.1)
Fried potatotes	Potatoes:fried	3.2 (3.2, 14.9)
Nuts		0.8 (0.8, 7.1)
Beer		168.6 (168.6, 1495.9)
Wine		55.5 (55.5, 321.1)
Spirits and brandy	Spirits/brandy	2.5 (2.5, 30.0)
Soft drinks		16.9 (16.9, 201.0)
Other animal fat	O/anim. fat	0.2 (0.2, 4.0)
Coffee		900.0 (900.0, 1600.0)

[a] Data are presented as median values (with 5th and 95th percentiles in parentheses).

Supplementary Table 3. Description of Food Groups in a Prospective Cohort Study of Dietary Patterns and Risk of Myocardial Infarction Among 26,155 Men, Denmark, 1993-2008.

Variable	Intake (g/day)
Fatty dairy products	Cheese, cream, whole milk, whole milk products
Lean dairy products	Low fat milk, low fat milk products
Vegetable juices	Carrot juice, tomato juice
Fruit juices	Orange juice, grapefruit juice, lemon juice
Tea	
Nonrefined cereals	Oatmeal, muesli, rye bread, rye meal, corn (cob/kernels)
Poultry	Chicken, turkey
Other root vegetables	Carrots, celeriac, ginger
Legumes	Beans, peas
Cabbages	Cauliflower, broccoli, red cabbage, white cabbage, borecole, brussel sprouts
Vegetable oils	
Leafy vegetables	Spinach, salads
Fruity vegetables	Tomatoes (incl. canned), cucumber, pepper, aubergine, squash, avocado, green beans
Mushrooms	Champignons (incl. preserved), other mushrooms
Onion and garlic	
Red meat	(Unprocessed) beef, pork, veal, lamb, entrails
Stalk vegetables	Leeks, chives, bean sprouts, asparagus, rhubarb, bamboo shots
Soya	Soy sauce
Fatty fish products	Fish and seafood (incl. preserved), >8 g fat/100 g
Lean fish products	Fish and seafood (incl. preserved), <2 g fat/100 g
Medium fat fish products	Fish and seafood (incl. preserved), 2-8 g fat/100 g
Citrus fruits	Oranges, grapefruits, mandarins
Other fruits	Apples, pears (incl. preserved), peaches (incl. preserved), prunes, nectarines, strawberries, bananas, kiwis, melon, pineapple (incl. preserved)
Refined cereals	White bread, wheat flour, pasta, rice, corn starch/meal, crisp bread
Butter	
Sugar, honey, jams, syrup	Desserts, cakes, honey, jam, syrup
Eggs	
Potatoes	Potatoes (non-fried)
Margarines	
Processed meat	Sausages, cold cuts, ham, bacon, liver paste
Mayonnaises	Mayonnaise, remoulade
Chocolate bars	Chocolate and chocolate bars
Other confectionaries	
Snacks	Chips, pork crackling
Fried potatotes	French fries, pan-fried potatoes
Nuts	
Beer	Low-alcohol, regular, and export beer
Wine	Red wine, white wine, port wine
Spirits and brandy	All kinds of spirits
Soft drinks	Carbonated/non-carbonated softdrinks
Other animal fat	Lard
Coffee	

Supplementary Table 4. Crude Hazard Ratios of Risk of Myocardial Infarction According to Quintiles of Factor Scores for 26,155 Men in a Prospective Cohort Study of Dietary Patterns and Risk of Myocardial Infarction, Denmark, 1993-2008[a].

	Quintile 1	Quintile 2		Quintile 3		Quintile 4		Quintile 5	
	HR	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
PCA									
Factor 1	1.0	0.93	0.78,1.10	1.07	0.91,1.26	1.21	1.03,1.41	1.22	1.04,1.43
Factor 2	1.0	1.17	0.99,1.39	1.17	0.99,1.39	1.41	1.20,1.66	1.60	1.36,1.88
Factor 3	1.0	1.03	0.87,1.21	0.94	0.80,1.11	1.11	0.95,1.31	1.07	0.91,1.25
Factor 4	1.0	1.15	0.97,1.35	1.15	0.97,1.35	1.20	1.02,1.42	1.21	1.03,1.42
Factor 5	1.0	1.04	0.88,1.22	0.95	0.81,1.12	1.00	0.85,1.18	1.02	0.87,1.20
Factor 6	1.0	0.97	0.83,1.13	0.87	0.74,1.02	0.97	0.83,1.14	1.02	0.88,1.20
Factor 7	1.0	1.03	0.88,1.22	0.95	0.80,1.12	1.11	0.95,1.30	1.22	1.05,1.43
TT									
Factor 1	1.0	1.00	0.86,1.16	0.78	0.67,0.92	0.78	0.66,0.91	0.82	0.70,0.96
Factor 2	1.0	1.10	0.93,1.30	1.25	1.06,1.48	1.24	1.05,1.46	1.39	1.19,1.64
Factor 3	1.0	0.94	0.79,1.12	1.19	1.01,1.40	1.28	1.09,1.50	1.40	1.19,1.64
Factor 4	1.0	1.00	0.86,1.17	0.89	0.76,1.04	0.92	0.78,1.08	1.02	0.87,1.19
Factor 5	1.0	1.05	0.89,1.23	1.08	0.92,1.27	1.03	0.87,1.21	0.98	0.83,1.15
Factor 6	1.0	1.09	0.93,1.28	1.01	0.86,1.19	0.99	0.84,1.16	0.89	0.76,1.05
Factor 7	1.0	0.75	0.64,0.87	0.71	0.61,0.82	0.63	0.54,0.74	0.71	0.61,0.82

Abbreviations:
PCA, principal component analysis; TT, treelet transform; HR, hazard ratio; CI, confidence interval.

Paper V

tt: Treelet Transform with Stata

Author list

Anders Gorst-Rasmussen
Aalborg University, Denmark

Summary

The treelet transform (TT) is a recent data reduction technique from the field of machine learning. Sharing many similarities with principal components analysis (PCA), TT can reduce a multidimensional data set to the projections on a small number of directions or components which account for much of the variation in the original data. However, in contrast to PCA, TT produces sparse components. This can greatly simplify interpretation. We describe the `tt` Stata add-on for performing TT. The add-on includes a Mata implementation of the TT algorithm, alongside other functionality to aid the practical application of TT. We show how a basic exploratory data analysis using the `tt` add-on might look.

Supplementary info

The software development related to this manuscript was done while based at Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital, as an affiliate of the Nordic Centre of Excellence 'SYSDIET' funded by NordForsk.

The manuscript has been published as:

> Gorst-Rasmussen A (2011). tt: Treelet Transform with Stata. *Technical report R-2011-09*. Department of Mathematical Sciences, Aalborg University.

It has moreover been submitted to *Stata Journal*.

1. Introduction

A common task in data analysis is to summarise a multidimensional data set. One popular and convenient approach is to find a few interesting directions in the data and use the corresponding linear projections of data as representatives of the original data in plots, regression models etc. This is known as dimension reduction. Principal components analysis (PCA) is a standard dimension reduction method which works by calculating the first few eigenvectors (components) of a covariance or correlation matrix and reducing the data set to a collection of component scores – the projection of data onto components. This strategy has the optimality property of explaining as much variation as possible in the original data using as few dimensions as possible. Often, entries of the components (loadings) are subjected to interpretation. Variables corresponding to 'large' loadings are interpreted as being important for describing the original data; variables corresponding to 'small' loadings can be discarded. Such interpretation is complicated by the fact that all component loadings are nonzero. Various cutoff rules, component rotation strategies etc. have been developed to simplify interpretation (Jolliffe, 2002) but these largely ad hoc procedures do not contribute to the transparency and objectivity of PCA.

In the machine learning community, there has been a growing interest in developing alternatives to PCA which offer more interpretable components by forcing loading patterns where many loadings are exactly zero, i.e. by forcing sparse components. For example, Zou *et al.* (2006) developed a variant of PCA where sparse components are estimated via penalised regression with automatic variable selection. The treelet transform (TT) proposed by Lee *et al.* (2008) is a similar recent alternative to PCA. TT introduces sparsity among component loadings in an elegant and simple fashion by combining ideas from hierarchical clustering analysis with ideas from PCA. This leads to sparse components which, similarly to PCA components, account for a large part of the variation in the original data and can be used in an analogous manner. In addition, it leads to an associated cluster tree which provides a concise visual representation of loading sparsity patterns and the general dependency structure of the data.

We describe in this paper the Stata add-on `tt` (Gorst-Rasmussen, 2011) which contains a Mata implementation of the TT algorithm. In addition to the TT algorithm itself, `tt` includes a number of other functions to aid in model selection and output analysis in practice. Using the `cars` data set which comes with Stata, we provide a small demonstration of how the various functions work together, and how a complete TT analysis using `tt` might look.

2. The treelet transform algorithm

This section provides a brief, nontechnical review of the TT algorithm. For a more formal derivation of TT and its properties, see the original paper by Lee *et al.* (2008).

Given a collection of p variables, the TT algorithm proceeds as follows:

Variable pairing. Locate the two variables with the largest correlation coefficient.

Local PCA. Merge these two variables by performing PCA on them. Keep the new variable/score with the largest variance (the 'sum' variable), discard the other new variable/score (the 'residual' variable).

This yields a new collection of $p-1$ variables, namely the sum variable and the remaining $p-2$ original variables, on which we then repeat the above two steps. The 'variable pairing'/'local PCA' scheme is repeated for a total of $p-1$ times until only a single sum variable is left. This in turn defines a basic hierarchical clustering algorithm, the output of which is conveniently represented as a binary tree with p levels (a cluster tree or cluster dendrogram). Variables that are 'close' in this cluster tree, and are merged early, represent groups of more highly correlated variables.

Hierarchical clustering is in itself a well-known technique. The novelty of TT is its use of PCA to merge variables since it enables us to construct, at each level of the TT cluster tree, a complete coordinate system for the data. Specifically, viewing TT in terms of its action on components rather than variables, we start out with a coordinate system consisting of the trivial, one-variable components (the standard coordinate system of $\mathbb{R}^p$). Each local PCA of two variables corresponds to performing an orthogonal rotation of two components. It follows that a coordinate system for the data at a given level of the TT cluster tree is given by the collection of:

1. components corresponding to sum variables available at the current level and;
2. components corresponding to all previously calculated residual variables and;
3. 'trivial' components for variables that have not yet joined the cluster tree.

The level- and data-specific coordinate system is thus comprised of 'sum' components which encode coarse-grained, low-resolution information about the dependency relationships between all variables included so far; alongside 'residual' components which encode information about the more local relationships between variables at an increasingly greater resolution. It can be shown that if TT is applied to a collection of variables with a covariance matrix featuring high intra-block correlation and low inter-block correlation then the loadings of sum components will be constant on variables within blocks (Lee *et al.*, 2008) in large samples. Hence, TT can help identify groups of correlated variables.

2.1. *Selecting a cut-level*

Application of TT to a data set yields, as its basic output, a cluster tree alongside a coordinate system for the data at each level of the cluster tree. As described above, the coordinate system combines coarse components not unlike components obtained from PCA, with higher-resolution components which reflect local dependency relationships. We seek to utilise this collection of coordinate systems for dimension reduction purposes.

If we knew which cluster tree level (cut-level) to use, we could calculate variances of the level-specific component scores and retain components corresponding to the highest-variance scores. This is the approach used in PCA with one difference: TT component scores are generally correlated and do not lead to a true decomposition of variance. This is a known issue in dimension reduction (Gervini and Rousson, 2004) since PCA is the only method yielding both orthogonal components and uncorrelated scores.

Selecting a cut-level for the TT cluster tree amounts to deciding the level of detail desired in the dimension reduction, i.e. the amount of regularisation. A coordinate system close to the leaves of the cluster tree contains mostly highly sparse components and may not be useful for dimension reduction in the sense that the high-resolution components are not much more informative than the original one-variable components. Conversely, a coordinate system close to the root includes coarse-grained, low-resolution components more suitable for dimension reduction but may be harder to interpret because of lacking sparsity. We usually prefer a data-driven choice of cut-level. Choosing a cut-level from data is not trivial since coordinate systems at different cut-levels are equally capable of describing the data if only we use a sufficiently large number of components. However, cross-validation can be used to find a cut-level at which we can describe the data using only few components. Suppose that we wish to describe the data using exactly m components. Then we determine an appropriate cut-level by using the following K-fold cross-validation strategy (Lee *et al.* (2008)):

1. Split the data randomly into K roughly equal-sized subsets. For each of these subsets, do the following:
    - For each cut-level $1,\dots,p-1$ calculate the m highest-variance components using all subsets of data *except* the current. Next, calculate the sum of variances of scores based on these components using only the *current* subset.
2. For each cut-level $1,\dots,p-1$, calculate a cross-validation score by averaging the K sums of component variances obtained in step 1.

A flowchart visualising step 1 of the cross-validation strategy is shown in Figure 1.

Once cross-validation scores have been obtained, a suitable cut-level can be found by locating a 'knee' on the graph of cross-validation scores against cut-level, i.e. a point at which increasing the cut-level does not substantially increase the cross-validation score. In other words, we select the cut-level at which we can explain almost as much variation as possible, using as low a cut-level as possible to simplify interpretation of components.

Note that the cross-validation strategy requires us to specify the number of components m to use. This is not much different from the corresponding problem of selecting the number of components to retain in PCA; or the number of clusters in a cluster analysis. In Section 4, we propose a simple data-driven strategy for selecting both cut-level and the number of components.

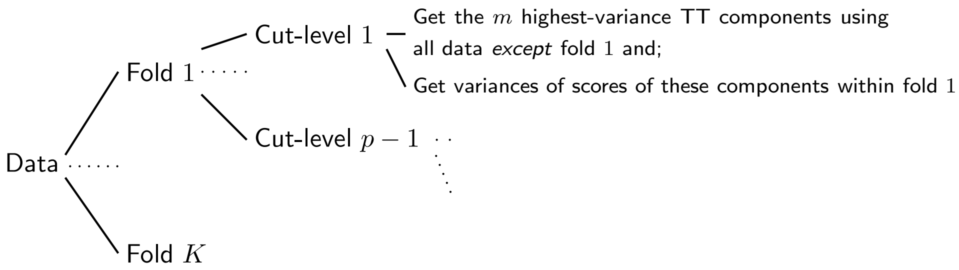
Data
Fold 1
Fold K
Cut-level 1
Cut-level $p-1$
Get the m highest-variance TT components using all data *except* fold 1 and;
Get variances of scores of these components within fold 1

Figure 1. Flow chart for the cross-validation strategy for deciding an optimal cut-level.

2.2. *Stability assessment*

A data analyst may wish to know how much trust to place in a collection of components obtained using TT. Since a key feature of TT is its ability to produce sparse components, it is of particular interest to assess the stability of loading sparsity patterns. This can be done by using a subsampling approach inspired by Ben-Hur *et al.* (2002).

We first specify a cut-level k and a number m of TT components to retain. Then we repeat the following subsampling scheme 100 times:

1. Randomly sample 80% of the data.
2. Within this subsample, calculate the m highest-variance TT components at cut-level k of the cluster tree. For each of these m components, do the following:
    - Calculate the sign pattern of the component. For example, a component $(-0.1, 0.2, 0, 0.1)$ corresponds to the sign pattern $(-,+,0,+)$.
    - Calculate the variance explained by the corresponding component.
    - Calculate the rank according to the variance explained by the corresponding component.

The collection of all $100 \cdot m$ sign patterns, alongside their variances and ranks, carries information about the stability and the importance of different sign patterns appearing in the subsampled TT analyses. As a measure of stability, we count the number of times we see a particular sign pattern among all $100 \cdot m$ patterns while using the average rank and average variance of the sign pattern as measures of importance. The final output of the stability analysis is the relative frequency, average variance, and average rank of each sign pattern occurring in more than 10 out of the 100 subsampled TT analyses. Note that this number is generally different from m.

3. The tt add-on

3.1. *Syntax*

The main function in the `tt` add-on (Gorst-Rasmussen, 2011) is implemented as a Mata function called via a Stata wrapper. It is loosely based on the R-code by Liu (2010) and has syntax:

```
tt varlist[if] [in] [weight] , cut(#) [options]
```

After calling `tt`, the user will typically call `ttcv` which uses the cross-validation strategy of Section 2.1 to select a cut-level for the TT cluster tree. It has the following syntax:

```
ttcv varlist[if] [in] [weight] , components(#) [options]
```

A range of different post-estimation commands is also available. As usual with post-estimation commands, they require an initial call to `tt`. Stability assessment as described in Section 2.2 is available in the command `ttstab` which has syntax:

```
ttstab [, options]
```

The TT cluster tree can be plotted by using the following command:

```
ttdendro [, dendro_options]
```

Scree plots of variances and 'skyscraper plots' of component loadings are implemented in the commands `ttscree` and `ttloading`, respectively, with syntax:

```
ttscree [, options scatter_options]
```

```
ttloading [, options scatter_options]
```

Finally, `ttpredict` implements prediction of component scores. As previously described, these are the projections of the original data onto the relevant TT components and can be informally interpreted as the degree of 'adherence' of a given observation vector to the given component. The `ttpredict` syntax is:

```
ttpredict [if] [in] {stub*|newvarlist}
```

3.2. *tt options*

`cut(#)` is required and specifies the cut-level of the TT cluster tree at which to extract components. The cut-level influences both the sparsity and composition of components.

`components(#)` sets the maximum number of components to be retained. `tt` displays the full set of components variances but displays loadings only for retained components. The default is the number of variables in *varlist*.

`corrrelation` or `covariance` specifies that TT cross-validation be based based on the correlation matrix or covariance matrix, respectively. The default is `correlation`. Usually, TT based on the covariance matrix will be meaningful only if variables are expressed in the same units.

`noblanks` display zero loadings as 0 instead of blanks; included for readability.

3.3. *ttcv options*

`components(#)` is required and sets the number of components to be retained. In practice, this number may not be known in advance; in which case one should investigate the output of `ttcv` for a range of different choices `components()`.

`folds(#)` specifies the number of folds (test samples) to use in cross-validation. The default is `folds(10)`.

`reps(#)` specifies the number of Monte-Carlo repetions of cross-validation. Default is `reps(5)`. Monte-Carlo repetitions reduce the sampling variation inherent in cross-validation; increase `reps(#)` if the output of `ttcv` appears unstable over different runs.

`percent(#)` specifies that a "knee" on the graph of cross-validation scores should be sought among cut-levels for which the score is within #percent of the cross-validation score associated with the maximal cut-level. Default is `percent(10)`.

`corrrelation` or `covariance` specifies that TT cross-validation be based based on the correlation matrix or covariance matrix, respectively. The default is `correlation`. Usually, TT based on the covariance matrix will be meaningful only if variables are expressed in the same units.

`force` try to force cross-validation even when zero-variance variables are detected in training samples. This is usually an indication that there is something wrong; use this option with caution.

3.4. *ttstab options*

`reps(#)` number of subsamples; default is `reps(100)`.

`subsample(#)` subsample size in percent of the original sample size; default is `subsample(80)`.

`keep(#)` keep sign patterns appearing in more than # percent of replications; default is `keep(20)`.

`force` tries to force subsampling even when zero-variance variables are found in subsamples. This is usually an indication that there is something wrong; use this option with caution.

3.5. *ttdendro options*

`dendro_options` are any of the options allowed by the `cluster dendrogram` command; see [MV] **cluster dendrogram**.

3.6. *ttscree and ttloading options*

`scatter_options` are any of the options allowed by the `graph twoway scatter` command; see [G] **graph twoway scatter**.

The following option applies to `ttscree` only:

`neigen` plot only the largest first # component variances; default is to plot all component variances

The following option applies to `ttloading` only:

`components` plot components in *numlist*; default is `components(1 2 3)`.

4. A data example

As a simple illustration of the proposed workflow when using the `tt` add-on, we consider the 1978 automobile toy data set which comes with Stata. This data set describes various characteristics of a total of 74 vehicles. We will use the 10 variables described below for the analysis; a total of 69 vehicles have complete observations for these variables.

```
. sysuse auto
(1978 Automobile Data)
. describe price-gear_ratio

              storage  display     value
variable name   type   format      label      variable label

-----------------------------------------------------------------------------
price           int    %8.0gc                 Price
mpg             int    %8.0g                  Mileage (mpg)
rep78           int    %8.0g                  Repair Record 1978
headroom        float  %6.1f                  Headroom (in.)
trunk           int    %8.0g                  Trunk space (cu. ft.)
weight          int    %8.0gc                 Weight (lbs.)
length          int    %8.0g                  Length (in.)
turn            int    %8.0g                  Turn Circle (ft.)
displacement    int    %8.0g                  Displacement (cu. in.)
gear_ratio      float  %6.2f                  Gear Ratio
```

4.1. *Step 1: running tt*

To get familiar with the data set, we first make a couple of preliminary runs of `tt` and the `tt_postestimation` plotting routines.

```
. tt price-gear_ratio, cor cut(3) components(3)

Treelet transform/correlation                    Number of obs    =        69
                                                 Number of comp.  =         3
                                                 Cut-level        =         3
    --------------------------------------------------------------------------
       Component |     Variance   Proportion   Cumulative    Adj. proportion
    -------------+------------------------------------------------------------
             TC1 |       3.6404       0.3640       0.3640             0.3640
             TC2 |       1.0000       0.1000       0.4640             0.0360
             TC3 |       1.0000       0.1000       0.5640             0.0746
             TC4 |       1.0000       0.1000       0.6640             0.0344
             TC5 |       1.0000       0.1000       0.7640             0.0787
             TC6 |       1.0000       0.1000       0.8640             0.0371
             TC7 |       1.0000       0.1000       0.9640             0.0652
             TC8 |       0.1875       0.0187       0.9828             0.0143
             TC9 |       0.1199       0.0120       0.9948             0.0086
            TC10 |       0.0522       0.0052       1.0000             0.0031
    --------------------------------------------------------------------------

Components

    -------------------------------------------
        Variable |      TC1        TC2        TC3
    -------------+-----------------------------
           price |
             mpg |
           rep78 |
        headroom |                         1.0000
           trunk |
          weight |   0.5080
          length |   0.5080
            turn |   0.4851
    displacement |   0.4985
      gear_ratio |              1.0000
    -------------------------------------------

. tt price-gear_ratio, cor cut(6) components(3)
```

— *Output continued on next page* —

```
Treelet transform/correlation                     Number of obs    =        69
                                                  Number of comp.  =         3
                                                  Cut-level        =         6
    --------------------------------------------------------------------------
        Component |     Variance   Proportion   Cumulative    Adj. proportion
    -------------+------------------------------------------------------------
              TC1 |       4.5497       0.4550       0.4550             0.4550
              TC2 |       1.6565       0.1657       0.6206             0.0432
              TC3 |       1.0000       0.1000       0.7206             0.0800
              TC4 |       1.0000       0.1000       0.8206             0.0717
              TC5 |       0.6353       0.0635       0.8842             0.0515
              TC6 |       0.4555       0.0455       0.9297             0.0328
              TC7 |       0.3435       0.0343       0.9640             0.0335
              TC8 |       0.1875       0.0187       0.9828             0.0143
              TC9 |       0.1199       0.0120       0.9948             0.0086
             TC10 |       0.0522       0.0052       1.0000             0.0031
    --------------------------------------------------------------------------

Components
    -------------------------------------------
         Variable |      TC1       TC2       TC3
    -------------+-----------------------------
            price |
              mpg |             0.7071
            rep78 |                       1.0000
         headroom |   0.3052
            trunk |   0.3639
           weight |   0.4471
           length |   0.4471
             turn |   0.4269
     displacement |   0.4387
       gear_ratio |             0.7071
    -------------------------------------------

. ttdendro
. ttscree
```

In both calls to `tt`, we retain 3 components but use different cut-levels 3 and 6, respectively. The relatively low cut-level of 3 in the first analysis yields more sparse components. In fact, components 2 and 3 in this first analysis are somewhat uninteresting for the purpose of dimension reduction since they contain only a single variable. The second analysis uses the cut-level 6 and leads to less sparse components.

The call to `tt` returns both the 'raw' variances explained by components and variances adjusted for correlation between scores using the conservative method of Gervini and Rousson (2004). For the present data, the first TT component explains the majority of the variation for both cut-levels 3 and 6, irrespective of the method used for variance calculation. In both analyses, this first component can be informally interpreted as measuring the overall 'size' of a vehicle.

The output of the call to `ttdendro` is shown in Figure 2. The TT cluster tree shows that `trunk`, `weight`, `length`, `displacement`, and `turn` form a tight cluster. With the addition of the variable `headroom`, it is this particular cluster that is reflected by the first TT component in the second call to `tt` above. It is a general feature of the TT algorithm that cluster membership in the cluster tree translates to nonzero loadings in some TT component. In other words, the cluster tree provides a concise visual representation of the possible TT components.

Figure 3 is obtained by calling `ttscree`. It is a graphical representation, similar to PCA scree plots, of the (unadjusted) variance explained by components. It is clear from this plot that a single component suffices to capture much of the variation in data.

The first TT component in the second call to `tt` above is very similar to the first component obtained from the corresponding PCA, as can be seen from the numerical loadings and Pearson correlation between scores calculated below. However, the first TT component is potentially simpler to interpret because of its sparsity.

```
. ttpredict tt1score
(9 components skipped)
. pca price-gear_ratio, cor components(2)

Principal components/correlation                  Number of obs    =        69
                                                  Number of comp.  =         2
                                                  Trace            =        10
    Rotation: (unrotated = principal)             Rho              =    0.7389

    --------------------------------------------------------------------------
       Component |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
           Comp1 |      6.31248      5.23618             0.6312       0.6312
           Comp2 |       1.0763     .0622654             0.1076       0.7389
           Comp3 |      1.01403      .583752             0.1014       0.8403
           Comp4 |      .430283     .0343745             0.0430       0.8833
           Comp5 |      .395908      .116712             0.0396       0.9229
           Comp6 |      .279196     .0229213             0.0279       0.9508
           Comp7 |      .256275      .130573             0.0256       0.9764
           Comp8 |      .125701     .0442338             0.0126       0.9890
           Comp9 |     .0814675     .0531123             0.0081       0.9972
          Comp10 |     .0283551            .             0.0028       1.0000
    --------------------------------------------------------------------------

Principal components (eigenvectors)

    -------------------------------------------------
        Variable |    Comp1     Comp2 | Unexplained
    -------------+--------------------+--------------
           price |   0.2074    0.3876 |       .5668
             mpg |  -0.3394    0.0520 |       .2699
           rep78 |  -0.1830    0.7639 |       .1606
        headroom |   0.2304    0.3049 |        .565
           trunk |   0.3003    0.3401 |       .3061
          weight |   0.3848    0.0095 |      .06535
          length |   0.3771    0.0432 |       .1003
            turn |   0.3542   -0.1831 |       .1719
    displacement |   0.3742   -0.0121 |       .1157
      gear_ratio |  -0.3306    0.1388 |       .2895
    -------------------------------------------------
. predict pc1score
. correlate tt1score pc1score

(obs=69)
             | tt1score pc1score
-------------+------------------
    tt1score |   1.0000
    pc1score |   0.9842   1.0000
```

4.2. *Step 2: running ttcv*

From the analysis in step 1, we found evidence that a single TT component suffices to describe the majority of variation in the data. It turns out that the optimal cut-level for a single-component solution is 9, the maximal possible; and that the single retained component has all nonzero loadings for this cut-level.

For the purpose of illustration, suppose instead that we decide to keep 3 components. We can then find a suitable cut-level by a call to `ttcv` as follows.

```
. ttcv price-gear_ratio, cor components(3)

 Cross-validation (10 folds, 5 repetitions)
0%            25%            50%            75%          100%
+--------------+--------------+--------------+--------------+
:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::

TT cross-validation/correlation        Number of obs   =      69
                                       Number of comp. =       3

                                       Number of folds =      10
                                       Number of reps  =       5
Cross-validation scores
```

— Output continued on next page —

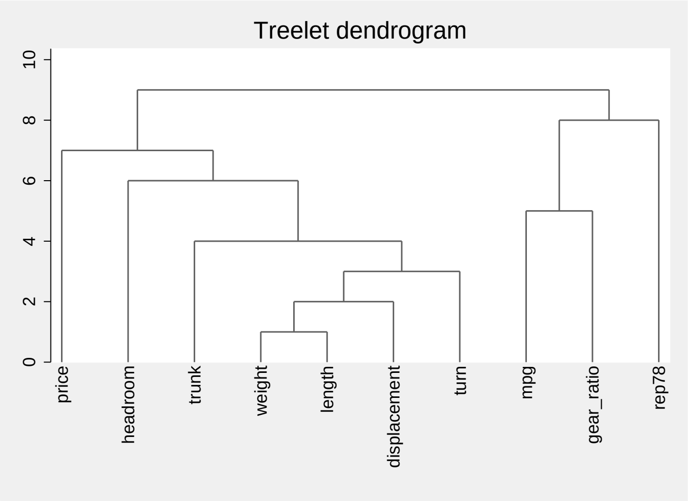

Figure 2. Cluster tree produced by `tt`.

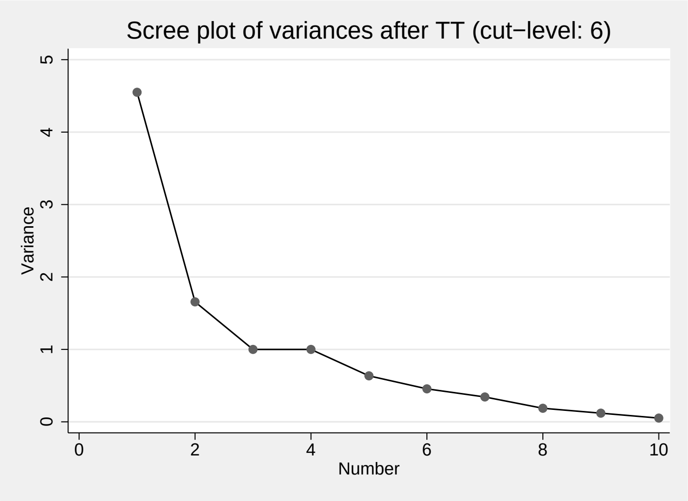

Figure 3. Scree plot of variances of TT component scores when the cut-level 6 is used.

```
-----------------------------------------
   Cut-level |       Score    Proportion
-------------+---------------------------
           1 |      5.3364       0.6509
           2 |      6.1585       0.7512
           3 |      6.9091       0.8428
           4 |      7.2181       0.8805
           5 |      7.5062       0.9156
           6 |      7.8466       0.9571
           7 |      7.7807       0.9491
           8 |      7.9603       0.9710
           9 |      8.1980       1.0000
-----------------------------------------
Estimated optimal cut-level = 6
(optimal cut-level sought within 10% of highest cut-level score)
```

Figure 4 shows a plot of the cross-validation scores generated when calling `ttcv`. Although not entirely convincing, a 'knee' in the graph seems to be located around level 6, indicating that increasing the cut-level beyond this level will not substantially improve the amount of variance explained by the 3 components. Thus, for a 3-component solution, a cut-level of 6 appears adequate.

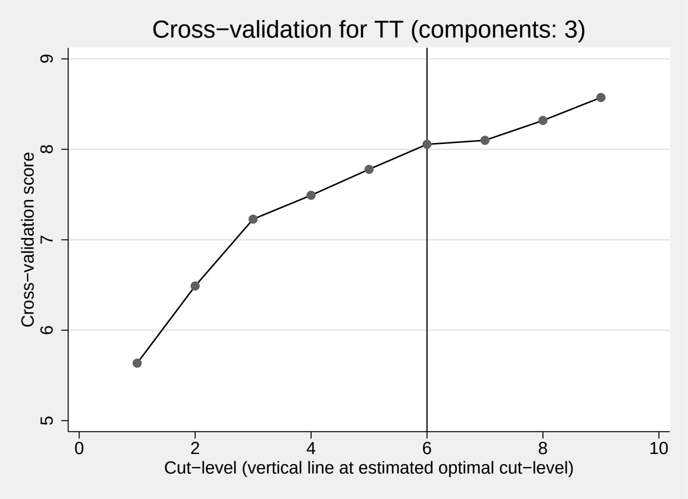

Figure 4. Graph of cross-validation scores for TT when 3 components are retained. The graph suggests that a 'knee' in the graph is located at cut-level 6.

Choosing simultaneously the number of components to retain *and* a cut-level is easy for the present data set since a single component-solution seems to be preferable at most nontrivial cut-levels. In situations where it is unclear how many components to retain, the choice can be more difficult. The following strategy is recommended:

- Decide on a range of different sensible values of `components()` in the call to `tt` via, for example, investigation of scree plots.
- Perform `ttcv` for each of these choices of `components()`.

In our experience, there will often be a reasonably small range of cut-levels that are universally preferable for the selected range of `components()`. A parsimonious solution is then to use the smallest acceptable cut-level among these.

4.3. *Step 3: running ttstab*

For the choice `cut(6)` and `components(3)` in the call to `tt`, we conclude our analysis by investigating the stability of the obtained solution via a call to `ttstab`.

```
. tt price-gear_ratio, cor cut(6) components(3)
  (output omitted)
. ttstab

Subsampling repetitions (100)
0%            25%           50%            75%          100%
+-------------+-------------+-------------+-------------+
:::::::::::::::::::::::::::::::::::::::::::::::::::::::::


Stability of TT/correlation                      Number of obs    =         69
                                                 Number of comp.  =          3
                                                 Cut-level        =          6

                                                 Subsample rep.   =        100
                                                 Subsample frac.  =       0.80
                                                 Subsample size   =         55

Average rank (by amount of variance explained) and frequency of sign patterns
Displaying results for patterns with frequency >=  10%

    ---------------------------------------------------------
    Sign pattern |    Avg. rank   Frequency   Avg. variance
    -------------+-------------------------------------------
               1 |        1.000       0.890           4.552
               2 |        2.000       0.990           1.656
               3 |        3.000       0.350           1.000
               4 |        3.000       0.610           1.000
    ---------------------------------------------------------

Structure of sign patterns

    ------------------------------------
        Variable |    1    2    3    4
    -------------+----------------------
           price |    0    0    +    0
             mpg |    0    +    0    0
           rep78 |    0    0    0    +
        headroom |    +    0    0    0
           trunk |    +    0    0    0
          weight |    +    0    0    0
          length |    +    0    0    0
            turn |    +    0    0    0
    displacement |    +    0    0    0
      gear_ratio |    0    +    0    0
    ------------------------------------
```

The call to `ttstab` performs 100 subsampling repetitions of TT, keeping the 3 highest-variance components in each subsampled analysis (at cut-level 6). It then transforms these into their corresponding sign patterns. Note that `ttstab` is set to return all sign patterns seen in more than 10% of the subsampling repetitions, here corresponding to 4 sign patterns. In the output, 'Avg. rank' is the the rank (according to explained variance of the corresponding component), averaged over the 100 subsamples. 'Frequency' is the relative frequency of the sign pattern among all $3 \cdot 100$ sign patterns returned. Lastly, 'Avg. variance' is the variance explained by the component corresponding to the sign pattern, averaged over the 100 subsamples.

We can see that sign patterns similar to those of the first two components from the original TT analysis with `components(3)` and `cut(6)` appear in almost all subsampling repetitions. If the first type of sign pattern appears, it corresponds to a component with rank 1. Moreover, the first component remains by far the most important one in terms of variance explained. Sign patterns 3 and 4, on the other hand, do not appear to be very stable. Increasing the number of retained components to 4 does lead to a greater stability in terms of frequency of inclusion but does not improve stability of the rank of the last two components.

5. Concluding remarks

The treelet transform can be viewed as an amalgamation of PCA and cluster analysis. It leads to components that are sparse and can be easier to interpret than their PCA counterparts. We have described the `tt` add-on for Stata which contains all the basic functionality needed to apply the treelet transform in practice, including an Mata implementation of the treelet transform algorithm. For a more advanced application example and a detailed comparison with the output produced by PCA, we refer to the paper by Gorst-Rasmussen *et al.* (2011).

Acknowledgements

I thank Søren Lundbye-Christensen and Christina C. Dahm for their helpful comments and suggestions when preparing this manuscript.

References

Ben-Hur, A., Elisseeff, A. A. and Guyon, I. (2002) A stability based method for discovering structure in clustered data. *Pacicific Symposium on Biocomputing*, 6–17.

Gervini, D. and Rousson, V. (2004) Criteria for evaluating dimension-reducing components for multivariate data. *American Statistician*, **58**, 72–76.

Gorst-Rasmussen, A. (2011) *tt – Stata add-on for performing treelet transformation*. URL `http://people.math.aau.dk/~gorst/software.htm`.

Gorst-Rasmussen, A., Dahm, C. C., Dethlefsen, C., Scheike, T. and Overvad, K. (2011) Exploring dietary patterns by using the treelet transform. *American Journal of Epidemiology*, **173**, 1097–1104.

Jolliffe, I. T. (2002) *Principal Components Analysis*. New York: Springer, second edn.

Lee, A. B., Nadler, B. and Wasserman, L. (2008) Treelets – an adaptive multi-scale basis for sparse unordered data. *Annals of Applied Statistics*, **2**, 435–471.

Liu, D. (2010) *treelet: Treelet*. URL `http://cran.r-project.org/package=treelet`. R-package version 0.2-0.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**, 265–286.

Paper VI

Coordinate Descent Methods for the Penalized Semiparametric Additive Hazards Model

Author list

Anders Gorst-Rasmussen
Aalborg University, Denmark

Thomas H. Scheike
University of Copenhagen, Denmark

Summary

For survival data with a large number of explanatory variables, lasso penalized Cox regression is a popular regularization strategy. However, a penalized Cox model may not always provide the best fit to data and can be difficult to estimate in high dimension because of its intrinsic nonlinearity. The semiparametric additive hazards model is a flexible alternative which is a natural survival analogue of the standard linear regression model. Building on this analogy, we develop a cyclic coordinate descent algorithm for fitting the lasso and elastic net penalized additive hazards model. The algorithm requires no nonlinear optimization steps and offers excellent performance and stability. An implementation is available in the R-package **ahaz** and we demonstrate this package in a small timing study and in an application to real data.

Supplementary info

This manuscript has been published as:

> Gorst-Rasmussen A, Scheike TH (2011). Coordinate Descent Methods for the Penalized Semiparametric Additive Hazards Model. *Technical report R-2011-10.* Department of Mathematical Sciences, Aalborg University.

It has moreover been submitted to *Journal of Statistical Software*.

1. Introduction

With the increasing interest in high-throughput biomarker research, there is a growing need for simple and efficient statistical methods for relating a survival time endpoint to a large number of explanatory variables. Variable selection methods such as lasso (Tibshirani, 1997) or SCAD (Fan and Li, 2001) offer convenient means of imposing additional regularity via penalization such that well-known regression models can be straightforwardly adapted to high-dimensional data. By now, many standard survival regression models have been subjected to various penalization strategies (Li, 2008), yet the Cox proportional hazards model continues to serve as a reference model and the main target of theoretical, algorithmic, and applied research on penalized survival regression. Although the Cox model is both flexible and simple to interpret, alternative modeling strategies deserve a wider appreciation for a number of reasons. For example, Ma *et al.* (2010) recently pointed out a fact which is well known from a lower-dimensional setting: that a Cox model may not always provide a satisfactory fit to a high-dimensional data set. Moreover, with the increasingly high-dimensional data available today, the intrinsically nonlinear Cox model is a peculiar reference model in terms of the computational efficiency and stability of fitting procedures. A range of algorithms have been developed for fitting penalized Cox models (Gui and Li, 2005; Park and Hastie, 2007; Sohn *et al.*, 2009; Goeman, 2010, and others) but their computational performance is limited by the use of costly Newton-Raphson iterations or similar to deal with the penalized partial likelihood.

Only recently did Simon *et al.* (2011) describe an impressively fast algorithm for fitting the penalized Cox model which combines iteratively reweighted least squares with cyclic coordinate descent. Cyclic coordinate descent optimizes a convex loss function by solving all coordinatewise optimization problems in an iterative manner. While not a new technique in the context of penalized regression (see the references in Friedman *et al.* (2010)), cyclic coordinate descent has recently been rediscovered for its ability to efficiently handle even very high-dimensional problems when carefully implemented. For generalized linear models and the Cox model, software for performing coordinate descent-based penalized estimation is available in the R-package **glmnet** (Friedman *et al.*, 2010).

In this paper, we develop a cyclic coordinate descent algorithm for the elastic net penalized variant of a flexible but less well-known alternative to the Cox model, the so-called semiparametric additive hazards model (Lin and Ying, 1994; McKeague and Sasieni, 1994). This model asserts a hazard function given by the sum of some baseline hazard function and a regression function of the explanatory variables. It is a survival analogue of the standard linear regression model and leads to natural estimating equations which are surprisingly similar to the normal equations. The flexibility and computational parsimony of the additive hazards model makes it a useful tool on which to base regularization methods for high-dimensional survival data (Ma *et al.*, 2006; Leng and Ma, 2007; Martinussen and Scheike, 2009, 2010). We describe how computational tricks for the penalized linear regression model can be adapted to obtain a very efficient and stable coordinate descent method for fitting the elastic net penalized additive hazards model. In contrast to coordinate descent methods for the penalized Cox model, convergence is theoretically guaranteed for our algorithm. The algorithm has been implemented in C to interface with the R-package **ahaz** (Gorst-Rasmussen, 2011), and we provide examples of its usage and performance on simulated and real data.

2. The semiparametric additive hazards model

Suppose that we observe $(T_1,\delta_1,Z_1),\dots,(T_n,\delta_n,Z_n)$ where T_i is a (right-censored) survival time, δ_i is the indicator which is 1 if subject i experiences an event at time T_i and 0 otherwise, and $Z_i \in \mathbb{R}^p$ is a vector of explanatory variables. To simplify notation, we will describe each pair (T_i,δ_i) via the counting process $N_i(t) := I(T_i \leqslant t)\delta_i$ and the at-risk-process $Y_i(t) := I(t \leqslant T_i)$ where I denotes the indicator function. The counting process integral $\int_0^t f(s)\mathrm{d}N_i(s)$ is then simply a notationally convenient way of writing $f(T_i)I(T_i \leqslant t)\delta_i$.

The semiparametric additive hazards model (Lin and Ying, 1994; McKeague and Sasieni, 1994) asserts a conditional hazard function of the form

$$\lambda(t|Z_i) = \lambda_0(t) + Z_i^\top \beta^0;$$

with λ_0 some unspecified baseline hazard constituting the nonparametric part of the model. Lin and Ying (1994) proposed to perform estimation in this model via estimating equations which mimic the score equations for the Cox model. Specifically, they proposed to estimate β^0 as the root of the pseudo-score function

$$U(\beta) := \int_0^\infty \sum_{i=1}^n Z_i\{\mathrm{d}N_i(t) - Y_i(t)\mathrm{d}\hat{\Lambda}_0(t;\beta) - Y_i(t)Z_i^\top \beta \mathrm{d}t\}, \tag{1}$$

where $\hat{\Lambda}_0$ is a Breslow-type estimator of the cumulative baseline hazard $\int_0^t \lambda_0(s)\mathrm{d}s$,

$$\hat{\Lambda}_0(t;\beta) := \int_0^t \frac{\sum_{i=1}^n \{\mathrm{d}N_i(s) - Y_i(s)Z_i^\top \beta \mathrm{d}s\}}{\sum_{i=1}^n Y_i(s)}.$$

Solving $U(\beta) = 0$ is equivalent to solving the $p \times p$ linear system of equations

$$D\beta = d, \tag{2}$$

taking

$$D := \int_0^\infty \sum_{i=1}^n \{Z_i - \bar{Z}(t)\}\{Z_i - \bar{Z}(t)\}^\top Y_i(t)\mathrm{d}t \tag{3}$$

$$d := \int_0^\infty \sum_{i=1}^n \{Z_i - \bar{Z}(t)\}\mathrm{d}N_i(t); \tag{4}$$

with $\bar{Z}(t) := \sum_{i=1}^n Z_i Y_i(t) / \sum_{i=1}^n Y_i(t)$ the at-risk-average of the Z_is. The estimator obtained from (2) can be shown root-n consistent by martingale arguments (Lin and Ying, 1994).

The estimating equation (2) is attractive for several reasons. Not only does it provide an explicitly calculable estimator in a flexible semiparametric model; it is also analytically very similar to the normal equations $(X^\top X)\beta = X^\top y$ for the classical linear regression model $y = X\beta^0 + \varepsilon$. In fact, defining 'responses' $y_i = \mathrm{d}N_i(t)$ and 'explanatory variables' $X_i = (Z_i - \bar{Z}(t))Y_i(t)$, it is seen that (2) is simply a time-averaged version of the normal equations. The similarity between (2) and the normal equations was exploited by Leng and Ma (2007) and Martinussen and Scheike (2009) to construct a lasso penalized estimator for the additive hazards model. They noted that solving (2) is equivalent to minimizing the loss function

$$L(\beta) = \beta^\top D\beta - 2\beta^\top d; \tag{5}$$

leading to a lasso penalized variant with a loss function of the form

$$L_{\mathrm{pen}}(\beta;\lambda) = L(\beta) + \lambda\|\beta\|_1. \tag{6}$$

Here $\|\cdot\|_1$ is the ℓ^1-norm while $\lambda \geqslant 0$ is a parameter controlling the degree of regularization. Because of geometric properties of the ℓ^1-norm, the lasso penalized estimator $\mathrm{argmin}_\beta L_{\mathrm{pen}}(\beta;\lambda)$ does shrinkage and variable selection simultaneously (Tibshirani, 1997). For large values of λ most lasso regression coefficients will be exactly zero – as λ grows smaller, the lasso regression coefficients become increasingly similar to their unpenalized counterparts.

Leng and Ma (2007) and Martinussen and Scheike (2009) proposed to use the lasso-LARS algorithm (Efron *et al.*, 2004) to calculate the lasso penalized estimator $\mathrm{argmin}_\beta L_{\mathrm{pen}}(\beta;\lambda)$. The lasso-LARS algorithm for the standard linear regression model is easily adapted to work with the additive hazards model by supplying pre-computed versions of D (in place of the covariance matrix) and d (in place of the covariate-response inner products). However, pre-computation of D may be unfeasible for large p. Even without pre-computation, the computational cost of lasso-LARS is similar to that of solving the unpenalized regression problem which is substantial for large p. In the following, we propose cyclic coordinate descent as a much more efficient alternative.

3. Model fitting via cyclic coordinate descent

Since the extension is straightforward, we will work with a variant of (6) which includes an ℓ^2-penalty term. That is, we consider the problem of minimizing the following penalized loss:

$$L_{\text{pen}}(\beta;\lambda,\alpha) := L(\beta) + \lambda\alpha\|\beta\|_1 + \frac{1}{2}\lambda(1-\alpha)\|\beta\|_2^2, \quad 0 < \alpha < 1. \tag{7}$$

We denote henceforth

$$\hat{\beta}(\lambda) := \operatorname{argmin}_{\beta} L_{\text{pen}}(\beta;\lambda,\alpha). \tag{8}$$

The ℓ^1/ℓ^2-penalization in (7) is known as elastic net penalization (Zou and Hastie, 2005). When $\alpha = 1$, the loss function reduces to lasso penalized loss. If $\alpha < 1$, the loss function favors joint selection of highly correlated variables. This follows by similar arguments as in Zou and Hastie (2005), utilizing the heuristic interpretation of D as a time-averaged covariance matrix. We have omitted the dependence of α in the left-hand side of (8) for notational simplicity.

Cyclic coordinate descent is a numerical optimization technique which approximates the minimum of a function $f : \mathbb{R}^p \to \mathbb{R}$ by iteratively for $k = 0, 1, 2, \ldots$ cycling through the p coordinatewise optimization problems

$$x_j^{(k)} := \operatorname{argmin}_{x_j} f\big(x_1^{(k)}, \ldots, x_{j-1}^{(k)}, x_j, x_{j+1}^{(k-1)}, \ldots, x_p^{(k-1)}\big), \quad j = 1, 2, \ldots, p; \tag{9}$$

fixing for the update of the jth coordinate all other coordinates at their most recent value. For a convex f satisfying certain separability conditions, the iterates $x^{(k)}$ converge to $\operatorname{argmin}_{x \in \mathbb{R}^p} f(x)$, irrespective of $x^{(0)}$ (Tseng, 1988). It suffices that f is a convex and continuously differentiable function subjected to elastic net penalization.

To use cyclic coordinate descent to calculate (8), simply observe that

$$\frac{\partial L_{\text{pen}}}{\partial \beta_j} = d_j - \sum_{i \neq j} \beta_i D_{ij} + \lambda\alpha \operatorname{sign}(\beta_j) + \lambda(1-\alpha)\beta_j.$$

It follows that the updating rule (9), for a given value of (λ, α), takes on the form

$$\beta_j^{(k)} := \frac{\mathscr{S}\Big(d_j - \sum_{i<j} \beta_i^{(k)} D_{ij} + \sum_{i>j} \beta_i^{(k-1)} D_{ij}, \lambda\alpha\Big)}{D_{jj} + \lambda(1-\alpha)}, \qquad j = 1, 2, \ldots, p;$$

where $\mathscr{S}$ denotes the soft-thresholding operator

$$\mathscr{S}(x, y) := \operatorname{sign}(x)(|x| - y)_+.$$

While convexity ensures theoretically that $\beta^{(k)}$ converges to $\hat{\beta}(\lambda)$, convergence can be very slow if $\beta^{(0)}$ is poorly chosen. Fundamental to ensuring rapid convergence and stability of coordinate descent are the following two structural properties of the elastic net problem:

1. If λ is sufficiently large then $\hat{\beta}(\lambda) = 0$ (sparsity).
2. If $\lambda_1 \approx \lambda_2$ then $\hat{\beta}(\lambda_1) \approx \hat{\beta}(\lambda_2)$ (continuity of regularization paths; Efron *et al.* (2004)).

Hence, $\hat{\beta}(\tilde{\lambda})$ for some $\tilde{\lambda}$ can be calculated efficiently and stably via coordinate descent by calculating a pointwise regularization path $\hat{\beta}(\lambda_{\max}),\ldots,\hat{\beta}(\tilde{\lambda})$ at a grid of closely spaced λ-values; starting out with some large $\lambda_{\max}$ so that $\hat{\beta}(\lambda_{\max}) = 0$ and using the most recent solution $\hat{\beta}(\lambda_{l-1})$ as the initial value in the coordinate descent algorithm for $\hat{\beta}(\lambda_l)$. This idea of using 'warm starts' was discussed in more detail by Friedman *et al.* (2007) and Friedman *et al.* (2010). For the penalized loss (7), it is easily seen that we obtain $\hat{\beta}(\lambda_{\max}) \equiv 0$ by taking

$$\lambda_{\max} := \max_{1\leqslant j\leqslant p} |d_j|.$$

Following Simon *et al.* (2011), we consider an exponentially decreasing sequence of regularization parameters of length m from $\lambda_{\max}$ to some $\lambda_{\min} < \lambda_{\max}$ such that

$$\lambda_l := \lambda_{\max}\left(\frac{\lambda_{\min}}{\lambda_{\max}}\right)^{l/m}, \quad l = 0,\ldots,m-1. \tag{10}$$

If we denote $\varepsilon := \lambda_{\min}/\lambda_{\max}$, a reasonable, although arbitrary, choice is to take $m = 100$ and $\varepsilon = 0.0001$ if $n < p$ and $\varepsilon = 0.05$ if $p \geqslant n$.

Naively, one would run the coordinate descent algorithm over all p coordinates (i.e. using all p variables) to obtain $\hat{\beta}(\lambda_0),\ldots,\hat{\beta}(\lambda_m)$. This is clearly undesirable for large p since it requires calculation of the entire matrix D. However, given $\hat{\beta}(\lambda)$ for some λ, the Karush-Kuhn-Tucker (KKT) conditions for the constrained optimization problem (7) imply that $\hat{\beta}_j(\lambda) = 0$ iff

$$\left| d_j - \sum_{i\neq j} \hat{\beta}_i(\lambda) D_{ij} \right| \leqslant \lambda\alpha. \tag{11}$$

This leads to the active set strategy (Friedman *et al.*, 2007): we maintain at all times a set A of 'active variables' which are included in the coordinate descent algorithm, starting out with $A := \emptyset$ at $\lambda_{\max}$. Upon convergence of coordinate descent among variables in A, we check (11) for each variable in $\{1,\ldots,p\}\backslash A$. If there are no violations, we have the final solution. If there are violations, we add the violators to A and restart the coordinate descent algorithm. With this approach, it is seen from (11) that we need only calculate rows $D_{j\bullet}$ for $j \in A$. A basic coordinate descent algorithm for the additive hazards model is thus the following.

Initialize $A := \emptyset$, $\lambda_{\max} := \max_{1\leqslant j\leqslant p} |d_j|$, and $\beta^{(0)}(\lambda_{\max}) := 0$. For $l = 0,\ldots,m-1$ do

1. Set $\lambda_l := \lambda_{\max}\varepsilon^{l/m}$. Do for $k = 0,1,\ldots,$ until convergence
    (a) For $j \in A$, update
    $$\beta_j^{(k)}(\lambda_l) := \frac{\mathscr{S}\left(d_j - \sum_{i\in A, i<j} \beta_i^{(k)}(\lambda_l) D_{ij} - \sum_{i\in A, i>j} \beta_i^{(k-1)}(\lambda_l) D_{ij}, \lambda_l\alpha\right)}{D_{jj} + \lambda_l(1-\alpha)}. \tag{12}$$
2. Set $\tilde{\beta} := \beta^{(k)}(\lambda_l)$ and for $j \in \{1,\ldots,p\}\backslash A$, calculate
    $$V := \left\{ j \notin A : \left| d_j - \sum_{i\in A} \tilde{\beta}_i D_{ij} \right| > \lambda_l\alpha \right\}.$$
    If $V \neq \emptyset$, calculate $D_{j1},\ldots,D_{jp}$ for $j \in A$, then adjoin V to A and go back to step 1, using $\tilde{\beta}$ as a warm start. Otherwise set $\hat{\beta}(\lambda_l) := \tilde{\beta}$ and $\beta^{(0)}(\lambda_{l+1}) := \tilde{\beta}$, and increment l.

Various stopping criteria can be used in step 1; either based on the relative change in the individual coefficient estimates or based on the relative change in the penalized loss function. We prefer the latter since the loss function is less susceptible to instabilities when many variable are included or when near a saturated fit. Specifically, we declare convergence when the relative change in $L_{\text{pen}}\{\beta^{(k)}(\lambda)\}$ from one value of k to the next is less than 10^{-5}.

Note that for $\alpha = 1$, at most $n-1$ variables can be included in the model, by the nature of the lasso penalized problem. In most cases, the user will specify some maximum number of variables to include which is strictly less than n.

3.1. *Efficient calculation of D*

The calculation of rows in the matrix D is the primary bottleneck of our basic coordinate descent algorithm for p large. In contrast to the partial likelihood in the Cox model, which essentially only depends on data at failure times, calculation of D uses data at both censoring and failure times. Fortunately, it turns out that (3) can still be evaluated rather efficiently.

Suppose that survival times are ordered such that $T_1 > T_2 > \cdots > T_n$, assuming no ties. Denote $\Delta_k := T_k - T_{k+1}$ (taking $T_{n+1} := 0$) and assume that variables are centered so that $\sum_{i=1}^n Z_i = 0$. By applying the summation by parts formula, we obtain

$$\begin{aligned} D_{ij} &= \sum_{k=1}^n Z_{jk}\left(Z_{ik}\int_0^\infty Y_k(t)\mathrm{d}t\right) - \int_0^\infty \bar{Z}_i(t)\sum_{k=1}^n Z_{jk}Y_k(t)\mathrm{d}t \\ &= \sum_{k=1}^n Z_{jk}(Z_{ik}T_k) + \sum_{k=1}^n \left(\Delta_k k^{-1}\sum_{h=1}^k Z_{ih}\right)\left(\sum_{h=1}^k Z_{jh}\right) \\ &= \sum_{k=1}^n Z_{jk}(Z_{ik}T_k) + \sum_{k=1}^{n-1}\left(\sum_{l=1}^k \Delta_l l^{-1}\sum_{m=1}^l Z_{im}\right)Z_{j,k+1} \\ &= \sum_{k=1}^n Z_{j,k}\tilde{Z}_{ik}; \end{aligned}$$

where we have defined

$$\tilde{Z}_{i1} := Z_{i1}T_1, \qquad \text{and } \tilde{Z}_{ik} := Z_{j,k}T_k + \sum_{l=1}^{k-1}\Delta_l l^{-1}\sum_{m=1}^l Z_{im}, \quad 2 \leqslant k \leqslant n.$$

Hence, if we pre-calculate and store $\tilde{Z}_{i1},\ldots,\tilde{Z}_{in}$, the subsequent calculation of each matrix element D_{ij} can be accomplished at the modest cost of $2n$ arithmetic operations.

3.2. *Increasing efficiency via improved KKT checks*

While our basic coordinate descent algorithm is already quite efficient, there is room for improvement. Denote by $\tilde{p}$ the size of the active set A at $\lambda_{\min}$. In retrospect, we need only $\tilde{p}^2$ entries in the matrix D to construct a regularization path; the remaining $(p-\tilde{p})\cdot\tilde{p}$ entries are used only for the KKT checks (11). A substantially more efficient KKT check can be devised by noting from (3) that

$$\sum_{i=1}^p D_{ij}\hat{\beta}_i(\lambda) = \sum_{i=1}^n Z_{ji}r_i(\lambda); \tag{13}$$

where

$$r_i(\lambda) := \int_0^\infty Y_i(t)\{R_i^\lambda - \bar{R}^\lambda(t)\}\mathrm{d}t, \tag{14}$$

taking $R_i^\lambda := Z_i^\top \hat{\beta}(\lambda)$ to be the linear risk score of the ith subject. Formulas as in Section 3.1 can be used for evaluating (14). Substituting (13) in (11), it follows that we can perform the necessary KKT checks by calculating the n-vector $r(\lambda)$ and subsequently evaluating $p - |A|$ inner products between n-vectors. Whenever a new variable j enters the model, symmetry of the matrix D implies that we need only calculate D_{ij} for $i \in A$ to be able to run the coordinate descent updates (12). This is a substantial improvement over the basic coordinate descent algorithm in which the entire row $D_{j\bullet}$ must be calculated for each new variable j.

An issue not addressed by this improved strategy is that KKT checks often fail. In fact, they fail at least whenever a new variable enters the model and in practice much more frequently. A failed check leads to a restart of the coordinate descent loop. Although another run of coordinate descent is rarely very expensive when using warm starts, calculating $p - |A|$ inner products between n-vectors for the next KKT check is costly. The cost could be reduced if we could first run the coordinate descent/check/restart procedure on a set of variables larger than the active set but still smaller than p; and outside which KKT violations are rare. Tibshirani *et al.* (2010) recently showed how to construct such a set. Adapting their formulas to the present problem, given some $\gamma > \lambda$, they proposed the following sequential strong condition

$$|d_j - Z_j^\top r(\gamma)| \leqslant \lambda - (\gamma - \lambda) = 2\lambda - \gamma; \tag{15}$$

and argued that if a variable j satisfies this condition then typically $\hat{\beta}_j(\lambda) = 0$. The sequential strong condition is not failsafe and (15) may hold true if $\hat{\beta}_j(\lambda) \neq 0$. The point is that this happens rarely. Consequently, by introducing the strong set

$$S := \left\{ j \notin A : |Z_j^\top r(\gamma)| > 2\lambda - \gamma \right\} \cup A,$$

we may further improve efficiency of coordinate descent via the following strategy for each λ:

1. Run coordinate descent for variables in A until convergence.
2. Check for violations of KKT conditions among variables in S only, using (13). If violations occur, add violators to A and go back to step 1, using the current solution as a warm start. Otherwise proceed to step 3.
3. Check for violation of KKT conditions among variables in $\{1,\ldots,p\}\backslash S$ using (13). If violations occur, add violators to A, update S, and go to step 1, using the current solution as a warm start. Otherwise we have the solution for this value of λ.

This strategy is an improvement since we tend to restart the algorithm fewer times in step 3. Accordingly, fewer inner products must be calculated. Other approximate discarding rules than (15) could be used instead since we always conclude by running a fail-safe check of KKT conditions among all variables.

3.3. *Implementation in ahaz*

The optimized version of the algorithm described in this section has been implemented in C to interface with the R-package **ahaz** (Gorst-Rasmussen, 2011) via the wrapper function `ahazpen`. Since all calculations are done in C, the code can easily be adapted to work with other front-ends than R.

The bottleneck of the algorithm is calculating the roughly p inner products between n-vectors. This can account for 50%-90% of the computation time. As also noted by Tibshirani *et al.* (2010), simultaneous inner product evaluations are embarrassingly parallel, suggesting good scalability of the algorithm. We have implemented the inner product evaluations via level 2 calls to the **BLAS** libraries linked to R, thus enabling the user to improve speed of `ahazpen` further by linking R against high-performance **BLAS** libraries such as **GotoBLAS** or **ATLAS**.

4. Additional details

The implementation of cyclic coordinate descent provided in `ahazpen` supports a similar set of options as **glmnet** (Friedman *et al.*, 2010); including observation weighting and differential penalization. Specifically, for nonnegative weights $w_1,\ldots,w_p$, `ahazpen` can accommodate a penalized loss function of the form

$$L(\beta)+\lambda\sum_{j=1}^{p} w_j|\beta_j|.$$

In the simplest case, differential penalization can be used to completely exclude a variable from penalization (by setting $w_j := 0$), offering a simple alternative to the more sophisticated approach of unpenalized adjustment discussed by Martinussen and Scheike (2009). Differential penalization also enables implementation of techniques such as adaptive lasso (Zou, 2006).

4.1. *Delayed entry*

An approach which is common in, for example, survival epidemiological studies is adjust for the age of study subjects by using it as a time axis in hazard regression models. This is popularly known as delayed entry (or left-truncation) and requires us to consider data of the form $(S_1,T_1,\delta_1),\ldots,(S_n,T_n,\delta_n)$ where $0 \leqslant S_i < T_i$ is the entry time of the ith individual. By keeping $N_i(t) = I(T_i \leqslant t \wedge \delta_i = 1)$ but setting $Y_i(t) = I(S_i \leqslant t \leqslant T_i)$, the regression models described in Section 2 extend straightforwardly to the delayed entry case.

Computer implementation of delayed entry is slightly more involved. Define for $i = 1,2,\ldots,n$ the following collection of ‘pseudo observations’:

$$\begin{aligned} Y_i^*(t) &:= I(0 \leqslant t \leqslant T_i), & Y_{i+n}^*(t) &:= -I(0 \leqslant t < S_i);\\ N_i^*(t) &:= N_i(t), & N_{i+n}^*(t) &:= 0;\\ Z_i^* &:= Z_i, & Z_{i+n}^* &:= Z_i. \end{aligned}$$

Since $Y_i(t) = Y_i^*(t) + Y_{i+n}^*(t)$, it follows that

$$D = \int_0^\infty \sum_{i=1}^{2n}\{Z_i^* - \bar{Z}^*(t)\}\{Z_i^* - \bar{Z}^*(t)\}^\top Y_i^*(t)\mathrm{d}t, \qquad \text{and } d = \int_0^\infty \sum_{i=1}^{2n}\{Z_i^* - \bar{Z}^*(t)\}\mathrm{d}N_i^*(t).$$

Hence we can deal with delayed entry by replacing the original n observations with $2n$ pseudo observations and using the algorithms developed for the case where $S_1 = S_2 = \cdots = S_n = 0$.

The support for delayed entry is not only useful for implementing nonstandard time axes. It can also be used to implement (piecewise constant) time-varying explanatory variables, as well as to implement observations from more general counting processes.

4.2. *Tuning parameter selection*

A complete lasso or elastic net regularization path is useful mainly for judging relative importance of variables. In practice, the experimenter typically seeks the solution for a single value of the regularization parameter λ which can then be used similarly to how a set of unpenalized regression coefficients would be used. To select a 'representative' value of λ, cross-validation is commonly employed. As argued in Martinussen and Scheike (2009), the loss function (5) for the additive hazards model can be interpreted as a 'prediction error' within a quite general setting. It follows that if $F_1,\dots,F_K$ is a partition of $\{1,\dots,n\}$, each F_i being roughly the same size, we may define a cross-validation score

$$\mathrm{CV}(\lambda_l) := \sum_{i=1}^{K} L^{(F_i)}\{\hat\beta^{(-F_i)}(\lambda_l)\}, \quad l = 0,1,\dots,m;$$

with $L^{(F_i)}$ the loss calculated using observations from F_i only, and $\hat\beta^{(-F_i)}(\lambda_l)$ the penalized regression coefficients based on observations in $\{1,\dots,n\}\backslash F_i$ only. We then select $\hat\lambda := \mathrm{argmin}_{\lambda_l}\mathrm{CV}(\lambda_l)$ as the optimal λ-value.

In **ahaz**, 5-fold cross-validation ($K = 5$) is the default and offers an acceptable compromise between accuracy and stability in moderately sized data sets. In small data sets cross-validation can be somewhat unstable. For this reason, **ahaz** also supports repeated cross-validation where $\mathrm{CV}(\lambda)$ is averaged over several independent splits of $\{1,\dots,n\}$ into folds.

It is also possible select λ via criteria similar to BIC (or AIC). Although the loss (5) is not based on a likelihood, we may still define the following analogue to BIC,

$$\mathrm{PBIC}(\lambda) := \kappa L(\beta) + \mathrm{df}\{\hat\beta(\lambda)\} f(n); \tag{16}$$

where κ is some scaling constant. A convenient estimate of $\mathrm{df}\{\hat\beta(\lambda)\}$ is $\|\hat\beta(\lambda)\|_0$, the number of nonzero variables in $\hat\beta(\lambda)$ (Zou *et al.*, 2007). Because the loss function L is of the least-squares type, the arguments of Wang and Leng (2007) can be used to show that for p fixed, if $n^{-1}f(n) \to 0$ and $f(n) \to \infty$, the choice $\hat\lambda := \mathrm{argmin}_\lambda \mathrm{PBIC}(\lambda)$ entails certain selection consistency properties, depending on the underlying penalization method. For example, we can take $f(n) := \log n$ (Gorst-Rasmussen and Scheike, 2011). In **ahaz**, we use

$$\kappa := \frac{d_{\mathscr{A}}^\top B_{\mathscr{A}}^- d_{\mathscr{A}}}{d_{\mathscr{A}}^\top D_{\mathscr{A}}^- d_{\mathscr{A}}};$$

where all quantities are calculated within the set $\mathscr{A}$ of nonzero variables at the smallest value of λ used, B is an estimator of the asymptotic covariance matrix of d, and X^- denotes the Moore-Penrose inverse. This choice of κ ensures that PBIC scales like a true BIC. Observe that (16), since it depends on the end point of the regularization path (through κ), is a sensible selection criterion primarily when $p < n$.

5. Timings and a data example

This section presents timing results for `ahazpen`, alongside an example of its usage on a real data set. We keep our timing study brief since previous work for other statistical models present a strong case that well-designed coordinate descent algorithms are universally faster than competing lasso fitting methods (Friedman *et al.*, 2010; Simon *et al.*, 2011),

5.1. *Timings*

Simon *et al.* (2011) used simulated data from a basic accelerated failure time model to assess runtimes of coordinate descent methods for the penalized Cox model. We adopt their simulation model for our runtime assessments and consider explanatory variables Z_i which are independent and identically distributed marginally standard Gaussian p-vectors satisfying $\mathrm{Cor}(Z_{1j}, Z_{1k}) = \rho$ for $j \neq k$. True survival times are generated conditionally on the Z_is as

$$\tilde{T}_i := \exp\left(\sum_{j=1}^{p} Z_{ij}\beta_j + W_i\right), \quad i = 1, 2, \ldots, n$$

where $\beta_j := (-1)^j \exp(-2(j-1)/20)$ and W_i is a mean zero Gaussian random variable with variance such that the signal-to-noise ratio is 3.0. Censoring times are generated as $C_i := \exp(W_i)$ and the observed survival times as $T_i := \min(C_i, \tilde{T}_i)$.

We compare the runtime of `ahazpen` with that of `surv.lars` from the R-package **timereg** (Scheike and Zhang, 2011) which is currently the only publicly available software for fitting the lasso penalized additive hazards model. The `surv.lars` function is a modified version of the `lars` function from the package **lars** and requires pre-calculation of the quantities D and d. We use a highly efficient C-routine for calculating D based on formulas as in Section 3.1 (function `ahaz` in the package **ahaz**). To make `ahazpen` and `surv.lars` reasonably comparable, we stop `surv.lars` after 100 steps, and use the corresponding smallest λ-value $\lambda_{\min}$ to construct an exponentially decreasing λ-sequence for `ahazpen` of length 100 as in (10).

Experiments were run on an Intel Core I7 2.93 GHz, 8 GB RAM system with standard **BLAS**.

Runtimes for different values of n, p, and ρ are shown in Table 1 (averaged over 3 repetitions). Table 2 shows the corresponding runtimes of the pre-calculation part of lasso-LARS (averaged over the three repetitions and ρ as well). Coordinate descent is universally faster than lasso-LARS, especially for large values of p. The bottleneck of lasso-LARS is obviously the pre-calculation of D which has complexity of order $O(np^2)$. Neither algorithm is much affected by large correlations. Table 3 shows runtimes of `ahazpen` for very large values of n or p (averaged over 3 repetitions), based on a path of 100 λ-values with $\lambda_{\min}$ chosen such that the maximal number of variables in the path is roughly 100. It is seen that the algorithm is fully capable of dealing with large amounts of data. It runs more slowly for very large n than for very large p since the calculations in the strong/active sets become costly as well for large n. In our detailed assessments (not shown), the runtime scaled approximately linearly in n for fixed p and vice versa. Similar behavior was reported by Simon *et al.* (2011) for their coordinate descent algorithm for the penalized Cox model. Finally, a negative effect of large correlations on runtimes starts to become apparent for these large problems.

It is tempting to compare the raw computational performance of `ahazpen` with that of the **glmnet** `coxnet` function for fitting lasso penalized Cox models (Simon *et al.*, 2011; Friedman *et al.*, 2010). Such a comparison can only be qualitative and superficial since the algorithms solve different problems, use different convergence criteria, and rely on completely independent implementations. Intuitively, one might expect `ahazpen`, which is based on a linear model, to be substantially faster than `coxnet`. This is not the case. In fact, our limited experiments with **glmnet** (version 1.7) suggest that the two methods often have surprisingly similar runtimes, both being roughly as fast as **glmnet** coordinate descent for the simple linear regression model for an equally sized problem. A plausible explanation is that comparatively little time is spent on nonlinear optimizations because of the efficient use of active-set calculations. On the other hand, for very large n, `ahazpen` can be more efficient than `coxnet` since the coordinate descent part of `ahazpen` is essentially 'kernelized' (via the use of D, d); whereas `coxnet` does coordinate descent via inner products between n-vectors. Also, cross-validation tends to be somewhat faster for `ahazpen` than for `coxnet` since it is based on the simple quadratic loss (5).

An appreciable and implementation-independent advantage of the linearity of the additive hazards model is the guaranteed convergence `ahazpen`. It is also our experience that the runtime of `ahazpen` is more predictable than that of `coxnet` where convergence of the nonlinear optimization part can be sensitive to the nature of the data considered.

5.2. *An example using real data*

To demonstrate the practical use of `ahazpen`, we consider the Sørlie data set (Sørlie *et al.*, 2003) which consists of 549 gene expression measurements and survival times for 115 women diagnosed with breast cancer. This data set was also used by Martinussen and Scheike (2009) to demonstrate the lasso penalized additive hazards model.

We consider the challenging problem of performing lasso penalized survival regression for both main effects and pairwise (multiplicative) interactions of gene expressions. The design matrix has $p = 549 + 549 \cdot (549-1)/2 = 150,975$ columns. We apply the additive hazards lasso directly to this design matrix, ignoring here the discussion whether it is sensible to allow for inclusion of interactions without the corresponding main effects.

We load and format the data as follows (note that generating `X` may take several minutes):

```
R> data("sorlie")
R> set.seed(10101)
R> surv <- Surv(sorlie$time + runif(nrow(sorlie)) * 1e-2, sorlie$status)
R> Z <- sorlie[,3:ncol(sorlie)]; p <- ncol(Z)
R> pw.comb <- combn(1:p,2)
R> X <- cbind(Z, Z[, pw.comb[1,]] * Z[, pw.comb[2,]])
```

It is common practice to put variables on the same scale before applying the lasso. In `ahazpen`, data is scaled by default (estimates are returned on the original scale) so it is not necessary standardize data manually. We make the following call to `ahazpen` to fit the lasso; the choice of `penalty` corresponds to the default value and is included here for completeness:

```
R> fit.init <- ahazpen(surv, X, dfmax = 50, penalty = lasso.control(alpha=1))
R> fit.init

Call:
ahazpen(surv = surv, X = X, dfmax = 50)
```

— Output continued on next page —

Table 1. Runtime (seconds) of `ahazpen` (CCD) and `surv.lars` (LAR) for the simulated data. Results are averaged over 3 repetitions.

		p =100		p =500		p =5,000		p =10,000	
n	ρ	CCD	LAR	CCD	LAR	CCD	LAR	CCD	LAR
200	0	0.02	0.06	0.02	0.14	0.13	4.81	0.28	16.28
	0.25	0.02	0.07	0.02	0.15	0.12	4.62	0.28	16.13
	0.5	0.03	0.07	0.03	0.17	0.12	4.37	0.28	15.93
	0.9	0.06	0.07	0.04	0.14	0.13	4.29	0.28	15.48
	0.95	0.05	0.06	0.04	0.14	0.13	4.14	0.30	15.41
500	0	0.02	0.07	0.03	0.18	0.27	8.40	0.55	30.08
	0.25	0.03	0.07	0.04	0.18	0.29	8.15	0.56	29.79
	0.5	0.03	0.07	0.04	0.18	0.27	7.82	0.58	29.58
	0.9	0.04	0.07	0.04	0.18	0.28	7.86	0.55	29.65
	0.95	0.06	0.07	0.06	0.18	0.32	7.91	0.56	29.40
1,000	0	0.04	0.08	0.06	0.22	0.59	13.95	1.16	51.82
	0.25	0.04	0.07	0.06	0.22	0.55	13.80	1.06	51.52
	0.5	0.04	0.07	0.06	0.24	0.56	13.50	1.12	51.22
	0.9	0.05	0.07	0.06	0.22	0.61	13.50	1.14	51.07
	0.95	0.06	0.07	0.06	0.23	0.56	13.37	1.21	51.13

Table 2. Time (seconds) spent calculating D, d for the simulated data of Table 1 (averaged over 3 repetitions and ρ).

	p			
n	100	500	5000	10,000
200	0.00	0.03	3.32	13.47
500	0.00	0.06	6.80	27.05
1,000	0.01	0.10	12.35	48.64

Table 3. Runtime (seconds) for `ahazpen` averaged over 3 repetitions.

	ρ				
$(n;p)$	0	0.25	0.5	0.90	0.95
(200; 40,000)	1.22	1.19	1.19	1.18	1.30
(200; 100,000)	2.97	2.95	2.99	2.90	3.13
(200; 250,000)	6.84	6.84	6.79	6.91	7.13
(40,000; 200)	1.55	1.59	1.57	1.52	2.37
(100,000; 200)	4.03	3.99	3.96	3.89	6.02
(250,000; 200)	10.59	10.48	10.51	10.19	15.91

```
* No. predictors:             150975
* No. observations:              115
* Max no. predictors in path:     53
* Penalty parameter lambda:
    -No. grid points:     32
    -Min value:       0.1057
    -Max value:       0.2700
```

To prevent `ahazpen` from calculating a complete regularization path, `dfmax` has been specified. This option is useful for reducing computation time since, in practice, the lasso often prefers rather sparse solutions.

Only 32 λ-values are used even though `ahazpen` is set to use 100 λ-values as default. This is because `ahazpen` cannot anticipate the λ-value at which `dfmax` is reached and hence simply truncates the default λ-sequence (10). A grid of λ-values with the desired density is easily obtained by a second call to `ahazpen`:

```
R> l <- range(fit.init$lambda)
R> fit <- ahazpen(surv, X, lambda.minf = l[1] / l[2])
R> plot(fit)
```

Evaluating this `ahazpen` call took roughly 9 seconds. A plot of the regularization path is shown in Figure 1 (left).

To determine an optimal value of λ, we use 5-fold cross-validation as follows:

```
R> set.seed(10101)
R> fit.tune <- tune.ahazpen(surv, X, lambda.minf = l[1] / l[2], tune = "cv")
R> fit.tune

Call:
tune.ahazpen(surv = surv, X = X, tune = "cv", lambda.minf = l[1]/l[2])

Cross-validation:  5  folds

Length of lambda sequence :  100
Optimal lambda            :  0.2051
d.f. at optimal lambda    :  4

R> plot(fit.tune)
```

Typically, K-fold cross-validation takes about as long as running `ahazpen` $K+1$ times. Figure 1 (right) shows the curve of cross-validation scores. Indices of the final nonzero regression coefficients are then obtained as follows

```
R> beta <- coef(fit.tune)
R> which(as.numeric(beta) != 0)

[1]  21 269 346 401
```

Apparently, the lasso prefers a model containing main effects only.

6. Discussion

Cyclic coordinate descent is a simple numerical optimization method which works exceptionally well for penalized regression problems with variable selection. We have developed a coordinate descent algorithm for the elastic net penalized additive hazards model and provided an implementation via the `ahazpen` function in the R-package **ahaz**. This function can handle very large amounts of data efficiently and is an important and flexible alternative to the more commonly used elastic net penalized Cox model. In terms of computational properties, the additive hazards model is intrinsically linear which implies theoretically guaranteed convergence of `ahazpen` and highly predictable runtimes in practice. Our specific implementation provides support for survival data with delayed entry which in turn enables the use of more complex data types such as nonstandard time axes, time-varying covariates, and general counting process data.

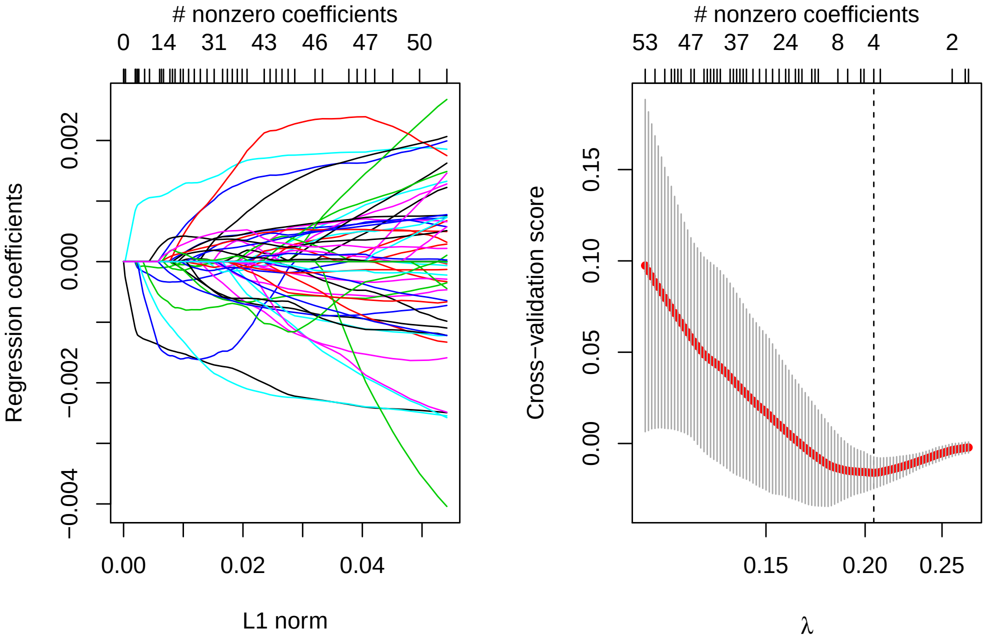

Figure 1. Plot of regularization path (left) and 5-fold cross-validation scores (right) for the additive hazards lasso applied to the Sørlie gene expression data with main effects and pairwise interactions.

References

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *The Annals of Statistics*, **32**, 407–499.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1**, 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22. URL `http://www.jstatsoft.org/v33/i01/`.

Goeman, J. J. (2010) L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, **52**, 70–84.

Gorst-Rasmussen, A. (2011) ***ahaz**: Regularization for semiparametric additive hazards regression*. URL `http://cran.r-project.org/package=ahaz`. R package.

Gorst-Rasmussen, A. and Scheike, T. H. (2011) Independent screening for single-index hazard rate models with ultra-high dimensional features. Tech. Rep. R-2011-06, Department of Mathematical Sciences, Aalborg University.

Gui, J. and Li, H. (2005) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**, 3001–3008.

Leng, C. and Ma, S. (2007) Path consistent model selection in additive risk model via lasso. *Statistics in Medicine*, **26**, 3753–3770.

Li, H. (2008) Censored data regression in high-dimensional and low-sample-size settings for genomic applications. In *Statistical Advances in Biomedical Sciences: State of the Art and Future Directions* (eds. A. Biswas, S. Datta, J. Fine and M. Segal). Wiley.

Lin, D. Y. and Ying, Z. (1994) Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61–71.

Ma, S., Huang, J., Shi, M., Li, Y. and Shia, B. (2010) Semiparametric prognosis models in genomic studies. *Briefings in Bioinformatics*, **11**, 385–393.

Ma, S., Kosorok, M. and Fine, J. P. (2006) Additive risk models for survival data with high-dimensional covariates. *Biometrika*, **62**, 202–210.

Martinussen, T. and Scheike, T. H. (2009) Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics*, **36**, 602–619.

Martinussen, T. and Scheike, T. H. (2010) The additive hazards model with high-dimensional regressors. *Lifetime Data Analysis*, **15**, 330–342.

McKeague, I. W. and Sasieni, P. D. (1994) A partly parametric additive risk model. *Biometrika*, **81**, 501–514.

Park, M. Y. and Hastie, T. (2007) L1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society B*, **69**, 659–677.

Scheike, T. H. and Zhang, M.-J. (2011) Analyzing competing risk data using the R **timereg** package. *Journal of Statistical Software*, **38**, 1–15. URL `http://www.jstatsoft.org/v38/i02/`.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011) Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, **39**, 1–13. URL `http://www.jstatsoft.org/v39/i05/`.

Sohn, I., Kim, J., Jung, S. and Park, C. (2009) Gradient lasso for Cox proportional hazards model. *Bioinformatics*, **25**, 1775–1781.

Sørlie, T., Partker, R., Hatie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Peour, C., Lønning, P., Brown, P., Børresen-Dale, A. and Botstein, D. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, **100**, 8418–8423.

Tibshirani, R. (1997) The lasso method for variable selection in the Cox model. *Statistics* *in Medicine*, **16**, 385–395.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J. and Tibshirani, R. (2010) Strong rules for discarding predictors in lasso-type problems. Tech. rep., Stanford University.

Tseng, P. (1988) Coordinate ascent for maximizing nondifferentiable concave functions. Tech. Rep. LIDS-P, 1840, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems.

Wang, H. and Leng, C. (2007) Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association*, **102**, 1039–1048.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, **67**, 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2007) On the “degrees of freedom” of the lasso. *The Annals of Statistics*, **35**, 2173–2192.

Paper VII

Independent Screening for Single-Index Hazard Rate Models with Ultra-High Dimensional Features

Author list

Anders Gorst-Rasmussen
Aalborg University, Denmark

Thomas H. Scheike
University of Copenhagen, Denmark

Summary

In data sets with many more features than observations, independent screening based on all univariate regression models leads to a computationally convenient variable selection method. Recent efforts have shown that in the case of generalized linear models, independent screening may suffice to capture all relevant features with high probability, even in ultra-high dimension. It is unclear whether this formal sure screening property is attainable when the response is a right-censored survival time. We propose a computationally very efficient independent screening method for survival data which can be viewed as the natural survival equivalent of correlation screening. We state conditions under which the method admits the sure screening property within a general class of single-index hazard rate models with ultra-high dimensional features. An iterative variant is also described which combines screening with penalized regression in order to handle more complex feature covariance structures. The methods are evaluated through simulation studies and through application to a real gene expression data set.

Supplementary info

This manuscript has been published as:

> Gorst-Rasmussen A, Scheike TH (2011). Independent screening for single-index hazard rate models with ultra-high dimensional features. *Technical report R-2011-06*. Department of Mathematical Sciences, Aalborg University

It has moreover been submitted to *Journal of the Royal Statistical Society, Series B*.

1. Introduction

With the increasing proliferation of biomarker studies, there is a need for efficient methods for relating a survival time response to a large number of features. In typical genetic microarray studies, the sample size n is measured in hundreds whereas the number of features p per sample can be in excess of millions. Sparse regression techniques such as lasso (Tibshirani, 1997) and SCAD (Fan and Li, 2001) have proved useful for dealing with such high-dimensional features but their usefulness diminishes when p becomes extremely large compared to n. The notion of NP-dimensionality (Fan and Lv, 2009) has been conceived to describe such ultra-high dimensional settings which are formally analyzed in an asymptotic regime where p grows at a non-polynomial rate with n. Despite recent progress (Bradic *et al.*, 2011), theoretical knowledge about sparse regression techniques under NP-dimensionality is still in its infancy. Moreover, NP-dimensionality poses substantial computational challenges. When for example pairwise interactions among gene expressions in a genetic microarray study are of interest, the dimension of the feature space will trouble even the most efficient algorithms for fitting sparse regression models. A popular ad hoc solution is to simply pretend that feature correlations are ignorable and resort to computationally swift univariate regression methods; so-called independent screening methods.

In an important paper, Fan and Lv (2008) laid the formal foundation for using independent screening to distinguish 'relevant' features from 'irrelevant' ones. For the linear regression model they showed that, when the design is close to orthogonal, a superset of the true set of nonzero regression coefficients can be estimated consistently by simple hard-thresholding of feature-response correlations. This sure independent screening (SIS) property of correlation screening is a rather trivial one, if not for the fact that it holds true in the asymptotic regime of NP-dimensionality. Thus, when the feature covariance structure is sufficiently simple, SIS methods can overcome the noise accumulation in extremely high dimension. In order to accommodate more complex feature covariance structures Fan and Lv (2008) and Fan *et al.* (2009) developed heuristic, iterated methods combining independent screening with forward selection techniques. Recently, Fan and Song (2010) extended the formal basis for SIS to generalized linear models.

In biomedical applications, the response of interest is often a right-censored survival time, making the study of screening methods for survival data an important one. Fan *et al.* (2010) investigated SIS methods for the Cox proportional hazards model based on ranking features according to the univariate partial log-likelihood but gave no formal justification. Tibshirani (2009) suggested soft-thresholding of univariate Cox score statistics with some theoretical justification but under strong assumptions. Indeed, independent screening methods for survival data are apt to be difficult to justify theoretically due to the presence of censoring which can confound marginal associations between the response and the features. Recent work by Zhao and Li (2010) contains ideas which indicate that independent screening based on the Cox model may have the SIS property in the absence of censoring.

In the present paper, we depart from the standard approach of studying SIS as a rather specific type of model misspecification in which the univariate versions of a particular regression model are used to infer the structure of the joint version of the same particular regression model. Instead, we propose a survival variant of independent screening based on a model-free statistic which we call the 'Feature Aberration at Survival Times' (FAST) statistic. The FAST statistic is a simple linear statistic which aggregates across survival times the aberration of each feature relative to its time-varying average. Independent screening based on this statistic can be regarded as a natural survival equivalent of correlation screening. We study the SIS property of FAST screening in ultra-high dimension for a general class of single-index hazard rate regression models in which the risk of an event depends on the features through some linear functional. A key aim has been to derive simple and operational sufficient conditions for the SIS property to hold. Accordingly, our main result states that the FAST statistic has the SIS property in an ultra-high dimensional setting under covariance assumptions as in Fan *et al.* (2009), provided that censoring is essentially random and that features satisfy a technical condition which holds when they follow an elliptically contoured distribution. Utilizing the fact that the FAST statistic is related to the univariate regression coefficients in the semiparametric additive hazards model (Lin and Ying (1994); McKeague and Sasieni (1994)), we develop methods for iterated SIS. The techniques are evaluated in a simulation study where we also compare with screening methods for the Cox model (Fan *et al.*, 2010). Finally, an application to a real genetic microarray data set is presented.

2. The FAST statistic and its motivation

Let T be a survival time which is subject to right-censoring by some random variable C. Denote by $N(t) := \mathbb{1}(T \wedge C \leqslant t \text{ and } T \leqslant C)$ the counting process which counts events up to time t, let $Y(t) := \mathbb{1}(T \wedge C \geqslant t)$ be the at-risk process, and let $\mathbf{Z} \in \mathbb{R}^p$ denote a random vector of explanatory variables or features. It is assumed throughout that $\mathbf{Z}$ has finite variance and is standardized, i.e. centered and with a covariance matrix Σ with unit diagonal. We observe n independent and identically distributed (i.i.d.) replicates of $\{(N_i, Y_i, \mathbf{Z}_i) : 0 \leqslant t \leqslant \tau\}$ for $i = 1, \ldots, n$ where $[0, \tau]$ is the observation time window.

Define the 'Feature Aberration at Survival Times' (FAST) statistic as follows:

$$\mathbf{d} := n^{-1} \int_0^\tau \sum_{i=1}^n \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\} \mathrm{d}N_i(t); \tag{1}$$

where $\bar{\mathbf{Z}}$ is the at-risk-average of the $\mathbf{Z}_i$s,

$$\bar{\mathbf{Z}}(t) := \frac{\sum_{i=1}^n \mathbf{Z}_i Y_i(t)}{\sum_{i=1}^n Y_i(t)}.$$

Components of the FAST statistic define basic measures of the marginal association between each feature and survival. In the following, we provide two motivations for using the FAST statistic for screening purposes. The first, being model-based, is perhaps the most intuitive – the second shows that, even in a model-free setting, the FAST statistic may provide valuable information about marginal associations.

2.1. *A model-based interpretation of the FAST statistic*

Assume in this section that the T_is have hazard functions of the form

$$\lambda_i(t) = \lambda_0(t) + \mathbf{Z}_i^\top \boldsymbol{\alpha}^0; \qquad i = 1, \ldots, n; \tag{2}$$

with λ_0 an unspecified baseline hazard rate and $\boldsymbol{\alpha}^0 \in \mathbb{R}^p$ a vector of regression coefficients. This is the so-called semiparametric additive hazards model (Lin and Ying (1994); McKeague and Sasieni (1994)), henceforth simply the Lin-Ying model. The Lin-Ying model corresponds to assuming for each N_i an intensity function of the form $Y_i(t)\{\lambda_0(t) + \mathbf{Z}_i^\top \boldsymbol{\alpha}^0\}$. From the Doob-Meyer decomposition $\mathrm{d}N_i(t) = \mathrm{d}M_i(t) + Y_i(t)\{\lambda_0(t) + \mathbf{Z}_i^\top \boldsymbol{\alpha}^0\}\mathrm{d}t$ with M_i a martingale, it is easily verified that

$$\sum_{i=1}^n \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\} \mathrm{d}N_i(t) = \left[\sum_{i=1}^n \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\}^{\otimes 2} Y_i(t) \mathrm{d}t \right] \boldsymbol{\alpha}^0 + \sum_{i=1}^n \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\} \mathrm{d}M_i(t), \quad t \in [0, \tau]. \tag{3}$$

This suggests that $\boldsymbol{\alpha}^0$ is estimable as the solution to the $p \times p$ linear system of equations

$$\mathbf{d} = \mathbf{D}\boldsymbol{\alpha}; \tag{4}$$

where

$$\mathbf{d} := n^{-1}\sum_{i=1}^{n}\int_0^\tau \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\}\mathrm{d}N_i(t), \quad \text{and } \mathbf{D} := n^{-1}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\mathrm{d}t. \tag{5}$$

Suppose $\hat{\boldsymbol{\alpha}}$ solves (4). Standard martingale arguments (Lin and Ying, 1994) imply root-n consistency of $\hat{\boldsymbol{\alpha}}$ so that $n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0)$ is asymptotically mean zero Gaussian with a covariance matrix which is consistently estimated by

$$\widehat{\mathrm{Var}}\{n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0)\} = \mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}. \tag{6}$$

For now, simply observe that the left-hand side of (4) is exactly the FAST statistic; whereas $d_j D_{jj}^{-1}$ for $j = 1,2,\dots,p$ estimate the regression coefficients in the corresponding p univariate Lin-Ying models. Hence we can interpret $\mathbf{d}$ as a (scaled) estimator of the univariate regression coefficients in a working Lin-Ying model.

A nice heuristic interpretation of $\mathbf{d}$ results from the pointwise signal/error decomposition (3) which is essentially a reformulated linear regression model $\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha}^0 + \mathbf{X}^\top\boldsymbol{\varepsilon} = \mathbf{X}^\top\mathbf{y}$ with 'responses' $y_i := \mathrm{d}N_i(t)$ and 'explanatory variables' $\mathbf{X}_i := \{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\}Y_i(t)$. The FAST statistic is given by the time average of $\mathbb{E}\{\mathbf{X}^\top\mathbf{y}\}$ and may accordingly be viewed as a survival equivalent of the usual predictor-response correlations.

2.2. *A model-free interpretation of the FAST statistic*

For a feature to be judged (marginally) associated with survival in any reasonable interpretation of survival data, one would first require that the feature is correlated with the probability of experiencing an event – second, that this correlation persists throughout the time window. The FAST statistic can be shown to reflect these two requirements when the censoring mechanism is sufficiently simple.

Specifically, assume that $C_1 \equiv \tau$ (administrative censoring at time τ). Set $V(t) := \mathrm{Var}\{F(t|\mathbf{Z}_1)\}^{1/2}$ where $F(t|\mathbf{Z}_1) := \mathbb{P}(T_1 \leqslant t|\mathbf{Z}_1)$ denotes the conditional probability of death before time t. For each j, denote by δ_j the population version of d_j (the in probability limit of d_j when $n \to \infty$). Then

$$\begin{aligned}
\delta_j &= \mathbb{E}\left(\int_0^\tau \left[Z_{1j} - \frac{\mathbb{E}\{Z_{1j}Y_1(t)\}}{\mathbb{E}\{Y_1(t)\}}\right]\mathbb{1}(T_1 \leqslant t\wedge\tau)\mathrm{d}t\right) \\
&= \mathbb{E}\{Z_{1j}F(\tau|\mathbf{Z}_1)\} - \int_0^\tau \frac{\mathbb{E}\{Z_{1j}Y_1(t)\}}{\mathbb{E}\{Y_1(t)\}}\mathbb{E}\{\mathrm{d}F(t|\mathbf{Z}_1)\} \\
&= V(\tau)\mathrm{Cor}\{Z_{1j}, F(\tau|\mathbf{Z}_1)\} + \int_0^\tau \mathrm{Cor}\{Z_{1j}, F(t|\mathbf{Z}_1)\}\frac{V(t)}{\mathbb{E}\{Y_1(t)\}}\mathbb{E}\{\mathrm{d}F(t|\mathbf{Z}_1)\}.
\end{aligned}$$

We can make the following observations:

(i). If $\mathrm{Cor}\{Z_{1j}, F(t|\mathbf{Z}_1)\}$ has constant sign on $[0,\tau]$, then $|\delta_j| \geqslant |V(\tau)\mathrm{Cor}\{Z_{1j}, F(\tau|\mathbf{Z}_1)\}|$.

(ii). Conversely, if $\mathrm{Cor}\{Z_{1j}, F(t|\mathbf{Z}_1)\}$ changes sign, so that the the direction of association with $F(t|\mathbf{Z}_1)$ is not persistent throughout $[0,\tau]$, then this will lead to a smaller value of $|\delta_j|$ compared to (i).

(iii). Lastly, if $\mathrm{Cor}\{Z_{1j}, F(t|\mathbf{Z}_1)\} \equiv 0$ then $\delta_j = 0$.

In other words, the sample version d_j estimates a time-averaged summary of the correlation function $t \mapsto \mathrm{Cor}\{Z_{1j}, F(t|\mathbf{Z}_1)\}$ which takes into account both magnitude and persistent behavior throughout $[0,\tau]$. This indicates that the FAST statistic is relevant for judging marginal association of features with survival beyond the model-specific setting of Section 2.1

3. Independent screening with the FAST statistic

In this section, we extend the heuristic arguments of the previous section and provide theoretical justification for using the FAST statistic to screen for relevant features when the data-generating model belongs to a class of single-index hazard rate regression models.

3.1. *The general case of single-index hazard rate models*

With the notation of Section 2, we assume survival times T_i to have hazard rate functions of single-index form:

$$\lambda_i(t) = \lambda\big(t, \mathbf{Z}_i^\top \boldsymbol{\alpha}^0\big), \quad j = 1, \dots, n. \tag{7}$$

Here $\lambda\colon [0,\infty) \times \mathbb{R} \to [0,\infty)$ is a continuous function, $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ are random vectors in $\mathbb{R}^{p_n}$, $\boldsymbol{\alpha}^0 \in \mathbb{R}^{p_n}$ is a vector of regression coefficients, and $\mathbf{Z}_i^\top \boldsymbol{\alpha}^0$ defines a risk score. We subscript p by n to indicate that the dimension of the feature space can grow with the sample size. Censoring will always be assumed at least independent so that C_i is independent of T_i conditionally on $\mathbf{Z}_i$. We impose the following assumption on the hazard 'link function' λ:

Assumption 1. The survival function $\exp\{-\int_0^t \lambda(s,\cdot)\mathrm{d}s\}$ is continuously differentiable and strictly monotonic for each $t \geqslant 0$.

Requiring the survival function to depend monotonically on $\mathbf{Z}_i^\top \boldsymbol{\alpha}^0$ is natural in order to enable interpretation of the components of $\boldsymbol{\alpha}^0$ as indicative of positive or negative association with survival. Note that it suffices that $\lambda(t,\cdot)$ is strictly monotonic (and continuously differentiable) for each $t \geqslant 0$. Assumption 1 holds for a range of popular survival regression models. For example, $\lambda(t,x) := \lambda_0(t) + x$ with λ_0 some baseline hazard yields the Lin-Ying model (2); $\lambda(t,x) := \lambda_0(t)\mathrm{e}^x$ is a Cox model; and $\lambda(t,x) := \mathrm{e}^x \lambda_0(t\mathrm{e}^x)$ is an accelerated failure time model.

Denote by $\boldsymbol{\delta}$ the population version of the FAST statistic under the model (7) which, by the Doob-Meyer decomposition $\mathrm{d}N_1(t) = \mathrm{d}M_1(t) + Y_1(t)\lambda(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\mathrm{d}t$ with M_1 a martingale, takes the form

$$\boldsymbol{\delta} = \mathbb{E}\left[\int_0^\tau \{\mathbf{Z}_1 - \mathbf{e}(t)\} Y_1(t) \lambda\big(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0\big)\mathrm{d}t\right]; \qquad \text{where } \mathbf{e}(t) := \frac{\mathbb{E}\{\mathbf{Z}_1 Y_1(t)\}}{\mathbb{E}\{Y_1(t)\}}. \tag{8}$$

Our proposed FAST screening procedure is as follows: given some (data-dependent) threshold $\gamma_n > 0$,

(i). calculate the FAST statistic $\mathbf{d}$ from the available data and
(ii). declare the 'relevant features' to be the set $\{1 \leqslant j \leqslant p_n : |d_j| > \gamma_n\}$.

By the arguments in Section 2, this procedure defines a natural survival equivalent of correlation screening. Define the following sets of features:

$$\begin{aligned}
\widehat{\mathrm{M}}_d^n &:= \big\{1 \leqslant j \leqslant p_n : |d_j| > \gamma_n\big\}, \\
\mathrm{M}^n &:= \big\{1 \leqslant j \leqslant p_n : \alpha_j^0 \neq 0\big\}, \\
\mathrm{M}_\delta^n &:= \big\{1 \leqslant j \leqslant p_n : \delta_j \neq 0\big\}.
\end{aligned}$$

The problem of establishing the SIS property of FAST screening amounts to determining when $\mathrm{M}^n \subseteq \widehat{\mathrm{M}}_\mathrm{d}^n$ holds with large probability for large n. This translates into two questions: first, when do we have $\mathrm{M}^n_\delta \subseteq \widehat{\mathrm{M}}^n_d$; second, when do we have $\mathrm{M}^n \subseteq \mathrm{M}^n_\delta$? The first question is essentially model-independent and requires establishing an exponential bound for $n^{1/2}|d_j - \delta_j|$ as $n \to \infty$. The second question is strongly model-dependent and is answered by manipulating expectations under the single-index model (7).

We state the main results here and relegate proofs to the appendix where we also state various regularity conditions. The following principal assumptions, however, deserve separate attention:

Assumption 2. There exists $\mathbf{c} \in \mathbb{R}^{p_n}$ such that $\mathbb{E}(\mathbf{Z}_1|\mathbf{Z}_1^\top \boldsymbol{\alpha}^0) = \mathbf{c}\mathbf{Z}_1^\top \boldsymbol{\alpha}^0$.

Assumption 3. The censoring time C_1 depends on $T_1, \mathbf{Z}_1$ only through Z_{1j}, $j \notin \mathrm{M}^n$.

Assumption 4. Z_{1j}, $j \in \mathrm{M}^n$ is independent of Z_{1j}, $j \notin \mathrm{M}^n$.

Assumption 2 is a ‘linear regression’ property which holds true for Gaussian features and, more generally, for features following an elliptically contoured distribution (Hardin, 1982). In view of Hall and Li (1993) which states that most low dimensional projections of high dimensional features are close to linear, Assumption 2 may not be unreasonable a priori even for general feature distributions when p_n is large.

Assumption 3 restricts the censoring mechanism to be partially random in the sense of depending only on irrelevant features. As we will discuss in detail below, such rather strong restrictions on the censoring mechanism seem necessary for obtaining the SIS property; Assumption 3 is both general and convenient.

Assumption 4 is the partial orthogonality condition also used by Fan and Song (2010). Under this assumption and Assumption 3, it follows from (8) that $\delta_j = 0$ whenever $j \notin \mathrm{M}^n$, implying $\mathrm{M}^n_\delta \subseteq \mathrm{M}^n$. Provided that we also have $\delta_j \neq 0$ for $j \in \mathrm{M}^n$ (that is, $\mathrm{M}^n \subseteq \mathrm{M}^n_\delta$), there exists a threshold $\zeta_n > 0$ such that

$$\min_{j \in \mathrm{M}^n} |\delta_j| \geqslant \zeta_n \qquad \max_{j \notin \mathrm{M}^n} |\delta_j| = 0.$$

Consequently, Assumptions 3-4 enable consistent model selection via independent screening. Although model selection consistency is not essential in order to capture just some superset of the relevant features via independent screening, it is pertinent in order to limit the size of such a superset.

The following theorem on FAST screening (FAST-SIS) is our main theoretical result. It states that the screening property $\mathrm{M}^n \subseteq \widehat{\mathrm{M}}^n_d$ may hold with large probability even when p_n grows exponentially fast in a certain power of n which depends on the tail behavior of features. The covariance condition in the theorem is analogous to that of Fan and Song (2010) for SIS in generalized linear models with Gaussian features.

THEOREM 1. *Suppose that Assumptions 1-3 hold alongside the regularity conditions of the appendix and that* $\mathbb{P}(|Z_{1j}| > s) \leqslant l_0 \exp(-l_1 s^\eta)$ *for some positive constants* l_0, l_1, η *and sufficiently large s. Suppose moreover that for some* $c_1 > 0$ *and* $\kappa < 1/2$,

$$\left|\mathrm{Cov}\left(Z_{1j}, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0\right)\right| \geqslant c_1 n^{-\kappa}, \quad j \in \mathrm{M}^n. \tag{9}$$

Then $\mathrm{M}^n \subseteq \mathrm{M}^n_\delta$. *Suppose in addition that* $\gamma_n = c_2 n^{-\kappa}$ *for some constant* $0 < c_2 \leqslant c_1/2$ *and that* $\log p_n = o\{n^{(1-2\kappa)\eta/(\eta+2)}\}$. *Then the SIS property holds,* $\mathbb{P}(\mathrm{M}^n \subseteq \widehat{\mathrm{M}}^n_d) \to 1$, $n \to \infty$.

Observe that with bounded features, we may take $\eta = \infty$ and handle dimension of order $\log p_n = o(n^{1-2\kappa})$.

We may dispense with Assumption 2 on the feature distribution by revising (9). By Lemma A5 in the appendix, taking $\tilde{e}_j(t) := \mathbb{E}\{Z_{1j}\mathbb{P}(T_1 \geqslant t|\mathbf{Z}_1)\}/\mathbb{E}\{\mathbb{P}(T_1 \geqslant t|\mathbf{Z}_1)\}$, it holds generally under Assumption 3 that

$$\delta_j = \mathbb{E}\{\tilde{e}_j(T_1 \wedge C_1 \wedge \tau)\}, \quad j \in \mathrm{M}^n.$$

Accordingly, if we replace (9) with the assumption that $\mathbb{E}|Z_{1j}\mathbb{P}(T_1 \geqslant t|\mathbf{Z}_1)| \geqslant c_1 n^{-\kappa}$ uniformly in t for $j \in \mathrm{M}^n$, the conclusions of Theorem 1 still hold. In other words, we can generally expect FAST-SIS to detect features which are 'correlated with the chance of survival', much in line with Section 2. While this is valuable structural insight, the covariance assumption (9) seems a more operational condition.

Assumption 3 is crucial to the proof of Theorem 1 and to the general idea of translating a model-based feature selection problem into a problem of hard-thresholding $\boldsymbol{\delta}$. A weaker assumption is not possible in general. For example, suppose that only Assumption 2 holds and that the censoring time also follows some single-index model of the form (7) with regression coefficients $\boldsymbol{\beta}^0$. Applying Lemma 2.1 of Cheng and Wu (1994) to (8), there exists finite constants ζ_1, ζ_2 (depending on n) such that

$$\boldsymbol{\delta} = \Sigma(\zeta_1\boldsymbol{\alpha}^0 + \zeta_2\boldsymbol{\beta}^0). \tag{10}$$

It follows that unrestricted censoring will generally confound the relationship between $\boldsymbol{\delta}$ and $\Sigma\boldsymbol{\alpha}^0$, hence $\boldsymbol{\alpha}^0$. The precise impact of such unrestricted censoring seems difficult to discern, although (10) suggests that FAST-SIS may still be able to capture the underlying model (unless $\zeta_1\boldsymbol{\alpha}^0 + \zeta_2\boldsymbol{\beta}^0$ is particularly ill-behaved). We will have more to say about unrestricted censoring in the next section.

Theorem 1 shows that FAST-SIS can consistently capture a superset of the relevant features. A priori, this superset can be quite large; indeed, 'perfect' screening would result by simply including all features. For FAST-SIS to be useful, it must substantially reduce feature space dimension. Below we state a survival analogue of Theorem 5 in Fan and Song (2010), providing an asymptotic rate on the FAST-SIS model size.

THEOREM 2. *Suppose that Assumptions 1-4 hold alongside the regularity conditions of the appendix and that* $\mathbb{P}(|Z_{1j}| > s) \leqslant l_0 \exp(-l_1 s^\eta)$ *for positive constants* l_0, l_1, η *and sufficiently large* s. *If* $\gamma_n = c_4 n^{-\kappa}$ *for some* $\kappa < 1/2$ *and* $c_4 > 0$, *there exists a positive constant* c_5 *such that*

$$\mathbb{P}\big[|\widehat{\mathrm{M}}_d^n| \leqslant O\{n^{2\kappa}\lambda_{\max}(\Sigma)\}\big] \geqslant 1 - O\big[p_n \exp\big\{-c_5 n^{(1-2\kappa)\eta/(\eta+2)}\big\}\big];$$

with $\lambda_{\max}(\Sigma)$ *the maximal eigenvalue of the feature covariance matrix* Σ.

Informally, the theorem states that, under similar assumptions as in Theorem 1 and the partial orthogonality condition (Assumption 4), if features are not too strongly correlated (as measured by the maximal eigenvalue of the covariance matrix) so that $n^{2\kappa}\lambda_{\max}(\Sigma)/p_n \to 0$, we can choose a threshold γ_n for hard-thresholding such that the false selection rate becomes asymptotically negligible.

Our theorems say little about how to actually select the hard-thresholding parameter γ_n in practice. Following Fan and Lv (2008) and Fan *et al.* (2009), we would typically choose γ_n such that $|\mathrm{M}_d^n|$ is of order $n/\log n$. Devising a general data-adaptive way of choosing γ_n is an open problem for independent screening methods in general and is beyond the scope of this paper. Suggestions were recently given by Zhao and Li (2010) and Fan *et al.* (2011) who described basic methods provide (asymptotic) control of false-positive rates. Their methods could be adapted to FAST screening as well.

3.2. *The special case of the Aalen model*

Additional insight into the impact of censoring on FAST-SIS is possible within the more restrictive context of the nonparametric Aalen model with Gaussian features (Aalen (1980); Aalen (1989)). This particular model asserts a hazard rate function for T_i of the form

$$\lambda_i(t) = \lambda_0(t) + \mathbf{Z}_i^\top \boldsymbol{\alpha}^0(t), \quad i = 1,\dots,n; \tag{11}$$

for some baseline hazard rate λ_0 and $\boldsymbol{\alpha}^0$ a vector of continuous regression coefficient functions. The Aalen model extends the Lin-Ying model of Section 2 by allowing time-varying regression coefficients. Alternatively, it can be viewed as defining an expansion to the first order of a general hazard rate function in the class (7) in the sense that

$$\lambda\big(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0\big) \approx \lambda(t,0) + \mathbf{Z}_1^\top \boldsymbol{\alpha}^0 \frac{\partial \lambda(t,x)}{\partial x}\bigg|_{x=0}. \tag{12}$$

For Aalen models with Gaussian features, the following analogue to Theorem 1 holds.

THEOREM 3. *Suppose that Assumptions 1-2 hold alongside the regularity conditions of the appendix. Suppose moreover that* $\mathbf{Z}_1$ *is mean zero Gaussian and that* T_1 *follows a model of the form* (11) *with regression coefficients* $\boldsymbol{\alpha}^0$. *Assume that* C_1 *also follows a model of the form* (11) *conditionally on* $\mathbf{Z}_1$ *and that censoring is independent. Let* $\mathbf{A}^0(t) := \int_0^t \boldsymbol{\alpha}^0(s)\mathrm{d}s$. *If for some* $\kappa < 1/2$ *and* $c_1 > 0$, *we have*

$$\big|\mathrm{Cov}\big[Z_{1j}, \mathbf{Z}_1^\top \mathbb{E}\{\mathbf{A}^0(T_1 \wedge C_1 \wedge \tau)\}\big]\big| \geqslant c_1 n^{-\kappa}, \quad j \in \mathrm{M}^n, \tag{13}$$

then the conclusions of Theorem 1 hold with $\eta = 2$.

In view of (12), Theorem 3 can be viewed as establishing, within the model class (7), conditions for first-order validity of FAST-SIS under a general (independent) censoring mechanism and Gaussian features. The expectation term in (13) is essentially the 'expected regression coefficients at the exit time' which is strongly dependent on censoring through the symmetric dependence on survival and censoring time.

In fact, general independent censoring is a nuisance even in the Lin-Ying model which would otherwise seem the 'natural model' in which to use FAST-SIS. Specifically, assuming only independent censoring, suppose that T_1 follows a Lin-Ying model with regression coefficients $\boldsymbol{\alpha}^0$ conditionally on $\mathbf{Z}_1$ and that C_1 also follows some Lin-Ying model conditionally on $\mathbf{Z}_1$. If $\mathbf{Z}_1 = \Sigma^{1/2}\tilde{\mathbf{Z}}_1$ where the components of $\tilde{\mathbf{Z}}_1$ are i.i.d. with mean zero and unit variance, there exists a $p_n \times p_n$ diagonal matrix $\mathbf{C}$ such that

$$\boldsymbol{\delta} = \Sigma^{1/2}\mathbf{C}\Sigma^{1/2}\boldsymbol{\alpha}^0. \tag{14}$$

See Lemma A6 in the appendix. It holds that $\mathbf{C}$ has constant diagonal iff features are Gaussian; otherwise the diagonal is non-constant and depends nontrivially on the regression coefficients of the censoring model. A curious implication is that, under Gaussian features, FAST screening has the SIS property for this 'double' Lin-Ying model irrespective of the (independent) censoring mechanism. Conversely, sufficient conditions for a SIS property to hold here under more general feature distributions would require the jth component of $\Sigma^{1/2}\mathbf{C}\Sigma^{1/2}\boldsymbol{\alpha}^0$ to be 'large' whenever α_j^0 is 'large'; hardly a very operational assumption. In other words, even in the simple Lin-Ying model, unrestricted censoring complicates analysis of FAST-SIS considerably.

3.3. *Scaling the FAST statistic*

The FAST statistic is easily generalized to incorporate scaling. Inspection of the results in the appendix immediately shows that multiplying the FAST statistic by some strictly positive, deterministic weight does not alter its asymptotic behavior. Under suitable assumptions, this also holds when weights are stochastic. In the notation of Section 2, the following two types of scaling are immediately relevant:

$$d_j^Z = d_j B_{jj}^{-1/2} \text{ (Z-FAST);} \tag{15}$$

$$d_j^{\mathrm{LY}} = d_j D_{jj}^{-1} \quad \text{(Lin-Ying-FAST).} \tag{16}$$

The Z-FAST statistic corresponds to standardizing $\mathbf{d}$ by its estimated standard deviation; screening with this statistic is equivalent to the standard approach of ranking features according to univariate Wald p-values. Various forms of asymptotic false-positive control can be implemented for Z-FAST, courtesy of the central limit theorem. Note that Z-FAST is model-independent in the sense that its interpretation (and asymptotic normality) does not depend on a specific model. In contrast, the Lin-Ying-FAST statistic is model-specific and corresponds to calculating the univariate regression coefficients in the Lin-Ying model, thus leading to an analogue of the idea of 'ranking by absolute regression coefficients' of Fan and Song (2010) .

We may even devise a scaling of $\mathbf{d}$ which mimics the 'ranking by marginal likelihood ratio' screening of Fan and Song (2010) by considering univariate versions of the natural loss function $\boldsymbol{\beta} \mapsto \boldsymbol{\beta}^\top \mathbf{D} \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{d}$ for the Lin-Ying model. The components of the resulting statistic are rather similar to (16), taking the form

$$d_j^{\mathrm{loss}} = d_j D_{jj}^{-1/2} \text{ (loss-FAST).} \tag{17}$$

Additional flexibility can be gained by using a time-dependent scaling where some strictly positive (stochastic) weight is multiplied on the integrand in (1). This is beyond the scope of the present paper.

4. Beyond simple independent screening – iterated FAST screening

The main assumption underlying any SIS method, including FAST-SIS, is that the design is close to orthogonal. This assumption is easily violated: a relevant feature may have a low marginal association with survival; an irrelevant feature may be indirectly associated with survival through associations with relevant features etc. To address such issues, Fan and Lv (2008) and Fan *et al.* (2009) proposed various heuristic iterative SIS (ISIS) methods which generally work as follows. First, SIS is used to recruit a small subset of features within which an even smaller subset of features is selected using a (multivariate) variable selection method such as penalized regression. Second, the (univariate) relevance of each feature not selected in the variable selection step is re-evaluated, adjusted for all the selected features. Third, a small subset of the most relevant of these new features is joined to the set of already selected features, and the variable selection step is repeated. The last two steps are iterated until the set of selected features stabilizes or some stopping criterion of choice is reached.

We advocate a similar strategy to extend the application domain of FAST-SIS. In view of Section 2.1, a variable step using a working Lin-Ying model is intuitively sensible. We may also provide some formal justification. Firstly, estimation in a Lin-Ying model corresponds to optimizing the loss function

$$L(\boldsymbol{\beta}) := \boldsymbol{\beta}^\top \mathbf{D}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{d}; \tag{18}$$

where $\mathbf{D}$ was defined in Section 2.1. As discussed by Martinussen and Scheike (2009), the loss function (18) is meaningful for general hazard rate models: it is the empirical version of the mean squared prediction error for predicting, with a working Lin-Ying model, the part of the intensity which is orthogonal to the at-risk indicator. In the present context, we are mainly interested in the model selection properties of a working Lin-Ying model. Suppose that T_1 conditionally on $\mathbf{Z}_1$ follows a single-index model of the form (7) and that Assumptions 3-4 hold. Suppose that $\Delta\boldsymbol{\beta}^0 = \boldsymbol{\delta}$ with Δ the in probability limit of $\boldsymbol{D}$. Then $\alpha_j^0 \equiv 0$ implies $\beta_j^0 = 0$ (Hattori, 2006) so that a working Lin-Ying model will yield conservative model selection in a quite general setting. Under stronger assumptions, the following result, related to work by Brillinger (1983) and Li and Duan (1989), is available (see the appendix for a proof).

THEOREM 4. *Assume that T_1 conditionally on $\mathbf{Z}_1$ follows a single-index model of the form* (7)*. Suppose moreover that Assumption 2 holds and that C_1 is independent of $T_1, \mathbf{Z}_1$ (random censoring). If $\boldsymbol{\beta}^0$ defined by $\Delta\boldsymbol{\beta}^0 = \boldsymbol{\delta}$ is the vector of regression coefficients of the associated working Lin-Ying model and Δ is nonsingular, then there exists a nonzero constant ν depending only on the distributions of $\mathbf{Z}_1^\top\boldsymbol{\alpha}^0$ and C_1 such that $\boldsymbol{\beta}^0 = \nu\boldsymbol{\alpha}^0$.*

Thus a working Lin-Ying model can consistently estimate regression coefficient signs under misspecification. From the efforts of Zhu *et al.* (2009) and Zhu and Zhu (2009) for other types of single-index models, it seems conceivable that variable selection methods designed for the Lin-Ying model will enjoy certain consistency properties within the model class (7). The conclusion of Theorem 4 continues to hold when Δ is replaced by any matrix proportional to the feature covariance matrix Σ. This is a consequence of Assumption 2 and underlines the considerable flexibility available when estimating in single-index models.

Variable selection based on the Lin-Ying loss (18) can be accomplished by optimizing a penalized loss function of the form

$$\boldsymbol{\beta} \mapsto L(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|); \tag{19}$$

where $p_\lambda : \mathbb{R} \to \mathbb{R}$ is some nonnegative penalty function, singular at the origin to facilitate model selection (Fan and Li, 2001) and depending on some tuning parameter λ controlling the sparsity of the penalized estimator. A popular choice is the lasso penalty (Tibshirani, 2009) and its adaptive variant (Zou, 2006), corresponding to penalty functions $p_\lambda(|\beta_j|) = \lambda|\beta_j|$ and $p_\lambda(|\beta_j|) = \lambda|\beta_j|/|\hat{\beta}_j|$ with $\hat{\boldsymbol{\beta}}$ some root n consistent estimator of $\boldsymbol{\beta}^0$, respectively. These penalties were studied by Leng and Ma (2007) and Martinussen and Scheike (2009) for the Lin-Ying model. Empirically, we have had better success with the one-step SCAD (OS-SCAD) penalty of Zou and Li (2008) than with lasso penalties. Letting

$$w_\lambda(x) := \lambda \mathbb{1}(x \leqslant \lambda) + \frac{(a\lambda - x)_+}{a-1}\mathbb{1}(x > \lambda), \quad a > 2, \tag{20}$$

an OS-SCAD penalty function for the Lin-Ying model can be defined as follows:

$$p_\lambda(|\beta_j|) := w_\lambda\big(\bar{D}|\hat{\beta}_j|\big)|\beta_j|. \tag{21}$$

Here $\hat{\boldsymbol{\beta}} := \operatorname{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$ is the unpenalized estimator and $\bar{D} := p^{-1}\mathrm{tr}(\mathbf{D})$ is the average diagonal element of $\mathbf{D}$; this particular re-scaling is just one way to lessen dependency of the penalization on the time scale. If $\mathbf{D}$ has approximately constant diagonal (which is often the case for standardized features), then re-scaling by $\bar{D}$ leads to a similar penalty as for OS-SCAD in the linear regression model with standardized features. The choice $a = 3.7$ in (20) was recommended by Fan and Li (2001). OS-SCAD has not previously been explored for the Lin-Ying model but its favorable performance in ISIS for other regression models is well known (Fan *et al.*, 2009, 2010). OS-SCAD can be implemented efficiently using, for example, coordinate descent methods for fitting the lasso (Gorst-Rasmussen and Scheike, 2011; Friedman *et al.*, 2007). For fixed p, the OS-SCAD penalty (21) has the oracle property if the Lin-Ying model holds true. A proof is beyond scope but follows by adapting Zou and Li (2008) along the lines of Martinussen and Scheike (2009).

In the basic FAST-ISIS algorithm proposed below, the initial recruitment step corresponds to ranking the regression coefficients in the univariate Lin-Ying models. This is a convenient generic choice because it enables interpretation of the algorithm as standard ‘vanilla ISIS’ (Fan *et al.*, 2009) for the Lin-Ying model.

ALGORITHM 1 (Lin-Ying-FAST-ISIS). Set $\mathrm{M} := \{1,\dots,p\}$, let $r_{\max}$ be some pre-defined maximal number of iterations of the algorithm.

1. (*Initial recruitment*). Perform SIS by ranking $|d_j D_{jj}^{-1}|$, $1 \le j \le p$, according to decreasing order of magnitude and retain the $k_0 \le d$ most relevant features $\mathrm{A}_1 \subseteq \mathrm{M}$.
2. For $r = 1,2,\dots$ do:
   (a) (*Feature selection*). Define $\omega_j := \infty$ if $j \notin \mathrm{A}_r$ and $\omega_j := 1$ otherwise. Estimate
   $$\hat{\boldsymbol{\beta}} := \operatorname{argmin}_{\boldsymbol{\beta}} \Big\{ L(\boldsymbol{\beta}) + \sum_{j=1}^{p} \omega_j p_{\hat{\lambda}}(|\beta_j|) \Big\};$$
   with p_λ defined in (21) for some suitable $\hat{\lambda}$. Set $\mathrm{B}_r := \{j : \hat{\beta}_j \neq 0\}$.
   (b) If $r > 1$ and $\mathrm{B}_r = \mathrm{B}_{r-1}$, or if $r = r_{\max}$; return B_r.
   (c) (*Re-recruitment*). Otherwise, re-evaluate relevance of features in $\mathrm{M}\backslash\mathrm{B}_r$ according to the absolute value of their regression coefficient $|\tilde{\beta}_j|$ in the $|\mathrm{M}\backslash\mathrm{B}_r|$ unpenalized Lin-Ying models including each feature in $\mathrm{M}\backslash\mathrm{B}_r$ and all features in B_r, i.e.
   $$(22) \qquad \tilde{\beta}_j := \hat{\beta}_1^{(j)}, \quad \text{where } \hat{\boldsymbol{\beta}}^{(j)} = \operatorname{argmin}_{\boldsymbol{\beta}_{\{j\}\cup\mathrm{B}_r}} L\big(\boldsymbol{\beta}_{\{j\}\cup\mathrm{B}_r}\big), \quad j \in \mathrm{M}\backslash\mathrm{B}_r.$$
   Take $\mathrm{A}_{r+1} := \mathrm{C}_r \cup \mathrm{B}_r$ where C_r is the set of the k_r most relevant features in $\mathrm{M}\backslash\mathrm{B}_r$, ranked according to decreasing order of magnitude of $|\tilde{\beta}_j|$.

Fan and Lv (2008) recommended choosing d to be of order $n/\log n$. Following Fan *et al.* (2009), we may take $k_0 = \lfloor 2d/3 \rfloor$ and $k_r = d - |\mathrm{A}_r|$ at each step. This k_0 ensures that we complete at least one iteration of the algorithm; the choice of k_r for $r > 0$ ensures that at most d features are included in the final solution.

Algorithm 1 defines an iterated variant of SIS with the Lin-Ying-FAST statistic (16). We can devise an analogous iterated variant of Z-FAST-SIS in which the initial recruitment is performed by ranking based on the statistic (15), and the subsequent re-recruitments are performed by ranking $|Z|$-statistics in the multivariate Lin-Ying model according to decreasing order of magnitude, using the variance estimator (6). A third option would be to base recruitment on (17) and re-recruitments on the decrease in the multivariate loss (18) when joining a given feature to the set of features picked out in the variable selection step.

The re-recruitment step (2c) in Algorithm 1 resembles that of Fan *et al.* (2009). Its naive implementation will be computationally burdensome when p is large, requiring a low-dimensional matrix inversion per feature. Significant speedup over the naive implementation is possible via the matrix identity

$$\text{(23)} \qquad \mathbf{D} = \begin{pmatrix} e & \mathbf{f}^\top \\ \mathbf{f} & \tilde{\mathbf{D}} \end{pmatrix} \Rightarrow \mathbf{D}^{-1} = \begin{pmatrix} k^{-1} & -k^{-1}\mathbf{f}^\top\tilde{\mathbf{D}}^{-1} \\ -k^{-1}\tilde{\mathbf{D}}^{-1}\mathbf{f} & (\tilde{\mathbf{D}} - e^{-1}\mathbf{f}\mathbf{f}^\top)^{-1} \end{pmatrix} \quad \text{where } k = e - \mathbf{f}^\top\tilde{\mathbf{D}}^{-1}\mathbf{f}.$$

Note that only the first row of $\mathbf{D}^{-1}$ is required for the re-recruitment step so that (22) can be implemented using just a single low-dimensional matrix inversion alongside $O(p)$ matrix/vector multiplications. Combining (23) with (6), a similarly efficient implementation applies for Z-FAST-ISIS.

The variable selection step (2a) of Algorithm 1 requires the choice of an appropriate tuning parameter. This is traditionally a difficult part of penalized regression, particularly when the aim is model selection where methods such as cross-validation are prone to overfitting (Leng *et al.*, 2007). Previous work on ISIS used the Bayesian information criterion (BIC) for tuning parameter selection (Fan *et al.*, 2009). Although BIC is based on the likelihood, we may still define the following ‘pseudo BIC’ based on the loss (18):

$$\text{(24)} \qquad \mathrm{PBIC}(\lambda) = \kappa\{L(\hat{\boldsymbol{\beta}}_\lambda) - L(\hat{\boldsymbol{\beta}})\} + n^{-1}\mathrm{df}_\lambda \log n.$$

Here $\hat{\boldsymbol{\beta}}_\lambda$ is the penalized estimator, $\hat{\boldsymbol{\beta}}$ is the unpenalized estimator, $\kappa > 0$ is a scaling constant of choice, and df_λ estimates the degrees of freedom of the penalized estimator. A computationally convenient choice is $\mathrm{df}_\lambda = \|\hat{\boldsymbol{\beta}}_\lambda\|_0$ (Zou *et al.*, 2007). It turns out that choosing $\hat{\lambda} = \mathrm{argmin}_\lambda \mathrm{PBIC}_\lambda$ may lead to model selection consistency. Specifically, the loss (18) for the Lin-Ying model is of the least-squares type. Then we can repeat the arguments of Wang and Leng (2007) and show that, under suitable consistency assumptions for the penalized estimator, there exists a sequence $\lambda_n \to 0$ yielding selection consistency for $\hat{\boldsymbol{\beta}}_{\lambda_n}$ and satisfying

$$\text{(25)} \qquad \mathbb{P}\Big\{\inf_{\lambda \in S} \mathrm{PBIC}(\lambda) > \mathrm{PBIC}(\lambda_n)\Big\} \to 1, \quad n \to \infty;$$

with S the union of the set of tuning parameters λ which lead to overfitted (strict supermodels of the true model), respectively underfitted models (any model which do not include the true model). While (25) holds independently of the scaling constant κ, the finite-sample behavior of PBIC depends strongly on κ. A sensible value may be inferred heuristically as follows: the range of a ‘true’ likelihood BIC is asymptotically equivalent to a Wald statistic in the sense that (for fixed p),

$$\text{(26)} \qquad \mathrm{BIC}(0) - \mathrm{BIC}(\infty) = \hat{\boldsymbol{\beta}}_{\mathrm{ML}}^\top \mathrm{I}(\boldsymbol{\beta}_0)\hat{\boldsymbol{\beta}}_{\mathrm{ML}} + o_p(n^{-1/2});$$

with $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ the maximum likelihood estimator and $\mathrm{I}(\boldsymbol{\beta}_0) \approx n^{-1}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}} - \boldsymbol{\beta}_0)^{-1}$ the information matrix. We may specify κ by requiring that $\mathrm{PBIC}(0) - \mathrm{PBIC}(\infty)$ admits an analogous interpretation as a Wald statistic. Since $\mathrm{PBIC}(0) - \mathrm{PBIC}(\infty) = \kappa\mathbf{d}^\top\mathbf{D}^{-1}\mathbf{d} + o_p(n^{-1/2})$, it follows from (6) that we should choose

$$\kappa := (\mathbf{d}^\top\mathbf{B}^{-1}\mathbf{d})/(\mathbf{d}^\top\mathbf{D}^{-1}\mathbf{d}).$$

This choice of κ also removes the dependency of PBIC on the time scale.

5. Simulation studies

In this section, we investigate the performance of FAST screening on simulated data. Rather than comparing with popular variable selection methods such as the lasso, we will compare with analogous screening methods based on the Cox model (Fan *et al.*, 2010). This seems a more pertinent benchmark since previous work has already demonstrated that (iterated) SIS can outperform variable selection based on penalized regression in a number of cases (Fan and Lv (2008); Fan *et al.* (2009)).

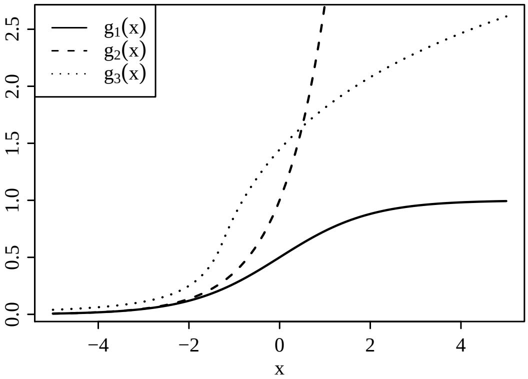

Figure 1. The three hazard rate link functions used in the simulation studies

For all the simulations, survival times were generated from three different conditionally exponential models of the generic form (7); that is, a time-independent hazard 'link function' applied to a linear functional of features. For suitable constants c, the link functions were as follows (see also Figure 1):

$$\begin{aligned} \text{Logit}: \lambda_{\text{logit}}(t,x) &:= \{1+\exp(c_{\text{logit}}x)\}^{-1} \\ \text{Cox}: \lambda_{\text{cox}}(t,x) &:= \exp(c_{\text{cox}}x) \\ \text{Log}: \lambda_{\text{log}}(t,x) &:= \log\{\mathrm{e}+(c_{\text{log}}x)^2\}\{1+\exp(c_{\text{log}}x)\}^{-1}. \end{aligned}$$

The link functions represent different characteristic effects on the feature functional, ranging from uniformly bounded (logit) over fast decay/increase (Cox), to fast decay/slow increase (log). We took $c_{\text{logit}} = 1.39$, $c_{\text{cox}} = 0.68$, and $c_{\text{log}} = 1.39$ and, unless otherwise stated, survival times were right-censored by independent exponential random variables with rate parameters 0.12 (logit link), 0.3 (Cox link) and 0.17 (log link). These constants were selected to provide a crude 'calibration' to make the simulation models more comparable: for a univariate standard Gaussian feature Z_1, a regression coefficient $\beta = 1$, and a sample size of $n = 300$, the expected $|Z|$-statistic was 8 for all three link functions with an expected censoring rate of 25%, as evaluated by numerical integration based on the true likelihood.

Methods for FAST screening have been implemented in the R-package 'ahaz' (Gorst-Rasmussen, 2011).

5.1. *Performance of FAST-SIS*

We first considered the performance of basic, non-iterated FAST-SIS. Features were generated as in scenario 1 of Fan and Song (2010). Specifically, let ε be standard

Gaussian. Define

$$Z_{1j} := \frac{\varepsilon_j + a_j\varepsilon}{\sqrt{1+a_j^2}}, \quad j = 1,\dots,p; \tag{27}$$

where ε_j is independently distributed as a standard Gaussian for $j = 1,2,\dots,\lfloor p/3 \rfloor$: independently distributed according to a double exponential distribution with location parameter zero and scale parameter 1 for $j = \lfloor p/3 \rfloor + 1,\dots,\lfloor 2p/3 \rfloor$; and independently distributed according to a Gaussian mixture $0.5N(-1,1)+0.5N(1,0.5)$ for $j = \lfloor 2p/3 \rfloor + 1,\dots,p$. The constants a_j satisfy $a_1 = \cdots = a_{15}$ and $a_j = 0$ for $j > 15$. With the choice $a_1 = \sqrt{\rho/(1-\rho)}$, $0 \leqslant \rho \leqslant 1$, we obtain $\mathrm{Cor}(Z_{1i}, Z_{1j}) = \rho$ for $i \neq j$, $i,j \leqslant 15$, enabling crude adjustment of the correlation structure of the feature distribution. Regression coefficients were chosen to be of the generic form $\boldsymbol{\alpha}^0 = (1,1.3,1,1.3,\dots)^\top$ with exactly the first s components nonzero.

For each combination of hazard link function, non-sparsity level s, and correlation ρ, we performed 100 simulations with p = 20,000 features and n = 300 observations. Features were ranked using the vanilla FAST statistic, the scaled FAST statistics (15) and (16), and SIS based on a Cox working model (Cox-SIS), the latter ranking features according their absolute univariate regression coefficient. Results are shown in Table 1. As a performance measure, we report the median of the minimum model size (MMMS) needed to detect all relevant features alongside its relative standard deviation (RSD), the interquartile range divided by 1.34. MMMS is a useful performance measure for this type of study since it eliminates the need to select a threshold parameter for SIS. The censoring rate in the simulations was typically 30%-40%.

Table 1. MMMS and RSD (in parentheses) for basic SIS with $n = 300$ and $p = 20{,}000$ (100 simulations).

		λ_{logit}			λ_{cox}			λ_{log}		
ρ		$s=3$	$s=6$	$s=9$	$s=3$	$s=6$	$s=9$	$s=3$	$s=6$	$s=9$
0	$\mathbf{d}$	3 (1)	32 (53)	530 (914)	3 (0)	7 (5)	45 (103)	3 (0)	22 (44)	202 (302)
	$\mathbf{d}^{\text{LY}}$	4 (1)	66 (95)	678 (939)	3 (0)	11 (14)	96 (176)	3 (1)	41 (87)	389 (466)
	$\mathbf{d}^{Z}$	3 (1)	40 (71)	522 (873)	3 (0)	7 (7)	48 (105)	3 (0)	22 (45)	262 (318)
	Cox	3 (1)	44 (68)	572 (928)	3 (0)	7 (4)	40 (117)	3 (0)	26 (51)	280 (306)
0.25	$\mathbf{d}$	3 (0)	6 (1)	11 (1)	3 (0)	6 (0)	9 (1)	3 (0)	6 (1)	10 (1)
	$\mathbf{d}^{\text{LY}}$	3 (0)	7 (1)	11 (2)	3 (0)	6 (1)	10 (1)	3 (0)	7 (1)	11 (1)
	$\mathbf{d}^{Z}$	3 (0)	6 (1)	11 (1)	3 (0)	6 (0)	10 (1)	3 (0)	6 (1)	10 (1)
	Cox	3 (0)	6 (1)	11 (1)	3 (0)	6 (0)	9 (1)	3 (0)	6 (1)	10 (1)
0.5	$\mathbf{d}$	3 (0)	7 (2)	12 (2)	3 (0)	6 (1)	10 (1)	3 (0)	7 (1)	11 (2)
	$\mathbf{d}^{\text{LY}}$	3 (0)	9 (3)	13 (1)	3 (0)	8 (2)	13 (2)	3 (0)	8 (2)	12 (2)
	$\mathbf{d}^{Z}$	3 (0)	8 (3)	12 (1)	3 (0)	7 (2)	12 (2)	3 (0)	7 (2)	12 (2)
	Cox	3 (1)	9 (3)	13 (2)	3 (0)	6 (1)	11 (2)	3 (0)	8 (2)	12 (2)
0.75	$\mathbf{d}$	3 (1)	9 (2)	13 (1)	3 (0)	8 (2)	12 (1)	3 (1)	9 (3)	12 (2)
	$\mathbf{d}^{\text{LY}}$	4 (2)	11 (3)	14 (2)	4 (1)	11 (3)	14 (1)	4 (2)	10 (2)	13 (1)
	$\mathbf{d}^{Z}$	4 (1)	10 (2)	13 (1)	3 (1)	10 (3)	13 (1)	3 (1)	9 (2)	13 (1)
	Cox	5 (3)	12 (2)	14 (1)	3 (0)	7 (2)	12 (2)	4 (1)	11 (3)	14 (2)

For all methods, the MMMS is seen to increase with feature correlation ρ and non-sparsity s. As also noted by Fan and Song (2010) for the case of SIS for generalized linear models, some correlation among features can actually be helpful since it increases the strength of marginal signals. Overall, the statistic $\mathbf{d}^{\text{LY}}$ seems to perform slightly worse than both $\mathbf{d}$ and $\mathbf{d}^{\text{Z}}$ whereas the latter two statistics perform similarly to Cox-SIS. In our basic implementation, screening with any of the FAST statistics was more than 100 times faster than Cox-SIS, providing a rough indication of the relative computational efficiency of FAST-SIS.

To gauge the relative difficulty of the different simulation scenarios, Figure 2 shows box plots of the minimum of the observed $|Z|$-statistics in the oracle model (the joint model with only the relevant features included and estimation based on the likelihood under the true link function) for the link function $\lambda_{\log}$. This particular link function represents an 'intermediate' level of difficulty; with $|Z|$-statistics for λ_{cox} generally being somewhat larger and $|Z|$-statistics for λ_{logit} being slightly smaller. Even with oracle information and the correct working model, these are evidently difficult data to deal with.

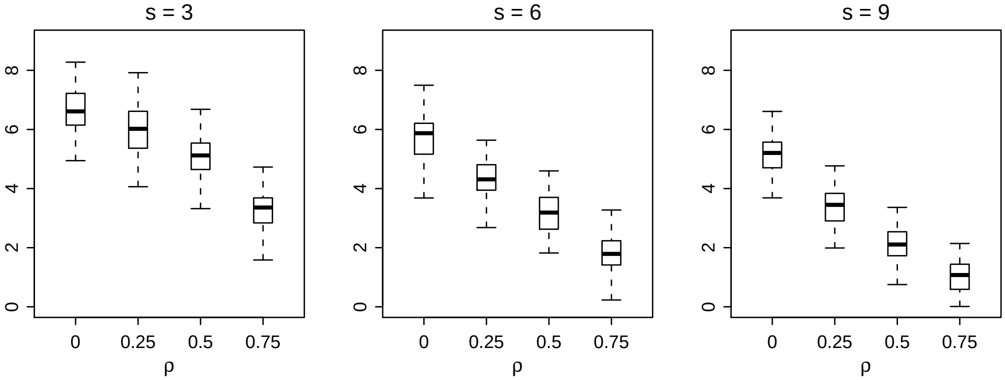

Figure 2. Minimum observed $|Z|$-statistics in the oracle model under $\lambda_{\log}$, for the SIS simulation study.

5.2. *FAST-SIS with non-Gaussian features and nonrandom censoring*

We next investigated FAST-SIS with non-Gaussian features and a more complex censoring mechanism. The simulation scenario was inspired by the previous section but with all features generated according to either a standard Gaussian distribution, a t-distribution with 4 degrees of freedom, or a unit rate exponential distribution. Features were standardized to have mean zero and variance one, and the feature correlation structure was such that $\text{Cor}(Z_{1i}, Z_{1j}) = 0.125$ for $i, j < 15$, $i \neq j$ and $\text{Cor}(Z_{1i}, Z_{1j}) = 0$ otherwise. Survival times were generated according to the link function $\lambda_{\log}$ with regression coefficients $\boldsymbol{\beta} = (1, 1.3, 1, 1.3, 1, 1.3, 0, 0, \ldots)$ while censoring times were generated according to the same model (link function $\lambda_{\log}$ and conditionally on the same feature realizations) with regression coefficients $\tilde{\boldsymbol{\beta}} = k\boldsymbol{\beta}$. The constant k controls the association between censoring and survival times, leading to a basic example of nonrandom censoring (competing risks).

Using $p = 20{,}000$ features and $n = 300$ observations, we performed 100 simulations under each of the three feature distributions, for different values of k. Table 2 reports the MMMS and RSD for the different screening methods of the previous section, and also for the statistic $\mathbf{d}^{\text{loss}}$ in (17). The censoring rate in all scenarios was around 50%.

From the column with $k = 0$ (random censoring), the heavier tailed t-distribution increases the MMMS, particularly for $\mathbf{d}^{\text{LY}}$. The vanilla FAST statistic $\mathbf{d}$ seems the least affected here, most likely because it does not directly involve second-order statistics which are poorly estimated due to the heavier tails. While $\mathbf{d}^Z$ and $\mathbf{d}^{\text{loss}}$ are also scaled by second-order statistics, the impact of the tails is dampened by the square-root transformation in the scaling factors. In contrast, the more distinctly non-Gaussian exponential distribution is problematic for $\mathbf{d}^Z$. Overall, the statistics $\mathbf{d}$ and $\mathbf{d}^{\text{loss}}$ seems to have the best and most consistent performance across feature distributions. Nonrandom censoring generally increases the MMMS and RSD, particularly for the non-Gaussian distributions. There appears to be no clear difference between the effect of positive and negative values of k. We found that the effect of $k \neq 0$ diminished when the sample size was increased (results not shown), suggesting that nonrandom censoring in the present example leads to a power rather than bias issue. This may not be surprising in view of the considerations below (14). However, the example still shows the dramatic impact of nonrandom censoring on the performance of SIS.

Table 2. MMMS and RSD (in parentheses) for SIS under non-Gaussian features/nonrandom censoring with $n = 300$ and $p = 20,000$ (100 simulations).

			k			
Feature distr.		$k = 0$	−0.5	−0.25	0.25	0.5
Gaussian	$\mathbf{d}$	6 (1)	8 (8)	7 (4)	6 (1)	7 (3)
	$\mathbf{d}^{LY}$	6 (1)	8 (6)	7 (3)	7 (2)	8 (5)
	$\mathbf{d}^Z$	6 (1)	7 (6)	7 (2)	6 (1)	7 (2)
	$\mathbf{d}^{\text{loss}}$	6 (1)	8 (6)	7 (3)	6 (1)	7 (3)
	Cox	6 (1)	8 (5)	7 (2)	6 (1)	7 (2)
t $(df = 4)$	$\mathbf{d}$	6 (1)	13 (17)	7 (5)	6 (1)	7 (3)
	$\mathbf{d}^{LY}$	11 (7)	12 (8)	9 (7)	48 (136)	99 (185)
	$\mathbf{d}^Z$	7 (3)	17 (20)	8 (5)	7 (2)	7 (3)
	$\mathbf{d}^{\text{loss}}$	6 (1)	8 (7)	7 (4)	8 (15)	10 (10)
	Cox	7 (4)	15 (23)	8 (10)	8 (4)	9 (5)
Exponential	$\mathbf{d}$	6 (1)	6 (2)	6 (1)	7 (4)	8 (7)
	$\mathbf{d}^{LY}$	6 (1)	11 (12)	7 (3)	6 (1)	6 (1)
	$\mathbf{d}^Z$	15 (10)	34 (36)	24 (17)	22 (28)	26 (29)
	$\mathbf{d}^{\text{loss}}$	6 (0)	7 (4)	6 (1)	6 (1)	6 (1)
	Cox	8 (4)	22 (31)	14 (11)	9 (6)	9 (8)

5.3. *Performance of FAST-ISIS*

We lastly evaluated the ability of FAST-ISIS (Algorithm 1) to cope with scenarios where FAST-SIS fails. As in the previous sections, we compare our results with the analogous ISIS screening method for the Cox model. To perform Cox-ISIS, we used the R package 'SIS', with (re)recruitment based on the absolute Cox regression coefficients and variable selection based on OS-SCAD. We also compared with Z-FAST-ISIS variant described below Algorithm 1 in which (re)recruitment is based on the Lin-Ying model $|Z|$-statistics (results for FAST-ISIS with (re)recruitment based on the loss function were very similar).

For the simulations, we adopted the structural form of the feature distributions used by Fan *et al.* (2010). We considered $n = 300$ observations and $p = 500$ features which were jointly Gaussian and marginally standard Gaussian. Only regression coefficients and feature correlations differed between cases as follows:

(a) The regression coefficients are $\beta_1 = -0.96$, $\beta_2 = 0.90$, $\beta_3 = 1.20$, $\beta_4 = 0.96$, $\beta_5 = -0.85$, $\beta_6 = 1.08$ and $\beta_j = 0$ for $j > 6$. Features are independent, $\mathrm{Cor}(Z_{1i}, Z_{1j}) = 0$ for $i \neq j$.

(b) Regression coefficients are the same as in (a) while $\mathrm{Corr}(Z_{1i}, Z_{1j}) = 0.5$ for $i \neq j$.

(c) Regression coefficients are $\beta_1 = \beta_2 = \beta_3 = 4/3$, $\beta_4 = -2\sqrt{2}$. The correlation between features is $\mathrm{Cor}(Z_{1,4}, Z_{1j}) = 1/\sqrt{2}$ for $j \neq 4$ and $\mathrm{Cor}(Z_{1i}, Z_{1j}) = 0.5$ for $i \neq j$, $i, j \neq 4$.

(d) Regression coefficients are $\beta_1 = \beta_2 = \beta_3 = 4/3$, $\beta_4 = -2\sqrt{2}$ and $\beta_5 = 2/3$. The correlation between features is $\mathrm{Cor}(Z_{1,4}, Z_{1j}) = 1/\sqrt{2}$ for $j \notin \{4,5\}$, $\mathrm{Cor}(Z_{1,5}, Z_{1j}) = 0$ for $j \neq 5$, and $\mathrm{Cor}(Z_{1i}, Z_{1j}) = 0.5$ for $i \neq j$, $i, j \notin \{4,5\}$.

Case (a) serves as a basic benchmark whereas case (b) is harder because of the correlation between relevant and irrelevant features. Case (c) introduces a strongly relevant feature Z_4 which is not marginally associated with survival; lastly, case (d) is similar to case (c) but also includes a feature Z_5 which is weakly associated with survival and does not 'borrow' strength from its correlation with other relevant features.

Following Fan *et al.* (2010), we took $d = \lfloor n/\log n/3 \rfloor = 17$ for the initial dimension reduction; performance did not depend much on the detailed choice of d of order $n/\log n$. For the three different screening methods, ISIS was run for a maximum of 5 iterations. (P)BIC was used for tuning the variable selection steps. Results are shown in Table 3, summarized over 100 simulations. We report the average number of truly relevant features selected by ISIS and the average final model size, alongside standard deviations in parentheses. To provide an idea of the improvement over basic SIS, we also report the median of the minimum model size (MMMS) for the initial SIS step (based on vanilla FAST-SIS only). The censoring rate in the different scenarios was 25%-35%.

Table 3. Simulation results for ISIS with $n = 300$, $p = 500$ and $d = 17$ (100 simulations). Numbers in parentheses are standard deviations (or relative standard deviation, for the MMMS).

			Average no. true positives (ISIS)			Average model size (ISIS)		
Link	Case	MMMS (RSD)	LY-FAST	Z-FAST	Cox	LY-FAST	Z-FAST	Cox
λ_{logit}	(a)	7 (3)	6.0 (0)	6.0 (0)	5.5 (1)	7.8 (1)	7.9 (2)	6.3 (2)
	(b)	500 (1)	5.5 (1)	5.5 (1)	3.4 (1)	7.0 (2)	6.7 (2)	4.3 (2)
	(c)	240 (125)	3.7 (1)	3.8 (1)	3.0 (2)	5.2 (2)	5.7 (3)	4.5 (4)
	(d)	230 (124)	4.8 (1)	4.7 (1)	3.5 (2)	5.9 (2)	6.2 (3)	4.9 (4)
λ_{cox}	(a)	7 (1)	6.0 (0)	6.0 (0)	6.0 (0)	7.5 (1)	7.5 (1)	6.2 (1)
	(b)	500 (1)	5.8 (1)	5.8 (1)	5.6 (1)	7.2 (2)	6.8 (1)	6.4 (2)
	(c)	218 (120)	3.7 (1)	3.6 (1)	3.0 (2)	5.1 (3)	5.3 (3)	4.9 (4)
	(d)	258 (129)	4.9 (1)	4.8 (1)	3.8 (2)	6.3 (2)	6.0 (2)	6.4 (5)
λ_{log}	(a)	6 (1)	6.0 (0)	6.0 (0)	6.0 (0)	7.3 (1)	7.4 (1)	6.3 (1)
	(b)	500 (1)	5.8 (1)	5.7 (1)	4.9 (1)	7.2 (2)	6.7 (1)	5.7 (2)
	(c)	252 (150)	3.9 (0)	3.9 (1)	3.4 (1)	5.3 (2)	4.9 (2)	5.5 (5)
	(d)	223 (132)	4.9 (1)	4.8 (1)	4.0 (2)	6.0 (2)	6.1 (2)	5.9 (5)

The overall performance of the three ISIS methods is comparable between the different cases. All methods deliver a dramatic improvement over non-iterated SIS, but no one method performs significantly better than the others. The two FAST-ISIS methods have a surprisingly similar performance. As one would expect, Cox-ISIS does particularly well under the link function λ_{cox} but does not appear to be uniformly better than the two FAST-ISIS methods even in this ideal setting. Under the link function

λ_{logit}, both FAST-ISIS methods outperform Cox-ISIS in terms of the number of true positives identified, as do they for the link function λ_{log}, although less convincingly. On the other hand, the two FAST-ISIS methods generally select slightly larger models than Cox-ISIS and their false-positive rates (not shown) are correspondingly slightly larger. FAST-ISIS was 40-50 times faster than Cox-ISIS, typically completing calculations in 0.5-1 seconds in our specific implementation. Figure 3 shows box plots of the minimum of the observed $|Z|$-statistics in the oracle model (based on the likelihood under the true model).

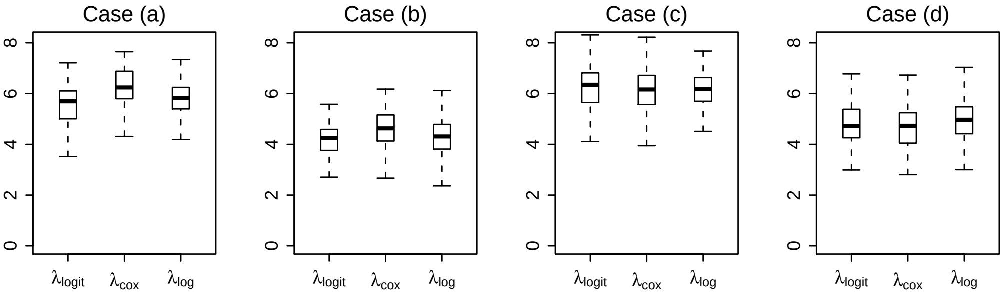

Figure 3. Minimum observed $|Z|$-statistics in the oracle models for the FAST-ISIS simulation study.

We have experimented with other link functions and feature distributions than those described above (results not shown). Generally, we found that Cox-ISIS performs worse than FAST-ISIS for bounded link functions. The observation from Table 3, that FAST-ISIS may improve upon Cox-ISIS even under the link function λ_{cox}, does not necessarily hold when the signal strength is increased. Then Cox-ISIS will be superior, as expected. Changing the feature distribution to one for which the linear regression property (Assumption 2) does not hold leads to a decrease in the overall performance for all three ISIS methods.

6. Application to AML data

The study by Metzeler *et al.* (2008) concerns the development and evaluation of a prognostic gene expression marker for overall survival among patients diagnosed with cytogenetically normal acute myeloid leukemia (CN-AML). A total of 44,754 gene expressions were recorded among 163 adult patients using Affymetrix HG-U133 A1B microarrays. Based the method of supervised principal components (Bair and Tibshirani, 2004), the gene expressions were used to develop an 86-gene signature for predicting survival. The signature was validated on an external test data set consisting of 79 patients profiled using Affymetrix HG-U133 Plus 2.0 microarrays. All data is publicly available on the Gene Expression Omnibus web site (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE12417. The CN-AML data was recently used by Benner *et al.* (2010) for comparing the performance of variable selection methods.

Median survival time was 9.7 months in the training data (censoring rate 37%) and 17.7 months in the test data (censoring rate 41%). Preliminary to analysis, we followed the scaling approach employed by Metzeler *et al.* (2008) and centered the gene expressions separately within the test and training data set, followed by a scaling of the training data with respect to the test data.

We first applied vanilla FAST-SIS to the $n = 163$ patients in the training data to reduce the dimension from $p = 44,754$ to $d = \lfloor n/\log(n) \rfloor = 31$. We then used OS-SCAD to select a final set among these 31 genes. Since the PBIC criterion can be somewhat conservative in practice, we selected the OS-SCAD tuning parameter using 5-fold cross-validation based on the loss function (18). Specifically, using a random split of $\{1,\dots,163\}$ into folds $F_1,\dots,F_5$ of approximately equal size, we chose λ as:

$$\hat{\lambda} = \mathrm{argmin}_\lambda \sum_{i=1}^{5} L^{(F_i)}\{\hat{\boldsymbol{\beta}}_{-F_i}(\lambda)\};$$

with $L^{(F_i)}$ the loss function using only observations from F_i and $\hat{\boldsymbol{\beta}}_{-F_i}(\lambda)$ the regression coefficients estimated for a tuning parameter λ, omitting observations from F_i. This approach yielded a set of 7 genes, 5 of which also appeared in the signature of Metzeler *et al.* (2008). For $\hat{\boldsymbol{\beta}}$ the estimated penalized regression coefficients, we calculated a risk score $\boldsymbol{Z}_j^\top \hat{\boldsymbol{\beta}}$ for each patient in the test data. In a Cox model, the standardized risk score had a hazard ratio of 1.69 ($p = 6 \cdot 10^{-4}$; Wald test). In comparison, lasso based on the Lin-Ying model (Leng *et al.* (2007); Martinussen and Scheike (2009)) with 5-fold cross-validation gave a standardized risk score with a hazard ratio of 1.56 ($p = 0.003$; Wald test) in the test data, requiring 5 genes; Metzeler *et al.* (2008) reported a hazard ratio of 1.85 ($p = 0.002$) for their 86-gene signature.

We repeated the above calculations for the three scaled versions of the FAST statistic (15)-(17). Since assessment of prediction performance using only a single data set may be misleading, we also validated the screening methods via leave-one-out (LOO) cross-validation based on the 163 patients in the training data. For each patient j, we used FAST-SIS as above (or Lin-Ying lasso) to obtain regression coefficients $\hat{\boldsymbol{\beta}}_{-j}$ based on the remaining 162 patients and defined the jth LOO risk score as the percentile of $\mathbf{Z}_j^\top \hat{\boldsymbol{\beta}}_{-j}$ among $\{\mathbf{Z}_i^\top \hat{\boldsymbol{\beta}}_{-j}\}_{i \neq j}$. We calculated Wald p-values in a Cox regression model including the LOO score as a continuous predictor. Results are shown in Table 4 while Table 5 shows the overlap between gene sets selected in the training data. There is seen to be some overlap between the different methods, particularly between vanilla FAST-SIS and the lasso, and many of the selected genes also appear in the signature of Metzeler *et al.* (2008). In the test data, the prediction performance of the different screening methods was comparable whereas the lasso had a slight edge in the LOO calculations. Lin-Ying SIS selected only a single gene in the test data and typically selected no genes in the LOO calculations. We found FAST screening to be slightly more sensitive to the cross-validation procedure than the lasso.

We next evaluated the extent to which iterated FAST-SIS might improve upon the above results. From our limited experience with applying ISIS to real data, instability can become an issue when several iterations of ISIS are run; particularly when cross-validation is involved. Accordingly, we ran only a single iteration of ISIS using Z-FAST-ISIS. The algorithm kept 2 of the genes from the first FAST-SIS round and selected 3 additional genes so that the total number of genes was 5. Calculating in the test data a standardized risk score based on the final regression coefficients, we obtained a Cox hazard ratio of only 1.06 ($p = 0.6$; Wald test) which is no improvement over non-iterated FAST-SIS. A similar conclusion was reached for the corresponding LOO calculations in the training data which gave a Cox Wald p-value of 0.001 for the LOO risk score, using a median of 4 genes. None of the other FAST-ISIS methods lead to improved prediction performance compared to their non-iterated counterparts. FAST-ISIS runs swiftly on this large data set: one iteration of the algorithm (re-recruitment and OS-SCAD feature selection with 5-fold cross-validation) completes in under 5 seconds on a standard laptop.

Altogether, the example shows that FAST-SIS can compete with a computationally more demanding full-scale variable selection method in the sense of providing similarly sparse models with competitive prediction properties. FAST-ISIS, while computationally very feasible, did not seem to improve prediction performance over simple independent screening in this particular data set.

Table 4. Prediction performance of FAST-SIS and Lin-Ying lasso in the AML data, evaluated in terms of the Cox hazard ratio for the standardized continuous risk score. The LOO calculations are based on the training data only.

		Screening method				
Scenario	Summary statistic	$\mathbf{d}$	$\mathbf{d}^{\mathrm{LY}}$	$\mathbf{d}^{\lvert Z \rvert}$	$\mathbf{d}^{\mathrm{loss}}$	Lasso
Test data	Hazard ratio	1.69	1.59	1.46	1.58	1.54
	p-value	$6 \cdot 10^{-4}$	0.0007	0.01	0.002	0.004
	No. predictors	7	1	3	7	5
LOO	p-value	$4 \cdot 10^{-7}$	0.16	$5 \cdot 10^{-5}$	$4 \cdot 10^{-4}$	$4 \cdot 10^{-8}$
	Median no. predictors	7	0	3	5	5

Table 5. Overlap between gene sets selected by the different screening methods and the signature of Metzeler *et al.* (2008).

Screening method	$\mathbf{d}$	$\mathbf{d}^{\mathrm{LY}}$	$\mathbf{d}^{\lvert Z \rvert}$	$\mathbf{d}^{\mathrm{loss}}$	Lasso	Metzeler
$\mathbf{d}$	7	0	1	2	4	5
$\mathbf{d}^{\mathrm{LY}}$		1	0	0	0	0
$\mathbf{d}^{\lvert Z \rvert}$			3	2	2	2
$\mathbf{d}^{\mathrm{loss}}$				7	2	5
Lasso					5	5
Metzeler						86

7. Discussion

Independent screening – the general idea of looking at the effect of one feature at a time – is a well-established method for dimensionality reduction. It constitutes a simple and excellently scalable approach to analyzing high-dimensional data. The SIS property introduced by Fan and Lv (2008) has enabled a basic formal assessment of the reasonableness of general independent screening methods. Although the practical relevance of the SIS property has been subject to scepticism (Roberts, 2008), the formal context needed to develop the SIS property is clearly useful for identifying the many implicit assumptions made when applying univariate screening methods to multivariate data.

We have introduced a SIS method for survival data based on the notably simple FAST statistic. In simulation studies, FAST-SIS performed on par with SIS based on the popular Cox model, while being considerably more amenable to analysis. We have shown that FAST-SIS may admit the formal SIS property within a class of single-index hazard rate models. In addition to assumptions on the feature distribution which are well known in the literature, a principal assumption for the SIS property to hold is that censoring times do not depend on the relevant features nor survival. While such partially random censoring may be appropriate to assume in many clinical settings, it indicates that additional caution is called for when applying univariate screening and competing risks are suspected.

A formal consistency property such as the SIS property is but one aspect of a statistical method and does not make FAST-SIS universally preferable. Not only is the SIS property unlikely to be unique to FAST screening, but different screening methods often highlight different aspects of data (Ma and Song, 2011), making it impossible and undesirable to recommend one generic method. We do, however, consider FAST-SIS a good generic choice of initial screening method for general survival data. Ultimately, the initial choice of a statistical method is likely to be made on the basis of parsimony, computational speed, and ease of implementation. The FAST statistic is about as difficult to evaluate as a collection of correlation coefficients while iterative FAST-SIS only requires solving one linear system of equations. This yields substantial computational savings over methods not sharing the advantage of linearity of estimating equations.

Iterated SIS has so far been studied to a very limited extent in an empirical context. The iterated approach works well on simulated data, but it is not obvious whether this necessarily translates into good performance on real data. In our example involving a large gene expression data set, ISIS did not improve results in terms of prediction accuracy. Several issues may affect the performance of ISIS on real data. First, it is our experience that the 'Rashomon effect', the multitude of well-fitting models (Breiman, 2001), can easily lead to stability issues for this type of forward selection. Second, it is often difficult to choose a good tuning parameter for the variable selection part of ISIS. Using BIC may lead to overly conservative results, whereas cross-validation may lead to overfitting when only the variable selection step – and not the recruitment steps – are cross-validated. He and Lin (2011) recently discussed how to combine ISIS with stability selection (Meinshausen and Bühlmann, 2010) in order to tackle instability issues and to provide a more informative output than the concise 'list of indices' obtained from standard ISIS. Their proposed scheme requires running many subsampling iterations of ISIS, a purpose for which FAST-ISIS will be ideal because of its computational efficiency. The idea of incorporating stability considerations is also attractive from a foundational point of view, being a pragmatic departure from the limiting *de facto* assumption that there is a single, true model. Investigation of such computationally intensive frameworks, alongside a study of the behavior of ISIS on a range of different real data sets, is a pertinent future research topic.

A number of other extensions of our work may be of interest. We have focused on the important case of time-fixed features and right-censored survival times but the FAST statistic can also be used with time-varying features alongside other censoring and truncation mechanism supported by the counting process formalism. Theoretical analysis of such extensions is a relevant future research topic, as is analysis of more flexible, time-dependent scaling strategies for the FAST statistic. Fan *et al.* (2011) recently discussed SIS where features enter in nonparametric, smooth manner, and an extension of their framework to FAST-SIS appears both theoretically and computationally feasible. Lastly, the FAST statistic is closely related to the univariate regression coefficients in the Lin-Ying model which is rather forgiving towards misspecification: under feature independence, the univariate estimator is consistent whenever the particular feature under investigation enters the hazard rate model as a linear function of regression coefficients (Hattori, 2006). The Cox model does not have a similar property (Struthers and Kalbfleisch, 1986). Whether such internal consistency under misspecification or lack hereof affects screening in a general setting is an open question.

Appendix: proofs

In addition to Assumptions 1-4 stated in the main text, we will make use of the following assumptions for the quantities defining the class of single-index hazard rate models (7):

A. $\mathbb{E}(Z_{1j}) = 0$ and $\mathbb{E}(Z_{1j}^2) = 1$, $j = 1, \ldots, p_n$.

B. $\mathbb{P}\{Y_1(\tau) = 1\} > 0$.

C. $\operatorname{Var}(\mathbf{Z}_1^\top \boldsymbol{\alpha}^0)$ is uniformly bounded above.

The details in Assumption A are included mainly for convenience; it suffices to assume that $\mathbb{E}(Z_{1j}^2) < \infty$.

Our first lemma is a basic symmetrization result, included for completeness.

LEMMA A1. *Let X be a random variable with mean μ and finite variance σ^2. For $t > \sqrt{8}\sigma$, it holds that $\mathbb{P}(|X-\mu| > t) \leqslant 4\mathbb{P}(|X| > t/4)$.*

Proof. First note that when $t > \sqrt{8}\sigma$ we have $\mathbb{P}(|X-\mu| > t/2) \leqslant 1/2$, by Chebyshev's inequality. Let X' be an independent copy of X. Then

$$2\mathbb{P}(|X| \geqslant t/4) \geqslant \mathbb{P}(|X'-X| > t/2) \geqslant \mathbb{P}(|X-\mu| > t \wedge |X'-\mu| \leqslant t/2). \tag{A1}$$

But

$$\mathbb{P}(|X-\mu| > t \wedge |X'-\mu| \leqslant t/2) = \mathbb{P}(|X-\mu| > t)\mathbb{P}(|X'-\mu| \leqslant t/2) \geqslant \frac{1}{2}\mathbb{P}(|X-\mu| > t).$$

Combining this with (A1), the statement of the lemma follows. ■

The next lemma provides a universal exponential bound for the FAST statistic and is of independent interest. It bears some similarity to exponential bounds reported by Bradic *et al.* (2010) for the Cox model.

LEMMA A2. *Under Assumptions A-B there exists constants $C_1, C_2 > 0$ independent of n such that for any $K > 0$ and $1 \leqslant j \leqslant p_n$, it holds that*

$$\mathbb{P}\{n^{1/2}|d_j - \delta_j| > C_1(1+t)\} \leqslant 10\exp\{-t^2/(2K^2)\} + C_2\exp(-n/2) + n\mathbb{P}(|Z_{1j}| > K).$$

Proof. Fix j throughout. Assume first that $|Z_{ij}| \leqslant K$ for some finite K. Define the random variables

$$A_n := n^{-1}\sum_{i=1}^n \int_0^\tau \{Z_{ij} - e_j(t)\}\mathrm{d}N_i(t), \quad B_n := \int_0^\tau \{e_j(t) - \bar{Z}_j(t)\}\mathrm{d}\bar{N}(t);$$

where $\bar{N}(t) := n^{-1}\{N_1(t) + \cdots + N_n(t)\}$ and $e_j(t) := \mathbb{E}\{\bar{Z}_j(t)\}$. Then we can write

$$n^{1/2}(d_j - \delta_j) = n^{1/2}\{A_n - \mathbb{E}(A_n)\} + n^{1/2}\{B_n - \mathbb{E}(B_n)\}.$$

We will deal with each term in the display separately. Since $\mathrm{d}N_i(t) \leqslant 1$, it holds that

$$|A_n| \leqslant \max_{1\leqslant i\leqslant n} |Z_{ij}| + \|e_j\|_\infty \leqslant 2K$$

and Hoeffding's inequality (Hoeffding, 1963) implies

$$\mathbb{P}(n^{1/2}|A_n - \mathbb{E}(A_n)| > t) \leqslant 2\exp\{-t^2/(2K^2)\}. \tag{A2}$$

Obtaining an analogous bound for $n^{1/2}\{B_n - \mathbb{E}(B_n)\}$ requires a more detailed analysis. Since $\mathrm{d}\bar{N}(t) \leqslant 1$,

$$|B_n| \leqslant \int_0^\tau |e_j(t) - \bar{Z}_j(t)|\mathrm{d}\bar{N}(t) \leqslant \|e_j - \bar{Z}_j\|_\infty. \tag{A3}$$

We will obtain an exponential bound for the right-hand side via empirical process methods. Define $E^{(k)}(t) := n^{-1}\sum_{i=1}^n Z_{ij}^k Y_i(t)$ and $e^{(k)}(t) := \mathbb{E}\{E^{(k)}(t)\}$ for $k = 0,1$. Set $\varepsilon := \inf_{t\in[0,\tau]} e^{(0)}(t)$ and observe that $0 < \varepsilon \leqslant 1$, by Assumption B. Moreover, by Cauchy-Schwartz's inequality,

$$\|e^{(1)}/e^{(0)}\|_\infty \leqslant m^{-1}\sqrt{\mathbb{E}|Z_{1j}|^2\|e^{(0)}\|_\infty} \leqslant \varepsilon^{-1}.$$

Define $\Omega_n := \{\inf_{t\in[0,\tau]} E^{(0)}(t) \geqslant \varepsilon/2\}$ and let 1_{Ω_n} be the indicator of this event. In view of the preceding display, we can write

$$|\bar{Z}_j(t) - e_j(t)|1_{\Omega_n} \leqslant \frac{1}{E^{(0)}(t)}\left\{\left|\frac{e^{(1)}(t)}{e^{(0)}(t)}\right||e^{(0)}(t) - E^{(0)}(t)| + |E^{(1)}(t) - e^{(1)}(t)|\right\}1_{\Omega_n} \tag{A4}$$

$$\leqslant 2\varepsilon^{-2}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_0} + \|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_1})1_{\Omega_n} \tag{A5}$$

with function classes $\mathrm{F}_k := \{t \mapsto Z^k 1(T \geqslant t \wedge C \geqslant t)\}$. We proceed to establish exponential bounds for the empirical process suprema in (A5). Each of the F_ks are Vapnik-Cervonenkis subgraph classes, and from Pollard (1989) there exists some finite constant ζ depending only on intrinsic properties of the F_ks such that

$$\mathbb{E}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_k}^2) \leqslant \zeta n^{-1}\mathbb{E}(Z_{1j}^2) = n^{-1}\zeta. \tag{A6}$$

In particular, it also holds that $\mathbb{E}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_k}) \leqslant n^{-1/2}\zeta^{1/2}$. Moreover,

$$|Z_{1j}^k 1(T_1 \geqslant t \wedge C_1 \geqslant t) - Z_{1j}^k 1(T_1 \geqslant s \wedge C_1 \geqslant s)|^2 \leqslant K^{2k}, \quad s,t \in [0,\tau], \quad k = 0,1.$$

With $k_1 := \zeta^{1/2}$, the concentration theorem of Massart (2000) implies

$$\mathbb{P}\{n^{1/2}\|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_k} > k_1(1+t)\} \leqslant \exp\{-t^2/(2K^2)\}, \quad k = 0,1. \tag{A7}$$

Combining (A3)-(A5), taking $k_2 := k_1\varepsilon^2/2$, we obtain

$$\mathbb{P}[\{n^{1/2}|B_n| > k_2(1+t)\} \cap \Omega_n] \leqslant 2\exp\{-t^2/(2K^2)\}. \tag{A8}$$

whereas (A5) and Cauchy-Schwarz's inequality imply

$$\mathbb{E}(B_n^2 1_{\Omega_n}) \leqslant \mathbb{E}(\|\bar{Z}_j - e_j\|_\infty^2 1_{\Omega_n}) \leqslant 4\varepsilon^{-4}\mathbb{E}\{(\|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_0} + \|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_1})^2 1_{\Omega_n}\} \leqslant 12\varepsilon^{-4}\zeta n^{-1}.$$

Combining Lemma A1 and (A8), there exists nonnegative k_3 (depending only on ε and ζ) such that

$$\mathbb{P}\{n^{1/2}|B_n - \mathbb{E}(B_n)| \geqslant k_3(1+t)\} \leqslant 8\exp\{-t^2/(2K^2)\} + \mathbb{P}(\Omega_n^c). \tag{A9}$$

To bound $\mathbb{P}(\Omega_n^c)$, recall that $e^{(0)}(t) \geqslant \varepsilon$ by assumption. Consequently,

$$\Omega_n^c \subseteq \{|E^{(0)}(t) - e^{(0)}(t)| > \varepsilon/2 \text{ for some } t\} \subseteq \{\|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_0} > \varepsilon/2\}.$$

By (A6), we have $\mathbb{E}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_0}) \leqslant \varepsilon/4$ eventually. By another application of the concentration theorem (Massart, 2000), there exists finite k_4 so that $\mathbb{P}\{\|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_0} > \varepsilon/4(1+t)\} \leqslant k_4 \exp(-nt^2/2)$. Setting $t = 1$,

$$\mathbb{P}(\Omega_n^c) \leqslant \mathbb{P}\{\|\mathbb{P}_n - \mathbb{P}\|_{\mathrm{F}_0} > \varepsilon/2\} \leqslant k_4 \exp(-n/2).$$

Substituting this bound in (A9) and combining with (A2), omitting now the assumption that Z_{ij} is bounded, it follows that there exists constants $C_1, C_2 > 0$ such that for any $K > 0$ and $t > 0$,

$$\mathbb{P}\{n^{1/2}|d_j - \delta_j| > C_1(1+t)\} \leqslant 10\exp\{-t^2/(2K^2)\} + C_2\exp(-n) + \mathbb{P}\left(\max_{1\leqslant i\leqslant n} |Z_{ij}| > K\right).$$

The statement of the lemma then follows from the union bound. ■

LEMMA A3. *Suppose that Assumptions A-B hold and that there exists positive constants l_0, l_1, η such that $\mathbb{P}(|Z_{1j}| > s) \leqslant l_0 \exp(-l_1 s^\eta)$ for sufficiently large s. If $\kappa < 1/2$ then for any $k_1 > 0$ there exists $k_2 > 0$ such that*

$$\mathbb{P}\left(\max_{1\leqslant j\leqslant p_n} |d_j - \delta_j| > k_1 n^{-\kappa}\right) \leqslant O[p_n \exp\{-k_2 n^{(1-2\kappa)\eta/(\eta+2)}\}]. \tag{A10}$$

Suppose in addition that $|\delta_j| > k_3 n^{-\kappa}$ whenever $j \in \mathrm{M}_\delta^n$ and that $\gamma_n = k_4 n^{-\kappa}$ where k_3, k_4 are positive constants and $k_4 \leqslant k_3/2$. Then

$$\mathbb{P}(\mathrm{M}_\delta^n \subseteq \widehat{\mathrm{M}}_d^n) \geqslant 1 - O[p_n \exp\{-k_2 n^{(1-2\kappa)\eta/(\eta+2)}\}]. \tag{A11}$$

In particular, if $\log p_n = o\{n^{(1-2\kappa)\eta/(\eta+2)}\}$ *then* $\mathbb{P}(\mathrm{M}_\delta^n \subseteq \widehat{\mathrm{M}}_d^n) \to 1$ *when* $n \to \infty$.

Proof. In Lemma A2, take $1 + t = k_1 n^{1/2-\kappa}/C_1$ and $K := n^{(1-2\kappa)/(\eta+2)}$. Then there exists positive constants $\tilde{k}_2, \tilde{k}_3$ such that for each $j = 1, \ldots, p_n$,

$$\mathbb{P}(|d_j - \delta_j| > k_1 n^{-\kappa}) \leqslant 10\exp\{-\tilde{k}_2 n^{(1-2\kappa)\eta/(\eta+2)}\} + C_2 \exp(-C_3 n) + n l_0 \exp\{-\tilde{k}_3 n^{(1-2\kappa)\eta/(\eta+2)}\}.$$

By the union bound, there exists $k_2 > 0$ such that

$$\mathbb{P}\left(\max_{1\leqslant j\leqslant p_n} |d_j - \delta_j| > k_1 n^{-\kappa}\right) \leqslant O[p_n \exp\{-k_2 n^{(1-2\kappa)\eta/(\eta+2)}\}];$$

which proves (A10). Concerning (A11), $k_3 n^{-\kappa} - |d_j| \leqslant |\delta_j - d_j|$ when $j \in \mathrm{M}_\delta^n$ by assumption and so

$$\mathbb{P}\left(\min_{j\in \mathrm{M}_\delta^n} |d_j| < \gamma_n\right) \leqslant \mathbb{P}\left(\max_{j\in \mathrm{M}_\delta^n} |d_j - \delta_j| \geqslant k_3 n^{-\kappa} - \gamma_n\right) \leqslant \mathbb{P}\left(\max_{j\in \mathrm{M}_\delta^n} |d_j - \delta_j| \geqslant n^{-\kappa} k_3/2\right);$$

where the last inequality follows since we assume $k_4 \leqslant k_3/2$. Taking $k_1 = k_3/2$ in (A10), we arrive at the desired conclusion:

$$\mathbb{P}(\mathrm{M}_\delta^n \subseteq \widehat{\mathrm{M}}_d^n) \geqslant 1 - \mathbb{P}\left(\min_{j \in \mathrm{M}_\delta^n} |d_j| < \gamma_n\right) \geqslant 1 - O[p_n \exp\{-k_2 n^{(1-2\kappa)\eta/(\eta+2)}\}].$$

Finally, $\mathbb{P}(\mathrm{M}_\delta^n \subseteq \widehat{\mathrm{M}}_d^n) \to 1$ when $n \to \infty$ follows immediately when $\log p_n = o\{n^{(1-2\kappa)\eta/(\eta+2)}\}$. ■

LEMMA A4. *Let* $\mathbf{Z} \in \mathbb{R}^p$ *be a random vector with zero mean and covariance matrix* Σ. *Let* $\mathbf{b} \in \mathbb{R}^p$ *and suppose that* $\mathbb{E}(\mathbf{Z}|\mathbf{Z}^\top \mathbf{b}) = \mathbf{c}\mathbf{Z}^\top \mathbf{b}$ *for some constant vector* $\mathbf{c} \in \mathbb{R}^p$. *Assume that* f *is some real function. Then*

$$\mathbb{E}\{\mathbf{Z} f(\mathbf{Z}^\top \mathbf{b})\} = \Sigma \mathbf{b} \frac{\mathbb{E}\{\mathbf{Z}^\top \mathbf{b} f(\mathbf{Z}^\top \mathbf{b})\}}{\mathrm{Var}(\mathbf{Z}^\top \mathbf{b})}; \tag{A12}$$

taking $0/0 := 0$. *If moreover* f *is continuously differentiable and strictly monotonic, there exists* $\varepsilon > 0$ *such that*

$$|\mathbb{E}\{Z_j f(\mathbf{Z}^\top \mathbf{b})\}| \geqslant \varepsilon |\mathrm{Cov}(Z_j, \mathbf{Z}^\top \mathbf{b})| / \mathrm{Var}(\mathbf{Z}^\top \mathbf{b}). \tag{A13}$$

In particular, $\mathbb{E}\{Z_j f(\mathbf{Z}^\top \mathbf{b})\} = 0$ *iff* $\mathrm{Cov}(Z_j, \mathbf{Z}^\top \mathbf{b}) = 0$.

Proof. Set $W := \mathbf{Z}^\top \mathbf{b}$. By standard properties of conditional expectations, it holds that

$$0 = \mathbb{E}\{W(\mathbf{Z} - \mathbb{E}(\mathbf{Z}|W))\} = \Sigma \mathbf{b} - \mathbb{E}\{W \mathbb{E}(\mathbf{Z}|W)\} = \Sigma \mathbf{b} - \mathbf{c}\mathbb{E}(W^2),$$

implying $\mathbb{E}(\mathbf{Z}|W) = \Sigma \mathbf{b} W / \mathrm{Var}(W)$. We then obtain (A12):

$$\mathbb{E}\{\mathbf{Z} f(\mathbf{Z}^\top \mathbf{b})\} = \mathbb{E}\{\mathbb{E}(\mathbf{Z}|W) f(W)\} = \Sigma \mathbf{b} \mathbb{E}\{W f(W)\} / \mathrm{Var}(W).$$

To show (A13), the mean value theorem implies the existence of some random variable $0 < \tilde{W} < W$ such that

$$\mathbb{E}\{W f(W)\} = \mathbb{E}[W\{f(0) + f'(\tilde{W})\}W] = \mathbb{E}\{W^2 f'(\tilde{W})\}.$$

Then

$$|\mathbb{E}\{W^2 f'(\tilde{W})\}| \geqslant |\mathbb{E}\{f'(\tilde{W}) W^2 1(W^2 \leqslant 1)\}| \geqslant \inf_{0 \leqslant x \leqslant 1} |f'(x)| \mathbb{E}\{W^2 1(W^2 \leqslant 1)\}.$$

Strict monotonicity of f then yields (A13). ■

LEMMA A5. *Assume that the survival time* T *has a general, continuous hazard rate function* $\lambda_T(t|Z)$ *depending on the random variable* $Z \in \mathbb{R}$ *and that the censoring time* C *is independent of* Z, T. *Then*

$$\delta = \int_0^\tau \tilde{e}(t) \mathrm{d}F(t) = \mathbb{E}\{\tilde{e}(T \wedge C \wedge \tau)\};$$

where $F(t) := \mathbb{P}(T \wedge C \wedge \tau \leqslant t)$ *and* $\tilde{e}(t) := \mathbb{E}\{Z \mathbb{P}(T \geqslant t|Z)\} / \mathbb{P}(T \geqslant t)$.

Proof. Let $S_T(\cdot|Z), S_C$ denote the (conditional) survival functions of T, C. Using the expression (8) for δ alongside the assumption of random censoring, we obtain

$$\delta = \mathbb{E}\left[\int_0^\tau \{Z - e(t)\} Y(t)\lambda_T(t|Z)\mathrm{d}t\right] \tag{A14}$$

$$= \int_0^\tau S_C(t)\mathbb{E}\{Z S_T(t|Z)\lambda_T(t|Z)\}\mathrm{d}t - \int_0^\tau \frac{\mathbb{E}\{Z S_T(t|Z)\}}{\mathbb{E}\{Y(t)\}} S_C(t)\mathbb{E}\{Y(t)\lambda_T(t|Z)\}\mathrm{d}t \tag{A15}$$

$$= -\int_0^\tau \frac{\mathrm{d}}{\mathrm{d}t}\tilde{e}(t)\mathbb{E}\{Y(t)\}\mathrm{d}t; \tag{A16}$$

where last equality follows since $S_T' = -\lambda_T S_T$. Integrating by parts, we obtain the statement of the lemma:

$$\delta = -\int_0^\tau \frac{\mathrm{d}}{\mathrm{d}t}\tilde{e}(t)\mathbb{E}\{Y(t)\}\mathrm{d}t = -\int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}\tilde{e}(t)\mathbb{E}\{\mathbb{P}(T \wedge C \wedge \tau \geqslant t|Z)\}\mathrm{d}t = \mathbb{E}\{\tilde{e}(T \wedge C \wedge \tau)\}.$$

■

Proof of Theorem 1. Set $\tilde{e}_j(t) := \mathbb{E}\{Z_{1j} S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}/\mathbb{E}\{S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}$ with

$$S_T(t,\cdot) = \exp\{-\int_0^t \lambda(s,\cdot)\mathrm{d}s\}.$$

Assumptions 1-2 and Lemma A4 imply that $\tilde{e}_j$ has constant sign throughout $[0,\tau]$. Invoking Lemma A5, (A12), and Assumption C, there exists a universal positive constant k_1 such that

$$|\delta_j| = \int_0^\tau |\tilde{e}_j(t)|\mathrm{d}F(t) \geqslant \int_0^\tau |\mathbb{E}\{Z_{1j} S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}|\mathrm{d}F(t) \geqslant k_1 |\mathrm{Cov}(Z_{1j}, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)|, \quad j \in \mathrm{M}^n.$$

Then $\mathrm{M}^n \subseteq \mathrm{M}^n_\delta$. The sure screening property follows from Lemma A3 and the assumptions. ■

Proof of Theorem 2. Suppose that

$$\|\boldsymbol{\delta}\|^2 = O\{\lambda_{\max}(\Sigma)\}. \tag{A17}$$

Set $\varepsilon := c_4/2$. On the set $B_n := \{\max_{1\leqslant j\leqslant p_n} |d_j - \delta_j| \leqslant \varepsilon n^{-\kappa}\}$, it then holds that

$$|\{j : |d_j| > 2\varepsilon n^{-\kappa}\}| \leqslant |\{j : |\delta_j| > \varepsilon n^{-\kappa}\}| \leqslant O\{n^{2\kappa}\lambda_{\max}(\Sigma)\}.$$

We then have

$$\mathbb{P}[|\widehat{\mathrm{M}}^n_d| \leqslant O\{n^{2\kappa}\lambda_{\max}(\Sigma)\}] = \mathbb{P}[|\{j : |d_j| > 2\varepsilon n^{-\kappa}\}| \leqslant O\{n^{2\kappa}\lambda_{\max}(\Sigma)\}] \geqslant \mathbb{P}(B_n).$$

By Lemma A3, with $k_1 = \varepsilon$, there exists c_5 such that $\mathbb{P}(B_n) \geqslant 1 - O[p_n \exp\{-c_5 n^{(1-2\kappa)\eta/(\eta+2)}\}]$ as claimed. So we need only verify (A17).

By Lemma A5, there exists a positive constant c_1 such that for $j \in \mathrm{M}^n$, it holds that $|\delta_j| \leqslant c_1 \int_0^\tau |\mathbb{E}\{Z_{1j} S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}|\mathrm{d}F(t)$ with F the unconditional distribution function of $T_1 \wedge C_1 \wedge \tau$. In contrast, $\delta_j = 0$ for $j \notin \mathrm{M}^n$, by Assumptions 3-4. It follows from Jensen's inequality that there exists a positive constant c_2 such that

$$\|\boldsymbol{\delta}\|^2 \leqslant c_2 \int_0^\tau \|\mathbb{E}\{\mathbf{Z}_1 S_T(t, \mathbf{Z}_1^\top \boldsymbol{\alpha}^0)\}\|^2 \mathrm{d}F(t). \tag{A18}$$

Lemma A4 implies

$$\mathbb{E}\{\mathbf{Z}_1 S_T(t,\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)\} = \frac{\mathbb{E}\{\mathbf{Z}_1^\top\boldsymbol{\alpha}^0 S_T(t,\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)\}}{\mathrm{Var}(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)}\Sigma\boldsymbol{\alpha}^0. \tag{A19}$$

By Cauchy-Schwartz's inequality, since $\|\Sigma\boldsymbol{\alpha}^0\|^2 \leqslant \|\Sigma^{1/2}\|^2\|\Sigma^{1/2}\boldsymbol{\alpha}^0\|^2 \leqslant \lambda_{\max}(\Sigma)\|\Sigma^{1/2}\boldsymbol{\alpha}^0\|^2$,

$$\|\mathbb{E}\{\mathbf{Z}_1 S_T(t,\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)\}\|^2 \leqslant \|\Sigma\boldsymbol{\alpha}^0\|^2/\mathrm{Var}(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0) \leqslant \lambda_{\max}(\Sigma).$$

Inserting this in (A18) then yields the desired result (A17). Note that this result does not rely on the uniform boundedness of $\mathrm{Var}(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)$ (Assumption C). ■

LEMMA A6. *Suppose that Assumption A holds and that both the survival time T_1 and censoring time C_1 follow a nonparametric Aalen model* (11) *with time-varying regression coefficients $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$, respectively. Suppose moreover that $\mathbf{Z}_1 = \Sigma^{1/2}\tilde{\mathbf{Z}}_1$ where $\tilde{\mathbf{Z}}_1$ has i.i.d. components and denote by $\phi(x) := \mathbb{E}\{\exp(\tilde{Z}_{1j}x)\}$ the moment generating function of $\tilde{Z}_{1j}$. Then*

$$\boldsymbol{\delta} = \Sigma^{1/2}\left[\int_0^\tau \mathrm{diag}\left\{\frac{\mathrm{d}}{\mathrm{d}x}\frac{\phi'(x)}{\phi(x)}\bigg|_{x=-\Gamma_j^0(t)}\right\}\mathbb{E}\{Y_1(t)\}\boldsymbol{\alpha}^0(t)^\top\mathrm{d}t\right]\Sigma^{1/2}; \tag{A20}$$

where $\Gamma^0(t) := \Sigma^{1/2}\int_0^t\{\boldsymbol{\alpha}^0(s)+\boldsymbol{\beta}^0(s)\}\mathrm{d}s$. *In particular, if* $\mathbf{Z}_1 \sim \mathrm{N}(0,\Sigma)$ *then*

$$\boldsymbol{\delta} = \Sigma\left\{\int_0^\tau \boldsymbol{\alpha}^0(t)\mathbb{E}\{Y_1(t)\}\mathrm{d}t\right\}. \tag{A21}$$

Proof. Let Λ_T and Λ_C denote the cumulative baseline hazard functions associated with T_1 and C_1. Combining (8) and (11), we get

$$\boldsymbol{\delta} = \mathbb{E}\left\{\int_0^\tau \mathbf{Z}_1\mathbf{Z}_1^\top Y_1(t)\boldsymbol{\alpha}^0(t)\mathrm{d}t\right\} - \int_0^\tau \mathbb{E}\{\mathbf{Z}_1 Y_1(t)\}^{\otimes 2}\mathbb{E}\{Y_1(t)\}^{-1}\boldsymbol{\alpha}^0(t)\mathrm{d}t \tag{A22}$$

$$= \int_0^\tau \Sigma^{1/2}\mathbf{H}(t)\Sigma^{1/2}\mathbb{E}\{Y_1(t)\}\boldsymbol{\alpha}^0(t)\mathrm{d}t; \tag{A23}$$

defining here

$$\mathbf{H}(t) := \frac{\mathbb{E}\{Y_1(t)\}\mathbb{E}\{\tilde{\mathbf{Z}}_1\tilde{\mathbf{Z}}_1^\top Y_1(t)\} - \mathbb{E}\{\tilde{\mathbf{Z}}_1 Y_1(t)\}^{\otimes 2}}{\mathbb{E}\{Y_1(t)\}^2}.$$

Since we have $\mathbb{E}\{Y_1(t)|\mathbf{Z}_1\} = \exp[-\{\Lambda_T(t)+\Lambda_C(t)+\tilde{\mathbf{Z}}_1^\top\Gamma^0(t)\}]$, independence of the components of $\tilde{\mathbf{Z}}_1$ clearly implies $[\mathbf{H}(t)]_{ij} \equiv 0$ for $i \neq j$. For $i = j$, factor the conditional at-risk indicator as $\mathbb{E}\{Y_1(t)|\mathbf{Z}_1\} = Y_1^{(j)}(t)Y_1^{(-j)}(t)$ where $Y_1^{(j)} := \exp\{-\tilde{Z}_{1j}\Gamma_j^0(t)\}$. Utilizing independence again, we get

$$[\mathbf{H}(t)]_{jj} = \frac{\mathbb{E}\{Y_1^{(j)}(t)\}\mathbb{E}\{\tilde{Z}_{1j}^2 Y_1^{(j)}(t)\} - \mathbb{E}\{Y_1^{(j)}(t)\tilde{Z}_{1j}\}^2}{\mathbb{E}\{Y_1^{(j)}(t)\}^2} = \frac{\mathrm{d}}{\mathrm{d}x}\frac{\phi'(x)}{\phi(x)}\bigg|_{x=-\Gamma_j^0(t)}$$

This proves (A20). To verify (A21), simply note that the moment generating function of a standard Gaussian is $\phi(x) = \exp(x^2/2)$ for which $\mathrm{d}/\mathrm{d}x(\phi'(x)\phi(x)^{-1}) = 1$. ■

From (A20), a 'simple' description of $\boldsymbol{\delta}$ (which does not involve factorizing a matrix in terms of $\Sigma^{1/2}$) is available exactly when features are Gaussian. Specifically, it holds for some fixed $K > 0$ that

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{\phi'(x)}{\phi(x)} = K, \qquad \text{and } \phi(0) = 1,$$

iff $\phi(x) = \exp(Kx^2/2)$, the moment generating function of a mean zero Gaussian random variable.

Proof of Theorem 3. We apply Lemma A6. Denote by $\mathbf{v}_j$ the jth canonical basis vector in $\mathbb{R}^{p_n}$. Integrating by parts in (A21), we obtain

$$\delta_j = \mathbf{v}_j^\top \Sigma \int_0^\tau \boldsymbol{\alpha}^0(t)\mathbb{E}\{Y_1(t)\}\mathrm{d}t = \mathbf{v}_j^\top \Sigma \int_0^\infty \boldsymbol{\alpha}^0(t)\mathbb{E}\{\mathbb{P}(T_1 \wedge C_1 \wedge \tau \geqslant t)\}\mathrm{d}t = \mathbf{v}_j^\top \Sigma\mathbb{E}\{\mathbf{A}^0(T_1 \wedge C_1 \wedge \tau)\}.$$

By the assumptions, $|\mathbf{v}_j^\top \Sigma\mathbb{E}\{\mathbf{A}^0(T_1 \wedge C_1 \wedge \tau)\}| \geqslant c_1 n^{-\kappa}$ whenever $j \in \mathrm{M}^n$. Thus $\mathrm{M}^n \subseteq \mathrm{M}^n_\delta$. For Gaussian Z_{1j}, we have $\mathbb{P}(|Z_{1j}| > s) \leqslant \exp(-s^2/2)$, and the SIS property then follows from Lemma A3. ■

Proof of Theorem 4. Recall that

$$\Delta = \mathbb{E}\left[\int_0^\tau \{\mathbf{Z}_1 - \mathbf{e}(t)\}^{\otimes 2} Y_1(t)\mathrm{d}t\right].$$

Then

$$\Delta\boldsymbol{\alpha}^0 = \int_0^\tau \frac{\mathbb{E}\{Y_1(t)\}\mathbb{E}\{Y_1(t)\mathbf{Z}_1\mathbf{Z}_1^\top\boldsymbol{\alpha}^0\} - \mathbb{E}\{Y_1(t)\mathbf{Z}_1^\top\boldsymbol{\alpha}^0\}\mathbb{E}\{Y_1(t)\mathbf{Z}_1\}}{\mathbb{E}\{Y_1(t)\}}\mathrm{d}t,$$

But by Lemma A4 and the assumption of random censoring,

$$\mathbb{E}\{Y_1(t)\mathbf{Z}_1\mathbf{Z}_1^\top\boldsymbol{\alpha}^0\} = \Sigma\boldsymbol{\alpha}^0\frac{\mathbb{E}\{(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)^2Y_1(t)\}}{\mathrm{Var}(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)}, \quad \text{and } \mathbb{E}\{\mathbf{Z}_1Y_1(t)\} = \Sigma\boldsymbol{\alpha}^0\frac{\mathbb{E}\{Y_1(t)\mathbf{Z}_1^\top\boldsymbol{\alpha}^0\}}{\mathrm{Var}(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0)}.$$

So we can construct a function ξ such that $\Delta\boldsymbol{\alpha}^0 = \Sigma\boldsymbol{\alpha}^0 \int_0^\tau \xi(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0, t)\mathrm{d}t$ where it holds that $\int_0^\tau \xi(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0, t)\mathrm{d}t \neq 0$, by nonsingularity of Δ. Similarly, using Lemma A5, we may construct a function ζ such that $\boldsymbol{\delta} = \Sigma\boldsymbol{\alpha}^0 \int_0^\tau \zeta(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0, t)\mathrm{d}t$. Taking $\nu := \int_0^\tau \zeta(\mathbf{Z}_1^\top\boldsymbol{\alpha}^0, t)\mathrm{d}t / \int_0^\tau \xi(t, \mathbf{Z}_1^\top\boldsymbol{\alpha})\mathrm{d}t$, $\boldsymbol{\beta}^0 = \nu\boldsymbol{\alpha}^0$ solves $\Delta\boldsymbol{\beta}^0 = \boldsymbol{\delta}$. ■

References

Aalen, O. O. (1980) A model for non-parametric regression analysis of counting processes. In *Lecture Notes on Mathematical Statistics and Probability 2* (eds. W. Klonecki, A. Kozek and J. Rosinski), 1–25. Springer-Verlag.

Aalen, O. O. (1989) A linear regression model for the analysis of lifetimes. *Statist. Med.*, **8**, 907–925.

Bair, E. and Tibshirani, R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, E108.

Benner, A., Zucknick, M., Hielscher, T. and et al. (2010) High-dimensional Cox models: The choice of penalty as part of the model building process. *Biom. J.*, **52**, 50–69.

Bradic, J., Fan, J. and Jiang, J. (2010) Regularization for Cox's proportional hazards model with NP-dimensionality. Tech. rep., Princeton University. `arXiv:1010.5233v2`.

Bradic, J., Fan, J. and Wang, W. (2011) Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Statist. Soc. B*, **73**, 325–349.

Breiman, L. (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.*, **16**, 199–231.

Brillinger, D. R. (1983) A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann* (eds. P. J. Bickel, K. A. Doksum and J. L. Hodges), 97–114. Wadsworth.

Cheng, K. F. and Wu, J. W. (1994) Adjusted least squares estimates for the scaled regression coefficients with censored data. *J. Amer. Statist. Assoc.*, **89**, 1483–1491.

Fan, J., Feng, Y. and Song, R. (2011) Nonparametric independence screening in sparse ultra-high dimensional additive models. Tech. rep., Princeton University. `arXiv:0912.2695v2`.

Fan, J., Feng, Y. and Wu, Y. (2010) *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, chap. High-dimensional variable selection for Cox's proportional hazards model. Institute of Mathematical Statistics.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348–1360.

Fan, J. and Lv, J. (2008) Sure independence screening for ultra-high dimensional feature space. *J. R. Statist. Soc. B*, **70**, 849–911.

Fan, J. and Lv, J. (2009) Non-concave penalized likelihood with NP-dimensionality. Tech. rep., Princeton University and University of South California. `arXiv:0910.1119v1`.

Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh dimensional feature selection: Beyond the linear model. *J. Machine Learning Res.*, **10**, 2013–2038.

Fan, J. and Song, R. (2010) Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, **38**, 3567–3604.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**, 302–332.

Gorst-Rasmussen, A. (2011) ***ahaz**: Regularization for semiparametric additive hazards regression*. URL `http://cran.r-project.org/package=ahaz`. R package.

Gorst-Rasmussen, A. and Scheike, T. H. (2011) Coordinate descent methods for the penalized semiparametric additive hazards model. Tech. Rep. R-2011-10, Department of Mathematical Sciences, Aalborg University.

Hall, P. and Li, K. (1993) On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.*, **21**, 867–889.

Hardin, C. D. (1982) On the linearity of regression. *Z. Wahrsch. verw. Gebiete*, **61**, 293–302.

Hattori, S. (2006) Some properties of misspecified additive hazards models. *Statist. Prob. Letters*, **76**, 1641–1646.

He, Q. and Lin, D. (2011) A variable selection method for genome-wide association studies. *Bioinformatics*, **27**, 1–8.

Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13–30.

Leng, C., Lin, Y. and Wahba, G. (2007) A note on the lasso and related procedures in model selection. *Statist. Sinica*, **16**, 1273–1284.

Leng, C. and Ma, S. (2007) Path consistent model selection in additive risk model via lasso. *Statist. Med.*, **26**, 3753–3770.

Li, K. and Duan, N. (1989) Regression analysis under link violation. *Ann. Statist.*, **17**, 1009–1052.

Lin, D. Y. and Ying, Z. (1994) Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61–71.

Ma, S. and Song, X. (2011) Ranking prognosis markers in cancer genomic studies. *Brief. Bioinform.*, **12**, 33–40.

Martinussen, T. and Scheike, T. H. (2009) Covariate selection for the semiparametric additive risk model. *Scand. J. Statist.*, **36**, 602–619.

Massart, P. (2000) About the constants in Talagrands concentration inequalities for empirical processes. *Ann. Prob.*, **28**, 863–884.

McKeague, I. W. and Sasieni, P. D. (1994) A partly parametric additive risk model. *Biometrika*, **81**, 501–514.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection. *J. R. Statist. Soc. B*, **72**, 417–473.

Metzeler, K. H., Hummel, M., Bloomfield, C. D. and et al. (2008) An 86 probe set gene expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*, **112**, 4193–4201.

Pollard, D. (1989) Asymptotics via empirical processes. *Statist. Sci.*, **4**, 341–354.

Roberts, C. P. (2008) Discussion of 'sure independence screening for ultrahigh dimensional feature space'. *J. R. Statist. Soc. B*, **70**, 901.

Struthers, C. A. and Kalbfleisch, J. D. (1986) Misspecified proportional hazards models. *Biometrika*, **73**, 363–369.

Tibshirani, R. (1997) The lasso method for variable selection in the Cox model. *Statist. Med.*, **16**, 385–395.

Tibshirani, R. (2009) Univariate shrinkage in the Cox model for high dimensional data. *Stat Appl Genet Mol Biol*, **8**. Article 21.

Wang, H. and Leng, C. (2007) Unified LASSO estimation by least squares approximation. *J. Amer. Statist. Assoc.*, **102**, 1039–1048.

Zhao, S. D. and Li, Y. (2010) Principled sure independence screening for Cox models with ultra-high-dimensional covariates. Tech. rep., Harvard School of Public Health. URL `http://www.bepress.com/harvardbiostat/paper111`.

Zhu, L., Qian, L. and Lin, J. (2009) Variable selection in a class of single-index models. *Ann. Inst. Statist. Math.* DOI: 10.1007/s10463-010-0287-4.

Zhu, L. and Zhu, L. (2009) Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *J. Multivariate Anal.*, **100**, 862–875.

Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418–1429.

Zou, H., Hastie, T. and Tibshirani, R. (2007) On the “degrees of freedom” of the lasso. *Ann. Statist.*, **35**, 2173–2192.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509–1533.