Julie Lyng Forman

Statistical Inference from Diffusion Driven Models

Ph.D. Thesis

Thesis advisor: Professor Michael Sørensen, University of Copenhagen

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

Preface

This dissertation concludes my work as a ph.d. student at the Department of Applied Mathematics and Statistics, today part of the larger Department of Mathematical Sciences at the University of Copenhagen. The work was funded by a grant from the University of Copenhagen beginning November 2003 and ending April 2007.

The topic of the thesis is statistical inference from diffusion driven models. The theory of estimating functions was introduced to me in a course taught by Michael Sørensen and later became the turning point of my master thesis. My research as a ph.d. student is a continuation of this work with the paper Forman (2005) as the foremost example. The interest in testing and comparing different diffusion models is my own agenda.

I would thank my colleagues for making my time as a ph.d. student worthwhile. Especially Michael Sørensen has been a forever enthusiastic supervisor. I have learned a lot from our discussions and from our joint work on the paper Forman & Sørensen (2006) in particular. Also thanks goes to Martin Jacobsen for our many interesting talks and to the other ph.d. students at the department for good commeradeship. In particular, Anders Tolver Jensen and Niels Richard Hansen have been good friends and role models offering guidance on my problems.
Special thanks goes to Helle Sørensen and Bo Markussen at the Department of Natural Sciences, LIFE, University of Copenhagen for sharing my interest in goodness of fit testing and for inviting me to work with them on the subject. Our collaboration has been truly fruitful with the paper Forman, Markusen & Sørensen (2007) as a natural conclusion.
Part of my time as a ph.d. student was spent at the Department of Statistics, University of California at Berkeley. I would like to thank professor David Brillinger and the people at the department for hosting my stay. Also thanks to the I-house community for making my time abroad enjoyable.

I dedicate this work to my friends and family. Without their love and support I would not have made it through the hard times.

Julie Lyng Forman
Copenhagen, April 2007

Summary

The topic of this dissertation is the statistical analysis of discretely observed diffusion driven models. Focus is on estimating and goodness of testing. Summed diffusions, integrated diffusions, and diffusion driven stochastic volatility models are explored in detail.

Only rarely the functional form of the likelihood function of a diffusion driven model is explicitly known. Moreover, the algorithms that can be used to simulate it are complicated from a mathematical as well as a computational point of view. Therefore alternative estimating schemes are often called upon when the models have to be fitted and validated. In particular, moment based methods such as general estimating functions and the generalized method of moments have proven themselves a successful means for making inference in discretely observed diffusion models. This thesis elaborates further on these ideas for specific diffusion driven models as well as for the general class driven by the so-called Pearson diffusions.

The thesis consists of two parts. The latter contains my research contributions in the form of three papers. The two first papers have been submitted for publication and very soon the third will follow. The first part serves as a general introduction to the analysis of diffusion driven models and to the papers in particular.

The introduction is likewise made out of two main chapters. Chapter 2 contains a survey on the diffusion driven models studied in the thesis and an account of the probabilistic features of scalar diffusions with emphasis on their statistical applications. Section 2.2.4 presents ongoing work on a new multivariate stochastic volatility model which has not yet converged to a form suitable for publication. Chapter 3 provides an overview of some existing statistical methods for diffusions and diffusion driven models including likelihood inference, the theory of estimating functions, the generalized method of moments, and nonparametric inference. The purpose of the chapter is to motivate and contrast the results found in the papers.

The first paper studies least squares estimators for the autocorrelation parameters in a summed diffusion process. The asymptotic theory is described in detail. Further a consistent procedure for selecting the number of underlying diffusions is presented. The results of the Monte Carlo simulations indicate that the optimally weighted least squares estimator is no less efficient than the maximum likelihood estimator.

The second paper introduces the term Pearson diffusions for the diffusion processes having a mean reverting linear drift and a quadratic squared diffusion coefficient. The classification of the models yields six more or less known classes of diffusion. Each class has it own distinct features. Treated as a whole the Pearson diffusions are highly tractable from a statistical point of view due to their explicitly computable polynomial eigenfunctions. It is further demonstrated that the tractability is inherited by summed Pearson diffusions, integrated Pearson diffusions, and Pearson stochastic volatility models.

The third paper is concerned with a new and generic goodness of fit test for stochastic process models that are fitted by means of general estimating functions. The basic idea is to compare the estimate obtained from the original sample to those obtained from downsampled data. The asymptotic theory for the test is derived and exemplified by linear drift diffusions. The small sample performance of the test is further explored through Monte Carlo simulations.

Dansk resumé

Den foreliggende afhandling omhandler statistisk inferens i diskrete observerede diffusions type modeller med fokus på estimation og goodness of fit test. Vi ser nærmere på summerede diffusioner, integrerede diffusioner og diffusionsdrevne stokastiske volatilitetsmodeller.

Kun i ganske få tilfælde har man eksplicitte udtryk for likelihood funktionen for en diffusion type model. Den kan findes ved simulation, men de forhåndenværende algoritmer er komplicerede fra et matematisk såvel som fra et programmeringsmæssigt perspektiv. Derfor benytter man ofte alternative statistiske metoder til at estimere modellernes parametre. Momentbaserede metoder såsom generelle estimationsfunktioner og GMM har vist sig at være særligt nyttige. Afhandlingen videreudvikler disse metoder for de diffusionsdrevne modeller hver især og for klassen af såkaldt Pearson-diffusionsdrevne modeller som helhed.

Afhandlingen består af to dele. Anden halvdel indeholder resultaterne af mit forskningsarbejde i form af tre artikler. De to første artikler er sendt afsted med henblik på publicering og den tredie vil meget snart gå samme vej. Første del af afhandlingen er en introduktion til den generelle statistiske teori for inferens i diffusionsdrevne modeller. Den er skrevet som oplæg til artiklernes resultater.

Introduktionsdelen består af to større kapitler. Kapitel 2 giver et overblik over de diffusionsdrevne modeller som behandles i afhandlingen samt en introduktion til sandsynhedsteorien for endimensionelle diffusioner med hovedvægt på dens statistiske anvendelser. Afsnit 2.2.4 præsenterer foreløbige resultater om en ny flerdimensionel volatilitetsmodel, der muligvis vil blive til en publicerbart manuskript engang i fremtiden. Kapitel 3 er en oversigt over en række eksisterende statistiske metoder for diffusioner og diffusionsdrevne modeller. Disse inkluderer likelihood baseret inferens, generelle estimationsfunktioner, GMM og ikke-parametriske metoder. Det overordnede formål med kapitlet er at sætte artiklernes resultater ind i et større perspektiv.

Den første artikel handler om mindste kvadraters metode for korrelationsparametrene i en summeret diffusion. Den asymptotiske teori er detaljeret beskrevet. Herudover præsenteres en konsistent procedure til valg af antallet af led i den underliggende sum. Simulationsstudiet viser at estimoren udledt fra den optimalt vægtede mindste kvadraters metode kan være ligeså efficient som maksimaliseringsestimator i de givne eksempler.

Den anden artikel handler om Pearson diffusionerne, hvor navnet Pearson diffusion introduceres som fællesbetegnelse for diffusioner med lineær drift og kvadratisk volatilitetskoefficient. Pearson diffusionerne kan inddeles i seks mere eller mindre kendte typer hver med sine særkender. Fra et statistisk synspunkt er Pearson diffusionerne under et nemme at analysere fordi de har polynomier som egenfunktion og disses koefficienter kan udregnes eksplicit. Endvidere diskuteres den statistiske analyse af Pearson diffusionsdrevne modeller som på mange punkter er ligeså nemme at gå til som Pearson diffusionerne selv.

Den tredie og sidste artikel handler om et nyt og generelt goodness of fit test for stokastiske processer. Testet bygger på en generel estimationsfunktion hvorfra parameterestimater kan udledes. Ideen går i sin enkelthed ud på at sammenligne parameterestimater for data udtaget ved forskellige frekvenser, en procedure vi betegner som downsampling. Den asymptotiske fordeling af testet er udledt. Som eksempel betragtes test af hypotesen om lineær drift i en diffusionsmodel. Testets egenskaber er yderligere belyst gennem en række Monte Carlo simulationer.

Contents

Part I

Diffusion driven models and their statistical inference

1 Introduction

Diffusion models provide a natural and flexible framework for modeling a large variety of phenomena that evolve continuously and randomly with time. Whereas financial time series tend dominate the applications, the models are of equal relevance in say, physical and biological applications. In this thesis diffusions take the form of solutions to the stochastic differential equation,

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t$$

where B_t is a Brownian motion (the source of randomness), $\mu(X_t)$ is the instantaneous mean and $\sigma^2(X_t)$ is the instantaneous variance of the diffusion. Given the past of the process up to time t the speed and direction of the diffusion only relies on the present state X_t, this is known as the Markov property. Loosely speaking the diffusion has no memory and therefore the plain diffusions cannot always account for the dependence structure found in real data. This has motivated the development of a large range of diffusions driven models with longer memory. For instance summed diffusions are formed by aggregating diffusions moving on different time scales. Integrated diffusions occur naturally when data is formed by measuring the average of a certain quantity over disjoint time intervals. The stochastic volatility models are meant to deal with the fact that the variance of for instance the log returns of a stock price varies randomly with time.

The topic of this thesis is statistical inference from discretely observed diffusion driven models. The focus will be on statistical methodology rather than on applications. The applications appear as no more (or no less one could say) than a strong motivating factor for the development of the preceding results.

Historically the statistical analysis of discretely observed diffusions has been challenging. Only rarely the functional form of the likelihood function is known (the likelihood function of a non-Markovian diffusion driven model is of course even more complicated). This has motivated the development of a large number of alternative estimating schemes. Many of these are based on the matching of moments through general estimating equations. Important directions are simulated and approximate likelihood inference, the theory of general estimating functions, the general method of moments, and nonparametric inference. All of these are reviewed in chapter 3. The method of indirect inference is briefly mentioned in section 2.2.3. A point made in this thesis is that the methods used for making inference in plain diffusion models are often (but not always) applicable to the more general diffusion driven models. In particular, our paper Forman & Sørensen (2006) shows that highly tractable diffusion type models can be build from a class of simple scalar diffusions the so-called Pearson diffusions. This idea is further elaborated on in section 2.2.4 where a new idea for modeling multivariate stochastic volatility models is presented.

1.1 Likelihood vs moment based inference

Throughout the thesis special attention is given to likelihood inference as the more or less unattainable ideal and to moment based inference which is exemplified by our papers Forman (2005), Forman & Sørensen (2006) and Forman, Markusen & Sørensen (2007).

Today specialized algorithms render likelihood inference applicable by means of computer intensive methods. In contrast to the pioneer algorithms the time spent on computing the maximum likelihood estimator with great accuracy is no longer devastating. The maximum likelihood estimator is efficient so is there any need for new moment based estimators today? Traditionally moment based estimators are promoted as being fast and tractable compared to the maximum likelihood estimator. This is more or less still the case. Even though general algorithms for computing the likelihood function are available the implementation is time consuming and demands a certain expertise on for instance Markov chain Monte Carlo methods to run smoothly. Misspecification may cause the algorithm to break down and most often a good initial guess is needed to find the maximum likelihood estimator. In contrary the moment based estimators often rely on explicit criteria and are thus very easy to handle. There exists several examples of simple moment based estimators that attain almost the same efficiency as the maximum likelihood estimator. It should be noted that the moment based estimators also apply to semiparametric models for which the likelihood function does not exist or cannot be simulated.

1.2 Specification testing

Another major topic of the thesis is specification testing for diffusion driven models. Only the imagination of the researchers limits the great variety of diffusion based model constructions. With a growing number of models at our disposal it becomes increasingly important to validate the choice of model made in a specific application. Compared to the vast literature on estimation in diffusion-type models, the material on goodness of fit is somewhat limited. A highly informative diagnostic is provided by the so-called uniform residuals obtained when applying the probability transform given past observations to each datum in turn. For most diffusion type models the probability transform is not explicitly known, but it can be simulated by means of the same computer intensive algorithms used in obtaining the likelihood function. In connection with moment based inference goodness of fit it typically based on checking excess moment conditions by use of the so-called overidentifying restrictions test. This is used in the model selection procedure considered in my paper Forman (2005). A new idea for checking the dependence structure in a stochastic process is presented in our paper Forman, Markusen & Sørensen (2007). The main idea is to compare the parameter estimates from the downsampled data. Downsampling can also be viewed as a generic way of creating excess moment conditions. Hence the test is closely related to the overidentifying restrictions test.

1.3 The structure of the thesis

The thesis consist of two parts. The second part is formed by the papers which contain my main contributions. The first part of the thesis serves as a general introduction to the statistical analysis of diffusion driven models and the papers in particular. Save from the multivariate stochastic volatility model considered in section 2.2.4 no new results are presented in this part of the thesis.

Chapter 2 of the introduction is concerned with the diffusion driven models. The first section provides a short crash course on scalar diffusions. A large number of basic features and probabilistic facts are summarized motivated by their use in statistics. The Pearson diffusions studied in our paper, Forman & Sørensen (2006) are used as an example. The second section reviews the three kinds of diffusion driven models studied in this thesis: summated diffusions, integrated diffusions, and diffusion driven stochastic volatility models. In addition a new idea for modeling multivariate volatility processes is presented in section 2.2.4.

Chapter 3 of the introduction outlines a number of statistical methods for analyzing discretely observed diffusions and diffusion driven models. Likelihood inference, general estimating functions, the generalized method of moments, and nonparametric statistics are considered. Also the so-called uniform residuals are discussed. The overall purpose of this chapter is to motivate and contrast the results found in my papers. Whenever relevant I use my own results as examples. The focus is on estimation and goodness of fit testing which are also the main topics of my papers.

The three papers are presented in a form that have been (or will soon be) submitted for publication. More detailed accounts of their contents are found in the preceding summary and the abstracts introducing each paper. The papers can be read independently from one another. An unfortunate consequence is that the notation differs in between papers. Hopefully this will not be the cause of confusion. The introduction has been written exclusively for this thesis and is also by and large self contained. In order to make a good reading of the individual sections the same information sometimes appear several different places in the introduction as well as in the papers. The list of references on the other hand is collected in a single bibliography at the end of the thesis.

2

Diffusion driven models

The following sections provide an overview of the various models encountered in the thesis. These are the plain scalar diffusions, the summated diffusions, the integrated diffusions and the diffusion driven stochastic volatility models. Section 2.2.4 presents a new idea for modeling multivariate volatility processes.

Preliminarily, section 2.1 gives a short account of the probabilistic features of scalar diffusion. The foremost purpose is to get the basic definitions in place. The invariance and mixing properties (sections 2.1.3 and 2.1.5) are of major importance in relation to statistical inference. Likewise the theory related to the infinitesimal generator (section 2.1.4) is indispensable not only for generating moment conditions to be used in estimation but for the deeper understanding of the theory of diffusions. All of the stated results can be found in the literature, hence no proofs are given. We consider as an example the class of Pearson diffusions the statistical analysis of which is the topic of our paper Forman & Sørensen (2006).

Section B.4 outlines the features of the distinctive diffusion driven models. A point made is that the diffusion driven models inherit many probabilistic features such as stationarity and mixing properties from the underlying diffusions. Pearson diffusion driven models are pointed out as they allow for the explicit computation of moments and mixed moments. This is one of the major results of our paper Forman & Sørensen (2006). Extensions to more general multivariate volatility model is discussed in section 2.2.4. In addition for each class of models a short survey on its statistical analysis is given. Further details on the statistical analysis of diffusion and diffusion driven model can be found in chapter 3 and in the quoted papers.

2.1 On scalar diffusions

We consider a scalar diffusion on the state space $I =]l; r[\subseteq \mathbb{R}$, the weak solution of a stochastic differential equation

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t. \tag{2.1}$$

where for convenience the drift $\mu\colon I \to \mathbb{R}$ and the diffusion coefficient $\sigma\colon I \to [0; \infty[$ are assumed to be continuous. Further we assume that σ is strictly positive on I. These assumptions ensure that for any initial distribution on I a weak solution exist, and that this solution is unique in the sense of probability law, see for instance Karatzas & Shreve (1991).

Definition 2.1.1 *Let $\{l_n\}$ and $\{r_n\}$ be sequences in I such that $l_1 < r_1$, $l_n \searrow l$, and $r_n \nearrow r$. The process $\{X_t\}_{t\geq 0}$ is a weak solution to (2.1) if there exists a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ admitting a $\{\mathcal{F}_t\}$-adapted Brownian motion $\{B_t\}$ such that*

- *$\{\mathcal{F}_t\}_{t\geq 0}$ satisfies the usual conditions*
- *$\{X_t\}_{t\geq 0}$ is a continuous, $\{\mathcal{F}_t\}$-adapted, $[l; r]$-valued process with $P(X_0 \in I) = 1$.*
- *For all $t \geq 0$ it holds that $\int_0^{t\wedge T_n} \{|b(X_s)| + \sigma^2(X_s)\}ds < \infty$ and*

$$X_{t\wedge T_n} = X_0 + \int_0^{t\wedge T_n} b(X_s)ds + \int_0^{t\wedge T_n} \sigma(X_s)dB_s$$

almost surely where $T_n = \inf\{t \geq 0\colon \ X_t \notin (l_n; r_n)\}$,

We refer to $T = \lim_{n\to\infty} T_n$ as the exit time from I. The process is said to be explosive if $P(T < \infty) > 0$ and non-explosive otherwise.

Simulation: The drift and diffusion coefficient has the interpretation as the instantaneous mean and standard deviation of the diffusion. For $\Delta \to 0$ the increment $X_{t+\Delta} - X_t$ is approximately normal with mean $\Delta\mu(X_t)$ and variance $\Delta\sigma^2(X_t)$. This property is used when simulating a diffusion process by means of the Euler scheme where the initial variable $\hat{X}_0$ is drawn from a relevant distribution and the following realizations are recursively drawn according to

$$\hat{X}_{(i+1)\Delta} = \hat{X}_{i\Delta} + \Delta\mu(\hat{X}_{i\Delta}) + \Delta^{1/2}\sigma(\hat{X}_{i\Delta})\varepsilon_{i+1}$$

where $\varepsilon_1, \varepsilon_2, \ldots$ are i.i.d. standard normal random numbers and Δ is a suitably small step size. See Kloeden & Platen (1999) for a thorough account on simulating solutions to the stochastic differential equation.

Transformation: The class of diffusions is closed under twice continously differentiable and invertible transformations. If g is such at transformation then by Ito's formula $Y_t = g(X_t)$ is a diffusion satisfying

$$dY_t = \mathcal{L}g(g^{-1}(Y_t))dt + g'(g^{-1}(Y_t))\sigma(g^{-1}(Y_t))dB_t$$

where $\mathcal{L}g = \mu g' + \sigma^2/2g''$ (more on the differential operator $\mathcal{L}$ in section ... below). The transformed diffusion inherits every important feature that $\{X_t\}$ might posses. In particular, whenever the diffusion $\{X_t\}$ is sufficiently simple that a successful statistical analysis can be carried out, the same holds for the transformed process $\{Y_t\}$.

2.1.1 Example: The Pearson diffusions

A Pearson diffusion is a stationary solution of a stochastic differential equation of the form

$$dX_t = -\theta(X_t - \mu)dt + \sqrt{2\theta(aX_t^2 + bX_t + c)}dB_t, \tag{2.2}$$

with mean reverting linear drift and squared diffusion coefficient which is a second order polynomial. The parameter $\theta > 0$ is a scaling of time that determines the speed of evolution of the diffusion. The parameters μ, a, b, and c determine the state space of the diffusion and the shape of its distribution. Our paper Forman & Sørensen (2006) identifies six subclasses of Pearson diffusions corresponding to whether the squared diffusion coefficient is constant, linear, a convex parabola with either zero, one or two roots, or a concave parabola with two roots. The Pearson class of diffusions is closed under translations and scale-transformations. Table 2.1 displays the state spaces and squared diffusion coefficients of the six standard-type Pearson diffusions considered in Forman & Sørensen (2006) together with their general rescaled form under the transformation $t(x) = \gamma x + \delta$ where $\gamma > 0$ (the case $\gamma < 0$ is similar).

	Standard $\sigma^2(x)$	I	Rescaled $\sigma^2(x)$	I
1	2θ	$\mathbb{R}$	$2\theta\gamma^2$	$\mathbb{R}$
2	$2\theta x$	$]0;\infty[$	$2\theta\gamma(x-\delta)$	$]\delta,\infty[$
3	$2\theta a(x^2+1)$	$\mathbb{R}$	$2\theta a(x^2 - 2\delta x + \delta^2 + \gamma^2)$	$\mathbb{R}$
4	$2\theta a x^2$	$]0;\infty[$	$2\theta a(x^2 - 2\delta x + \delta^2)$	$]\delta;\infty[$
5	$2\theta a x(x+1)$	$]0;\infty[$	$2\theta a\{x^2 - (\gamma - 2\delta)x + \delta(\delta - \gamma)\}$	$]\delta;\infty[$
6	$2\theta a x(x-1)$	$]0;1[$	$2\theta a\{x^2 - (\gamma + 2\delta)x + \delta(\delta + \gamma\}$	$]\delta;\gamma+\delta[$

Table 2.1: Squared diffusion coefficients and state spaces of the Pearson subclasses. The standard drift is $\mu(x) = -\theta(x - \mu)$ and the rescaled drift is $\mu(x) = -\theta(x - \gamma\mu - \delta)$.

The first and most simple subclass is formed by the well known Ornstein-Uhlenbeck processes sometimes referred to as the Vasiček model in the finance literature. The Ornstein-Uhlenbeck process is a Gaussian continuous time autoregression. Due to its tractability the process is often used as benchmark in e.g. simulation studies. The second subclass contains the square-root processes which are also known as the Cox-Ingersoll-Ross processes in the finance literature. Feller (1951) used this process as a model of population growth, whereas Cox, Ingersoll & Ross (1985) used it as a model for the term structure of interest rates. Just like the Ornstein-Uhlenbeck process, the square-root process is well understood and often serves as a testing case for statistical methods in survey. The third subclass to my knowledge is new to the literature, it is thus exemplified in our paper Forman & Sørensen (2006). The fourth type of Pearson diffusion is known as the GARCH-diffusion in the finance literature as Nelson (1990) showed that it is the continuous-time limits of the GARCH(1,1) process. The fifth class of Pearson diffusions have not received much attention. The sixth class is formed by the so-called Jacobi diffusions used by De Jong, Drost & Werker (2001) and Larsen & Sørensen (2003) to model exchange rates in a target zone. Most of the Pearson diffusions were derived by Wong (1964) and are also among the diffusion models studied in Bibby, Skovgaard & Sørensen (2005). See Nagahara (1996) for an application.

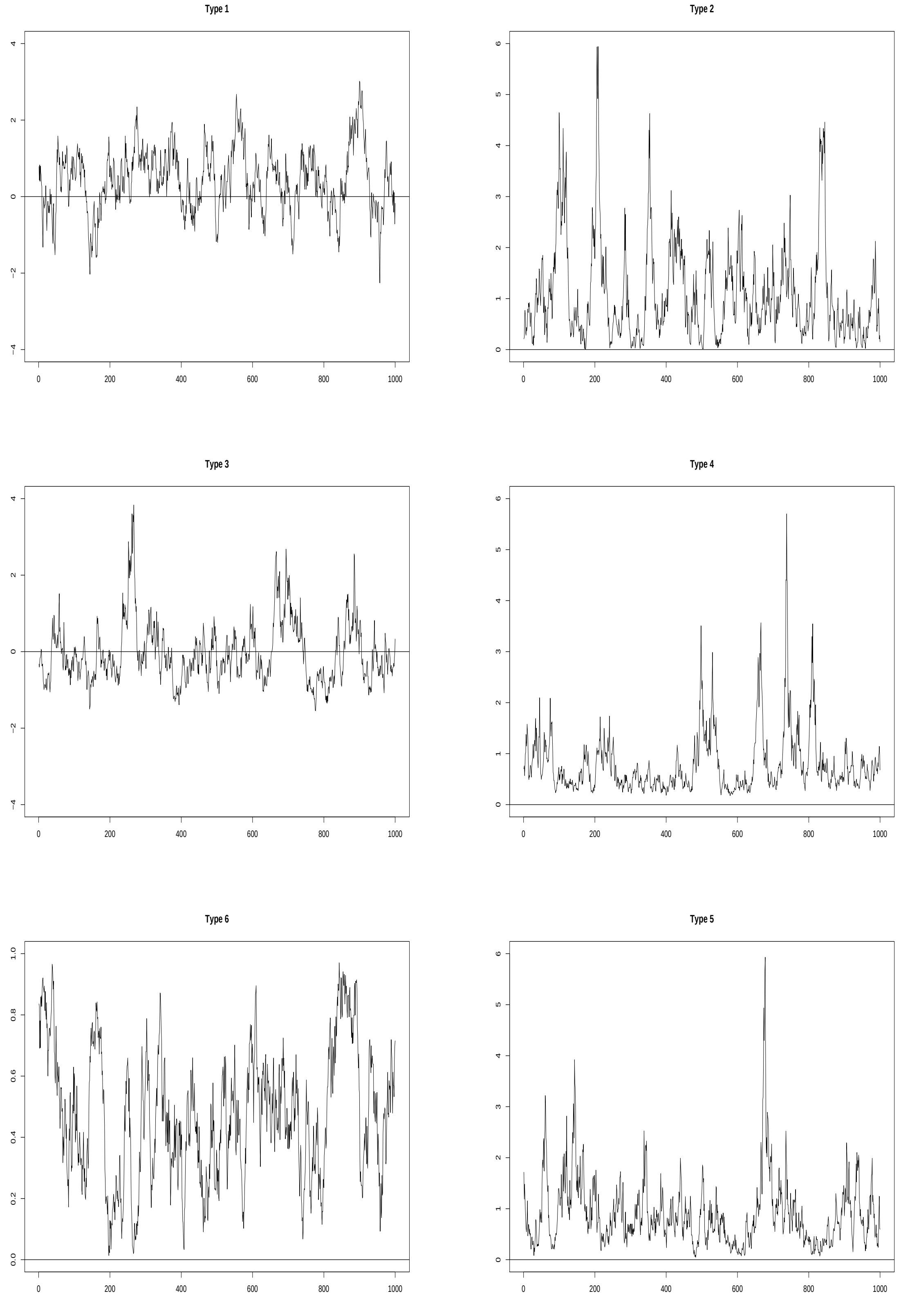

Figure 2.1: Sample paths of simulated standard Pearson diffusions. For all of the realizations $\theta = 0.05$, The means are $\mu = 0$ for types 1 and 3, $\mu = 1$ for types 2, 4, and 5, $\mu = 0.5$ for type 6. For the sixth case $a = -0.25$. In case 1 through 5 the remaining parameter was chosen to match a unit variance.

2.1.2 Scale function and speed measure

The scale and the speed measure densities of the diffusion (2.1) are defined by

$$s(x) = \exp\left(-2\int_{x_0}^{x} \frac{\mu(u)}{\sigma^2(u)}du\right) \quad \text{and} \quad m(x) = \frac{1}{s(x)\sigma^2(x)}$$

where x_0 is a fixed point in I, the exact value is not important. The scale function S is defined as an antiderivative of s. Note that the scale and speed densities uniquely determines the drift and diffusion coefficients. When studying the behavior of scalar diffusions both measures are indispensable. For instance the non-explosive diffusions can be characterized in terms of their scale function and speed measure.

Theorem 2.1.1 *(Fellers test for explosion) The diffusion (2.1) is non-explosive if and only if* $\lim_{a\to r}\int_{x_0}^{a}\{S(a)-S(x)\}m(x)dx = \infty$ *and* $\lim_{b\to l}\int_{b}^{x_0}\{S(x)-S(b)\}m(x)dx = \infty$.

An often employed sufficient condition for non-explosiveness is

$$S(l) = -\infty \textit{ and } S(r) = \infty. \tag{2.3}$$

This is also known as the recurrence condition as it implies $P(\inf_t X_t = l, \sup_t X_t = r) = 1$. If condition (2.3) holds true, the boundaries cannot even be reached in the limit as $t \to \infty$, while on the contrary, if for instance $S(l) > -\infty$, then $P(\lim_{t\to T} X_t = l) > 0$ and the boundary l is said to be attracting.

Example 2.1.1 *The scale and speed densities of the Pearson diffusions (B.1) are*

$$s(x) = \exp\left(\int_{x_0}^{x}\frac{u-\mu}{au^2+bu+c}du\right) \quad \text{and} \quad m(x) = \frac{1}{2\theta s(x)(ax^2+bx+c)}$$

where x_0 is a fixed point such that $ax_0^2 + bx_0 + c > 0$. The speed and scale densities of the individual subclasses are given in table 2.2. △

	scale density $s(x)$	speed measure density $m(x)$
1	$\exp(\frac{(x-\mu)^2}{2})$	$\exp(-\frac{(x-\mu)^2}{2})$
2	$x^{-\mu}\exp(x)$	$x^{\mu-1}\exp(-x)$
3	$(x^2+1)^{\frac{1}{2a}}\exp(-\frac{\mu}{a}\tan^{-1}x)$	$(x^2+1)^{-\frac{1}{2a}-1}\exp(\frac{\mu}{a}\tan^{-1}x)$
4	$x^{\frac{1}{a}}\exp(\frac{\mu}{ax})$	$x^{-\frac{1}{a}-2}\exp(-\frac{\mu}{ax})$
5	$(1+x)^{\frac{\mu+1}{a}}x^{-\frac{\mu}{a}}$	$(1+x)^{-\frac{\mu+1}{a}-1}x^{\frac{\mu}{a}-1}$
6	$(1-x)^{\frac{1-\mu}{a}}x^{\frac{\mu}{a}}$	$(1-x)^{-\frac{1-\mu}{a}-1}x^{-\frac{\mu}{a}-1}$

Table 2.2: Scale and speed measure densities of the standard type Pearson diffusions.

2.1.3 Boundary classification and stationarity

In statistical applications the stationary diffusions are considered the most tractable. If for instance the diffusion gets absorbed at the right boundary at time $T = k$ the consecutive observations are totally uninformative. Fortunately simple and explicit condition characterize the stationary scalar diffusions. At first we briefly summarize the boundary classification scheme that characterizes the behavior of the diffusion near its right boundary, see Karlin & Taylor (1981) for details. The conditions for the left boundary are analogous.
Define $S(r) = \int_{x_0}^{r} s(x)dx$, $M(r) = \int_{x_0}^{r} m(x)dx$, $\Sigma(r) = \lim_{a\to r} \int_{x_0}^{a} \{S(a) - S(x)\}m(x)dx$, and $N(r) = \int_{x_0}^{r} \{S(x) - S(x_0)\}m(x)dx$, where x_0 is some interior point in I. The behavior of the diffusion near its right boundary is given by one of the following exclusive categories, see table 2.3 for a brief resume.

A regular boundary can be reached and left again in finite time. We consider only the case where the regular boundary is made instantaneously reflecting. The boundary r is regular if and only if $S(r) < \infty$ and $M(r) < \infty$.

Please note that a diffusion with instantaneously reflecting boundaries may be ergodic even though it hits the boundary in finite time. An example of such a process is the square root process with $\alpha \leq 1$. Many papers are overly restrictive when focusing solely on non-explosive diffusions.

An exit boundary can with positive probability be reached in finite time but never left again. The boundary r is exit if and only if $S(r) < \infty$, $M(r) = \infty$, and $\Sigma(r) < \infty$.

An entrance boundary can never be reached from within I. However the diffusion may be initialized at the entrance whereupon it leaves never to return again. The boundary r is entrance if and only if $S(r) = \infty$, $M(r) < \infty$, and $N(r) < \infty$.

The final category covers any other case.

A natural boundary cannot be reached in finite time and cannot be used as starting point for the diffusion. However, it may happen that the boundary is attained as limit as $t \to \infty$. The boundary r is natural if and only if $\Sigma(r) = \infty$ and $N(r) = \infty$. Note that $P(\lim_{t\to\infty} X_t = r) > 0$ if and only if $S(r) < \infty$.

$S(r)$	$M(r)$	boundary classification
finite	finite	r is regular
finite	∞	r is exit if $\Sigma(r) < \infty$ and otherwise natural
∞	finite	r is entrance if $N(r) < \infty$ and otherwise natural
∞	∞	r is natural

Table 2.3: This table summarizes the classification of the right boundary. The quantities $S(r)$, $M(r)$, $\Sigma(r)$, and $N(r)$ are defined in the above. Similar criteria are valid for the left boundary.

Note that the diffusions with natural boundaries can behave quite differently according to further subclassifications. A diffusion with both boundaries natural can be positive recurrent ($S(x) = \infty$ and $M(x) < \infty$ for $x = l, r$), null-recurrent ($S(x) = \infty$ and $M(x) = \infty$ for $x = l, r$), or non-recurrent ($S(x) < \infty$ for $x = l, r$).

A key result is that a stationary scalar diffusion has invariant distribution that is proportional to the speed measure. This for instance is used by Bibby, Skovgaard & Sørensen (2005) to construct diffusion models with pre-specified marginals.

Theorem 2.1.2 *Suppose that the diffusion (2.1) has boundaries that are entrance, natural or regular with instantaneous reflection, then an invariant distribution exists if and only if the speed measure is finite. Furthermore the invariant distribution is unique with density given by $m(x)/\int_l^r m(x)dx$.*

Example 2.1.2 *The invariant densities of the stationary Pearson diffusion are specified by table 2.4. Most of the invariant distributions are well known. The name Pearson diffusion is due to the fact that the invariant densities all belong to the Pearson system, Pearson (1895), as*

$$\frac{dm(x)}{dx} = -\frac{(2a+1)x - \mu + b}{ax^2 + bx + c} m(x).$$

Just like the Pearson densities the Pearson diffusions can be positive, negative, real valued, or bounded, symmetric or skewed, and heavy- or light-tailed. The class 3 marginals have a type IV Pearson distribution which is a skewed kind of t-distribution. The class may thus be of interest in say financial applications. △

	speed measure density $m(x)$	integrable for	type
1	$\exp(-\frac{(x-\mu)^2}{2})$	all	normal
2	$x^{\mu-1}\exp(-x)$	$\mu > 0$	Gamma
3	$(x^2+1)^{-\frac{1}{2a}-1}\exp(\frac{\mu}{a}\tan^{-1}x)$	$a > 0$	(skewed) t
4	$x^{-\frac{1}{a}-2}\exp(-\frac{\mu}{ax})$	$a, \mu > 0$	inverse Gamma
5	$(1+x)^{-\frac{\mu+1}{a}-1}x^{\frac{\mu}{a}-1}$	$a, \mu > 0$	(scaled) F
6	$(1-x)^{-\frac{1-\mu}{a}-1}x^{-\frac{\mu}{a}-1}$	$a < 0$ and $0 < \mu < 1$	Beta

Table 2.4: Types of integrable speed measure densities of the standard type Pearson diffusions.

2.1.4 The transition probabilities and their generator

The scalar diffusion (2.1) is a strong Markov chain. As the diffusion is regular (see Karatzas & Shreve (1991), pg. 344) the transition probabilities have continuous densities, say $p(t, x, y)$ satisfying Kolmogorov's backward equation

$$\frac{\partial p(t,x,y)}{\partial t} = \mu(x)\frac{\partial p(t,x,y)}{\partial y} + \frac{\sigma^2(x)}{2}\frac{\partial^2 p(t,x,y)}{\partial y^2}$$

and the Kolmogorov forward or Fokker Planck equation

$$\frac{\partial p(t,x,y)}{\partial t} = \frac{\partial p(t,x,y)\mu(y)}{\partial y} + \frac{1}{2}\frac{\partial^2 p(t,x,y)\sigma^2(y)}{\partial y^2}.$$

Save from a few simple diffusions like the Ornstein-Uhlenbeck and square root process, the functional form of the transition densities are hardly ever known. As a consequence likelihood inference for diffusion models is typically only feasible through computer intensive methods. Simulated likelihood inference is discussed in section 3.1 below.

The infinitesimal generator

In discrete time the distribution of a stationary Markov chain is uniquely determined by its one-step transition operator which is therefore an important object to study. In continuous time the zero-step transition operator is trivial, but the derivative of the transition operators at time zero is highly informative. This mapping is known as the infinitesimal generator of the Markov process.

For a strictly stationary diffusion let $P_t f(x) = E\{f(X_t)|X_0 = x\}$ define the t-step transition operator on $L_2(\pi)$ where π is the invariant distribution. The infinitesimal generator is defined as $\mathcal{A}f = \lim_{t\to 0}(P_t f - f)/t$ whenever the limit exists. The infinitesimal generator uniquely determines the transition semi-group and hence the distribution of the Markov chain. In case of a stationary scalar diffusion it is well known the generator coincides with the differential operator

$$\mathcal{L}f = \mu f' + (1/2)\sigma^2 f''.$$

In fact $\mathcal{A}$ is the restriction of $\mathcal{L}$ to the domain consisting of all functions $\psi \in L_2(\pi)$ for which ψ' is absolutely continuous, $\mathcal{L}\psi \in L_2(\pi)$, and

$$\lim_{x\to l}\frac{\psi'(x)}{s(x)} = 0 \text{ and } \lim_{x\to r}\frac{\psi'(x)}{s(x)} = 0.$$

The last condition is automatically fulfilled in case the boundary is either natural or instantaneously reflecting. Only entrance boundaries need additional checking, see Hansen, Scheinkman & Touzi (1998). Note that the generator of a strictly stationary scalar diffusion is self adjoint which in turn implies that these diffusions are time reversible, i.e. (X_s, X_t) has the same distribution as (X_t, X_s) for all $s, t \geq 0$, see Ritz (2000).

The spectrum of the generator plays a central part when studying the mixing properties of diffusions, see section 2.1.5, and in several applications where it is used to generate moment conditions for estimating the parameters, see sections 3.2 and 3.3. An eigenfunction of the generator is a function $\phi \in \mathcal{D}$ satisfying $\mathcal{A}\phi = -\lambda\phi$ for some eigenvalue $-\lambda \leq 0$ (the generator is negative semidefinite, hence all eigenvalues are non-positive). The infinitesimal generator and the transition operators share their eigenfunction and the eigenvalues are linked by the exponential function. I.e. if ϕ is an eigenfunction of the generator with eigenvalue $-\lambda$, then $E\{\phi(X_t)|X_0\} = e^{-\lambda t}\phi(X_0)$. This is used by Kessler & Sørensen (1999) in the construction of the martingale estimating functions which we further explored in Forman & Sørensen (2006) in case of the Pearson diffusions.

Example 2.1.3 *The generator of the Pearson diffusion (B.1) is given by*

$$\mathcal{L}f(x) = -\theta(x-\mu)f'(x) + \theta(ax^2 + bx + c)f''(x).$$

In particular the generator maps a square integrable polynomial to a polynomial of at most the same degree. Recursive formula for the polynomial eigenfunctions can be found in our paper Forman & Sørensen (2006). △

An alternative specification

An alternative specification for scalar diffusions was suggested by Hansen, Scheinkman & Touzi (1998). It is given by a triple (q, ψ, κ) where κ is a positive constant and q and ψ are functions on $I =]l, r[$ satisfying that

- q is a strictly positive continuous density on $I =]l, r[$.
- $\psi \in C^2(I)$ with $\psi' > 0$, $\int_l^r \psi(x)^2 q(x)dx < \infty$, and $\int_l^r \psi(x)q(x)dx = 0$.

Hansen, Scheinkman & Touzi (1998) show that (q, ψ, κ) determines a unique stationary diffusion on I with scale density $s(x) = \psi'(x)/\{2\kappa \int_l^r \psi(y)q(y)dy\}$ and speed measure density $m(x) = q(x)$. The drift and diffusion coefficients of the diffusion are thus given by

$$\sigma^2(x) = \frac{2\kappa \int_x^r \psi(y)q(y)dy}{\psi'(x)q(x)}, \quad \mu(x) = -\frac{\sigma^2(x)\psi''(x) + \kappa\psi(x)}{2\psi'(x)}.$$

By construction ψ is an eigenfunction of the generator of the diffusion and $-\kappa$ is the corresponding eigenvalue which is also the maximum non-zero eigenvalue the so-called spectral gap. In particular the diffusion is ρ-mixing, see section 2.1.5 below.

Example 2.1.4 *The triple $\{\pi(x), -\theta(x-\mu), \theta\}$ where π has mean μ and finite second order moment corresponds to a diffusion with linear drift $-\theta(x-\mu)$ and invariant density π. Diffusions of this kind were studied in Aït-Sahalia (1996a) and Bibby, Skovgaard & Sørensen (2005).* △

Spectral representation of the transition probabilities

The spectral representation of the transition probabilities is outlined in Karlin & Taylor (1981). In order to find an explicit expression of the function $u(t, x) = E\{f(X_t)|X_0 = x\}$ where f is continuous and bounded on I, it is useful to note that u satisfies the partial differential equation

$$\frac{\partial u(t, x)}{\partial t} = \mathcal{L}u(t, x) \tag{2.4}$$

with initial condition $u(0, x) = f(x)$, and the further restriction that $\partial u(t, l)/\partial x = 0$ if l is a reflecting boundary, and similarly if r is reflecting. By separations of variables a solution is sought out among the functions of the form $u(t, x) = c(t)\phi(x)$ where $dc(t)/dt = -\lambda c(t)$ and $d\phi(x)/dx = -\lambda\mathcal{L}\phi(x)$ for some $\lambda \geq 0$. Obviously this implies that $c(t) = ce^{-\lambda t}$ and $\phi(x)$ is an eigenvalue of the generator. Assuming an entirely discrete spectrum $\{\lambda_n\}_{n\in\mathbb{N}}$ with associated eigenfunctions $\{\phi_n(x)\}_{n\in\mathbb{N}}$ the solution is given by

$$u(t, x) = \sum_{n=0}^{\infty} c_n e^{-\lambda_n t}\phi_n(x), \quad c_n = \int_l^r f(x)\phi_n(x)m(x)dx \cdot \left(\int_l^r \phi_n(x)^2 m(x)dx\right)^{-1},$$

where $m(x)$ is the speed measure density. An additional argument is needed to show that this is the unique solution of (2.4 under the given boundary conditions.
The spectral representation of the transition densities,

$$p(t,x,y) = m(y) \cdot \sum_{n=0}^{\infty} e^{-\lambda_n t} \phi_n(x)\phi_n(y) \left(\int_l^r \phi_n(u)^2 m(u) du \right)^{-1}.$$

is obtained applying the above to $f(x) = (b-a)^{-1} 1_{(a;b)}(x)$ and letting (a,b) shrink to $\{y\}$. If the spectrum has a continuous component, the desired expansion takes a more general form

$$m(y) \cdot \int_0^\infty e^{-\lambda t} \phi_\lambda(x)\phi_\lambda(y) d\psi(\lambda)$$

where ψ is a measure on $[0,\infty[$. Wong (1964) derived the spectral representations of the transition probabilities for most of the Pearson diffusions.

2.1.5 Mixing

Recall that the α and ρ mixing coefficients of a stationary continuous time Markov process are given by

$$\begin{aligned} \alpha_t(X) &= \sup_{A,B \in \mathbb{B}} |P(X_0 \in A, X_t \in B) - P(X_0 \in A)P(X_t \in B)| \\ \rho_t(X) &= \sup_{f,g \in L_2(\pi)} |\operatorname{Cor}(f(X_0), g(X_t))| \end{aligned}$$

and that $\alpha_t(X) \leq 4\rho_t(X) \leq 1$. If $\alpha_t(X) \to 0$ ($\rho_t(X) \to 0$) as $t \to \infty$ then the diffusion is α-mixing (ρ-mixing) and in particular ergodic. The ρ-mixing coefficients of a scalar diffusion are determined by the spectrum of the infinitesimal generator, see Genon-Catalot, Jeantheau & Laredo (2000).

Theorem 2.1.3 *Suppose that the scalar diffusion (2.1) is strictly stationary, then the ρ-mixing coefficients are given by $\rho_t(X) = e^{-\lambda_0 t}$ where*

$$\lambda_0 = \sup\{ \textstyle\int_l^r f(x)\mathcal{A}f(x)\pi(x)dx / \int_l^r f(x)^2 \pi(x) dx \ : \ f \in \mathcal{D}(\mathcal{A}),\ \int_l^r f(x)\pi(x)dx = 0 \}.$$

Further $\lambda_0 > 0$ if and only if zero is an isolated point of the spectrum of the generator. This being the case λ_0 is the so-called spectral gap,

All stationary Pearson diffusions with second order moment are ρ-mixing.

Example 2.1.5 *Suppose that the diffusion (2.1) is strictly stationary with second order moment and linear drift $\mu(x) = -\theta(x-\mu)$. Then $\rho_t(X) = e^{-\theta t}$. This follows by appealing to the alternative specification of Hansen, Scheinkman & Touzi (1998), see section 2.1.4 above.* △

Please note that the mixing coefficients are shared by the discretely sampled diffusion $\{X_{i\Delta}\}_{i\in\mathbb{N}}$ for any $\Delta > 0$. The mixing coefficients play a key role if the following central limit theorem is to apply, see Doukhan (1994).

Theorem 2.1.4 *Suppose that the stochastic process* $\{Y_t\}_{t\in\mathbb{N}}$ *is stationary and* α*-mixing with* $\sum_{t=1}^{\infty}\alpha_t(Y)^{\delta/(2+\delta)} < \infty$ *for some* $\delta > 0$ *such that* $E|Y_t|^{2+\delta} < \infty$*, then*

$$n^{-1/2}\sum_{t=1}^{n}(Y_t - EY_t) \overset{\mathcal{D}}{\to} \mathcal{N}(0, \tau^2)$$

where $\tau^2 = \mathrm{Var}(Y_1) + 2\sum_{t=2}^{\infty}\mathrm{Cov}(Y_1, Y_t)$*.*

The mixing properties of scalar diffusions carry over to non-Markovian transformations such as the diffusion type models considered in the coming sections. For instance the mixing properties of a set of independent diffusions is inherited by their sum since $\alpha_t(X^{(1)} + \ldots + X^{(m)}) \le \alpha_t(X^{(1)}) + \ldots + \alpha_t(X^{(1)})$, see Doukhan (1994). For these models martingale estimating functions are no longer available. Hence, the more general central limit theorems is needed for proving asymptotic normality of their estimators.
The following results are deduced from Genon-Catalot, Jeantheau & Laredo (2000) and Rogers & Williams (1987).

Theorem 2.1.5 *Assume that the diffusion (2.1) is strictly stationary and satisfies the recurrence condition (2.3). Then the diffusion* $\{X_t\}_{t\ge 0}$ *is ergodic. If further* $\mu \in C^1(I)$ *and* $\sigma^2 \in C^2(I)$ *satisfies that for some constant* $K > 0$

$$|\mu(x)| \le K(1 + |x|) \text{ and } \sigma^2(x) \le K(1 + x^2) \text{ for all } x \in I,$$

Then the diffusion is α*-mixing. Finally, if in addition* $\lim_{x\to l,r} m(x)\sigma(x) = 0$ *and both of the limits*

$$\lim_{x\to l}\{\sigma'(x) - 2\mu(x)/\sigma(x)\}^{-1} \text{ and } \lim_{x\to r}\{\sigma'(x) - 2\mu(x)/\sigma(x)\}^{-1}$$

exist and are finite, then $\{X_t\}_{t\ge 0}$ *is* ρ*-mixing.*

2.2 Diffusion driven models

The scalar diffusion processes can be used as building blocks to obtain more general diffusion-type models which typically are not Markovian. In what follows we consider integrated diffusions, summed diffusions, and stochastic volatility models. A new idea for modeling multivariate stochastic volatility based on scalar diffusions is presented in section 2.2.4. We outline the distinctive features of the models and briefly discus how they can be analyzed. Further details on the various statistical methods are found in the quoted papers and in chapter 3 and the references therein.

2.2.1 Summed diffusions

Sums of mean reverting linear drift diffusions constitute a flexible class of stochastic process models having a particularly nice and explicit autocorrelation function. Bibby, Skovgaard & Sørensen (2005) show that the summed diffusions fit turbulence data well. Also the summed diffusions are appropriate for modeling stochastic volatility in analogy with the superpositions of Barndorff-Nielsen & Shephard (2001a).

The construction is as follows. Let $X_t = X_{1,t} + \ldots + X_{m,t}$ where $\{X_{1,t}\}_{t\geq 0}, \ldots, \{X_{m,t}\}_{t\geq 0}$ are independent diffusions, solving

$$dX_{i,t} = -\theta_i(X_{i,t} - \mu_i) + \sigma_i(X_{i,t})dB_{i,t}, \quad i = 1, \ldots, m \tag{2.5}$$

where $\theta_1, \ldots, \theta_m > 0$ and the diffusion coefficients $\sigma_1, \ldots, \sigma_m$ are continuous and strictly positive. If all of the underlying diffusions are stationary with finite second moment, then so is the summed diffusion $\{X_t\}_{t\geq 0}$ and its autocorrelation function is given by

$$\rho(t) = \phi_1 \exp(-\theta_1 t) + \ldots + \phi_M \exp(-\theta_M t) \tag{2.6}$$

with $\phi_i = \mathrm{Var}(X_{i,t})/\{\mathrm{Var}(X_{1,t}) + \cdots + \mathrm{Var}(X_{m,t})\}$. Note that $\phi_1 + \ldots + \phi_m = 1$. The expectation of X_t is $\mu_1 + \cdots + \mu_m$. The joint moments of the summed diffusion are linked to those of the underlying diffusions as for instance,

$$E(X_s^k X_t^\ell) = \sum\sum \binom{k}{k_1, \ldots, k_m}\binom{\ell}{\ell_1, \ldots, \ell_m} E(X_{1,s}^{k_1} X_{1,t}^{\ell_1}) \cdots E(X_{m,s}^{k_m} X_{m,t}^{\ell_m})$$

where the summation is over $k_1, \ldots, k_m \geq 0$ such that $k_1 + \ldots + k_m = k$ and similarly for the ℓ's. The diffusion coefficients can be chosen to accommodate a vide range of marginal distributions. Sums of diffusions with a pre-specified marginal distribution were considered by Bibby & Sørensen (2003) and Bibby, Skovgaard & Sørensen (2005). It was shown that for any $\theta_i > 0$ the mean μ_i and the diffusion coefficient σ_i can be chosen to match continuous, strictly positive, and bounded density f_i with second order moment. To be specific the selection

$$\mu_i = \int_l^u f_i(x)dx \quad \text{and} \quad \sigma_i^2(x) = f_i(x)^{-1} 2\theta_i \int_l^x (\mu_i - y) f_i(y) dy,$$

implies that (2.5) has a unique stationary weak solution with marginal density f_i and autocorrelation function $\rho_i(t) = \exp(-\theta_i t)$. It follows that the summed diffusion $\{X_t\}_{t\geq 0}$

among many unknown mixture distributions admits any infinitely divisible marginal density subject to some regularity conditions.

It is important to notice that the time changed process $\{X_{i,\delta t}\}_{t \geq 0}$ where $\delta > 0$ has the same marginal distribution as $\{X_{i,t}\}_{t \geq 0}$ and autocorrelation function $\tilde{\rho}_i(t) = e^{-\theta_i \delta t}$. This implies that θ_i measures the speed at which the underlying diffusion evolves with time. Hence, the summed diffusion process can be interpreted as an aggregation over multiple time scales. If for instance $\{X_t\}_{t \geq 0}$ is the sum of two independent diffusions we can think of the slower moving diffusion as a stochastic trend and the faster moving diffusion as noise.

Statistical analysis of summed diffusion models can in principle be based on the likelihood function which being far from explicit can be simulated by use of some of the algorithms described in section 3.1 below. The same algorithms can be modified to output the uniform residuals which can be used for diagnostics, see section 3.5. How well the algorithms in fact perform is an open question; Simulated likelihood has never been attempted for summed diffusions.

Moment based estimation is far more tractable due to the explicit autocorrelation function. My paper Forman (2005) investigate least squares estimators for the autocorrelation parameters and find these to behave well in theory as well as in practice. The related overidentifying restrictions test, section 3.3.3, can be used to asses the goodness of fit for the autocorrelation function. In particular, it demonstrated that a certain forward selection procedure yields consistent estimates of the number of underlying diffusions. For fitting a full model the parameters of the pre-specified marginal distribution can be estimated for instance by means of marginal estimating functions, see section 3.2.3, or by means of the nonparametric methods of Aït-Sahalia (1996a) which can also be used to asses the fit of the marginal density, see section 3.4.3.

In our paper Forman & Sørensen (2006) we consider sums of Pearson diffusions and show how these can be fitted using suitable prediction based estimating functions, see section 3.2.2. If the predictors and targeted variables are chosen among powers of the observations, we obtain explicit expressions of an optimal prediction based estimating function. Goodness of fit can be based on the overidentifying restrictions test, section 3.3.3 below, or on the downsampled estimating function as in Forman, Markusen & Sørensen (2007), which is shown to be successful in distinguishing a plain diffusions from a sum.

The summed diffusions are related to the Ornstein-Uhlenbeck type processes studied in Barndorff-Nielsen, Jensen & Sørensen (1998), Barndorff-Nielsen & Shephard (2001b), and Barndorff-Nielsen & Shephard (2001a). These models are based on the solutions of the stochastic differential equation

$$dX_{i,t} = -\theta_i X_{i,t} dt + dZ_{i,t} \quad i = 1, \ldots, m \tag{2.7}$$

where the $(Z_{i,t})_{t \geq 0}$'s are independent homogenous Levy process. The summed Ornstein-Uhlenbeck type models share the flexibility of the summed diffusions in having the same form of autocorrelation function (2.6) and a large range of admissible marginal distributions depending on the choice of underlying Levy processes. A noteworthy difference between the two classes of models is that save from the ordinary Ornstein-Uhlenbeck process driven by Brownian motion, all other Ornstein-Uhlenbeck type processes have jumps. The Ornstein-Uhlenbeck type models can be analyzed by the same means as the summed diffusions, see for instance Forman (2005). It is worth noting that the discrete time process $(X_{i,t})_{t\in\mathbb{N}}$ forms an auto-regression

$$X_{i,t+1} = \lambda_i \cdot X_{i,t} + \varepsilon_{i,t}$$

where $(\varepsilon_{i,t})_{t\in\mathbb{N}}$ are i.i.d. In particular the conditional moments of $\{X_{i,t}\}_{t\in\mathbb{N}}$ up to any order can easily be calculated. Masuda (2004) study the mixing properties of the Ornstein-Uhlenbeck type processes.

2.2.2 Integrated diffusions

Integrated observations occur when a diffusion cannot be observed directly, for instance if the diffusion $\{X_t\}_{t\geq 0}$ is observed after passage through an electronic filter. Ditlevsen, Ditlevsen & Andersen (2002) makes inference for the paleo-temperature by use of an integrated Ornstein-Uhlenbeck model. The paleo-temperature cannot be observed directly, but the isotope ratio $^{18}O/^{16}O$ measured as an average in pieces from the ice core serves as a proxy. Another important example is realized volatility, see Andersen & Bollerslev (1998), Andersen et al. (2001), Barndorff-Nielsen & Shephard (2002), and section 2.2.3 below. Daily realized volatility is computed by summing squared intraday returns. Andersen et al. (2001) argue that for practical purposes realized volatility based on high frequency data is free of measurement error. Hence, integrated volatility can be treated as observed and analyzed by means of integrated diffusion models.

To be specific, let the stationary diffusion, and the integrated observations be given by

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t, \quad Y_i = \frac{1}{\Delta}\int_{(i-1)\Delta}^{i\Delta} X_s\, ds$$

for some fixed Δ. Since $\{X_t\}_{t\geq 0}$ is stationary, the integrated observations $\{Y_i\}_{i\in\mathbb{N}}$ form a stationary process with the same mixing properties as $\{X_t\}_{t\geq 0}$. The mean of the integrated observations is identical to that of the underlying diffusion. The joint moments of the integrated observations are linked to the joint moments of the underlying diffusion by

$$E(Y_i^k Y_j^\ell) = \Delta^{-(k+\ell)} \int_{[(i-1)\Delta,i\Delta]^k\times[(j-1)\Delta,j\Delta]^\ell} E\{X_{s_1}\cdots X_{s_k} X_{t_1}\cdots X_{t_\ell}\} ds_1 \dots ds_k dt_1 \dots dt_\ell.$$

Please note that the domain of integration can be reduced considerably by symmetry arguments. If the underlying diffusion has a mean reverting linear drift yielding its autocorrelation function to be exponentially decreasing with coefficient θ, then the autocovariance function of the integrated observations is given by

$$\mathrm{Var}(Y_i) = \frac{2\,\mathrm{Var}(X_t)(\theta\Delta + e^{-\theta\Delta} - 1)}{(\theta\Delta)^2}, \quad \mathrm{Cov}(Y_i, Y_{i+j}) = \frac{\mathrm{Var}(X_t)(1-e^{-\theta\Delta})^2 e^{-(j-1)\theta\Delta}}{(\theta\Delta)^2}.$$

Save from the integrated Ornstein-Uhlenbeck process which is Gaussian, the invariant distribution of the integrated diffusion does not have a simple closed form expression. Gloter (2001) considers estimation in the integrated Ornstein-Uhlenbeck process. Observing that the integrated observations form a Gaussian ARMA(1,1) process he shows that the model can be efficiently estimated by using the Whittle approximation to the likelihood function. For most other integrated diffusions likelihood inference is only feasible through simulation. The algorithms of Pitt, Chib & Shephard (2006) and Durham & Gallant (2002) made to form inference for stochastic volatility models should work just as well when applied to integrated diffusions, see section 3.1. When adequately modified the same algorithms output uniform residuals which can be used for diagnostics, see section 3.5.

Moment based estimation is considered by Bollerslev & Zhou (2002) who find the first two conditional moments of an integrated square root model and use these to construct GMM-estimators. Ditlevsen & Sørensen (2004) propose prediction based estimating functions where predictors and targeted variables are found among the powers of past and present observations. Integrated Ornstein-Uhlenbeck and square root processes are exemplified. These results are further generalized in Forman & Sørensen (2006) where explicit and optimal prediction-based estimating functions are found for a general underlying Pearson diffusion.

In the high-frequency setting the integrated observations approaches the underlying diffusion. Hence, the parameters can be estimated by use of an Euler-type approximation. Gloter (2006) provides the appropriate approximation accounting for the fact that the integrated diffusion is no longer a Markov chain.

A more general model with similar inference is obtained when the underlying diffusion is replaced by the sum of independent diffusions or by an Ornstein-Uhlenbeck type process. See Sørensen (2000), Barndorff-Nielsen & Shephard (2001a), Bollerslev & Zhou (2002), Barndorff-Nielsen & Shephard (2002), and Forman & Sørensen (2006) for relevant examples.

2.2.3 Diffusion driven stochastic volatility.

Stochastic volatility models are mainly used in financial economics as e.g. models for exchange rates and stock prices. Here we focus on continuous-time, diffusion driven stochastic volatility models solely, see Shephard (2005) for a general introduction. A stochastic volatility model is a generalization of the Black-Scholes model for the logarithm of an asset price that takes into account the empirical finding that the variance varies randomly over time. Following Hull & White (1987) the variance process or the volatility is often modeled as a diffusion. Thus the stochastic volatility model is a partially observed two dimensional diffusion evolving according to,

$$dX_t = (\kappa + \beta v_t)dt + \sqrt{v_t}dW_t, \quad dv_t = \mu(v_t) + \sigma(v_t)dB_t$$

where $\{W_t\}_{t\geq 0}$ and $\{B_t\}_{t\geq 0}$ are independent Brownian motions. The volatility $\{v_t\}_{t\geq 0}$ by assumption cannot be observed directly. Given the volatility $\{v_t\}_{t\geq 0}$, the observed returns

$Y_i = X_{i\Delta} - X_{(i-1)\Delta}$ are independent and normally distributed with mean M_i and variance S_i given by

$$M_i = \kappa\Delta + \beta S_i, \quad S_i = \int_{(i-1)\Delta}^{i\Delta} v_t dt.$$

Please note that the sequence of conditional variances the so-called actual volatility process $\{S_i\}_{i\in\mathbb{N}}$ is an integrated diffusion, see section 2.2.2 above. Simple examples of volatility models specify $\{v_t\}_{t\geq 0}$ as a square root process or as the exponential of an Ornstein-Uhlenbeck model. Another simple specification models $\{v_t\}_{t\geq 0}$ as a Pearson diffusion from the fourth class of Forman & Sørensen (2006) which can be interpreted as a continuous time analogue to the GARCH(1,1) model, see Nelson (1990).
We assume that $\{v_t\}_{t\geq 0}$ is stationary, thus so is the return process $\{Y_i\}_{i\in\mathbb{N}}$. The returns inherit the mixing properties of the volatility process, see Genon-Catalot, Jeantheau & Laredo (2000). The invariant distribution is the normal mixture with respect to the invariant distribution of the integrated volatility process which is typically unknown. The means and variances of returns are given by $E(Y_i) = \kappa\Delta + \beta E(S_i)$ and $\operatorname{Var}(Y_i) = E(S_i) + \beta^2 \operatorname{Var}(S_i)$. The joint moments can be calculated by

$$E(Y_i^k Y_j^\ell) = \sum\sum \binom{k}{k_1\ k_2\ k_3}\binom{\ell}{\ell_1\ \ell_2\ \ell_3}(\kappa\Delta)^{k_1+\ell_1}\beta^{k_2+\ell_2}\zeta_{k_3}\zeta_{\ell_3}E(S_i^{k_2+k_3/2}S_j^{\ell_2+\ell_3/2}),$$

where the sum is over integers $k_1, k_2, k_3 \geq 0$ such that $k_1+k_2+k_3 = k$ and similarly for the ℓ's. The constant ζ_m is the m'th order moment of the standard normal distribution. Note that $\zeta_m = 0$ when m is odd. Hence, the problems reduces to finding the joint moments of an integrated diffusion, see section 2.2.2 above. For instance the covariances are given by $\operatorname{Cov}(Y_i, Y_j) = \beta^2 \operatorname{Cov}(S_i, S_j)$ for $i \neq j$. If $\beta = 0$, then the returns are uncorrelated and more can be learned from the squared returns for which $\operatorname{Cov}(Y_i^2, Y_j^2) = \operatorname{Cov}(S_i, S_j)$ for $i \neq j$ (assuming $\beta = 0$).

As the likelihood function is not readily available, the statistical analysis of volatility models has been a challenge through the last two or three decades resulting in a vast number of papers, see Shephard (2005) for a selective overview. It is important to notice that most of the econometric papers from the 1980's and 1990's are concerned with discrete time stochastic volatility models. The derived estimating schemes should be applied with caution to the continuous time models as the discretization scheme may be the source of bias. A classical approach suggested by Harvey, Ruiz & Shephard (1994) is to apply the Gaussian quasi-likelihood to the log-transformed returns.

Today likelihood inference is indeed applicable using suitable simulation schemes, see section 3.1 below. We emphasize the Markov chain Monte Carlo algorithm, e.g. Pitt, Chib & Shephard (2006), and the importance sampler, e.g. Durham & Gallant (2002). In addition these algorithms can output one-step ahead predictions and uniform residuals to be used for diagnostics, see section 3.5. The efficiency gain of the maximum likelihood estimator on the quasi-likelihood and moment based estimators can be substantial, see Jacquier, Polson & Rossi (1994).

Moment based estimation is considered by for instance Melino & Turnbull (1990), Andersen & Sørensen (1996), Sørensen (2000), and Genon-Catalot, Jeantheau & Laredo (2000).

For continuous time models moment conditions can be found for instance by aid of the above formulae. In our paper Forman & Sørensen (2006) we find the explicit optimal estimating function based on prediction of powers of returns for the stochastic volatility models driven by a Pearson diffusion. In connection with moment based estimation it is natural to base goodness of fit on the overidentifying restrictions test, section 3.3.3 below, or by downsampling the estimating function, see Forman, Markusen & Sørensen (2007).

Another influential approach is that of indirect inference also known as the efficient method of moments, see Gourieroux, Monfort & Renault (1993) and Gallant & Tauchen (1996). An estimate is obtained by first introducing an auxiliary model, for instance a GARCH model, for which the maximum likelihood estimator is easy to compute. The second step is to simulate long time series from the stochastic volatility model searching for a set of parameter values that will match the auxiliary estimator obtained from the simulation with the one obtained from the data. At best the indirect estimator attains the efficiency of the intractable maximum likelihood estimator with considerably less computational effort, but much depends on the choice of auxiliary model. Gallant & Tauchen (1996) have particular recipes for making a sensible selection.

Recently, high frequency data has rendered the integrated volatilities almost observable through the so-called realized volatilities, which are estimates of the actual volatilities. These are easily derived by observing that the quadratic variation of $\{X_t\}_{t\geq 0}$ is given by

$$[X]_t = \int_0^t v_s ds.$$

For instance daily realized volatility is computed by summing squared intraday returns. Hence, high frequency stochastic volatility models can be analyzed by means of integrated diffusion models as suggested in Andersen et al. (2001). The statistical analysis of volatility models based on high frequency data is further discussed in Genon-Catalot, Jeantheau & Laredo (1999), Hoffmann (2002), and Barndorff-Nielsen & Shephard (2002).

A more general model with similar inference is obtained when the underlying diffusion is replaced by the sum of independent diffusions or by an Ornstein-Uhlenbeck type process, see section 2.2.1 above. Barndorff-Nielsen & Shephard (2001a) demonstrated that the autocorrelation function (2.6) of the summed diffusions fits empirical autocorrelation functions of volatility well, while an autocorrelation function like that of a single linear drift, mean reverting diffusion is too simple to obtain a good fit. In our paper Forman & Sørensen (2006) we derive explicit prediction based estimating functions for a volatility model where the underlying volatility is the sum of independent Pearson diffusions.

2.2.4 The construction of a multivariate volatility process

This section presents a new idea for modeling multivariate stochastic volatility based on scalar diffusions. In section 2.2.3 only univariate stochastic volatility models were considered. A general multivariate stochastic volatility process is the solution of the stochastic differential equation,

$$dX_t = (A + C\Sigma_t\Sigma_t^T) + \Sigma_t dB_t \tag{2.8}$$

where $\{B_t\}_{t\geq 0}$ is a k-dimensional Brownian motion and $\{\Sigma_t\}_{t\geq 0}$ is a k by k matrix valued stochastic process so that $V_t = \Sigma_t\Sigma_t^T$ is positive semidefinite for all t. In most of the existing models the spot volatility matrix is determined by a lower dimensional structure. That is $\Sigma_t\Sigma_t^T$ does not vary freely in the space of positive definite matrices. Some classical examples are found in Harvey, Ruiz & Shephard (1994), Danielsson (1998), Pitt & Shephard (1999b), Aguilar & West (2000), and Liesenfeld & Richard (2003). Most of these models are constricted in the sense that the conditional correlations are constant over time. However, there is evidence of time-varying correlations in multivariate financial time series, see Yu & Meyer (2004) for a comparative study of two-dimensional stochastic volatility models for exchange rates. Recent models for the volatility process are the Wishart diffusions studied in Philipov & Glickman (2006) and Gourieroux, Jasiak & Sufana (2004) and the Ornstein-Uhlenbeck type processes on the space of positive semidefinite matrices considered by Barndorff-Nielsen & Stelzer (2006).

We suggest modeling the matrix valued volatility process through its diagonal representation,

$$\Sigma_t\Sigma_t^T = O_t\Lambda_t O_t^T \tag{2.9}$$

where $\Lambda_t = \text{diag}\{\lambda_{j,t}\}$ contains the eigenvalues of $\Sigma_t\Sigma_t^T$ and O_t is the orthogonal matrix the columns of which contain the eigenvectors. Note that if $\{O_t\}_{t\geq 0}$ is assumed constant, the model of Harvey, Ruiz & Shephard (1994) is recovered. However, we aim at a random specification of O_t with the potential of hitting any value in the set of orthogonal matrices. To this end we appeal to the decomposition

$$O_t = \prod_{1\leq i<j\leq k} \Phi_{i,j,t}. \tag{2.10}$$

where $\Phi_{i,j,t}$ is the k by k matrix with elements equal to those of the identity save from the (i,j)-submatrix which has the form

$$\begin{pmatrix} \cos(\phi_{i,j,t}) & -\sin(\phi_{i,j,t}) \\ \sin(\phi_{i,j,t}) & \cos(\phi_{i,j,t}) \end{pmatrix} \tag{2.11}$$

Note that $\Phi_{i,j,t}$ is the matrix representing a turn in k-space. Visually speaking the standard base in $\mathbb{R}^k$ (represented by the identity matrix) is mapped into an other orthonormal base (represented by O_t) by performing a series of turns. Sequentially each pair of basis vectors is turned counter clockwise in the plane they span while the other basis vectors are fixed. The angles of the consecutive turns are $\phi_{1,2,t}, \ldots, \phi_{1,k,t}, \ldots, \phi_{k-1,k,t}$.

For a full model specification we need to model the diagonal elements $\lambda_1, \ldots, \lambda_k$ and the turning angles $\phi_{1,2}, \ldots, \phi_{1,k}, \ldots, \phi_{k-1,k}$. For instance the λ's could be specified as independent square root processes or the exponentials of independent Ornstein-Uhlenbeck processes. The model is completed by taking the angles to be a set of stochastic processes. Diffusions on $]-\pi/2;\pi/2[$ seem the natural choice for the $\phi_{i,j}$'s. A particularly tractable process occur when both angles and diagonal elements are modeled by suitable transformed Pearson diffusions. For the angles we can assume for instance the Ornstein-Uhlenbeck process on $]-\pi/2;\pi/2[$ introduced in Kessler & Sørensen (1999) or its asymmetric generalization derived by Larsen & Sørensen (2003). For the diagonal elements we can assume plain Pearson diffusions as long as these are non-negative. The non-negative Pearson diffusions are those from the second, fourth, and fifth class of Forman & Sørensen

(2006), i.e. the square root processes, the GARCH diffusions, and the Pearson diffusions with marginal F-distributions. All of these selections allow for explicit moment calculations, which in turn yield explicit estimating functions for fitting the model.

The 2D model

In two dimensions the volatility matrix takes the form (suppressing the dependence on t)

$$V_t = \begin{pmatrix} \cos(\phi)^2\lambda_1 + \sin(\phi)^2\lambda_2 & \cos(\phi)\sin(\phi)(\lambda_1 - \lambda_2) \\ \cos(\phi)\sin(\phi)(\lambda_1 - \lambda_2) & \sin(\phi)^2\lambda_1 + \cos(\phi)^2\lambda_2 \end{pmatrix} \tag{2.12}$$

In particular the conditional correlation is given by

$$\rho_{2\mathrm{D}} = \frac{\cos(\phi)\sin(\phi)(\lambda_1 - \lambda_2)}{\sqrt{(\cos(\phi)^2\lambda_1 + \sin(\phi)^2\lambda_2)(\sin(\phi)^2\lambda_1 + \cos(\phi)^2\lambda_2)}}. \tag{2.13}$$

Note that the correlation coefficient depends on λ_1 and λ_2 only through their quotient. For $\phi = 0$ and $\phi = \pm\pi/2$ the correlation equals zero. Otherwise $\rho_{2\mathrm{D}} = 0$ only if $\lambda_1 = \lambda_2$ and for $\frac{\lambda_1}{\lambda_2}$ tending to zero or infinity $|\rho_{2\mathrm{D}}| \to 1$. Given λ_1 and λ_2 maximum numerical correlation is attained for $|\cos(\phi)| = |\sin(\phi)| = \frac{1}{\sqrt{2}}$ in which case $|\rho_{2\mathrm{D}}| = \frac{|\lambda_1 - \lambda_2|}{\lambda_1 + \lambda_2}$. The sign of the correlation depends on the signs of $\sin(\phi)$ and $\lambda_1 - \lambda_2$. A positive correlation can be forced by choosing $\phi \in]0; \pi/2[$ and $\lambda_1 \geq \lambda_2$. For instance the latter is obtained by replacing λ_1 with $\lambda_1 + \lambda_2$. All in all we get an very flexible dynamic conditional correlation.

Example 2.2.1 *An example of a specific two-dimensional model is given by the choice of λ_1 and λ_2 being stationary square root processes,*

$$d\lambda_{i,t} = -\theta_i(\lambda_{i,t} - \alpha_i\beta_i)dt + \sqrt{2\theta_i\beta_i}dW_{i,t}$$

where $\alpha_i, \beta_i, \theta_i > 0$ for $i = 1, 2$ and $\{W_{1,t}\}_{t\geq 0}$ and $\{W_{1,t}\}_{t\geq 0}$ are independent Brownian motions. Further define ϕ to be an arcsine-transformed Jacobi diffusion as in Larsen & Sørensen (2003). As $\cos(\phi_t)$ is determined by $\cos(\phi_t) = \sqrt{1 - \sin(\phi_t)^2}$ we might as well model $Z_i = \sin(\phi_t)$ directly as the translated and rescaled Jacobi diffusion,

$$dZ_t = -\theta_3(Z_t - \mu)dt + \sqrt{2\theta_3\gamma(1 - Z_t^2)}dW_{3,t}$$

on $]-1; 1[$ where $-1 < \mu < 1$, $\gamma, \theta_3 > 0$, ... , and $\{W_{3,t}\}_{t\geq 0}$ is another Brownian motion independent of $\{W_{1,t}\}_{t\geq 0}$ and $\{W_{2,t}\}_{t\geq 0}$. Both the square root processes and the rescaled Jacobi diffusion are Pearson diffusions. Hence, recursive formula for computing explicit moments and conditional moments or any order are found in Forman & Sørensen (2006). $\triangle$

Assume for simplicity that $A = C = 0$, then given the volatility process $\{V_t\}_{t\geq 0}$ the two-dimensional returns $Y_i = X_{i\Delta} - X_{(i-1)\Delta}$ are independent and normal with mean zero and covariance matrix given by

$$S_i = \begin{pmatrix} \int_{(i-1)\Delta}^{i\Delta}\{\cos(\phi_t)^2\lambda_{1,t} + \sin(\phi_t)^2\lambda_{2,t}\}dt & \int_{(i-1)\Delta}^{i\Delta}\cos(\phi_t)\sin(\phi_t)\{\lambda_{1,t} - \lambda_{2,t}\}dt \\ \int_{(i-1)\Delta}^{i\Delta}\cos(\phi_t)\sin(\phi_t)\{\lambda_{1,t} - \lambda_{2,t}\}dt & \int_{(i-1)\Delta}^{i\Delta}\{\sin(\phi_t)^2\lambda_{1,t} + \cos(\phi_t)^2\lambda_{2,t}\}dt \end{pmatrix}$$

It follows that the mean return is $E(Y_i) = 0$ and the covariance matrix is $\mathrm{Cov}(Y_i) = E(S_i)$ implying that

$$\begin{aligned}
\mathrm{Var}(Y_{1,i}) &= \Delta[E\{1-\sin(\phi)^2\}E(\lambda_1) + E\{\sin(\phi)^2\}E(\lambda_2)] \\
\mathrm{Var}(Y_{1,i}) &= \Delta[E\{\sin(\phi)^2\}E(\lambda_1) + E\{1-\sin(\phi)^2\}E(\lambda_2)] \\
\mathrm{Cov}(Y_{1,i}, Y_{2,i}) &= \Delta E\{\sin(\phi)\sqrt{1-\sin(\phi)^2}\}\{E(\lambda_1) - E(\lambda_2)\}.
\end{aligned}$$

Save from the mean of $\sin(\phi)\sqrt{1-\sin(\phi)^2}$ all of these moments are explicitly known in the above example, and the problematic term can be found by numerical integration as the invariant distribution of $\sin(\phi)$ is merely a rescaled Beta distribution. In particular, if the Beta distribution is symmetric, then $E(\sin(\phi)\sqrt{1-\sin(\phi)^2}) = 0$. Further, let ζ_m denote the m'th order moment of the standard normal distribution, then for $i \neq j$ the joint moment $E(Y_{1,i}^k Y_{1,j}^\ell)$ is given by

$$\zeta_k\zeta_\ell \cdot \int_{[(i-1)\Delta;i\Delta]^k \times [(j-1)\Delta;j\Delta]^\ell} E\{f(s_1)\cdots f(s_k)g(t_1)\cdots g(t_\ell)\}ds_1 \ldots ds_k dt_1 \ldots dt_\ell,$$

where

$$\begin{aligned}
f(s) &= \{1-\sin(\phi_s)^2\}\lambda_{1,s} + \sin(\phi_s^2)\lambda_{2,s} \\
g(t) &= \sin(\phi_t)^2\lambda_{1,s} + \{1-\sin(\phi_t^2)\}\lambda_{2,t}
\end{aligned}$$

and similar equations hold for the joint moments $E(Y_{2,i}^k Y_{2,j}^\ell)$ and $E(Y_{1,i}^k Y_{2,j}^\ell)$. After some lengthy calculations following the lines of Forman & Sørensen (2006), we obtain explicit expressions for the stochastic volatility model driven by Pearson diffusion as in the above example.

All in all moment based estimation is feasible for the two-dimensional volatility model, and for the Pearson driven model in particular explicit expressions of moments and joint moments can be found. We emphasize the prediction based estimating functions of Sørensen (2000). For the univariate stochastic volatility models driven by a Pearson diffusion the optimal prediction based estimating functions based on powers of returns were derived by Forman & Sørensen (2006).

Alternatively the model can for a wide range of underlying diffusions be analyzed by simulated likelihood or Markov chain Monte Carlo methods, see section 3.1 below and the papers by Durham & Gallant (2002) and Pitt, Chib & Shephard (2006).

The 3D model

The three dimensional model displays the same features as in the two dimensional setting, only now there are three covariances/correlations in play. The volatility matrix is given by

$$
\begin{aligned}
V_{11} &= c_{12}^2 c_{13}^2 \lambda_1 + (s_{12}c_{23} - c_{12}s_{13}s_{23})^2 \lambda_2 + (s_{12}s_{23} + c_{12}s_{13}c_{23})^2 \lambda_3 \\
V_{22} &= s_{12}^2 c_{13}^2 \lambda_1 + (c_{12}c_{23} - s_{12}s_{13}s_{23})^2 \lambda_2 + (c_{12}s_{23} + s_{12}s_{13}c_{23})^2 \lambda_3 \\
V_{33} &= s_{13}^2 \lambda_1 + c_{13}^2 s_{23}^2 \lambda_2 + c_{13}^2 c_{23}^2 \lambda_3 \\
V_{12} &= -c_{12}s_{12}c_{13}^2 \lambda_1 + (s_{12}c_{23} - c_{12}s_{13}s_{23})(c_{12}c_{23} - s_{12}s_{13}s_{23})\lambda_2 \\
&\quad +(s_{12}s_{23} + c_{12}s_{13}c_{23})(c_{12}s_{23} + s_{12}s_{13}c_{23})\lambda_3 \\
V_{13} &= c_{12}s_{13}c_{13}^2 \lambda_1 + (s_{12}c_{23} - c_{12}s_{13}s_{23})c_{13}s_{23}\lambda_2 + (s_{12}s_{23} + c_{12}s_{13}c_{23})c_{13}c_{23}\lambda_3 \\
V_{23} &= s_{12}c_{13}s_{13}\lambda_1 + (c_{12}c_{23} - s_{12}s_{13}s_{23})c_{13}s_{23}\lambda_2 + (c_{12}s_{23} + s_{12}s_{13}c_{23})c_{13}c_{23}\lambda_3
\end{aligned}
$$

Where we abbreviate $c_{ij} = \cos(\phi_{ij})$ and $s_{ij} = \sin(\phi_{ij})$. Regrettably, as turning angles are defined relative to previous turns, the model lack symmetry in the variance and covariance formulae. Hence if coordinates are interchanged we may have to redefine the angle processes.

3

Statistical inference

In this chapter we review some important methods for making statistical inference in diffusion driven models which motivate and contrast the results presented in our papers.

Throughout the chapter we consider inference from the stationary continuous time process $\{X_t\}_{t\geq 0}$ based on discrete time observations defined by $Y_i = X_{i\Delta}$, $i = 1, \ldots, N$ where Δ^{-1} is the sampling frequency. Unless otherwise stated we are concerned with the low frequency setting where Δ is fixed and asymptotic results are proven as the number of observations tend to infinity.
For the unknown distribution of $\{X_t\}_{t\geq 0}$ we assume a (semi) parametric model parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$, where d is the dimension of the parameter. Two main problems are addressed. Firstly, how can the parameter θ be estimated? Secondly, how do we assess whether or not the fitted model provide an acceptable description of the data?

Ideally we would base inference on the likelihood function. However, most diffusion-type models do not admit an explicit likelihood function and alternative estimating schemes are thus often called upon. We emphasize the theory of general estimating functions as a suitable means which often yield simple and explicit criteria. In section 3.1 we review some important approximations of the log-likelihood function. The evolution of the last decade implies that today likelihood inference is in fact feasible. However, the algorithms are still demanding from a computational as well as from a mathematical point of view. In comparison the general estimating functions considered in section 3.2 are often far more tractable. We summarize some important results on how to construct estimating equations and how to combine these in an optimal way. The theory is exemplified by the martingale estimating functions and the prediction based estimating functions encountered in our paper Forman & Sørensen (2006). Section 3.3 covers GMM, i.e. the generalized method of moments. Similar to the general estimating functions the GMM criterion from which the estimators are obtained is based on moment conditions. Indeed the two fields intersect in many regards. We highlight some of the GMM specific results on covariance estimation and goodness of fit testing. The goodness of fit test developed in our paper Forman, Markusen & Sørensen (2007) is closely related to the GMM overidentifying restrictions test and so is the models selection procedure of my first paper Forman (2005). Section 3.4 reviews nonparametric inference from diffusion processes. We consider the nonparametric setting a natural framework for goodness of fit testing. However, a similar theory for diffusion-type models is yet to be explored. Finally, section 3.5 is concerned with the uniform residuals. We believe that these provide an excellent diagnostic not just for diffusion models but for stochastic process models in general.

The overall focus of this chapter will be on the basic ideas and results. Hence, regularity conditions and other technical details must be looked up in the relevant papers. An exception to the rule are the asymptotic results of section 3.2.5 which are presented with all regularity conditions included.

Please note that this introduction by no means claim to be exhaustive. For instance we do not discuss the methodology of indirect inference Gourieroux, Monfort & Renault (1993) nor the efficient method of moments Gallant & Tauchen (1996) which also provide popular ways of forming statistical inference for diffusion-type models. See instead Gallant & Tauchen (2004) for a review.

3.1 Likelihood based inference

Likelihood inference is the preferred means by which statistical models are analyzed due to its generic form and the asymptotic efficiency of the maximum likelihood estimator. But when it comes to diffusion type models the functional form of the likelihood function is hardly ever known. This complicates the evaluation of estimates and likelihood ratios. For instance the log-likelihood function of a discretely observed diffusion is given by

$$l_N(\theta) = \sum_{i=1}^{N-1} \log\{p_\Delta(Y_{i+1}|Y_i,\theta)\}$$

where $p_\Delta(Y_{i+1}|Y_i,\theta)$ is the, typically unknown, Δ-step transition probability. As a consequence: If likelihood inference is to be carried out for a diffusion model the transition density will have to be replaced by a suitable analytical or numerical approximation. In the coming sections we review some of the most important approximation schemes emphasizing basic ideas and computations. In either case likelihood inference is computationally demanding.

Note that at least for some diffusions likelihood inference is uncomplicated. The fact that it has simple Gaussian transition densities makes the Ornstein-Uhlenbeck process the most celebrated diffusion and the most used test case in theoretical examples as well as in simulation studies. Pedersen (1995b) summarizes the maximum likelihood estimators. By similar arguments the square root process is often used as benchmark. The remaining Pearson diffusions are potentially accessible for likelihood inference as their transition probabilities have fairly explicit spectral representations. How to make proper use of these is a subject for future research.

3.1.1 Simulated likelihood

The method of simulated likelihood estimation for discretely observed diffusion processes was proposed independently by Pedersen (1995b) and Brandt & Santa-Clara (2002). The basic idea is to approximate the unknown transition densities with mixtures of normal densities corresponding to a suitably fine Euler scheme and evaluate these by means of Monte Carlo simulations. To be specific the M'th order approximation is defined by

$$p^{(M)}(Y_{i+1}|Y_i,\theta) = \int \prod_{t=0}^{M} \phi\{\hat{Y}_{i,t+1}|\,\hat{Y}_{i,t} + \delta\mu(\hat{Y}_{i,t},\theta),\, \delta\sigma^2(\hat{Y}_{i,t},\theta)\}d\hat{Y}_{i,1}\cdots d\hat{Y}_{i,M} \tag{3.1}$$

where $\delta = \Delta/(M+1)$, $\hat{Y}_{i,0} = Y_i$, $\hat{Y}_{i,M+1} = Y_{i+1}$, and $\phi(x|\,\kappa,\,\tau^2)$ denotes the normal density with mean κ and variance τ^2. The variable $\hat{Y}_{i,t}$ is interpreted as a missing data point at time $\{i+t/(M+1)\}\Delta$. If $\{\hat{Y}^{(s)}_{i,1},\ldots,\hat{Y}^{(s)}_{i,M}\}_{s=1,\ldots,S}$ are i.i.d. realizations of the Euler scheme started at $\hat{Y}_{i,0} = Y_i$, the integral can be approximated by

$$\frac{1}{S}\sum_{s=1}^{S} \phi\{Y_{i+1}|\,\hat{Y}^{(s)}_{i,M} + \delta\mu(\hat{Y}^{(s)}_{i,M},\theta),\, \delta\sigma^2(\hat{Y}^{(s)}_{i,M},\theta)\}. \tag{3.2}$$

There are two sources of error in the approximation of the log-likelihood; Bias adhering from the approximation of the transition density (3.1) and simulation error from the Monte Carlo evaluation (3.2). Both can be reduced at the expense of increased computations; As $M, S \to \infty$ with $S^{1/2}/M \to 0$ the simulated log-likelihood function converges to the true one, and so does the simulated maximum likelihood estimator. It is important to notice that the same random numbers must be recycled when computing the transition density for different values of θ in order to obtain a smooth log-likelihood function. Further note that no numerical differentiation is needed for the optimization as explicit expressions of the gradient and Hessian of the simulated log-likelihood are obtained from differentiation of the Euler scheme.

It has been pointed out for instance by Durham & Gallant (2002) that the simulation scheme of Pedersen (1995b) and Brandt & Santa-Clara (2002) is highly inefficient and can be improved dramatically by use of a suitable importance sampler. The main reason is that the auxiliary observations $\hat{Y}_{i,1}, \dots, \hat{Y}_{i,M}$ are simulated without using the information contained in Y_{i+1}, i.e. for many sampled $\hat{Y}_{i,M}$'s the transition to Y_{i+1} is very unlikely. In importance sampling the auxiliary data is sampled according to an adapted density $r_i(\hat{Y}_{i,1}, \dots, \hat{Y}_{i,M})$, and the transition density approximated by

$$\frac{1}{S}\sum_{s=1}^{S} \frac{\prod_{t=0}^{M} \phi\{\hat{Y}^{(s)}_{i,t+1} | \, \hat{Y}^{(s)}_{i,t} + \delta\mu(\hat{Y}^{(s)}_{i,t}, \theta), \, \delta\sigma^2(\hat{Y}^{(s)}_{i,t}, \theta)\}}{r_i(\hat{Y}^{(s)}_{i,1}, \dots, \hat{Y}^{(s)}_{i,M})}.$$

The importance sampler works well if the sample density is close to the target density. Durham & Gallant (2002) obtain good results sampling the auxiliary data from a tied down diffusion - the so-called modified Brownian bridge. Also good results are obtained from Richard & Zhang (1998)'s efficient importance sampler where the sampling density is taken from a family of densities indexed by a high dimensional parameter and preliminarily fitted to the data. Recently Beskos & Roberts (2005) proved that exact simulation of a diffusion is feasible not just in theory but also in practice, see in addition Beskos et al. (2006).

Simulated likelihood works also for non-equidistant samples, for non-stationary diffusions and for time-inhomogeneous diffusions. Extension to diffusion type models is possible though more complicated as outlined by Durham & Gallant (2002) in the case of a stochastic volatility model.

3.1.2 An analytical approximation

A closed form approximation to the log-likelihood function based on its Hermite polynomial expansion was developed by Aït-Sahalia (2002). The resulting approximate likelihood function is computationally less demanding than simulating the likelihood function and also more accurate, see Jensen & Poulsen (2002) for a comparative study.
First assume that the diffusion coefficient is constantly equal to one. Then the K'th order Hermite expansion of the transition density around the normal density takes the form

$$p^{(K)}_{\Delta}(x|x_0, \theta) = \frac{1}{\Delta^{1/2}}\phi\left(\frac{x - x_0}{\Delta^{1/2}}\right) \exp\left(\int_{x_0}^{x} \mu(u, \theta)du\right) \sum_{k=0}^{K} c_k(x|x_0, \theta)\frac{\Delta^k}{k!}$$

where the correction terms have be grouped in orders of Δ. The coefficients are recursively defined by $c_0(x|x_0,\theta) = 1$ and

$$c_j(x|x_0,\theta) = j(x-x_0)^{-j}\int_{x_0}^{y}(u-x_0)^{j-1}\left\{\lambda(u,\theta)c_{j-1}(u|x_0,\theta) + \frac{1}{2}\frac{\partial^2}{\partial u^2}c_{j-1}(u|x_0,\theta)\right\}du$$

for $j \geq 1$ where $\lambda(x,\theta) = -(1/2)\{\mu(x,\theta)^2 + \partial_x\mu(x,\theta)\}$. Making use of the Taylor expansion of the logarithm the expansion translates into a closed form approximation of the log-density function

$$\{\log p_\Delta\}^{(K)}(x|x_0,\theta) = -\frac{\log(2\pi\Delta)}{2} + C_{-1}(x|x_0,\theta)\frac{1}{\Delta} + \sum_{k=0}^{K} C_k(x|x_0,\theta)\frac{\Delta^k}{k!}.$$

The coefficient are most easily found by substituting the above into the Kolmogorov forward and backward equations for the log-transition density. This in turn yields a set of differential equations for the $C_k(x|x_0,\theta)$'s which can be solved explicitly, see Aït-Sahalia (2003) for details.

If the diffusion coefficient differs from one, the diffusion preliminarily has to be transformed. The Lamperti transformation, $\gamma(x,\theta) = \int^x \sigma(u,\theta)^{-1}du$ achieves the goal as $Z_t = \gamma(X_t,\theta)$ satisfies $dz_t = \mu_Z(Z_t,\theta)dt + dB_t$ with

$$\mu_Z(z,\theta) = \mu\{\gamma^{-1}(z,\theta),\theta\}/\sigma\{\gamma^{-1}(z,\theta),\theta\} - (1/2)\partial_x\sigma\{\gamma^{-1}(z,\theta),\theta\}.$$

Loosely speaking the transformation brings the diffusion closer to being Gaussian which is needed for the Hermite expansion to converge. The desired approximation to the log-likelihood function of $\{X_t\}_{t\geq 0}$ is obtained as $p_\Delta(x|x_0,\theta) = \sigma(x,\theta)^{-1}p_{\Delta,Z}\{\gamma(x,\theta)|\gamma(x_0,\theta),\theta\}$.

There has been some effort made to extend the analytical expansion to multivariate diffusions and stochastic volatility models, see Aït-Sahalia (2003) and Aït-Sahalia & Kimmel (2004). The lack of a multivariate analogue to the Lamperti transform makes the extension a more complicated matter.

3.1.3 Bayesian inference.

Elerian, Chib & Shephard (2001) suggested Markov Chain Monte Carlo algorithms for performing Bayesian inference for Diffusion processes. The same idea was launched independently by Eraker (2001).

From a purist Bayesian point of view the one goal of the statistical analysis is to find the posterior distribution of the parameter. From a pragmatic point of view the posterior mean is just another estimator which can be used for various statistical purposes. Based on moderate size datasets the results are largely determined by the information in the likelihood function not in the prior. Hence, suitably flat priors leads to inference similar to simulated likelihood. As in simulated likelihood the basic idea is to augment the data with the set of latent variables $\{\hat{Y}_{i,t}\}_{i=1,\dots,N-1,t=1,\dots,M}$. If M is sufficiently large the distribution of the augmented dataset is well approximated by the Euler scheme. Thus the joint posterior of $\{\hat{Y}_{i,t}\}_{i=1,\dots,N-1,t=1,\dots,M}$ and θ can be resolved using the following Gibbs sampler:

1. Initially select a prior on θ along with sensible values of θ and $\{\hat{Y}_{i,t}\}_{i=1,\dots,N-1,t=1,\dots,M}$.
2. For $i = 1, \dots, N-1$ update $\hat{Y}_{i,1}, \dots, \hat{Y}_{i,M}$ by sampling from the conditional distribution given Y_i, Y_{i+1}, and θ.
3. Update θ by sampling from the conditional distribution given the data $\{Y_i\}_{i=1,\dots,N}$ and the augmented variables $\{\hat{Y}_{i,t}\}_{i=1,\dots,N-1,t=1,\dots,M}$.
4. Repeat the updating schemes of step 2 and 3 for a large number of sweeps.

As the conditional distributions hardly ever are explicitly known, separate Metropolis-Hastings algorithms are needed to perform steps 2 and 3. The same proposal distributions used in the importance sampler for simulated likelihood are of relevance when sampling the latent variables in step 2. Further, if M is large it may be difficult to sample all of $\hat{Y}_{i,1}, \dots, \hat{Y}_{i,M}$ at a time. Elerian, Chib & Shephard (2001) suggest to sample blocks of length m sequentially in each step using a Metropolis-Hastings algorithm to sample from the approximate density

$$f(\hat{Y}_{i,k}, \dots, \hat{Y}_{i,k+m} | \hat{Y}_{i,k-1}, \hat{Y}_{i,k+m+1}, \theta) \propto \prod_{t=k}^{k+m} \phi\{\hat{Y}_{i,t+1} | \, \hat{Y}_{i,t} + \mu(\hat{Y}_{i,t}, \theta)\delta, \, \sigma^2(\hat{Y}_{i,t}, \theta)\delta\}$$

with a data based Laplace approximation as proposal density. Please note that there is a trade-of in the choice of block size. A small m simplifies the sampling of $\hat{Y}_{i,1}, \dots, \hat{Y}_{i,M}$ but a the same time implies a slow mixing rate for the Markov chain as neighbor $Y_{i,k}$'s are highly dependent. In simulation studies conducted by Elerian, Chib & Shephard (2001) the best performance occurs for a random (Poisson distributed) block size.

Similarly, in step 3 the parameter must be sampled according to the density

$$f(\theta | \{\hat{Y}_{i,t}\}_{i=1,\dots,N-1,t=0,\dots,M+1}) \propto f_0(\theta) \prod_{i=1}^{N-1} \prod_{t=0}^{M} \phi\{\hat{Y}_{i,t+1} | \, \hat{Y}_{i,t} + \mu(\hat{Y}_{i,t}, \theta)\delta, \, \sigma^2(\hat{Y}_{i,t}, \theta)\delta\},$$

where we let $\hat{Y}_{i,0} = Y_i$ and $\hat{Y}_{i,M+1} = Y_{i+1}$, and f_0 denotes the prior. If θ is high dimensional, it is suggestible to sample the parameters block-wise. Note that in the limit as $M \to \infty$ the parameters in the diffusion coefficient are fully determined by the quadratic variation of the underlying path. As pointed out by Roberts & Stramer (2001) this may cause the algorithm to slow down considerably for large values of M. In order to avoid this problem Roberts & Stramer (2001) further suggests transforming the latent variables prior to sampling θ.

Once the estimate is obtained the algorithm can be altered to compute likelihood ratios, filtered or smoothed values, and residuals and predictions for model diagnostics, see Pitt & Shephard (1999a), Elerian, Chib & Shephard (2001) and Chib & Jeliazkov (2001) for details.

The major drawback of the Bayesian approach it that it is computationally demanding and requires good programming skill as well as expertise on Markov chain Monte Carlo methods. To evaluate the performance of the algorithm one can plot the sample path and autocorrelation function of each simulated parameter. A useful diagnostic is the simulation inefficiency factor which is defined as the ratio of the variance of estimate to the variance of estimate from a hypothetical i.i.d. sampler. The latter can be estimated by the posterior variance divided by the total number of sweeps, whereas the former can be estimated by a HAC estimator, see section 3.3.2 below. An inefficiency factor of say 200 has the interpretation that the algorithm has to be run for 200 as many sweeps as the i.i.d. sampler to obtain the same numerical precision. The serial correlation can be quite high for badly behaved algorithms.

The algorithm easily extends to non-equidistant samples, non-stationary diffusions, and time inhomogeneous diffusions. Furthermore, the Bayesian approach extends to diffusion type models in generality as described in Pitt, Chib & Shephard (2006).

3.1.4 Further topics

Lo (1988) suggested computing the likelihood function by solving the Kolmogorov forward equation numerically. See Jensen & Poulsen (2002) for further results on the approximation.

Goodness of fit and hypothesis testing can be based on the likelihood ratio statistic. Durham (2003) evaluates various nested models of the short-term interest rate by performing simulated likelihood ratio tests. When comparing non-nested models the distribution of the likelihood ratio will have to be simulated. Repeating the approximate maximum likelihood estimation for a large number of datasets seems a lengthy project, though. Likelihood ratio testing of non-nested models is reviewed in Gourieroux & Monfort (1994).

Asymptotic theory for the maximum likelihood estimator of a discretely observed diffusion process can be found in for instance Billingsly (1961), Dacunha-Castelle & Florens-Zmirou (1986), Pedersen (1995a), and Aït-Sahalia (2002). The asymptotic behavior of the maximum likelihood estimators in other diffusion type models such as stochastic volatility models is yet to be resolved. See, however Sørensen (2003) for an approximation to the likelihood function of a volatility model resulting in an estimator with tractable asymptotics.

3.2 General estimating functions.

A general estimating function is a function of the parameter and the data,

$$F_N(\theta) = F_N(Y_1, \ldots, Y_N, \theta) \in \mathbb{R}^d.$$

The related estimator is obtained by solving the estimating equation, $F_N(\theta) = 0$. As a general frame the theory of estimating functions covers virtually any estimating scheme. For instance the prime example of an estimating function is the score function for which the maximum likelihood estimator is a zero point. However, in connection with diffusion-type models we are mainly interested in estimating functions which are fairly explicit as they posses the analytical tractability lacked by the likelihood.

Estimating functions are often obtained by combining relationships between consecutive observations that are informative about the parameter. I.e. if $h_{ij}(Y_1, \ldots, Y_i, \theta)$ where $j = 1, \ldots, m$ are real valued functions for, then an estimating function is given by,

$$F_N(\theta) = \sum_{i=1}^{N} w_i(Y_1, \ldots, Y_{i-1}, \theta) h_i(Y_1, \ldots, Y_i, \theta)$$

where $h_i = (h_{i1}, \ldots, h_{im})^T$ and $w_1, \ldots, w_N$ are possibly random d by m weight matrix.
Two natural questions arise in connection with diffusion-type models. How do we derive useful relations? How do we combine these relations in the best possible way?
In what follows we summarize some important classes of estimating functions, review the theory on optimal estimating functions and the general asymptotic theory. See also Bibby, Jacobsen & Sørensen (2004) and the references therein for a thorough introductions and various examples.

3.2.1 Martingale estimating functions

The score function is usually a martingale. Hence, it is natural to try to approximate it with a simpler estimating function which shares this property. The estimating function F_n is a martingale estimating function if

$$E_\theta\{F_N(\theta)|Y_1, \ldots, Y_{N-1}\} = F_{N-1}(\theta) \quad N = 1, 2, \ldots$$

where $F_0 = 0$. Moreover, martingale estimating functions are particularly tractable due to the well developed martingale limit theory, see Hall & Heyde (1980).
Martingale estimating functions have turned out to be very useful for estimating the parameters of discretely observed diffusions. If $\{Y_i\}_{i \in \mathbb{N}}$ is a Markov chain, then a generic martingale estimating function is given by

$$F_N(\theta) = \sum_{i=2}^{N} \sum_{j=1}^{m} w_j(Y_{i-1}, \theta)[f_j(Y_i) - E_\theta\{f_j(Y_i)|Y_{i-1}\}] \tag{3.3}$$

where $w_1, \ldots, w_m$ are $d \times 1$ weight functions.

Example 3.2.1 *The linear martingale estimating function studied by Bibby & Sørensen (1995) takes the form*

$$F_N(\theta) = \sum_{i=2}^{N} w(Y_{i-1}, \theta)\{Y_i - E_\theta(Y_i|Y_{i-1})\}$$

where w is a d by one weight function. If the underlying diffusion has a mean reverting linear drift, then the conditional mean is explicitly known. More generally the conditional means can be simulated. Forman, Markusen & Sørensen (2007) exemplify the down sampled linear estimating function as a means for testing the goodness of fit of mean reverting linear drift diffusions. △

The martingale estimating function 3.3 is particularly tractable if the functions $f_1, \dots, f_m$ are chosen such that the conditional means are explicitly known.

Example 3.2.2 *Kessler & Sørensen (1999) suggested basing martingale estimating functions on eigenfunctions $\phi_1, \dots, \phi_m$ of the infinitesimal generator. If ϕ_j is an eigenfunction of the generator with corresponding eigenvalue $-\lambda_j$, then $E\{\phi_j(Y_i)|Y_{i-1}\} = e^{-\lambda_j \Delta}\phi_j(Y_{i-1})$. Our paper Forman & Sørensen (2006) gives a thorough treatment of the martingale estimating functions based on the polynomial eigenfunctions of the Pearson diffusions.* △

For non-Markovian diffusion-type models one could in principle apply a martingale estimating function of the form

$$F_N(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{m} w_{ij}(Y_1, \dots, Y_{i-1}, \theta)[f_j(Y_i) - E_\theta\{f(Y_i)|Y_1, \dots, Y_{i-1}\}]. \tag{3.4}$$

However, save for trivial cases the conditional moments are not explicitly known and the computational burden involved in simulating them usually is not worth the effort (one might as well attack the score function directly). In comparison the prediction based estimating functions considered below are much more tractable as only unconditional moments need to be computed.

3.2.2 Prediction based estimating functions

When the data generating process is no longer Markovian, martingale estimating functions are hard to come by. Sørensen (2000) developed prediction based estimating functions as a generic alternative. Here we briefly outline the basic ideas and computations.

In order to construct a set of relations the functions $f_1(Y_i), \dots, f_m(Y_i)$ of the present observation are targeted. Each of them we predict based on the functions of past observations $Z_{jk} = h_{jk}(Y_{i-r+1}, \dots, Y_{i-1})$, $k = 1, \dots, q_j$. The best linear predictor is the L_2 projection $\hat{\pi}_j^{(i-1)}$ of $f_j(Y_i)$ onto the prediction space $\mathcal{P}_{i-1,j}$ spanned by the basic predictors $1, Z_{j1}, \dots, Z_{jq_j}$. The prediction based estimating function takes the form,

$$F_N(\theta) = \sum_{i=r}^{N} \sum_{j=1}^{m} w_{ij}(\theta)\{f_j(Y_i) - \hat{\pi}_j^{(i-1)}(\theta)\} \tag{3.5}$$

where $w_{ij}(\theta)$ is a d-dimensional data dependent vector of weights, the coordinates of which belong to $\mathcal{P}_{i-1,j}$. The best linear predictor is given by $\hat{\pi}_j^{(i-1)}(\theta) = \hat{a}_j(\theta)^T Z_j^{(i-1)}$ with $Z_j^{(i-1)} = (Z_{j1}^{(i-1)}, \ldots, Z_{jq_j}^{(i-1)})^T$ and $\hat{a}_j(\theta)^T = \{\hat{a}_{j0}(\theta), \ldots, \hat{a}_{jq_j}(\theta)\}$ defined as

$$\{\hat{a}_{j1}(\theta), \ldots, \hat{a}_{jq_j}(\theta)\} = C_j(\theta)^{-1} b_j(\theta), \quad \hat{a}_{j0}(\theta) = E_\theta(Y_1^j) - \sum_{k=1}^{q_j} \hat{a}_{jk}(\theta) E_\theta(Z_{jk}^{(r)})$$

where $C_j(\theta)$ is the covariance matrix of $(Z_{j1}^{(r)}, \ldots, Z_{jq_j}^{(r)})^T$ and $b_j(\theta)$ is the covariance vector $(\mathrm{Cov}_\theta\{Z_{j1}^{(r)}, f_j(Y_{r+1})\}, \ldots, \mathrm{Cov}_\theta\{Z_{jq_j}^{(r)}, f_j(Y_{r+1})\})^T$. Thus to find $\hat{\pi}_j^{(i-1)}(\theta)$, $j = 1, \ldots, m$, we need to calculate the covariances in $C_j(\theta)$ and $b_j(\theta)$. In practice, the best linear predictor can be found by means of the Durbin-Levinson algorithm or the innovations algorithm, see Brockwell & Davis (1991).

Example 3.2.3 *Our paper Forman & Sørensen (2006) considers prediction based estimating functions for diffusion-type models driven by Pearson diffusion. It is demonstrated that predicting powers of the observations Y_i^j in terms of powers of past observations $\{Y_{i-\ell}^\kappa \mid \ell = 1, \ldots, r, \kappa = 0, \ldots, j\}$ yields explicit estimating functions. Calculating the best linear predictors essentially amounts to finding the joint moments $E_\theta(Y_1^\kappa Y_\ell^j)$ for $0 \le \kappa \le j \le m$ and $\ell = 1, \ldots, r$, which can be explicitly derived for integrated and summated Pearson diffusions as well as Pearson stochastic volatility models.* △

3.2.3 Simple estimating functions

A simple estimating function is an estimating function of the form

$$F_N(\theta) = \sum_{i=1}^N f(Y_i, \theta)$$

where $E_\theta f(Y_i, \theta) = 0$. Simple estimating functions do not take into account the relations between consecutive observations and thus can only identify parameters in the invariant distribution.

Example 3.2.4 *In case $\{Y_i\}_{i \in \mathbb{N}}$ has invariant density $\mu(\cdot, \theta)$ the simple estimating function with $f(y, \theta) = \partial_{\theta^T} \log \mu(y, \theta)$ yields estimators of the parameters of the invariant distribution. Note that if the observations were independent this would be score function. Together with other simple estimating functions it was studied by Kessler (2000).* △

Another class of simple estimating functions are derived from the first moment condition of Hansen & Scheinkman (1995).

Example 3.2.5 *Suppose that $\{Y_i\}_{i \in \mathbb{N}}$ is a discretely observed scalar diffusion with drift $\mu(\cdot, \theta)$ and diffusion coefficient $\sigma^2(\cdot, \theta)$. Hansen & Scheinkman (1995) show that for all functions g in the domain of the infinitesimal generator, see section 2.1.4, it holds that $E_\theta\{\mu(Y_i, \theta) h'(Y_i) + \sigma^2(Y_i, \theta) h''(Y_i)/2\} = 0$. Hence, a simple estimating function is given by*

$$F_N(\theta) = \sum_{i=1}^N \sum_{j=1}^m w_j(\theta) \{\mu(Y_i, \theta) h_j'(Y_i) + \sigma^2(Y_i, \theta) h_j''(Y_i)/2\}$$

where $w_1, \ldots, w_m$ are d by one weight functions. △

3.2.4 Optimal estimating functions

The problem of finding the best possible weights for the various estimating function is better understood in the general framework of Godambe & Heyde (1987). Recall that the Godambe information of an unbiased estimating function F_N is the d by d matrix,

$$K_{F_N}(\theta) = S_{F_N}(\theta)^T E_\theta\{F_N(\theta)F_N(\theta)^T\}^{-1} S_{F_N}(\theta)$$

where $S_{F_N}(\theta) = E_\theta\{\partial_{\theta^T} F_N(\theta)\}$ is the sensitivity function. The estimating function $F_N^\star$ is Godambe optimal within a class of unbiased estimating functions $\mathcal{F}_N$, if it attains maximal Godambe information w.r.t. the partial ordering of positive semidefinite matrices. It is important to notice that $K_{F_N}(\theta)^{-1}$ converges to the asymptotic variance of the related estimator $\hat{\theta}_N$ as $N \to \infty$, see theorem 3.2.1 below. Hence, the Godambe optimal estimating function also attains minimum asymptotic variance among the estimators obtained from $\mathcal{F}_N$. Further the optimal estimating functions can be interpreted as approximations to the score function. If $\mathcal{F}_N$ is a closed subspace in L_2, then the optimal estimating function $F_N^\star$ is the L_2 projection of the score function onto $\mathcal{F}_N$, see Heyde (1997).

For the generic estimating functions of the previous subsections it is possible to find optimal weights. If the estimating function takes the form

$$F_N(\theta) = w(\theta) H_N(\theta), \quad H_N(\theta) = \textstyle\sum_{i=1}^N h_i(Y_1, \ldots, Y_i, \theta),$$

then the optimal weights are given by $w^\star(\theta) = -E_\theta\{\partial_{\theta^T} H_N(\theta)\} E_\theta\{H_N(\theta)H_N(\theta)^T\}$.
The optimal choice of weights in the prediction based estimating function (B.15) were derived in Sørensen (2000).

Example 3.2.6 *In case of the prediction based estimating functions for integrated, summated, and stochastic volatility Pearson diffusion-models studied in Forman & Sørensen (2006) the optimal estimating functions are explicit. Forman & Sørensen (2006) show that in order to calculate the optimal weights all that needs to be found are the mixed moments* $E_\theta(Y_1^{\kappa_1} Y_{\ell_1}^{\kappa_2} Y_{\ell_2}^{\kappa_3} Y_{\ell_3}^{\kappa_4})$ *for* $1 \le \ell_1 \le \ell_2 \le \ell_3$ *and* $\kappa_1 + \kappa_2 + \kappa_3 + \kappa_4 \le 4m$.

For the martingale estimating functions (3.3) and (3.4) the optimal weights are given by

$$\begin{aligned} w_i^\star(X_1, \ldots, X_{i-1}, \theta) \quad &= \quad -E_\theta\{\partial_{\theta^T} h_i(X_1, \ldots, X_i, \theta) | X_1, \ldots, X_{i-1}\} \cdot \\ & \qquad E_\theta\{h_i(X_1, \ldots, X_i, \theta) h_i(X_1, \ldots, X_i, \theta)^T | X_1, \ldots, X_{i-1})\}. \end{aligned}$$

where $h_{i,j}(Y_1, \ldots, Y_i, \theta) = f_j(Y_i) - E_\theta\{f(Y_i) | Y_1, \ldots, Y_{i-1}\}$.
Forman & Sørensen (2006) provides explicit formulae for computing the optimal martingale estimating functions for the plain Pearson diffusions based on their polynomial eigenfunctions.
Note that in the generic examples optimality is only attained within the class generated from the initial set of relations. The overall efficiency depends crucially on the choice of relations, see for instance Kessler (2000).
The choice of relations at the same time is the major strength and the major weakness of the estimating function approach. Computationally a simple explicit estimating function is a gain on the inaccessible score function. However there might be a price to pay; the moment based estimators are often less efficient than the maximum likelihood estimator.

3.2.5 Asymptotic theory

To simplify matters we consider only unbiased estimating functions of the form

$$F_N(\theta) = \sum_{i=r}^{N} f(Y_{i-r+1}, \dots, Y_i, \theta)$$

which appeals directly to the central limit theorem and the law of large numbers. F_N is unbiased if and only if $E_\theta f(Y_1, \dots, Y_r, \theta) = 0$. Denote by $\hat{\theta}_N$ a solution to the estimating equation $F_N(\theta) = 0$. We briefly outline the regularity conditions ensuring that $\hat{\theta}_N$ eventually exists and is a consistent and asymptotically normal estimator of the true parameter θ_0.

R1: θ_0 belongs to the interior of Θ.

R2: The process $\{Y_i\}_{i\in\mathbb{N}}$ is stationary and ergodic.

R3: There exists a neighborhood $U(\theta_0)$ of θ_0 such that $E_{\theta_0} f(Y_1, \dots, Y_r, \theta)$ is finite for all $\theta \in U(\theta_0)$ and $E_{\theta_0} f(Y_1, \dots, Y_r, \theta_0) = 0$.

R4: $f(y_1, \dots, y_r, \theta)$ is continuously differentiable w.r.t. θ for all $(y_1, \dots, y_r)$.

R5: For $i, j = 1, \dots, d$ each family $\{\partial_{\theta_i} f_j(Y_1, \dots, Y_r, \theta)\}_{\theta\in U(\theta_0)}$ is dominated by an integrable random variable.

R6: The matrix $S(\theta_0) = E\{\partial_{\theta^T} f(Y_1, \dots, Y_r, \theta_0)\}$ is invertible.

R7: $N^{-1/2} F_N(\theta_0) \to \mathcal{N}(0, V(\theta_0))$ where $V(\theta_0) = \lim_{N\to\infty} N^{-1} E\{F_N(\theta_0) F_N(\theta_0)^T\}$.

Theorem 3.2.1 *If* **R1** - **R7** *hold true, then with probability tending to one as $N \to \infty$ a solution $\hat{\theta}_N$ to the estimating equations exists such that $\hat{\theta}_N \to \theta_0$ in probability and*

$$n^{1/2}(\hat{\theta}_N - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, S(\theta_0)^{-1} V(\theta_0) \{S(\theta_0)^{-1}\}^T).$$

If in addition $E_{\theta_0} f(Y_1, \dots, Y_r, \theta) \neq 0$ when $\theta \neq \theta_0$ and each family $\{f_j(Y_1, \dots, Y_r, \theta)\}_{\theta\in U(\theta_0)}$ is dominated by an integrable random variable, then $\hat{\theta}_N$ is eventually the unique zero point of F_N on every compact subset of Θ containing θ_0.

In a forthcoming paper by Jacod & Sørensen (2007) this result it is proved for martingale estimating functions. As the martingale property is needed only for establishing the central limit theorem for $F_N(\theta_0)$, we have reinforced condition **R7** instead of the weaker assumption that $E_{\theta_0} f(Y_1, \dots, Y_r, \theta) f(Y_1, \dots, Y_r, \theta)^T$ be well defined.

It should be noted that the estimator $\hat{\theta}_N$ is eventually identical to the generalized method of moments estimator obtained by minimizing the criterion $F_N(\theta) W_N F_N(\theta)$ for any positive definite weight matrix W_N. Consistency and asymptotic normality of the generalized method of moments-estimators was proven by Hansen (1982) under slightly different regularity conditions. We review the generalized method of moments in section ... below.

In practice, it is usually a good idea to replace for instance the optimal weights $w_i^*(\theta)$ by estimated weights $w_i^*(\tilde{\theta}_N)$, where $\tilde{\theta}_N$ is a $\sqrt{N}$-consistent estimator of θ. This has the advantages that the weight matrices need only be evaluated once for every datum and that a simpler estimating equation is hereby obtained. There is no loss in efficiency by doing so as the asymptotic variance of the estimator is preserved. The following corollary to appear in Jacod & Sørensen (2007) establish the desired asymptotic behavior of the estimating function.

$$\tilde{F}_N(\theta) = \sum_{i=r}^{N} w(Y_{i-r+1}, \ldots, Y_{i-1}, \tilde{\theta}_N) h(Y_{i-r+1}, \ldots, Y_i, \theta)$$

where w is the d by m weight function.

Corollary 3.2.1 *Suppose that the above regularity conditions* **R1** – **R3** *and* **R6** – **R7** *hold true for* $f(y_1, \ldots, y_r, \theta) = w(y_1, \ldots, y_{r-1}, \theta_0) h(y_1, \ldots, y_r, \theta)$ *and that*

R8: *The functions* $\theta \mapsto w(y_1, \ldots, y_{r-1}, \theta)$ *and* $\theta \mapsto h(y_1, \ldots, y_r, \theta)$ *are continuously differentiable for all possible outcomes of* $(y_1, \ldots, y_r)$.

R9: *The family* $\{\partial_{\theta_i} w(Y_1, \ldots, Y_{r-1}, \theta)_{jk} f_k(Y_1, \ldots, Y_r, \theta)\}_{\theta \in U(\theta_0)}$, *is dominated by an integrable random variable, and so are* $\{\partial_{\theta_i} w(Y_1, \ldots, Y_{r-1}, \theta)_{jk} \partial_{\theta_i} f_k(Y_1, \ldots, Y_r, \theta)\}_{\theta \in U(\theta_0)}$ *and* $\{w(Y_1, \ldots, Y_{r-1}, \theta)_{jk} \partial_{\theta_i} f_k(Y_1, \ldots, Y_r, \theta)\}_{\theta \in U(\theta_0)}$, *for every* $i, j = 1, \ldots, d$ *and* $k = 1, \ldots, m$.

Then with probability tending to one as $N \to \infty$ *a solution* $\hat{\theta}_N$ *to the estimating equation* $\tilde{F}_N(\theta) = 0$ *exists such that* $\hat{\theta}_N \to \theta_0$ *in probability and*

$$n^{1/2}(\hat{\theta}_N - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, S(\theta_0)^{-1} V(\theta_0) \{S(\theta_0)^{-1}\}^T),$$

where $S(\theta_0)$ *and* $V(\theta_0)$ *are defined as in conditions* **R6** *and* **R7** *above. Moreover,* $\hat{\theta}_N$ *is eventually unique on every compact subset of* Θ *containing* θ_0 *under the additional assumptions of theorem 3.2.1.*

The estimator $\tilde{\theta}_N$ can, for instance, be obtained from a similar estimating function, where the estimated weights have been replaced by suitable simple matrices independent of θ, but such that the estimating equation has a solution.

3.2.6 Further topics

A goodness of fit test can be based on the surplus relations as prescribed by the overidentifying restrictions test, section 3.3.3 below. Our paper Forman, Markusen & Sørensen (2007) suggest down-sampling the estimating function in order to check that the parameter estimates are consistent with varying sampling frequencies.

Criteria for estimating functions being optimal in a high-frequency asymptotics were developed by Jacobsen (2001). See also Jacobsen (2002).

3.3 The Generalized Method of Moments

The generalized method of moments - GMM in brief - was introduced by Hansen (1982) as a general theory concerning estimation in stochastic process models. Among practitioners it has become a popular statistical tool for making inference from economic and financial time series. See for instance Melino & Turnbull (1990) on estimation in a stochastic volatility model.

GMM inference is based on a q-dimensional set of moment conditions given by the function $f : \mathcal{Y}^r \times \Theta \mapsto \mathbb{R}^q$ satisfying that for the true parameter θ_0,

$$E\{f(Y_{i-r+1}, \dots, Y_i, \theta_0)\} = 0.$$

For instance Hansen & Scheinkman (1995) have shown how the infinitesimal generator of a scalar diffusion can be used to generate moment conditions for the discretely observed process. The generalized method of moments estimator is the minimum chi square-type estimator defined by

$$\hat{\theta}_N = \arg\min_{\theta \in \Theta} \left\{ \frac{1}{N} \sum_{i=r}^{N} f(Y_{i-r+1}, \dots, Y_i, \theta) \right\} W_N \left\{ \frac{1}{N} \sum_{i=r}^{N} f(Y_{i-r+1}, \dots, Y_i, \theta) \right\}^T , \quad (3.6)$$

where $\{W_n\}$ is an possibly data dependent sequence of positive semi definite $q \times q$ weight matrices converging to a deterministic positive definite matrix W as $N \to \infty$.

The GMM framework includes for instance maximum likelihood and least squares estimators. In many ways it parallels the setting of general estimating function. Important development within the GMM-theory includes covariance estimation, hypothesis testing, goodness of fit testing, and the behavior of estimates under misspecification. We would like to point out that many of these results applies readily to the general estimating functions.

While, depending on the moment condition, the GMM-estimators sometimes have a poor small sample performance, there are still good reasons why GMM is often preferred to the more efficient maximum likelihood estimators. As pointed out by Hall (2005), GMM does not require a full model specification and is thus less sensitive to model misspecification. Furthermore, GMM is computationally far less burdensome if the criterion is chosen to be simple and explicit. The same of course can be said about estimators obtained from general estimating functions.

3.3.1 GMM and estimating functions

Note that minimizing the GMM criterion is equivalent to solving

$$\left\{ \frac{1}{N} \sum_{i=r}^{N} \partial_{\theta^T} f(Y_{i-r+1}, \dots, Y_i, \theta) \right\}^T W_N \left\{ \frac{1}{N} \sum_{i=r}^{N} f(Y_{i-r+1}, \dots, Y_i, \theta) \right\} = 0$$

which makes the connection to generalized estimating functions obvious. Indeed, in its most general form Hansen (1982) defines the GMM estimator as a solution to an estimating equation of the above form allowing also an additional term of order $o_P(N^{-1/2})$ to be added.

Example 3.3.1 *My paper Forman (2005) considers the estimation of the sums of linear drift diffusions modeled in Bibby, Skovgaard & Sørensen (2005). The autocorrelation function of these processes take the form*

$$\rho(t, \lambda, \phi) = \sum_{i=1}^{m} \phi_i \exp(-\lambda_i t).$$

where $\lambda_1 > \ldots > \lambda_m > 0$ *and* $\phi_1 + \ldots + \phi_m = 1$. *It is demonstrated that when the number of underlying diffusions, m is known, the correlation parameters can be estimated by least squares estimation,*

$$(\hat{\lambda}_N, \hat{\phi}_N) = \arg \min_{(\lambda,\phi)\in\Theta} \begin{pmatrix} \rho(\lambda, \phi, 1) - r_N(1) \\ \vdots \\ \rho(\lambda, \phi, k) - r_N(k) \end{pmatrix}^T W_n \begin{pmatrix} \rho(\lambda, \phi, 1) - r_N(1) \\ \vdots \\ \rho(\lambda, \phi, k) - r_N(k) \end{pmatrix} \tag{3.7}$$

where $k \geq 2m - 1$, $r_N(1), \ldots, r_N(k)$ *are the empirical correlations, and* $(W_n)_{n\in\mathbb{N}}$ *is a sequence of k by k weight matrices. The least squares estimator is not a GMM estimator in the strict sense (3.6) but in the more general sense of Hansen (1982) as*

$$\rho(j, \lambda, \phi) - r_N(j) = \frac{1}{N} \sum_{i=k+1}^{N} \frac{\rho(j, \lambda, \phi)(Y_i - \mu)^2 - (Y_i - \mu)(Y_{i-j} - \mu)}{\sigma^2} + o_P(N^{-1/2})$$

where μ *and* σ^2 *are the mean and variance of the* Y_i*'s. Hence, the asymptotic behavior of the least squares estimator can be verified along the lines of GMM.* △

3.3.2 Asymptotics and covariance estimation

Hansen (1982) derived the asymptotics for the GMM estimator. Note that the regularity conditions for asymptotic normality are similar to those given for the general estimating functions. Assuming $\{Y_i\}_{i\in\mathbb{N}}$ to be ergodic, consistency of the GMM estimator can often be verified directly by showing that the criterion converges to a function which has a unique zero point at θ_0. If further $N^{-1/2} \sum_{i=r}^{N} f(Y_{i-r+1}, \ldots, Y_i, \theta_0)$ admits a central limit theorem, with asymptotic variance Σ, then $\hat{\theta}_N$ is asymptotically normal,

$$N^{1/2}(\hat{\theta}_N - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, M(\theta_0)^T \Sigma(\theta_0) M(\theta_0))$$

where $M(\theta) = \{F(\theta)^T W F(\theta)\}^{-1} F(\theta)^T W$ with $F(\theta) = E\{\partial_{\theta^T} f(Y_{i-r+1}, \ldots, Y_i, \theta)\}$. The weights are optimal in the sense of minimum asymptotic variance if they converge to $W = \Sigma(\theta_0)^{-1}$. This being the case the asymptotic variance is $F(\theta_0)^T \Sigma(\theta_0)^{-1} F(\theta_0)$. The optimal estimator is typically found using a two-step procedure where $W_N = \Sigma(\tilde{\theta}_N)^{-1}$ is estimated using an initial estimate of θ.

For forming asymptotic confidence intervals for the parameters and for computing optimal weights we need to estimate $\Sigma(\theta_0)$ (obviously $F(\theta_0)$ is easy to estimate). In general $\Sigma(\theta)$ takes the form,

$$\Sigma(\theta) = \Sigma_0(\theta) + \sum_{j=1}^{\infty} \{\Sigma_j(\theta) + \Sigma_j(\theta)^T\} \tag{3.8}$$

where $\Sigma_j(\theta)$ is the j'th order autocovariance of the process $\{f(Y_i,\dots,Y_{i+r-1},\theta)\}_{i\in\mathbb{N}}$, namely

$$E_\theta(\{f(Y_1,\dots,Y_r,\theta)-E_\theta f(Y_1,\dots,Y_r,\theta)\}^T\{f(Y_{j+1},\dots,Y_{j+r},\theta)\}-E_\theta f(Y_{j+1},\dots,Y_{j+r},\theta)\}).$$

Example 3.3.2 *If $\{f(Y_i,\dots,Y_{i+r-1},\theta_0)\}_{i\in\mathbb{N}}$ is serially uncorrelated, then $\Sigma(\theta_0)=\Sigma_0(\theta_0)$ which is easily estimated. In particular, if $\{f(Y_i,\dots,Y_{i+r-1},\theta_0)\}_{i\in\mathbb{N}}$ is a martingale difference, i.e. $E\{f(Y_{i-r+1},\dots,Y_i,\theta_0)|Y_1,\dots,Y_{i-1})=0$, then $\{f(Y_i,\dots,Y_{i+r-1},\theta_0)\}_{i\in\mathbb{N}}$ is serially uncorrelated and $E\{f(Y_i,\dots,Y_{i+r-1},\theta_0)\}=0$.* △

For non-Markovian diffusion type models the series (3.8 typically has an infinite number of non-zero terms. The individual autocovariances can be estimated by

$$\hat\Sigma_j=\frac{1}{N-j}\sum_{i=r}^{N-j}\{f(Y_{i-r+1},\dots,Y_i,\hat\theta_N)-\overline{f}_N\}^T\{f(Y_{i-r+j},\dots,Y_{i+j},\hat\theta_N)-\overline{f}_N\}.$$

but it is a non-trivial task to combine these into an estimate of Σ which is both consistent and positive semidefinite. The heteroscedasticity and autocorrelation covariance matrix is given by

$$\hat\Sigma_{\text{HAC}}=\hat\Sigma_0+\sum_{i=1}^{N}\omega_{i,N}(\hat\Sigma_i+\hat\Sigma_i^T)$$

where ω is a suitable kernel ensuring that $\hat\Sigma_{\text{HAC}}$ is positive semidefinite. For instance a suitable choice could be the Parzen kernel

$$\omega_{i,N}=(1-6a_i^2+6a_i^3)I\{0\le a_i\le 1/2\}+2(1-a_i)^3I\{1/2\le a_i\le 1\}\quad a_i=i/(b_N+1)$$

with bandwidth $b_N\in\mathbb{N}$ or the quadratic spectral kernel

$$\omega_{i,N}=\frac{25b_N^2}{12\pi^2i^2}\left\{\frac{\sin(m_i)}{m_i}-\cos(m_i)\right\}\text{ with }m_i=6\pi i/(5b_N)$$

with bandwidth $b_N>0$. For each or the above kernels the HAC estimator is consistent under suitable regularity conditions, see Andrews (1991). The optimal bandwidth in terms of asymptotic mean squared error is $b_N=O(n^{1/5})$. A procedure for bandwidth selection was proposed by Newey & West (1994).

The finite sample performance of the HAC estimator can be quite poor if the autocorrelation dies out slowly. This for instance is the case if the process $\{f(Y_i,\dots,Y_{i+r-1},\theta_0)\}_{i\in\mathbb{N}}$ has a substantial autoregressive component. Andrews & Monahan (1992) propose a modification known as pre-whitening and recoloring to account for the problem. The idea is to fit a first order autoregression to $\{f(Y_i,\dots,Y_{i+r-1},\theta_0)\}_{i\in\mathbb{N}}$, compute the HAC estimator for the residuals, and re-transform to obtain an estimate of Σ. However, if an explicit expression of $\Sigma(\theta)$ can be found together with an initial estimate $\tilde\theta_N$, then $\Sigma(\tilde\theta_N)$ usually is the more precise estimate, whereas $\hat\Sigma_{\text{HAC}}$ by construction is more robust.

Example 3.3.3 *My paper Forman (2005) presents explicit expressions of the covariance matrix for the least squares estimator (3.7) when the underlying diffusions belong to the Pearson family.* △

3.3.3 The overidentifying restrictions test

Only d equations are needed to identify the parameter. A model specification test can thus be based on the remaining $q-d$ equations by checking that $N^{-1}\sum_{i=r}^{N} f(Y_{i-r+1},\dots,Y_i,\hat{\theta}_N)$ is close to zero. Hansen (1982) introduced the overidentifying restrictions test for testing $H_0 : Ef(Y_i,\dots,Y_{i+r-1},\theta_0) = 0$ using the test statistic

$$J_N = N\left\{\frac{1}{N}\sum_{i=r}^{N} f(Y_{i-r+1},\dots,Y_i,\hat{\theta}_N)\right\}^T \hat{\Sigma}_N^{-1}\left\{\frac{1}{N}\sum_{i=r}^{N} f(Y_{i-r+1},\dots,Y_i,\hat{\theta}_N)\right\},$$

which converges to a χ^2_{q-d} distribution under the null. Newey (1985) proposed basing the goodness of fit test on a subset of moment conditions (the ones believed to fail) to enhance the power.

Example 3.3.4 *Forman (2005) uses the overidentifying restrictions test based on the criterion (3.7) to estimate the number of terms in the summed diffusion model of Bibby, Skovgaard & Sørensen (2005). The suggested estimator is the smallest number of terms for which the model passes the test. By appealing to the law of the iterated logarithm, it is demonstrated that the estimator is consistent when the level of the test tends to zero at a sufficiently slow rate as $N \to \infty$.* △

Hall (2000) considered the behavior of the GMM estimator and the HAC estimators under misspecification, and proved that overidentifying restrictions test is consistent assuming that under the alternative $Ef(Y_{i-r+1},\dots,Y_i,\theta) = \mu(\theta)$ where $||\mu(\theta)|| > 0$ for all $\theta \in \Theta$, that $\hat{\theta}_N \to \overline{\theta}$ as $N \to \infty$ for some $\overline{\theta} \in \Theta$, and that $\hat{\Sigma}_N^{-1}$ converges to a positive definite matrix. Forman, Markusen & Sørensen (2007) prove the consistency of their goodness of fit tests under similar conditions. Further Hall (2000) established the convergence of the centered HAC estimator under certain regularity conditions.

3.3.4 Further topics

The use of moment conditions in statistical analysis dates back to the 1890's where Pearson introduced it as a mean to estimate the parameters of for instance the Pearson distributions, Pearson (1895).

Nested hypothesis can be tested in the GMM framework by mimicking the Wald, Lagrange Multiplier, and Likelihood Ratio tests from likelihood theory. The Likelihood function is simply replaced by the GMM criterion, Newey & West (1987a). Hall & Inoue (2003) investigate the behavior of these tests under misspecification.

The HAC estimator is known in other fields as the empirical covariance estimator or the sandwich covariance estimator. A general concern is the overall efficiency which can be very poor, see for instance, Liang, Zeger & Qaqish (1992) and Kauermann & Carroll (2004).

3.4 Nonparametric inference

Nonparametric diffusion-type models provides a natural framework for goodness of fit testing. Here we briefly outline some of the existing schemes for testing a parametric diffusion model against the nonparametric alternative

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t \tag{3.9}$$

where μ and σ satisfy suitable regularity conditions to ensure the existence and uniqueness of a stationary weak solution. Further we consider nonparametric estimation of the drift, diffusion coefficient, the invariant density and the transition densities of a plain stationary diffusion. The basic ideas are taken from standard nonparametric theory on density estimation and regression analysis. See for instance Bosq (1996) for a general introduction in the stochastic process setting.

Whereas the nonparametric theory for scalar diffusions is well developed, little has been said so far about the non-Markovian diffusion-type models. To my knowledge no literature exist on nonparametric inference for integrated and summated diffusions. In particular, the estimation of the drift and diffusion coefficients of the underlying diffusions from a discretely observed diffusion-type model is a topic yet to be explored. Note that in case of a sum-of-diffusions model this clearly presents an ill-posed problem unless the model is restricted somehow.

3.4.1 Estimation of the drift and the diffusion coefficient

The nonparametric estimators of the drift and the diffusion coefficient can be derived from the approximate regression

$$X_{i\Delta} - X_{(i-1)\Delta} \approx \mu(X_{(i-1)\Delta})\Delta + \sigma(X_{(i-1)\Delta})\Delta^{1/2}\varepsilon_i \tag{3.10}$$

where $\{\varepsilon_i\}$ are i.i.d. $\mathcal{N}(0,1)$ variables. As

$$\begin{aligned} \Delta^{-1}E(X_{i\Delta} - X_{(i-1)\Delta}|X_{(i-1)\Delta} = x) &= \mu(x) + o(\Delta) \\ \Delta^{-1}E(\{X_{i\Delta} - X_{(i-1)\Delta}\}^2|X_{(i-1)\Delta} = x) &= \sigma^2(x) + o(\Delta) \end{aligned}$$

approximate kernel regression estimates of μ and σ^2 are given by,

$$\begin{aligned} \hat{\mu}(x) &= \sum_{i=2}^{N} \frac{Y_i - Y_{i-1}}{\Delta \cdot b_N} K\left(\frac{Y_{i-1} - x}{b_N}\right) \cdot \left\{\sum_{i=2}^{N} \frac{1}{b_N} K\left(\frac{Y_{i-1} - x}{b_N}\right)\right\}^{-1} \\ \hat{\sigma}^2(x) &= \sum_{i=2}^{N} \frac{(Y_i - Y_{i-1})^2}{\Delta \cdot b_N} K\left(\frac{Y_{i-1} - x}{b_N}\right) \cdot \left\{\sum_{i=2}^{N} \frac{1}{b_N} K\left(\frac{Y_{i-1} - x}{b_N}\right)\right\}^{-1} \end{aligned}$$

where K is the kernel and b_N is the bandwidth. See Florens-Zmirou (1993) and Bandi & Phillips (2003) for results on the consistency and asymptotic normality of two similar kernel regression estimates as, $N \to \infty$, $b_N \to 0$, $\Delta = \Delta_N \to 0$ and $N\Delta_N \to \infty$. For fixed Δ the estimates are inconsistent due to the approximation error. Later studies indicate that the bias is particularly bad near the boundaries. The boundary bias is reduced when the conditional moments are estimated by local polynomial regression, Fan & Gijbels (1996). See also Hoffmann (1999) on wavelet based estimators. The local polynomial estimators are defined as follows. Consider the Taylor approximation $\mu(x) \approx \mu(x_0) + \beta_1(x - x_0) + \ldots + \beta_q(x - x_0)^q$, then $\hat{\mu}_{b_N}(x_0) = \hat{\beta}_0$ where $(\hat{\beta}_0, \ldots, \hat{\beta}_q)$ are the least squares estimators obtained by minimizing the criterion,

$$\sum_{i=2}^{N} \{Y_i - Y_{i-1} - \beta_0 - \ldots - \beta_q(Y_{i-1} - x_0)^q\}^2 \frac{1}{b_N} K\left(\frac{Y_{i-1} - x_0}{b_N}\right)$$

for a suitable choice of kernel K and bandwidth b_N. Fan & Gijbels (1996) recommends the local linear estimator; $q = 1$. The diffusion estimator $\hat{\sigma}^2_{b_N}(x_0)$ is obtained in a similar way when replacing $Y_i - Y_{i-1}$ with $\{Y_i - Y_{i-1}\}^2$ in the above. Fan & Yao (1998) show that the bias of $\hat{\sigma}^2_{b_N}$ is further reduced when the squared increments are replaced by the squared residuals $\{Y_i - Y_{i-1} - \hat{\mu}_{b_N}(Y_{i-1})\}^2$. At times the local linear estimators can also be improved by weighting the least squares estimators.

Nonparametric estimators based on higher order differences are suggested by Stanton (1997) who points of their potential in bias reduction but fail to recognize the accompanying variance inflation later documented by Fan & Zhang (2003).

A general problem is that in practice it is hard to quantify the approximation error. See the discussion following Fan (2005) for a discouraging example in the parametric case. The approximation error can also be sidestepped by recognizing that the conditional moments $m(x) = E(X_{i\Delta} - X_{(i-1)\Delta} | X_{(i-1)\Delta} = x)$ and $s^2(x) = E(\{X_{i\Delta} - X_{(i-1)\Delta}\}^2 | X_{(i-1)\Delta} = x)$ are being estimated rather than the drift and diffusion coefficients in themselves.

3.4.2 The generalized likelihood ratio test

The generalized likelihood ratio test of Fan, Zhang & Zhang (2001) can be applied to test that the drift or diffusion coefficient have a specific parameterized form. For testing

$$H_0: \ \mu(x) = \mu_0(x, \theta) \ \text{ for some } \theta \in \Theta \text{ against } H_A: \ \mu(x) \neq \mu_0(x, \theta) \ \forall \theta \in \Theta$$

the generalized likelihood ratio statistic is given by

$$\lambda_N(b_N) = \frac{N-1}{2} \log\left\{\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1(b_N)}\right\} \tag{3.11}$$

where

$$\mathrm{RSS}_0 = \sum_{i=2}^{N} \frac{\{Y_i - Y_{i-1} - \Delta\mu_0(Y_{i-1}, \hat{\theta}_N)\}^2}{\tilde{\sigma}^2_N(Y_{i-1})\Delta^{1/2}}, \quad \mathrm{RSS}_1(b_N) = \sum_{i=2}^{N} \frac{\{Y_i - Y_{i-1} - \Delta\hat{\mu}_{b_N}(Y_{i-1})\}^2}{\tilde{\sigma}^2_N(Y_{i-1})\Delta^{1/2}}.$$

Here $\hat{\mu}_{b_N}$ is the weighted local linear estimator and $\hat{\theta}_N$ is the weighted least squares estimator from the approximate regression (3.10). The weights are given by $\tilde{\sigma}^2_N(Y_{i-1})^{-1}$ where $\tilde{\sigma}^2_N$ is an initial estimate of the diffusion coefficient. In order to test that

$$H_0': \ \sigma^2(x) = \sigma_0^2(x, \theta) \ \text{ for some } \theta \in \Theta \text{ against } H_A: \ \sigma^2(x) \neq \sigma_0^2(x, \theta) \ \forall \theta \in \Theta$$

consider the log-transformed and centered residuals of the approximate regression (3.10),

$$Z_i = \log\{\Delta^{-3/2}(Y_i - Y_{i-1}) - \Delta^{-1/2}\hat{\mu}_{b_N}(Y_{i-1})\} - E(\log\varepsilon_i).$$

The (unweighted) local linear estimator $\widehat{\log(\sigma^2)}_{b_N}$ and the (unweighted) least squares estimator $\hat{\theta}_N$ are obtained from the regression of residuals

$$Z_i \approx \log\{\sigma_0^2(Y_{i-1},\theta)\} + \eta_i, \quad \eta_i = \log(\varepsilon_i) - E\{\log(\varepsilon_i)\}. \tag{3.12}$$

Hence, the same form of statistic (3.11) applies with $\mathrm{RSS}_0 = \sum_{i=2}^N\{Z_i - \log(\sigma_0^2)(Y_{i-1},\hat{\theta})\}^2$ and $\mathrm{RSS}_1(b_N) = \sum_{i=2}^N\{Z_i - \widehat{\log(\sigma^2)}_{b_N}(Y_{i-1})\}^2$. Fan & Zhang (2003) indicate a limit distribution of $\lambda_N(b_N)$, which in both cases is independent of nuisance parameters and approximates a χ^2 distribution with increasing degrees of freedom. The proof is left for future research. To be specific, the presumed limit distribution is

$$\frac{r_K\lambda_N(b_N) - d_N(b_N)}{\sqrt{2d_N(b_N)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ as } b_N \to 0 \text{ and } Nb_N^{3/2} \to \infty$$

where $d_N(b_N) = c_K r_K b_N^{-1}|\Omega|$, r_k and c_K are constants depending on the kernel K, see table 2 in Fan, Zhang & Zhang (2001), and $|\Omega|$ is the length of the support of $\{Y_i\}_{i\in\mathbb{N}}$. In finite samples we suggest replacing the support of the data with the range. It seems a reasonable thing to do, in particular if the support is unlimited as in the examples of Fan & Zhang (2003). We presume that the asymptotic result should be considered in a high frequency asymptotics as $\Delta = \Delta_N \to 0$, otherwise the estimators obtained from the approximate regression would be inconsistent. In practice Fan & Zhang (2003) recommends bootstrapping the p-value under the null (nuisance parameters can be fixed at whatever point of interest), which probably improves on possible biases.

3.4.3 Inference based on the invariant density

Aït-Sahalia (1996b) is one of the first to consider goodness of fit for diffusion models. To this end he compares the invariant density implied by a parametric model to a nonparametric estimator

$$\hat{\pi}(x) = \frac{1}{N}\sum_{i=1}^N \frac{1}{b_N}K\left(\frac{Y_i - x}{b_N}\right)$$

where K is a kernel, say a symmetric density function subject to certain regularity conditions, and b_N is the bandwidth. The goodness of fit statistic is given by

$$\hat{M} = \frac{1}{N}\sum_{i=1}^N\{\pi(Y_i,\hat{\theta}) - \hat{\pi}(Y_i)\}^2 \text{ with } \hat{\theta}_N = \arg\min_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^N\{\pi(Y_i,\theta) - \hat{\pi}(Y_i)\}^2.$$

Assuming the process to be mixing at a sufficiently fast rate and the density $\pi(x,\theta)$ to be smooth it is demonstrated that $\hat{\theta}_N$ is asymptotically normal and

$$b_N^{-1/2}\{\hat{M} - E_M\} \to \mathcal{N}(0, V_M)$$

where $E_M = \int K(x)^2dx \cdot \int \pi_0(x)^2dx$ and $V_M = 2\int \pi_0(x)^4dx \cdot \int\{\int K(u)K(u+x)du\}^2dx$. Further Aït-Sahalia (1996b) applies the test to a number of existing diffusion models for interest rates, firmly rejecting every one of them and pointing to nonlinearity of drift as the main reason of failure. However, later simulation studies conducted by Pritsker (1998) and Chapman & Pearson (2000) questions these finding as the small sample behavior of the test is often quite bad. Indeed Pritsker (1998) finds that it takes 2755 years of daily data for the asymptotics to work adequately in an empirically relevant Vasicek model.

The relation between the drift, the diffusion coefficient and the the invariant density can be inverted to yield a non-parametric estimate of either the drift or the diffusion coefficient in a semiparametric diffusion model. For instance Aït-Sahalia (1996a) transforms the nonparametric density estimate into a nonparametric estimate of the diffusion coefficient in a semiparametric diffusion model with linear drift. This estimator is consistent in the low frequency setting.

3.4.4 Inference based on the transition densities

Please note that two different diffusions may have the same invariant density. Hence, the test of Aït-Sahalia (1996b) is not omnibus for the class of diffusion models. Goodness of fit could instead be based on the difference between nonparametric and parametric estimates of the transition densities. In comparison to the test based on the invariant density these tests are consistent against a larger family of alternatives.
Fan, Yao & Tong (1998) suggest estimating the transition density by use of a double kernel method. Note that

$$E\left\{\frac{1}{b_N}K\left(\frac{Y_i-x}{b_N}\right)|Y_{i-1}=x_0\right\}\to p_\Delta(x|x_0) \text{ as } b_N\to 0.$$

Thus, $p_\Delta(x|x_0)$ can for instance be estimated by

$$\hat{p}_\Delta(x|x_0)=\sum_{i=2}^N\frac{1}{b_N^2}K\left(\frac{Y_i-x}{b_N}\right)K\left(\frac{Y_{i-1}-x_0}{b_N}\right)\cdot\left\{\sum_{i=2}^N\frac{1}{b_N}K\left(\frac{Y_{i-1}-x_0}{b_N}\right)\right\}^{-1}$$

or more generally by a local linear or polynomial estimator. A goodness of fit test can be based on the log-likelihood ratio of the parametric and nonparametric estimates of the transition densities,

$$\lambda(b_N)=\sum_{i=2}^N\log\hat{p}_\Delta(Y_i|Y_{i-1})-\sum_{i=2}^N\log p_\Delta(Y_i|Y_{i-1},\hat{\theta})$$

where $\hat{\theta}_N$ is the maximum likelihood estimator and $\hat{p}_\Delta$ is the local linear transition density estimator. This is a generalized likelihood ratio test, see Fan, Zhang & Zhang (2001).

The Kolmogorov forward and backward equations implies the relation

$$\frac{1}{2}\frac{\partial^2}{\partial x^2}\{\sigma^2(x,\theta)p_\Delta(x|x_0)\}-\frac{\partial}{\partial x}\{\mu(x,\theta)p_\Delta(x|x_0)\}=\mu(x_0,\theta)\frac{\partial}{\partial x_0}p_\Delta(x|x_0)-\frac{1}{2}\sigma^2(x_0,\theta)\frac{\partial^2}{\partial x_0^2}p_\Delta(x|x_0)$$

between the drift, the diffusion coefficient, and the transition probabilities of a diffusion. Aït-Sahalia (1996b) propose a goodness of fit test based on this relation where nonparametric estimates are inserted for the transition probabilities.

In principle, the transition densities determines the drift and the diffusion coefficient of the diffusion. Estimates of the drift and the diffusion coefficients can be obtained from the estimates of the transition probabilities and the invariant density as follows. We appeal to the alternative parameterization of Hansen, Scheinkman & Touzi (1998), see section 2.1.4. Assuming that the diffusion is ρ-mixing we look for the eigenfunction ϕ_1 of the infinitesimal generator attaining maximal non-zero eigenvalue $-\lambda_1 < 0$. At the same time ϕ_1 is the the eigenfunction of the Δ-step transition operator attaining maximal correlation. That, is

$$\exp(-\lambda_1\Delta) = \mathrm{Cor}\{\phi_1(Y_i), \phi_1(Y_{i+1})\} = \sup_{f\colon Ef(Y_i)=0} \mathrm{Cor}\{f(Y_i), f(Y_{i+1})\}.$$

In practice, ϕ_1 and λ_1 can be estimated by means of wavelet methods. Gobet, Hoffmann & Reiß (2004) finds the optimal rate of convergence for this scheme which is consistent in the low frequency setting.

The usefulness of the estimators and tests based on the transition densities is somewhat questionable as the small sample behavior of the nonparametric estimates of the transition densities undoubtedly is even worse than for the invariant density.

3.4.5 Further topics

Goodness of fit for multivariate continuous-time models can also be based on the conditional characteristic functions as demonstrated by Chen & Hong (2005).

Aït-Sahalia, Hansen & Scheinkman (2003) suggest testing that a process is Markovian by testing that nonparametric estimates of the one- and two-step transition probabilities behave as prescribed by the Chapman-Kolmogorov equations.

Recent results on the estimation of the invariant density of the volatility process of a stochastic volatility model are given by van Es, Spreij & van Zanten (2003) and Comte & Genon-Catalot (2006).

3.5 Diagnostics: the uniform residuals

A powerful diagnostic for stochastic process models is provided by the uniform residuals. In a univariate stochastic process model the uniform residuals $U_1(\theta), \ldots, U_N(\theta)$ are obtained when applying the conditional probability transform given past observations to the present observations, i.e.

$$U_i(\theta) = F_\theta(Y_i|Y_1, \ldots, Y_{i-1}).$$

If the model is correctly specified for $\theta_0 \in \Theta$, then $U_1(\theta_0) \ldots, U_N(\theta_0)$ are i.i.d. uniform variables. This result goes back to Rosenblatt (1952). As a tool for model checking in diffusion models it appears in Pedersen (1994).

In practice, the parameter must be estimated prior to computing the residuals. If the estimator $\hat{\theta}_N$ is consistent, then the residuals $\hat{U}_i = U_i(\hat{\theta}_N)$ are approximately i.i.d. uniform as $N \to \infty$. Moreover the probability transform must be found by simulation as its functional form usually is not explicitly known. For a plain diffusion model for instance the Euler scheme applies. For a non-Markovian diffusion-driven model the probability transform can be simulated by use of a suitable importance sampler or a Metropolis-Hastings algorithm, see for instance Elerian, Chib & Shephard (2001). We refer to sections 3.1.1 and 3.1.3 above and the references therein for further details.

In higher dimensions the probability transform is useless for model checking as it does not yield i.i.d. uniform random variables when applied to a multivariate process. Hence, in order to obtain residuals for a multivariate process, the coordinates of the process must be sequentialized. Note that the values of the residuals depend on the ordering.

3.5.1 Goodness of fit testing

Hong & Li (2005) propose a goodness of fit test based on the uniform residuals of a parametric model for which a $\sqrt{N}$-consistent estimate of the parameter exists. It applies to a vide range of stationary stochastic process models including diffusions and the diffusion-type models considered in chapter 2. The test statistic is given by

$$Q_{b_N}(j) = \frac{b_N(N-j)\int_0^1\int_0^1 |\hat{g}_{b_N}(j,u_1,u_2)-1|^2 du_1 du_2 - A_{b_N}}{V^{1/2}}$$

where $\hat{g}(j)$ is a boundary modified kernel density estimator of the joint density of (U_i, U_{i-j}) based on the kernel k and bandwidth b_N, see Hong & Li (2005) for the exact form. The normalizing constants are given by

$$\begin{aligned} A_{b_N} &= \left[(b_N^{-1}-2)\int_{-1}^1 k(u)^2 du + 2\int_0^1\int_{-1}^w \left\{\frac{k(w)}{\int_{-1}^w k(v)dv}\right\}^2 du dw\right]^2 - 1 \\ V &= 2\left(\int_{-1}^1\left\{\int_{-1}^1 k(u+v)k(v)dv\right\}^2 du\right)^2. \end{aligned}$$

In case the model is correct $\{Q_{b_N}(1), \ldots, Q_{b_N}(k)\}$ converge in distribution to the k-dimensional normal distribution as $N \to \infty$, $b_N \to 0$, and $Nb_N^5 \to \infty$. The bandwidth $b_N = s_N(\hat{U}) \cdot N^{-1/6}$ where $s_N(\hat{U})$ is the empirical standard deviation of the residuals is recommended. The advantage of the test is that the probability transform removes the dependence in the data, so that the asymptotics work faster than for the nonparametric specification tests of Aït-Sahalia (1996b) considered in sections 3.4.3 and 3.4.4 above. Further Hong & Li (2005) show that their test is omnibus for univariate models subject to certain regularity conditions. I.e. it is consistent against any other stationary stochastic process model satisfying the regularity conditions. The same is not true for multivariate models. See Chen & Hong (2005) for an example of a misspecified multivariate model for which the residuals are i.i.d. uniform whatever ordering is chosen.

In addition to the overall goodness of fit tests Hong & Li (2005) suggest diagnostics based on weighted averages of the empirical correlations of U_i^m and U_{i-j}^l for varying lags j. These provide some insight as to how the model could be misspecified. In essence the tests of Hong & Li (2005) captures the information contained in the QQ-plot and the lag-plot of $\hat{U}_i$ vs $\hat{U}_{i-j}$. Other indications of misspecification might display in the plots of residuals vs time and residuals vs past observations.

Part II

Papers

A

Least Squares Estimation for Autocorrelation Parameters with Applications to Sums of Ornstein-Uhlenbeck Type Processes

Least Squares Estimation for Autocorrelation Parameters with Applications to Sums of Ornstein-Uhlenbeck Type Processes

Julie Lyng Forman

Department of Applied Mathematics and Statistics, University of Copenhagen,
Universitetsparken 5, DK-2100 Copenhagen Ø.

Email: julief@math.ku.dk

Abstract

Least squares estimators are developed for the parameters in the autocorrelation function of a stationary process. Regularity conditions for consistency and asymptotic normality are given, and optimal weights are derived. It is shown how goodness of fit and model selection can be based on the distance between empirical and fitted autocorrelations. Examples of sums of Ornstein-Uhlenbeck type processes and sums of linear drift diffusions are studied in greater detail. The performance of the estimators and the goodness of fit test is evaluated through Monte Carlo simulations.

Key words: asymptotic normality, consistency, goodness of fit, Levy process, model selection, optimal estimation, stochastic differential equation.

A.1 Introduction

When dealing with data series sampled over time measurements typically are dependent and quantifying dependence is thus an important part of the statistical analysis. One aspect often focused upon is that of simple linear correlation as summarized by the autocorrelation function of a stationary stochastic process. Many statistical models in longitudinal data analysis (see Diggle et al. (2002)) and time series analysis involve parameters specifying the correlation structure. Matching the correlation structure found in real data is an issue in model building, model selection, and in assesing goodness of fit. Stochastic processes with a more delicate correlation structure have recently been constructed in Barndorff-Nielsen, Jensen & Sørensen (1998) and Bibby, Skovgaard & Sørensen (2005). The purpose being to better model the autocorrelation found in high frequency data in the fields of turbulence and finance, see also Barndorff-Nielsen & Shephard (2001a). In this paper we consider inference for the correlation parameters in models such as these. It is often the case that the likelihood function is intractable, hence other means of estimating the parameters must be sought for. A straightforward and thus often used way of estimating the parameters is by least squares estimation based on the empirical autocorrelation function. The estimates in the examples of Bibby, Skovgaard & Sørensen (2005) and Barndorff-Nielsen & Shephard (2001a) are least squares estimates. The advantages to this approach are obvious. Least squares estimation is simple from a theoretical as well as a computational point of view. As demonstrated in section A.2 below the least squares estimator is strongly consistent and eventually unique under mild regularity conditions. In practice, estimates can be calculated with any statistical standard software, and the adequacy of the estimates calculated can be checked by comparing the estimated autocorrelation function with the empirical one. Nevertheless, least squares estimation of correlation parameters is not just a matter of simple curve fitting. Contrary to the case of simple regression the empirical autocorrelations typically have unequal variances and are mutually dependent. This should be taken into account when looking at the deviation between empirical and estimated autocorrelation functions.

The structure of the paper is as follows. In section A.2 we consider least squares estimation for autocorrelation parameters in a general setup. The asymptotic theory for the least squares estimator is developed, an optimal weight is derived together with a goodness of fit statistic and a model selection strategy. In section 3 we apply the results of section A.2 to the processes constructed in Barndorff-Nielsen, Jensen & Sørensen (1998) and Bibby, Skovgaard & Sørensen (2005). That is stationary stochastic processes with autocorrelation functions of the form $\rho(t) = \sum_{j=1}^{m} \phi_j \lambda_j^t$. In particular, we show that the integer valued parameter m can be estimated consistently. The examples are concluded by a thorough simulation study.

A.2 LSE for autocorrelation parameters

In this section we consider least squares estimation for the parameter in an autocorrelation function of a stationary process. Large sample results are derived under standard regularity conditions. As usual the conditions form a compromise between the demands for generality and simplicity; They are well suited for several classes of processes as the ones considered in section A.3 although not necessary in a strict mathematical sense. For technical details we refer to the proofs in appendix A.

Recall that the *autocorrelation function* of a stationary stochastic process $(Y_i)_{i\in\mathbb{N}}$ is defined by

$$\rho(t) = \mathrm{Cor}(Y_i, Y_{i+t}) = \{E(Y_i Y_{i+t}) - \mu^2\} \cdot \sigma^{-2}, \quad t \in \mathbb{N}_0$$

where μ is the mean of $(Y_i)_{i\in\mathbb{N}}$ and σ^2 is the variance. Based on a the sample $Y_1, \ldots, Y_n$ the *t-lag correlation* $\rho(t)$, where $t < n$, can be estimated by a moment estimator such as

$$r_{n,t} = \frac{c_{n,t}}{s_n^2} = \frac{\frac{1}{n-t}\sum_{i=1}^{n-t} Y_i Y_{i+t} - \{\frac{1}{n-t}\sum_{i=1}^{n-t} Y_i\}\{\frac{1}{n-t}\sum_{i=1}^{n-t} Y_{i+t}\}}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \{\frac{1}{n}\sum_{i=1}^{n} Y_i\}^2}, \tag{A.1}$$

where the nominator $c_{n,t}$ is the empirical t-lag covariance and the denominator $s_n^2 = c_{n,0}$ is the empirical variance of $Y_1, \ldots, Y_n$. We term $r_{n,t}$ *the empirical t-lag correlation.* If $(Y_i)_{i\in\mathbb{N}}$ is ergodic, then $r_{n,t}$ is a strongly consistent estimator of $\rho(t)$ for any fixed t.

Remark A.2.1 Other estimators are referred to as empirical t-lag correlations. The correlation of the empirical distribution on $\{(Y_1, Y_{t+1}), \dots, (Y_{n-t}, Y_n)\}$ contrary to $r_{n,t}$ attains values in $[-1, 1]$ almost surely. Another often employed empirical correlation is obtained by replacing the empirical covariance in (A.1) with $\frac{1}{n-t}\sum_{i=1}^{n-t} Y_i Y_{i+t} - \{\frac{1}{n}\sum_{i=1}^{n} Y_i\}^2$. As it is based on the pooled estimate of μ it is argued to be a better estimate than $r_{n,t}$. However, it is not invariant under translation.

For the present exposure we stick to the empirical correlations defined by (A.1) noting that the large sample results given below are valid for other empirical correlations, such as the ones of remark A.2.1, as well.

A.2.1 The least squares estimator

Suppose that the autocorrelation function ρ of $(Y_i)_{i\in\mathbb{N}}$ belongs to a parameterized class $\{\rho(\theta)\colon \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$. We denote the true parameter by $\theta^\star$ so that $\rho = \rho(\theta^\star)$. For a fixed $k \in \mathbb{N}$ we consider the one through k-lag correlations using bold face letters to indicate vectors of such, e.g.

$$\begin{aligned} \mathbf{r}_{n,k} &= \{r_{n,1}, \dots, r_{n,k}\}^T \\ \boldsymbol{\rho}_k(\theta) &= \{\rho(\theta, 1), \dots, \rho(\theta, k)\}^T. \end{aligned}$$

A least squares estimator of θ is obtained by minimizing the criterion

$$l_n^W(\theta) = \{\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}\}^T \cdot W_n \cdot \{\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}\} \qquad \text{(A.2)}$$

where $(W_n)_{n\in\mathbb{N}}$ is a sequence of positive definite k by k weight matrices which may depend on the data.

Definition A.2.1 *Let $\Theta^\star$ be a subset of Θ. The least squares estimator of θ on $\Theta^\star$ is a sequence of $\Theta^\star$-valued random variables $(\hat{\theta}_n)_{n\in\mathbb{N}}$ satisfying that*

$$l_n^W(\hat{\theta}_n) = \inf_{\theta\in\Theta^\star} l_n^W(\theta)$$

eventually with probability one.

For calculations it is useful to note that l_n^W has the same minima as

$$\tilde{l}_n^W = \{s_n^2 \cdot \boldsymbol{\rho}_k(\theta) - \mathbf{c}_{n,k}\}^T \cdot W_n \cdot \{s_n^2 \cdot \boldsymbol{\rho}_k(\theta) - \mathbf{c}_{n,k}\}. \qquad \text{(A.3)}$$

Re-expressing the least squares estimator in terms of empirical covariances is beneficial since the means of the empirical covariances can be calculated whereas the means of the empirical correlations usually cannot.

The least squares estimator resembles the GMM estimators of Hansen (1982), but is not a GMM estimator in itself. Note for instance that

$$\begin{aligned} E(s_n^2 \cdot \rho(t) - c_{n,t}) &= \rho(t)\sigma^2\{\tfrac{1}{n} - \tfrac{1}{n-t}\} - 2\rho(t)\sigma^2\tfrac{1}{n}\textstyle\sum_{s=1}^{n} \tfrac{n-s}{n}\rho(s) \\ &\quad + \sigma^2\tfrac{1}{n-t}\textstyle\sum_{s=1}^{n-t} \tfrac{n-t-s}{n-t}\{\rho(t+s) + \rho(|t-s|)\} \end{aligned}$$

need not equal zero. The least squares estimator can also be regarded from an estimating function point of view. By differentiating $\tilde{l}_n^W$ we obtain an estimating function for θ which is typically biased but nevertheless gives rise to consistent estimates.

A.2.2 Consistency and uniqueness

Only mild regularity conditions are needed to ensure the eventual existence and strong consistency of least squares estimators on every compact subset of Θ containing $\theta^\star$.

A1: The process $(Y_i)_{i\in\mathbb{N}}$ is stationary and ergodic.

A2: For all $\theta_1, \theta_2 \in \Theta$ we have that $\boldsymbol{\rho}_k(\theta_1) = \boldsymbol{\rho}_k(\theta_2)$ if and only if $\theta_1 = \theta_2$.

A3: $(W_n)_{n\in\mathbb{N}}$ converge almost surely to a non-random, regular matrix W_0.

The strategy for proving consistency is standard. One shows that l_n^W converge uniformly to a deterministic function l^W having a unique global minimum at $\theta^\star$. In addition to **A1**-**A3**, we need an assumption to guarantee that l^W only attains values close to the minimum in a small neighborhood of $\theta^\star$. If Θ is compact, it suffices that $\boldsymbol{\rho}_k$ be continuous. For non-compact Θ we also have to rule out approximate minimum points on the boundary of Θ.

Theorem A.2.1 (strong consistency.)
Under **A1**-**A3** *the following hold:*

1. *If $\theta \mapsto \boldsymbol{\rho}_k(\theta)$ is continuous, then for every compact subset $\Theta^\star$ of Θ such that $\theta^\star \in \Theta^\star$ a least squares estimator of θ on $\Theta^\star$ exist, and any least squares estimator of θ on $\Theta^\star$ converge almost surely to $\theta^\star$.*

2. *If $\boldsymbol{\rho}_k$ satisfies that*
$$\inf\{|\boldsymbol{\rho}_k(\theta) - \boldsymbol{\rho}_k|^2 \colon |\theta - \theta^\star| \geq \varepsilon\} > 0 \tag{A.4}$$
for all $\varepsilon > 0$ then a least squares estimator on Θ exists, and any least squares estimator on Θ converge almost surely to $\theta^\star$.

Note that if $\boldsymbol{\rho}_k$ is continuous and the parameter space Θ is compact, then by **A2** condition (A.4) holds and case *1.* and *2.* coincide.
The least squares estimators of theorem A.2.1 can be shown to be unique eventually with probability one if the limit criterion is convex in a neighborhood of $\theta^\star$. The following assumptions ensure that this is the case. The same assumption is used when demonstrating asymptotic normality of the least squares estimator.

A4: For $t = 1, \ldots, k$ the mappings $\theta \mapsto \rho(\theta, t)$ are twice continously differentiable with respect to θ and the first order derivative $V = \partial_{\theta^T} \boldsymbol{\rho}_k(\theta)|_{\theta=\theta^\star}$ evaluated at $\theta^\star$ satisfy that for some d-subset $\{t_1, \ldots, t_d\}$ of $\{1, \ldots, k\}$ the rows $(V)_{t_1}, \ldots, (V)_{t_d}$ are linearly independent.

It is illuminating to make the following observation. The second order derivative of the limit criterion l^W at $\theta^\star$ equals $2 \cdot V^T W_0 V$ which by the regularity of W_0 is positive definite if and only if V has rank d. Namely if the requirements of **A4** are satisfied. Consequently, if **A4** is to hold, the number of autocorrelations considered must be greater than or equal to the number of parameters in the autocorrelation function, $k \geq d$ that is.

Theorem A.2.2 (uniqueness)
Suppose that **A1**-**A4** *hold. Then in both cases 1. and 2. of theorem A.2.1 the least squares estimator is unique eventually with probability one.*

A.2.3 Asymptotic normality and optimal weights

In order to show root n consistency of the least squares estimator a few more assumptions are needed. The definition of α-mixing can be found in Doukhan (1994) or Nahapetian (1991).

A5: $\theta^\star$ is an interior point of Θ.

A6: There exists a $\delta > 0$ such that $E(Y_1^{2\cdot(\delta+2)})$ is finite and such that $(Y_t)_{t\in\mathbb{N}}$ is α-mixing with mixing coefficients satisfying that $\sum_{i=1}^\infty \alpha_i^{\delta\cdot(\delta+2)^{-1}} < \infty$.

The second assumption strengthens **A1**. It allows us to apply the central limit theorem in the proof of theorem A.2.3 below and could be replaced by any other condition with the same impact. Please note that the processes considered in section A.3.1 are α-mixing with exponentially decaying mixing coefficients and hence comply with **A6**.

Theorem A.2.3 (asymptotic normality.)
Assume that **A1**-**A6** *hold. If* $\hat{\theta}_n$ *is a consistent least squares estimator, then*

$$\sqrt{n}\cdot\{\hat{\theta}_n - \theta^\star\} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\Lambda)$$

where $\Lambda = (V^T W_0 V)^{-1}\cdot V^T W_0 \Sigma W_0 V\cdot(V^T W_0 V)^{-1}$ *and* Σ *is the* k *by* k *matrix with entries given by*

$$\sigma_{s,t} = \rho(s)\rho(t)S_{0,0} - \rho(s)S_{0,t} - \rho(t)S_{0,s} + S_{s,t}$$

where $S_{s,t}$ *for* $s,t\in\{0,\dots,k\}$ *denotes the series*

$$\mu_{1,1,s+1,t+1} - \rho(s)\rho(t) + \sum_{i=1}^\infty\{\mu_{1,s+1,i+1,i+t+1} + \mu_{1,t+1,i+1,i+s+1} - 2\rho(s)\rho(t)\}$$

of standardized joint moments

$$\mu_{i_1,i_2,i_3,i_4} = E\left(\{(Y_{i_1}-\mu)(Y_{i_2}-\mu)(Y_{i_3}-\mu)(Y_{i_4}-\mu)\}\cdot\sigma^{-4}\right).$$

If confidence sets for θ are to be calculated from theorem A.2.3, a consistent estimate of the variance matrix Λ is in demand. To this end we consider the estimation of each of the components V, W_0, and Σ. A consistent estimator of V is given by $\hat{V}_n = \partial_{\theta^T}\boldsymbol{\rho}_k(\theta)|_{\theta=\hat{\theta}_n}$, whereas W_0 by assumption can be estimated by W_n. The matrix Σ forms a real challenge as the standardized joint moments need not be known let alone the series of such. Moreover, as the model is semiparametric, Σ may not be fully specified as a function of θ. Nevertheless, some cases including example A.3.1 yield simple closed form expressions of the moment series leading to explicit estimators of Σ. In connection with the examples of section A.3.1 it is advantageous to restate the matrix Σ in terms of cumulants.

Remark A.2.2 Following Barndorff-Nielsen & Cox (1989) but with a slightly different annotation we define the standardized joint cumulants by

$$\kappa_{i_1,i_2,i_3,i_4} = \mu_{i_1,i_2,i_3,i_4} - \rho|i_2-i_1|\rho|i_4-i_3| - \rho|i_3-i_1|\rho|i_4-i_2| - \rho|i_4-i_1|\rho|i_3-i_2|.$$

For $s, t \in \{0, \ldots, k\}$ let $K_{s,t}$ denote the cumulant series

$$\kappa_{1,1,s+1,t+1} + \sum_{i=1}^{\infty} \{\kappa_{1,s+1,i+1,i+t+1} + \kappa_{1,t+1,i+1,i+s+1}\},$$

then

$$\begin{aligned} \sigma_{s,t} &= \rho(s)\rho(t)K_{0,0} - \rho(s)K_{0,t} - \rho(t)K_{0,s} + K_{s,t} \\ &\quad +\rho(s)\rho(t)R_{0,0} - \rho(s)R_{0,t} - \rho(t)R_{0,s} + R_{s,t} \end{aligned}$$

where

$$\begin{aligned} R_{s,t} &= \rho(|s-t|) + \rho(s)\rho(t) + 2 \cdot \textstyle\sum_{i=1}^{\infty}\rho(i)\{\rho(i + |s-t|) + \rho(i+s+t)\} \\ &+ \textstyle\sum_{i=1}^{|s-t|}\rho(i)\rho(|s-t|-i) + \sum_{i=1}^{s}\rho(i+t)\rho(s-i) + \sum_{i=1}^{t}\rho(i+s)\rho(t-i). \end{aligned}$$

is a remainder consisting of essentially known terms.

In general closed form expressions of Σ are rare. Still Σ can be estimated by a kernel estimator, the so-called *autocorrelation consistent covariance matrix estimator* going back to Newey & West (1987b) and White (1984). We recommend Hansen (1992) for an introduction, Andrews & Monahan (1992) and Jansson (2002) for further reading. It is not obvious from the statement of theorem A.2.3 that Σ is a long run covariance matrix. However, from the proof we get that

$$\Sigma = \lim_{n\to\infty} \sum_{i=1}^{n} \sum_{i'=1}^{n} E(Z_i Z_{i'}^T)$$

where $Z_i = \{\rho(1)Y_i^2 - Y_iY_{i+1}, \ldots, \rho(k)Y_i^2 - Y_iY_{i+k}\}^T$ defines a k dimensional stationary process with the same mixing properties as $(Y_i)_{i\in\mathbb{N}}$.

An important issue when applying the least squares estimator is to choose the weights so that the resulting estimate is likely to be as close to $\theta^\star$ as possible. Let us consider the class of least squares estimators induced by sequences of weight matrices $(W_n)_{n\in\mathbb{N}}$ satisfying **A3**. We term a sequence of weight matrices *optimal* if the asymptotic variance of the induced estimator $\hat{\theta}_n$ is minimal with respect to the partial ordering of positive semidefinite d by d matrices. Please note that optimality of a sequence of weight matrices depends only on the limit W_0. The following result carries over directly from Hansen (1982) theorem 3.2 and is therefore stated without proof.

Theorem A.2.4 (optimal weights)
Suppose that **A1**-**A6** *hold and that the matrix Σ is regular. The weight matrices $(W_n)_{n\in\mathbb{N}}$ form an optimal sequence if and only if*

$$V^T W_0 = B \cdot V^T \Sigma^{-1}$$

for some regular d by d matrix B. Moreover, all optimal least squares estimators have asymptotic variance $\Lambda = (V^T\Sigma^{-1}V)^{-1}$.

Hence, a sequence of optimal weights is given by $W_n = \hat{\Sigma}_n^{-1}$ where $\hat{\Sigma}_n$ is a consistent estimator of Σ. In practice, we need an initial estimate of θ to calculate the weight matrix. Such an estimate could be the ordinary (i.e. identity weight) least squares estimator.

A.2.4 Misspecification and goodness of fit

Even when the model is misspecified the least squares estimator has a meaningful interpretation. As long as **A1** and **A3** are in power, it holds that

$$\inf_{\theta\in\Theta}\{\boldsymbol{\rho}_k(\theta)-\mathbf{r}_{n,k}\}^T\cdot W_n\cdot\{\boldsymbol{\rho}_k(\theta)-\mathbf{r}_{n,k}\}\xrightarrow{\text{a.s.}}\inf_{\theta\in\Theta}\{\boldsymbol{\rho}_k(\theta)-\boldsymbol{\rho}_k\}^T\cdot W_0\cdot\{\boldsymbol{\rho}_k(\theta)-\boldsymbol{\rho}_k\}$$

Asymptotically speaking, the autocorrelation function induced by the least squares estimator is as close to the true autocorrelation function as possible in the sense that its one through k-lag correlations attains minimum distance with respect to W_0. In particular, if there exists a unique $\overline{\theta}\in\Theta$ such that for all $\varepsilon>0$

$$\inf_{|\theta-\overline{\theta}|\geq\varepsilon}\{\boldsymbol{\rho}_k(\theta)-\boldsymbol{\rho}_k\}^T\cdot W_0\cdot\{\boldsymbol{\rho}_k(\theta)-\boldsymbol{\rho}_k\}>\{\boldsymbol{\rho}_k(\overline{\theta})-\boldsymbol{\rho}_k\}^T\cdot W_0\cdot\{\boldsymbol{\rho}_k(\overline{\theta})-\boldsymbol{\rho}_k\},$$

then a least squares estimator exists eventually with probability one and any least squares estimator converge almost surely to $\overline{\theta}$. To verify this claim recycle the proof of theorem A.2.1.

In order to check the adequacy of the model, $\{\rho(\theta)\colon\theta\in\Theta\}$, we can apply a goodness of fit test like the one of Hansen (1982) lemma 4.2, which we refer to for proof.

Theorem A.2.5 (goodness of fit)
Suppose that **A1**-**A6** *hold, that $k>d$, and that the matrix Σ defined in theorem A.2.3 is regular. If $\hat{\theta}_n$ is an optimal least squares estimator, then*

$$n\cdot\{\boldsymbol{\rho}_k(\hat{\theta}_n)-\mathbf{r}_{n,k}\}^T\cdot W_n\cdot\{\boldsymbol{\rho}_k(\hat{\theta}_n)-\mathbf{r}_{n,k}\}\xrightarrow{\mathcal{D}}\chi^2_{k-d}$$

for any sequence of optimal weights $(W_n)_{n\in\mathbb{N}}$.

The ultimate goal when assessing the fit of the model is to ascertain whether or not $\rho\in\{\rho(\theta)\colon\theta\in\Theta\}$. To this end the goodness of fit test induced by theorem A.2.5 is likely but not sure to be successful; Whenever testing at a fixed level p_0 there will be a small probability of approximately p_0 of rejecting the model even though it is true. One could try to mend this by decreasing p_0 as the number of observations increase. That is, with $(p_n)_{n\in\mathbb{N}}$ a sequence decreasing to zero

$$\text{Accept }\{\rho(\theta)\colon\theta\in\Theta\}\text{ if }\{\boldsymbol{\rho}_k(\hat{\theta}_n)-\mathbf{r}_{n,k}\}^TW_n\{\boldsymbol{\rho}_k(\hat{\theta}_n)-\mathbf{r}_{n,k}\}\leq n^{-1}\cdot\chi^2_{k-d,1-p_n}.$$

$$\text{Reject }\{\rho(\theta)\colon\theta\in\Theta\}\text{ if }\{\boldsymbol{\rho}_k(\hat{\theta}_n)-\mathbf{r}_{n,k}\}^TW_n\{\boldsymbol{\rho}_k(\hat{\theta}_n)-\mathbf{r}_{n,k}\}>n^{-1}\cdot\chi^2_{k-d,1-p_n}.$$

Where $\chi^2_{k-d,1-p_n}$ denotes the $1-p_n$ quantile in the χ^2 distribution with $k-d$ degrees of freedom. Let $\varepsilon_n=n^{-1}\cdot\chi^2_{k-d,1-p_n}$. Our next result, inspired by Dembo & Peres (1994), gives conditions on $(\varepsilon_n)_{n\in\mathbb{N}}$ ensuring that the above scheme is successful eventually with probability one.

In case the model is misspecified it may be problematic to show that an estimated sequence of weights converge almost surely. It is thus comforting to find that theorem A.2.6 still hold when **A3** is replaced by the weaker condition **A7**. For the second part of the theorem we need a slightly strengthened version of condition **A6**.

A7: $(W_n)_{n\in\mathbb{N}}$ is contained in a compact set of positive definite k by k matrices.

A8: There exists constants $\Delta, \delta, \delta' > 0$ such that $E(Y_1^{2\cdot(\delta+2)})$ is finite and such that $(Y_t)_{t\in\mathbb{N}}$ is α-mixing with mixing coefficients satisfying that $\sum_{i=n}^{\infty} \alpha_i^{\delta\cdot(\delta+2)^{-1}} \leq \Delta\cdot(\log n)^{-(3+\delta')}$ eventually.

Theorem A.2.6 (discernibility)
Let ρ be the autocorrelation function of the stationary process $(Y_i)_{i\in\mathbb{N}}$ and $\{\rho(\theta)\colon \theta\in\Theta\}$ a family of autocorrelation functions.

1. *Suppose that* $\inf_{\theta\in\Theta} |\boldsymbol{\rho}_k(\theta) - \boldsymbol{\rho}_k|^2 > 0$. *If* **A1** *and* **A7** *hold, then for every sequence* $(\varepsilon_n)_{n\in\mathbb{N}}$ *such that* $\varepsilon_n \to 0$

$$\inf_{\theta\in\Theta}\{\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}\}^T \cdot W_n \cdot \{\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}\} > \varepsilon_n$$

holds eventually with probability one.

2. *Suppose that* $\rho \in \{\rho(\theta)\colon \theta\in\Theta\}$. *If* **A7** *and* **A8** *hold, then for every sequence* $(\varepsilon_n)_{n\in\mathbb{N}}$ *such that* $\varepsilon_n \cdot n \cdot \{\log(\log n)\}^{-1} \to \infty$

$$\inf_{\theta\in\Theta}\{\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}\}^T \cdot W_n \cdot \{\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}\} \leq \varepsilon_n$$

holds eventually with probability one.

Please note that **A3** and **A7** are guaranteed to hold for any constant deterministic weight sequence. For a sequence of estimated optimal weights **A7** has to be established under misspecification. For instance, the behavior of the optimal weights $W_n = \hat{\Sigma}_n^{-1}$ depend on the estimator $\hat{\Sigma}_n$.

Remark A.2.3 In simple settings such as in example A.3.1 of section A.3 we might find $\hat{\Sigma}_n = \Sigma(\tilde{\theta}_n, \tilde{\tau}_n)$ with $\tilde{\theta}_n$ an initial estimate of θ, $\tilde{\tau}_n$ an almost surely convergent statistic, and Σ a continuous function of (θ,τ). If $\Sigma(\theta,\tau)$ is positive definite for all (θ,τ), then for **A7** to hold it suffices that $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ stays within a compact subset of Θ.

Theorem A.2.6 is particularly useful when facing a complicated model selection problem like the one described in section A.3.2. Given a hierarchic sequence of models $(\{\rho(\theta_m)\colon \theta_m\in\Theta_m\})_{m=1,2,\ldots}$ the theorem suggests that the smallest m for which ρ belongs to $\{\rho(\theta_m)\colon \theta_m\in\Theta_m\}$ can be picked out eventually with probability one. The hierarchy should not be taken literally; We need not assume that the sequence of models is increasing with respect to inclusion. Rather we shall think of the ordering as a matter of preference; Models appearing in the beginning of the sequence are preferred to models indexed by larger numbers as long as they display a satisfactory fit of the data. The hierarchy could be induced by an increasing complexity of the models in the sense that $\dim(\Theta_m)$ increases with m, but we might as well deal with the problem of choosing between competing models where the ordering relies on vague arguments.
To be more specific, fix $k\in\mathbb{N}$ and let $(W_{m,n})_{n\in\mathbb{N}}$ where $m = 1, 2, \ldots$ be sequences of k by k weight matrices. As above, goodness of fit is measured in terms of the distances

$$d_{n,m} = \inf_{\theta_m\in\Theta_m}\{\boldsymbol{\rho}_k(\theta_m) - \mathbf{r}_{n,k}\}^T \cdot W_n \cdot \{\boldsymbol{\rho}_k(\theta_m) - \mathbf{r}_{n,k}\}$$

between the empirical autocorrelations and the least squares induced estimates. For $m = 1, 2, \dots$ let $(\varepsilon_{m,n})_{n\in\mathbb{N}}$ be sequences of maximal acceptable distances. Then model number $\widehat{m}_n$ is selected where

$$\widehat{m}_n = \inf\{m = 1, 2, \dots : d_{n,m} \leq \varepsilon_{n,m}\}.$$

Theorem A.2.6 translates directly into consistency of $\widehat{m}_n$. We denote by $m^\star$ the smallest m for which the true autocorrelation function belongs to $\{\rho(\theta_m) \colon \theta_m \in \Theta_m\}$.

Corollary A.2.1 (model selection)
Assume that **A8** *holds, that* **A7** *holds for the weights indexed by* $m = 1, \dots, m^\star$*, and that* $\inf_{m=1,\dots,m^\star-1} \inf_{\theta_m\in\Theta_m} |\boldsymbol{\rho}_k(\theta_m) - \boldsymbol{\rho}_k|^2 > 0$*. If the sequences* $(\varepsilon_{n,m})_{n\in\mathbb{N}}$ *satisfy that* $\varepsilon_{n,m} \to 0$ *and* $\varepsilon_{n,m} \cdot n \cdot \{\log(\log n)\}^{-1} \to \infty$ *for* $m = 1, \dots, m^\star$*, then* $\widehat{m}_n = m^\star$ *eventually with probability one.*

A.3 Examples

In this section we consider the specific class of autocorrelation functions given by

$$\rho(t) = \sum_{j=1}^{m} \phi_j \lambda_j^t \tag{A.5}$$

where $m \in \mathbb{N}$, $\lambda_1, \dots, \lambda_m \in [0;1]$, $\phi_1, \dots, \phi_m > 0$, and $\sum_{j=1}^m \phi_m = 1$. The autocorrelations of form (B.32) have been studied by Barndorff-Nielsen, Jensen & Sørensen (1998) and Bibby, Skovgaard & Sørensen (2005). Both papers are devoted to the construction of classes of stationary processes which allow for flexibility in the choice of marginal distribution as well as in the autocorrelation function. Their construction is reviewed in section A.3.1 below. Next, in section A.3.2 we apply the results of section A.2 to derive consistent estimators for the parameters in (B.32). In particular we demonstrate that the integer valued m can be estimated consistently.

A.3.1 Modeling autocorrelation

In this section we review the construction of stationary processes with autocorrelation function of the form (B.32) as given by Barndorff-Nielsen, Jensen & Sørensen (1998) and Bibby, Skovgaard & Sørensen (2005). In addition, we study the mixing properties of these processes as well as the problem of calculating their quadro-variate joint moments.
The basic idea is as follows. Suppose that $(X_t^{(1)})_{t\geq 0}, \dots, (X_t^{(m)})_{t\geq 0}$ are independent stationary processes each having a well defined autocorrelation function $\rho_j(t) = \lambda_j^t$, then

$$Y_t = X_t^{(1)} + \dots + X_t^{(m)},$$

defines a stationary process which has autocorrelation function of form (B.32) with $\phi_j = \mathrm{Var}(X^{(j)}) \cdot \{\mathrm{Var}(X^{(1)}) + \dots + \mathrm{Var}(X^{(m)})\}^{-1}$.
The constructions of Barndorff-Nielsen, Jensen & Sørensen (1998) and Bibby, Skovgaard & Sørensen (2005) differ only in their choices of underlying processes. We summarize these in examples A.3.1 and A.3.2 below.

Remark A.3.1 The models constructed above are sometimes referred to as multiple time scale models. This is due to the following property. Let $\delta \geq 0$, then the time changed process $(X_{\delta t}^{(j)})_{t\geq 0}$ has the same marginal distribution as $(X_t^{(j)})_{t\geq 0}$ and autocorrelation function $\tilde{\rho}(t) = (\lambda_j^{\delta})^t$. Hence, the λ_j's measure the speed at which the underlying processes evolve with time. Moreover, the λ_j's can be chosen independently of any admissible marginal distribution.

Example A.3.1 (Ornstein-Uhlenbeck type processes.)
Barndorff-Nielsen, Jensen & Sørensen (1998) take their underlying processes to be Ornstein-Uhlenbeck type processes, i.e. solutions of the stochastic differential equation

$$dX_t = \theta X_t dt + dZ_t \tag{A.6}$$

driven by a homogenous Levy process $(Z_t)_{t\geq 0}$. In case $(Z_t)_{t\geq 0}$ is a Brownian motion, we recover the ordinary Ornstein-Uhlenbeck process; the only Ornstein-Uhlenbeck type process which does not have jumps.
Given a $\theta > 0$ and a characteristic function C of some probability distribution satisfying

A9: C is self-decomposable, differentiable at any point other than zero, and yields a continuous extension of $\xi \cdot \frac{d}{d\xi} \log C(\xi)$ at zero

then with $(Z_t)_{t\geq 0}$ specified by the characteristic function $C_{Z_1}(\xi) = \exp\{\theta\xi \cdot \frac{d}{d\xi} \log C(\xi)\}$, a stationary solution of (A.6) is given by

$$X_t = \exp(-\theta t)X_0 + \int_0^t \exp\{-\theta(t-s)\}dZ_s.$$

where X_0 is distributed according to C and independent of $(Z_t)_{t\geq 0}$. If the marginal distribution has second order moment, then $(X_t)_{t\geq 0}$ has autocorrelation function $\rho(t) = \exp(-\theta t)$. Moreover, it follows from Masuda (2004) theorem 4.3 that $(X_t)_{t\geq 0}$ is strongly mixing with exponentially decaying mixing coefficients.
It is worth noting that the discrete time process $(X_t)_{t\in\mathbb{N}}$ forms an autoregression

$$X_{t+1} = \lambda \cdot X_t + \varepsilon_t$$

where $\lambda = \exp(-\theta)$ and $(\varepsilon_t)_{t\in\mathbb{N}}$ are i.i.d. It is thus straightforward to calculate conditional moments of any order. Let μ and σ^2 denote the mean and variance of $(X_t)_{t\geq 0}$, then

$$\begin{aligned} &E(X_{s+t} - \mu\ |X_s) = \lambda^t(X_s - \mu), \\ &E(\{X_{s+t} - \mu\}^2|X_s) = \lambda^{2t}(X_s - \mu)^2 + (1 - \lambda^{2t})\sigma^2, \end{aligned}$$

and with ζ_3 denoting the third order standardized moment of $(X_t)_{t\geq 0}$

$$E(\{X_{s+t} - \mu\}^3|X_s) = \lambda^{3t}(X_s - \mu)^3 + 3\lambda^t(1 - \lambda^{2t})\sigma^2(X_s - \mu) + (1 - \lambda^{3t})\zeta_3\sigma^3.$$

Further calculations based on successive conditioning lead to explicit formulae for standardized joint moments and cumulants. The latter are given by

$$\kappa_{i_1,i_2,i_3,i_4} = \lambda^{i_2+i_3+i_4-3i_1}(\zeta_4 - 3)$$

for $i_1 \leq i_2, i_3, i_4$ and with ζ_4 denoting the fourth order standardized moment of X_0. (See theorem A.2.3 and remark A.2.2 for the definitions of standardized joint moments and cumulants.)

Example A.3.2 (Diffusion type processes.)
Bibby, Skovgaard & Sørensen (2005) consider linear drift diffusions, i.e. solutions of the stochastic differential equation

$$dX_t = -\theta(X_t - \mu)dt + \sigma(X_t)dB_t. \tag{A.7}$$

Given a $\theta > 0$ and a density f on the interval $]l;u[$ satisfying

A10: f is continuous, strictly positive, and bounded on $]l;u[$ with second order moment,

they show that with $\mu = \int_l^u f(x)dx$ and $\sigma^2(x) = f(x)^{-1}\int_l^x (\mu - y)f(y)dy$ a unique stationary weak solution with marginal density f exists. It is an ergodic, time reversible diffusion and has autocorrelation function $\rho(t) = \exp(-\theta t)$. Well known examples of linear drift diffusions are the Ornstein-Uhlenbeck processes (Gaussian marginal) and the Cox-Ingersoll-Ross processes (Gamma marginal).
The mixing properties of diffusions are reviewed in Genon-Catalot, Jeantheau & Laredo (2000). Their theorem 2.6 in combination with the re-parameterization of Hansen, Scheinkman & Touzi (1998) section 5 yield that $(X_t)_{t\geq 0}$ is strongly mixing with mixing coefficients $\alpha_t \leq \frac{1}{4}\exp(-\theta t)$ decaying exponentially fast.
Bibby, Skovgaard & Sørensen (2005) derive a formula for the conditional first order moment

$$E(X_{s+t} - \mu \,|X_s) = \lambda^t(X_s - \mu)$$

where $\lambda = \exp(-\theta)$. Higher order conditional moments in general are hard to come by. In case the diffusion coefficient is quadratic and the process has fourth order moment, an explicit expression of the conditional second order moment can be found. Suppose that $\theta^{-1}\sigma^2(x) = a(x-\mu)^2 + b(x-\mu) + c$. Itô's formula yields an evolution equation for $(X_t - \mu)^2$ from which we deduce that the variance of $(X_t)_{t\geq 0}$ is given by $\sigma^2 = c(2-a)^{-1}$ and moreover that,

$$\begin{aligned} E(\{X_{s+t} - \mu\}^2 | X_s) &= \lambda^{(2-a)t}(X_s - \mu)^2 + \lambda^t\{1 - \lambda^{(1-a)t}\}b(1-a)^{-1}(X_s - \mu) \\ &\quad + \{1 - \lambda^{(2-a)t}\}\sigma^2. \end{aligned}$$

Note that in case $(X_t)_{t\geq 0}$ has sixth order moment, $b(1-a)^{-1} = E(X_t - \mu)^3\sigma^{-2}$. Further calculations based on successive conditioning and exploiting the time reversibility reveal the joint moments and finally the joint cumulants which are given by

$$\begin{aligned} \kappa_{i_1,i_2,i_3,i_4} &= \lambda^{i_4+(1-a)i_3-(1-a)i_2-i_1}\{\zeta^4 - b(1-a)^{-1}\sigma^{-1}\zeta^3 - 1\} \\ &\quad + \lambda^{i_4-i_1}b(1-a)^{-1}\sigma^{-1}\zeta_3 - 2\lambda^{i_4+i_3-i_2-i_1} \end{aligned}$$

for $i_1 \leq i_2 \leq i_3 \leq i_4$ and with ζ_3 denoting the standardized third order moment of $(X_t)_{t\geq 0}$. Examples of diffusions with quadratic diffusion coefficients can be found in table 1 of Bibby, Skovgaard & Sørensen (2005).

In both of the examples the $(X_t^{(j)})_{t\geq 0}$'s are Markov processes. Typically when adding the processes, the Markov property is destroyed, and statistical inference is thus complicated. However, mixing properties caries over from the underlying processes to their sum. If each $(X_t^{(j)})_{t\geq 0}$ is α-mixing, then so is $(Y_t)_{t\geq 0}$ as $\alpha_t(Y) \leq \sum_{j=1}^m \alpha_t(X^{(j)})$.[1]

[1] Doukhan (1994) theorem 1.1.1

Furthermore, once the joint moments of the underlying processes are found, the joint moments of the sum process can easily be calculated. The standardized joint cumulants satisfy

$$\kappa_{i_1,i_2,i_3,i_4}(Y) = \sum_{j=1}^{m} \phi_j^2 \cdot \kappa_{i_1,i_2,i_3,i_4}(X^{(j)}).$$

Remark A.3.2 In the above examples $\lambda_j = \exp(-\theta_j)$ with $\theta_j > 0$ implying that $0 < \lambda_j < 1$. However, the autocorrelation function (B.32) does make sense even if $\lambda_j = 0$ or $\lambda_j = 1$; Any i.i.d. sequence with second order moment has autocorrelation function $\rho(t) = \lambda^t$ with $\lambda = 0$, whereas $\lambda = 1$ defines the autocorrelation function of a non-degenerate, constant process.

A.3.2 Estimating the parameters

Suppose that $Y_1, \ldots, Y_n$ is a sample from a stationary process with autocorrelation function given by (B.32). We consider the estimation problem. It is important to notice that m determines the dimensions of the remaining parameters. In case m is known, least squares estimation can be employed as described in section A.2. The large sample results are summarized below. The parameter m cannot be estimated by least squares estimation the problem being that the distance between the empirical and the fitted autocorrelations decreases with m. Rather, the estimation of m should be viewed as a model selection problem as in section A.2.4.
To be more specific, we consider for $m \in \mathbb{N}$ the model parameterized by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{m-1})$ belonging to the parameter space

$$\Theta_m = \{(\boldsymbol{\lambda}, \boldsymbol{\phi}) \in [0;1]^{2m-1} \colon \lambda_1 > \ldots > \lambda_m,\ \phi_1, \ldots \phi_{m-1} > 0,\ \textstyle\sum_{j=1}^{m-1} \phi_j < 1\}.$$

Note that $\phi_m = 1 - \sum_{j=1}^{m-1} \phi_j$ is implicitly given. The restrictions $\lambda_1 > \ldots > \lambda_m$ and $\phi_1, \ldots, \phi_m > 0$ are needed to make the parameters identifiable. Properties of the parameterization are listed in the following lemma. Part 1 ensures that the parameters can be recognized from the sequence $\{\rho(t)\}_{t \in \mathbb{N}}$. The remaining parts of the lemma refer to the regularity conditions of section A.2; Provided that a sufficient number of lags are considered **A2**, **A4**, and the additional assumptions demanded by theorem A.2.1 and corollary A.2.1 hold true. As in section A.2 the vector of one- through k-lag correlations is denoted by $\boldsymbol{\rho}_k$.

Lemma A.3.1 (properties of the parameterization)

1. *Let $(\boldsymbol{\lambda}, \boldsymbol{\phi}) \in \Theta_m$, $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\phi}}) \in \Theta_{\tilde{m}}$, and $k \geq m + \tilde{m} - 1$. If $\boldsymbol{\rho}_k(\boldsymbol{\lambda}, \boldsymbol{\phi}) = \boldsymbol{\rho}_k(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\phi}})$, then $m = \tilde{m}$ and $(\boldsymbol{\lambda}, \boldsymbol{\phi}) = (\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\phi}})$.*

2. *If $k \geq 2m-1$, then $V(\boldsymbol{\lambda}, \boldsymbol{\phi}) = \partial_{(\boldsymbol{\lambda}, \boldsymbol{\phi})^T} \boldsymbol{\rho}_k(\boldsymbol{\lambda}, \boldsymbol{\phi})$ has full rank $2m-1$ for all $(\boldsymbol{\lambda}, \boldsymbol{\phi}) \in \Theta_m$.*

3. *Let $(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star) \in \Theta_m$ and $\varepsilon > 0$. If $k \geq 2m-1$, then*

$$\inf\{|\boldsymbol{\rho}_k(\boldsymbol{\lambda}, \boldsymbol{\phi}) - \boldsymbol{\rho}_k(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star)|^2 \colon (\boldsymbol{\lambda}, \boldsymbol{\phi}) \in \Theta_m,\ |(\boldsymbol{\lambda}, \boldsymbol{\phi}) - (\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star)| \geq \varepsilon\} > 0.$$

4. *Let* $(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star) \in \Theta_{m^\star}$. *If* $k \geq 2m^\star - 1$, *then*

$$\inf_{m<m^\star} \inf_{(\boldsymbol{\lambda},\boldsymbol{\phi})\in\Theta_m} |\boldsymbol{\rho}_k(\boldsymbol{\lambda}, \boldsymbol{\phi}) - \boldsymbol{\rho}_k(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star)|^2 > 0.$$

The proof can be found in Appendix A.3.3.

Estimating $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$

In case m is known, the remaining parameters can be identified from the first $2m-1$ autocorrelations. Hence, fix $k \geq 2m - 1$ and let $(W_n)_{n\in\mathbb{N}}$ be a pre-specified sequence of k by k weight matrices. The least squares estimator of $(\boldsymbol{\lambda}, \boldsymbol{\phi})$ is given by

$$(\hat{\boldsymbol{\lambda}}_n, \hat{\boldsymbol{\phi}}_n) = \arg \min_{(\boldsymbol{\lambda},\boldsymbol{\phi})\in\Theta_m} \{\boldsymbol{\rho}_k(\boldsymbol{\lambda}, \boldsymbol{\phi}) - \mathbf{r}_{n,k}\}^T W_n \{\boldsymbol{\rho}_k(\boldsymbol{\lambda}, \boldsymbol{\phi}) - \mathbf{r}_{n,k}\}$$

whenever a minimum is attained on Θ_m. The large sample results carry over from section A.2 assuming that the process $(Y_i)_{i\in\mathbb{N}}$ satisfies suitable mixing and moment conditions. Note that for instance the processes of examples A.3.1 and A.3.2 satisfy the mixing condition. Moreover, the weights $(W_n)_{n\in\mathbb{N}}$ must converge almost surely to a regular matrix W_0. Let $\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star$ denote the true parameter values and V the derivative of $\boldsymbol{\rho}_k$ evaluated at $(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star)$.

- If $(Y_i)_{i\in\mathbb{N}}$ is ergodic, then a unique least squares estimator exists eventually with probability one, and it is strongly consistent.

- If $(Y_i)_{i\in\mathbb{N}}$ is strongly mixing and has $4+\delta$ order moment for some $\delta > 0$ such that **A6** holds and $\lambda_m^\star > 0$, then

$$\sqrt{n} \cdot \{(\hat{\boldsymbol{\lambda}}_n, \hat{\boldsymbol{\phi}}_n)^T - (\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star)^T\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Lambda)$$

  where $\Lambda = (V^T W_0 V)^{-1} \cdot V^T W_0 \Sigma W_0 V \cdot (V^T W_0 V)^{-1}$ with the matrix Σ defined as in theorem A.2.3.

- If $(Y_i)_{i\in\mathbb{N}}$ is strongly mixing and has $4+\delta$ order moment for some $\delta > 0$ such that **A6** holds, $\lambda_m^\star > 0$, and Σ is regular, then optimal sequences of weights exist and are characterized by having $V^T W_0 = B V^T \Sigma^{-1}$ for some regular matrix B. In particular, if $\widehat{\Sigma}_n$ is a consistent estimator of Σ, then $W_n = \widehat{\Sigma}_n^{-1}$ is optimal.

- If $(Y_i)_{i\in\mathbb{N}}$ is strongly mixing and has $4+\delta$ order moment for some $\delta > 0$ such that **A6** holds, $k \geq 2m$, $\lambda_m^\star > 0$, Σ is regular, and $(W_n)_{n\in\mathbb{N}}$ is an optimal sequence of weights, then

$$n \cdot \{\boldsymbol{\rho}_k(\hat{\boldsymbol{\lambda}}_n, \hat{\boldsymbol{\phi}}_n) - \mathbf{r}_{n,k}\}^T \cdot W_n \cdot \{\boldsymbol{\rho}_k(\hat{\boldsymbol{\lambda}}_n, \hat{\boldsymbol{\phi}}_n) - \mathbf{r}_{n,k}\} \xrightarrow{\mathcal{D}} \chi^2_{k-d}.$$

Explicit expressions of the matrix Σ are available for many of the processes considered section A.3.1. Confidence sets and optimal weights are thus easy to come by in these examples; We need only insert consistent estimates of the parameters and moments in

the expression of Σ.
Recall that by remark A.2.2 the entries of Σ are given by

$$\begin{aligned}\sigma_{s,t} &= \rho(s)\rho(t)K_{0,0} - \rho(s)K_{0,t} - \rho(t)K_{0,s} + K_{s,t} \\ &\quad + \rho(s)\rho(t)R_{0,0} - \rho(s)R_{0,t} - \rho(t)R_{0,s} + R_{s,t}\end{aligned}$$

where the $K_{s,t}$'s are series of joint cumulants and the $R_{s,t}$'s are remainder terms depending only on ρ. In case ρ takes the form (B.32), we find that

$$\begin{aligned}R_{s,t} &= \rho(|s-t|) + \rho(s)\rho(t) + \textstyle\sum_{j=1}^{m}\phi_j^2\{|s-t|\lambda_j^{|s-t|} + (s+t)\lambda_j^{s+t}\} \\ &\quad + \textstyle\sum_{j\neq j'}\phi_j\phi_{j'}\{\lambda_{j'}^{|s-t|} - \lambda_j^{|s-t|} + \lambda_j^s\lambda_{j'}^t + \lambda_j^t\lambda_{j'}^s - 2\lambda_j^{s+t}\}\cdot\lambda_j(\lambda_{j'}-\lambda_j)^{-1} \\ &\quad + 2\cdot\textstyle\sum_{j=1}^{m}\sum_{j'=1}^{m}\phi_j\phi_{j'}\{\lambda_j^{|s-t|} + \lambda_j^{s+t}\}\cdot\lambda_j\lambda_{j'}(1-\lambda_j\lambda_{j'})^{-1}.\end{aligned}$$

Additional assumptions are needed to calculate the cumulant series.

Example A.3.1 continued: If $(Y_t)_{t\geq 0}$ is the sum of independent stationary Ornstein-Uhlenbeck type processes with fourth order moment, then

$$K_{s,t} = \textstyle\sum_{j=1}^{m}\phi_j^2\lambda_j^{s+t}\{1 + 2\lambda_j^2(1-\lambda_j^2)^{-1}\}(\zeta_{4,j} - 3)$$

where $\zeta_{4,j}$ denotes the fourth order standardized moment of the j'th underlying process. A particularly simple case occurs when the underlying processes are ordinary Ornstein-Uhlenbeck processes. Then $\zeta_{4,j} = 3$ and consequently $K_{s,t} = 0$.

Example A.3.2 continued: If $(Y_t)_{t\geq 0}$ is the sum of independent stationary diffusions with linear drifts and quadratic diffusion coefficients

$$\theta_j^{-1}\sigma_j^2(x) = a_j(x-\mu_j)^2 + b_j(x-\mu_j) + c_j,$$

define constants A_j and B_j by

$$A_j = \zeta_{4,j} - b_j(1-a_j)^{-1}\sigma^{-1}\zeta_{3,j} - 1, \quad B_j = b_j(1-a_j)^{-1}\sigma^{-1}\zeta_{3,j}$$

where $\zeta_{3,j}$ and $\zeta_{4,j}$ denote the standardized moments. Note that if the underlying process has sixth order moment, $B_j = \zeta_{3,j}^2$. The cumulant series are given by

$$K_{s,t} = \textstyle\sum_{j=1}^{m}\phi_j^2 K_{s,t,j}$$

where for $a_j \neq 0$, $t \leq s$,

$$\begin{aligned}K_{s,t,j} &= A_j\lambda_j^{s+t}\{(s-t)\lambda_j^{-a_jt} + 2\lambda_j^{a_j}(1-\lambda_j^{a_j})^{-1} + 2\lambda_j^{2-a_j}(1-\lambda_j^{2-a_j})^{-1}\} \\ &\quad + B_j\lambda_j^s\{s-t+2\lambda_j(1-\lambda_j)^{-1}\} - 2\lambda_j^{s+t}\{s+t+2\lambda_j^2(1-\lambda_j^2)^{-1}\},\end{aligned}$$

and for $a_j = 0$, $t \leq s$,

$$K_{s,t,j} = (A_j - 2)\lambda_j^{s+t}\{s+t+2\lambda_j^2(1-\lambda_j^2)^{-1}\} + B_j\lambda_j^s\{s-t+2\lambda_j(1-\lambda_j)^{-1}\}.$$

In, particular for the sum of independent Cox-Ingersoll-Ross processes having marginal distributions with shape parameters $\alpha_1,\ldots,\alpha_m > 1$ and joint scale parameter we find that

$$K_{s,t} = 2\cdot\textstyle\sum_{j=1}^{m}\phi_j^2\alpha_j^{-1}[\lambda_j^{s+t}\{s+t+2\lambda_j^2(1-\lambda_j^2)^{-1}\} + 2\lambda_j^s\{s-t+2\lambda_j(1-\lambda_j)^{-1}\}].$$

Estimating m

The above is still valid when the known m is replaced by a consistent estimator. In practice only a finite number of different m can be distinguished. The problem is that the whole sequence of autocorrelations $\{\rho(t)\}_{t\in\mathbb{N}}$ is needed to identify an arbitrary $m \in \mathbb{N}$, and for a fixed sample only a limited number of empirical autocorrelations are available. However, it does not seem unreasonable to bound m. One reason for doing so is that the dimension of the parameter space increases with m. Another is that the classes of autocorrelation functions are practically indistinguishable even for moderate values of m. In order to estimate $m \in \{1, \ldots, M\}$, one needs a fixed $k \geq 2M$ and sequences $(\varepsilon_{m,n})_{n\in\mathbb{N}}$ satisfying that $\varepsilon_{m,n} \to 0$ and $\varepsilon_{m,n} \cdot n \cdot \{\log(\log n)\}^{-1} \to \infty$. The estimator $\widehat{m}_n$ is the smallest number in $\{1, \ldots, M\}$ for which

$$\{\boldsymbol{\rho}_k(\hat{\boldsymbol{\lambda}}_{m,n}, \hat{\boldsymbol{\phi}}_{m,n}) - \mathbf{r}_{n,k}\}^T \cdot W_{m,n} \cdot \{\boldsymbol{\rho}_k(\hat{\boldsymbol{\lambda}}_{m,n}, \hat{\boldsymbol{\phi}}_{m,n}) - \mathbf{r}_{n,k}\} \leq \varepsilon_{m,n}$$

where $(\hat{\boldsymbol{\lambda}}_{m,n}, \hat{\boldsymbol{\phi}}_{m,n})$ is the least squares estimator in Θ_m. Let $m^\star$ denote the true value of m.

- If $(Y_i)_{i\in\mathbb{N}}$ is strongly mixing and has $4+\delta$ order moment for some $\delta > 0$ such that **A8** hold true and the weights indexed by $m = 1, \ldots, m^\star$ are bounded in the sense of **A7**, then $\widehat{m}_n = m^\star$ eventually with probability one.

Section A.2.4 suggests taking the weights to be optimal and maximal acceptable distances of the form $\varepsilon_{m,n} = n^{-1}\chi^2_{k-2m+1,1-p_n}$ where $\chi^2_{k-2m+1,1-p_n}$ is the $1 - p_n$ quantile of the χ^2 distribution and $(p_n)_{n\in\mathbb{N}}$ decreases slowly to zero. In doing so we hope to bound the probability of overestimating m with a probability of approximately p_n. However, **A7** still needs to be checked. When it comes to the probability of underestimating m, it cannot be bounded as the true autocorrelation function could have a component with a indefinitely tiny ϕ-value.

A.3.3 Numerical results

In this section we investigate the small-sample behaviour of the least squares estimators for sums of independent Ornstein-Uhlenbeck type processes (example A.3.1). The purpose is to asses whether the asymptotic results are useful in practice and to compare the ordinary and optimally weighted least squares estimators. To faciliate simulation the underlying processes are taken to be ordinary Ornstein-Uhlenbeck processes. The optimal weights are estimated from the formula of section A.3.2 inserting the ordinary least squares estimates as initial estimates. The standard errors of estimates are obtained from the same formulae.

Single Ornstein-Uhlenbeck process

First we consider a single Ornstein-Uhlenbeck process with zero mean, unit variance, and correlation parameter θ, i.e. autocorrelation function $\rho(t) = \exp(-\theta t)$. Table A.1 compares the ordinary and optimally weighted least squares estimates for varying numbers of lags, sample sizes, and three different values of θ. The table displays the sample mean (Mean) and the standard errors (SSE) of the simulated estimates together with the sample mean of their estimated standard errors (SEE) and the covering frequency of the approximate 95% confidence intervals (CP$_{95}$).

		OLSE				Optimally weighted LSE			
n	k	Mean	SSE	SEE	CP$_{95}$	Mean	SSE	SEE	CP$_{95}$
$\theta = 1$									
1,000	5	1.01	0.10	0.10	0.96	1.01	0.08	0.08	0.96
	25	1.01	0.10	0.10	0.96	1.01	0.08	0.08	0.96
	50	1.01	0.10	0.10	0.96	1.01	0.08	0.08	0.96
5,000	5	1.00	0.04	0.04	0.95	1.00	0.04	0.04	0.95
	25	1.00	0.04	0.04	0.95	1.00	0.04	0.04	0.95
	50	1.00	0.04	0.04	0.95	1.00	0.04	0.04	0.95
10,000	5	1.00	0.03	0.03	0.95	1.00	0.03	0.03	0.95
	25	1.00	0.03	0.03	0.95	1.00	0.03	0.03	0.95
	50	1.00	0.03	0.03	0.95	1.00	0.03	0.03	0.95
$\theta = 0.1$									
1,000	5	0.105	0.017	0.017	0.95	0.105	0.016	0.015	0.94
	25	0.109	0.024	0.024	0.96	0.107	0.016	0.015	0.94
	50	0.110	0.026	0.027	0.95	0.107	0.016	0.015	0.94
5,000	5	0.101	0.007	0.007	0.96	0.101	0.007	0.007	0.96
	25	0.102	0.010	0.011	0.95	0.101	0.007	0.007	0.95
	50	0.102	0.012	0.012	0.95	0.101	0.007	0.007	0.95
10,000	5	0.101	0.005	0.005	0.95	0.100	0.005	0.005	0.95
	25	0.101	0.007	0.007	0.95	0.101	0.005	0.005	0.95
	50	0.101	0.008	0.008	0.95	0.101	0.005	0.005	0.95
$\theta = 0.01$									
1,000	5	0.0148	0.0064	0.0054	0.91	0.0148	0.0063	0.0054	0.9
	25	0.0155	0.0071	0.0060	0.92	0.0153	0.0065	0.0055	0.89
	50	0.0163	0.0078	0.0068	0.96	0.0156	0.0066	0.0055	0.88
5,000	5	0.0108	0.0022	0.0021	0.94	0.0108	0.0022	0.0021	0.94
	25	0.0109	0.0023	0.0022	0.95	0.0109	0.0022	0.0021	0.94
	50	0.0110	0.0025	0.0024	0.95	0.0109	0.0022	0.0021	0.94
10,000	5	0.0104	0.0015	0.0015	0.94	0.0104	0.0015	0.0014	0.94
	25	0.0105	0.0016	0.0015	0.94	0.0105	0.0015	0.0015	0.94
	50	0.0105	0.0017	0.0016	0.95	0.0105	0.0015	0.0015	0.94

Table A.1: Comparison of least squares estimates for Ornstein-Uhlenbeck processes with autocorrelation function $\rho(t) = \exp(-\theta t)$. Sample size is denoted by n, whereas k denotes the number of lags. Based on 10,000 time series.

As the discretely observed Ornstein-Uhlenbeck process is merely an AR(1) process, closed form maximum likelihood estimates of mean, variance, and correlation parameter can be found when conditioning on the initial observation. For a comparison maximum likelihood estimates of θ are summarized in table A.2. Mean and variance estimates are not reported as these parameters are treated as nuisance.

		MLE			
θ	n	Mean	SSE	SEE	CP_{95}
1	1,000	1.01	0.08	0.08	0.95
	5,000	1.00	0.04	0.04	0.95
	10,000	1.00	0.02	0.02	0.95
.1	1,000	0.104	0.016	0.015	0.95
	5,000	0.101	0.007	0.007	0.96
	10,000	0.100	0.005	0.005	0.95
.01	1,000	0.0145	0.0062	0.0053	0.91
	5,000	0.0108	0.0022	0.0021	0.94
	10,000	0.0104	0.0015	0.0014	0.94

Table A.2: Summary of maximum likelihood estimates for Ornstein-Uhlenbeck processes with autocorrelation function $\rho(t) = \exp(-\theta t)$.

In all but one scenario the estimators behave as predicted by the asymptotic theory. For $\theta = 0.01$ the sample size $n = 1000$ yields estimators more variable than predicted and both estimators are biased. However, the maximum likelihood estimator is likewise flawed. As the parameter $\theta = 0.01$ is close to the boundary of the parameter space a better normal approximation might be obtained for $\log(\theta)$. When the sample size is increased all of the estimators behave nicely. The efficiency of the weighted least squares estimator matches the maximum likelihood estimator. The variance of the ordinary least squares estimator tend to increase with the number of lags. For five lags it is almost the same as for the weighted least squares estimator. For $\theta = 0.1$ the standard error of the ordinary least squares estimator nearly doubles when the number of lags is increased to fifty. The weighted least squares estimator remains unaffected when the number of lags is increased.

Sums of two Ornstein-Uhlenbeck processes

Next, we consider the weighted sum of two Ornstein-Uhlenbeck processes with mean zero, unit variance and correlation parameters θ_1, θ_2, and ϕ_1. That is, the autocorrelation function is given by $\rho(t) = \phi_1 \exp(-\theta_1 t) + (1 - \phi_1) \exp(-\theta_2 t)$. We fix $\theta_1 = 0.1$ and $\theta_2 = 1$. For three different values of the weight parameter ϕ_1 tables A.3 through A.5 compare the least squares estimators. Please note that $\theta_2 - \theta_1$ have been estimated instead of θ_2 in order for the parameters to vary freely. The tables are similar to those of the previous section. The only difference is that in this case a number of estimates are missing due to the fact that the least squares estimators only exist with a probability tending to one. The frequency on non-existing estimates are reported as P_{ne}.

			OLSE				Optimally weighted LSE			
n	k	P_{ne}	Mean	SSE	SEE	CP_{95}	Mean	SSE	SEE	CP_{95}
Estimates of $\theta_1 = 0.1$ when $\theta_2 = 1$, $\phi_1 = 0.2$										
1,000	5	0.42	0.199	0.130	0.214	0.96	0.197	0.125	0.222	0.96
	25	0.08	0.134	0.091	0.105	0.94	0.128	0.086	0.092	0.93
	50	0.06	0.136	0.089	0.109	0.92	0.131	0.085	0.093	0.93
5,000	5	0.18	0.129	0.075	0.096	0.96	0.129	0.075	0.095	0.95
	25	0.00	0.106	0.036	0.036	0.94	0.106	0.034	0.032	0.94
	50	0.00	0.106	0.038	0.038	0.92	0.107	0.034	0.032	0.94
10,000	5	0.09	0.112	0.059	0.068	0.96	0.112	0.058	0.067	0.96
	25	0.00	0.103	0.025	0.024	0.94	0.103	0.022	0.022	0.94
	50	0.00	0.103	0.026	0.026	0.93	0.104	0.022	0.022	0.94
Estimates of $\theta_1 = 0.1$ when $\theta_2 = 1$, $\phi_1 = 0.5$										
1,000	5	0.14	0.113	0.059	0.072	0.97	0.115	0.058	0.071	0.96
	25	0.06	0.114	0.050	0.054	0.94	0.111	0.044	0.042	0.93
	50	0.10	0.118	0.052	0.061	0.93	0.113	0.044	0.043	0.94
5,000	5	0.00	0.100	0.030	0.030	0.96	0.100	0.030	0.030	0.95
	25	0.00	0.103	0.022	0.022	0.95	0.103	0.017	0.017	0.95
	50	0.02	0.103	0.025	0.026	0.94	0.103	0.017	0.017	0.95
10,000	5	0.00	0.099	0.021	0.021	0.96	0.100	0.021	0.021	0.95
	25	0.00	0.101	0.015	0.015	0.95	0.101	0.012	0.012	0.95
	50	0.00	0.101	0.018	0.018	0.94	0.101	0.012	0.012	0.95
Estimates of $\theta_1 = 0.1$ when $\theta_2 = 1$, $\phi_1 = 0.8$										
1,000	5	0.08	0.100	0.033	0.038	0.98	0.101	0.032	0.037	0.98
	25	0.41	0.096	0.034	0.043	0.95	0.100	0.027	0.029	0.94
	50	0.55	0.097	0.037	0.051	0.94	0.102	0.029	0.030	0.94
5,000	5	0.00	0.100	0.015	0.015	0.95	0.100	0.015	0.015	0.95
	25	0.09	0.100	0.018	0.019	0.96	0.101	0.013	0.012	0.95
	50	0.24	0.099	0.023	0.023	0.96	0.101	0.013	0.012	0.94
10,000	5	0.00	0.100	0.010	0.010	0.95	0.100	0.010	0.010	0.95
	25	0.03	0.101	0.013	0.013	0.96	0.101	0.009	0.009	0.95
	50	0.13	0.100	0.017	0.017	0.96	0.101	0.009	0.009	0.95

Table A.3: Comparison of estimates for the sum of two Ornstein-Uhlenbeck processes with autocorrelation function $\rho(t) = \phi_1 \exp(-\theta_1 t) + (1 - \phi_1) \exp(-\theta_2 t)$ for varying sample size n and number of lags k. Based on 10,000 time series.

			OLSE				Optimally weighted LSE			
n	k	P_{ne}	Mean	SSE	SEE	CP_{95}	Mean	SSE	SEE	CP_{95}
Estimates of $\theta_2 - \theta_1 = 0.9$ when $\theta_1 = 0.1$, $\phi_1 = 0.2$										
1,000	5	0.42	1.100	0.372	0.694	1.00	1.070	0.345	0.512	1.00
	25	0.08	1.010	0.270	0.403	0.98	0.964	0.233	0.240	0.98
	50	0.06	1.020	0.276	0.421	0.98	0.962	0.238	0.252	0.98
5,000	5	0.18	0.928	0.079	0.085	0.98	0.927	0.082	0.082	0.98
	25	0.00	0.913	0.078	0.082	0.95	0.907	0.058	0.057	0.95
	50	0.00	0.915	0.086	0.092	0.94	0.907	0.058	0.057	0.95
10,000	5	0.09	0.912	0.046	0.050	0.98	0.912	0.045	0.048	0.98
	25	0.00	0.906	0.054	0.055	0.95	0.903	0.039	0.039	0.95
	50	0.00	0.906	0.060	0.062	0.95	0.903	0.039	0.039	0.95
Estimates of $\theta_2 - \theta_1 = 0.9$ when $\theta_1 = 0.1$, $\phi_1 = 0.5$										
1,000	5	0.14	1.040	0.335	0.376	0.97	1.040	0.340	0.358	0.97
	25	0.06	1.130	0.525	1.030	0.93	0.964	0.266	0.252	0.96
	50	0.10	1.200	0.600	1.420	0.91	0.963	0.273	0.261	0.96
5,000	5	0.00	0.918	0.113	0.112	0.95	0.919	0.110	0.107	0.94
	25	0.00	0.939	0.178	0.193	0.93	0.911	0.083	0.083	0.96
	50	0.02	0.956	0.240	0.278	0.91	0.912	0.084	0.083	0.96
10,000	5	0.00	0.907	0.077	0.077	0.94	0.907	0.074	0.074	0.95
	25	0.00	0.917	0.119	0.124	0.94	0.904	0.057	0.057	0.95
	50	0.00	0.924	0.164	0.178	0.92	0.904	0.057	0.057	0.95
Estimates of $\theta_2 - \theta_1 = 0.9$ when $\theta_1 = 0.1$, $\phi_1 = 0.8$										
1,000	5	0.08	1.070	0.542	0.726	0.94	1.050	0.498	0.606	0.95
	25	0.14	1.200	0.874	16.10	0.85	0.951	0.432	0.466	0.92
	50	0.55	1.300	0.956	14.80	0.86	0.945	0.428	0.468	0.91
5,000	5	0.00	0.926	0.194	0.193	0.95	0.927	0.187	0.184	0.95
	25	0.09	1.120	0.665	1.350	0.87	0.917	0.160	0.157	0.95
	50	0.24	1.240	0.828	2.630	0.88	0.910	0.160	0.155	0.95
10,000	5	0.00	0.915	0.133	0.133	0.95	0.915	0.128	0.126	0.95
	25	0.03	1.040	0.489	0.672	0.90	0.913	0.111	0.109	0.95
	50	0.13	1.160	0.687	1.390	0.89	0.911	0.110	0.109	0.95

Table A.4: Comparison of estimates for the sum of two Ornstein-Uhlenbeck processes with autocorrelation function $\rho(t) = \phi_1 \exp(-\theta_1 t) + (1 - \phi_1) \exp(-\theta_2 t)$ for varying sample size n and number of lags k. Based on 10,000 time series.

			OLSE				Optimally weighted LSE			
n	k	P_{ne}	Mean	SSE	SEE	CP_{95}	Mean	SSE	SEE	CP_{95}
Estimates of $\phi_1 = 0.2$ when $\theta_1 = 0.1$, $\theta_2 = 1$										
1,000	5	0.42	0.336	0.160	0.261	0.95	0.329	0.155	0.264	0.95
	25	0.08	0.248	0.127	0.159	0.96	0.231	0.116	0.123	0.94
	50	0.06	0.250	0.128	0.167	0.94	0.231	0.117	0.124	0.93
5,000	5	0.18	0.237	0.083	0.100	0.97	0.238	0.083	0.098	0.97
	25	0.00	0.207	0.053	0.054	0.94	0.205	0.047	0.045	0.94
	50	0.00	0.208	0.057	0.059	0.93	0.206	0.047	0.045	0.94
10,000	5	0.09	0.216	0.060	0.068	0.98	0.216	0.059	0.066	0.98
	25	0.00	0.203	0.037	0.037	0.95	0.202	0.031	0.031	0.95
	50	0.00	0.203	0.040	0.041	0.93	0.202	0.031	0.031	0.94
Estimates of $\phi_1 = 0.5$ when $\theta_1 = 0.1$, $\theta_2 = 1$										
1,000	5	0.14	0.518	0.121	0.153	0.94	0.520	0.120	0.150	0.93
	25	0.06	0.516	0.132	0.156	0.96	0.498	0.098	0.104	0.94
	50	0.10	0.525	0.147	0.184	0.95	0.496	0.099	0.105	0.94
5,000	5	0.00	0.498	0.067	0.068	0.95	0.498	0.066	0.067	0.94
	25	0.00	0.503	0.066	0.067	0.95	0.500	0.044	0.044	0.95
	50	0.02	0.502	0.085	0.085	0.94	0.500	0.044	0.044	0.95
10,000	5	0.00	0.497	0.048	0.048	0.95	0.498	0.047	0.047	0.95
	25	0.00	0.501	0.046	0.047	0.95	0.499	0.031	0.031	0.95
	50	0.00	0.500	0.060	0.061	0.94	0.499	0.031	0.031	0.95
Estimates of $\phi_1 = 0.8$ when $\theta_1 = 0.1$, $\theta_2 = 1$										
1,000	5	0.08	0.778	0.101	0.115	0.94	0.781	0.097	0.110	0.94
	25	0.41	0.744	0.158	0.192	0.97	0.768	0.081	0.088	0.95
	50	0.55	0.738	0.182	0.251	0.95	0.765	0.084	0.091	0.95
5,000	5	0.00	0.794	0.046	0.045	0.95	0.795	0.044	0.043	0.94
	25	0.09	0.786	0.089	0.084	0.97	0.795	0.035	0.036	0.94
	50	0.24	0.771	0.143	0.120	0.96	0.793	0.036	0.036	0.94
10,000	5	0.00	0.797	0.031	0.031	0.95	0.798	0.030	0.030	0.95
	25	0.00	0.795	0.061	0.058	0.95	0.799	0.025	0.025	0.95
	50	0.13	0.784	0.113	0.086	0.97	0.798	0.025	0.025	0.95

Table A.5: Comparison of estimates for the sum of two Ornstein-Uhlenbeck processes with autocorrelation function $\rho(t) = \phi_1 \exp(-\theta_1 t) + (1 - \phi_1) \exp(-\theta_2 t)$ for varying sample size n and number of lags k. Based on 10,000 time series.

Many data are missing among the smallest samples, among moderate samples for $\phi = 0.2$ when five lags are considred, and for $\phi = 0.8$ when twenty five or fifty lags are considered. In practice one can usually find an existing estimate by varying the number of lags. In cases of many missing data the estimators appear biased and less variable than predicted. In the remaining cases the least squares estimators are well behaved and the weighted least squares estimator is the most precise. The variance of the estimators depend on the number of lags. From five to twenty five lags the variances of both estimators tend to decrease. From twenty five to fifty the variance of the ordinary least squares estimator tend to increase whereas the weighted least squares estimator is unaffected. The optimal number of lags depend on the parameters. For $\phi = 0.2$ it lies between twenty five and fifty, for $\phi = 0.5$ around twenty five, and for $\phi = 0.8$ between five and twenty five.

Goodness of fit

Finally, we apply the goodness of fit test of section A.2.4 to the sums of Ornstein-Uhlenbeck processes considered in the previous simulation studies. To be specific, we test whether one or two underlying processes describe the autocorrelation function of the data in a satisfactory manner. Empirical levels and powers of the test are reported in tables A.6 and A.7 for nominal levels of one and five percent, respectively.

$\alpha = 1\%$										
		$m = 1$ $(m = 1)$			$m = 2$ $(m = 2)$			$m = 1$ $(m = 2)$		
		θ_1			ϕ_1			ϕ_1		
n	k	1	0.1	0.01	0.2	0.5	0.8	0.2	0.5	0.8
1,000	5	0.01	0.01	0.02	0.01	0.01	0.01	0.63	0.99	0.99
1,000	25	0.02	0.07	0.08	0.01	0.02	0.04	0.49	0.98	0.98
1,000	50	0.02	0.14	0.16	0.02	0.05	0.10	0.54	0.99	0.99
5,000	5	0.01	0.01	0.01	0.01	0.01	0.01	1.00	1.00	1.00
5,000	25	0.01	0.02	0.03	0.01	0.01	0.02	1.00	1.00	1.00
5,000	50	0.01	0.05	0.06	0.01	0.02	0.06	0.98	1.00	1.00
10,000	5	0.01	0.01	0.01	0.01	0.01	0.01	1.00	1.00	1.00
10,000	25	0.01	0.02	0.02	0.01	0.01	0.02	1.00	1.00	1.00
10,000	50	0.01	0.03	0.04	0.01	0.02	0.05	1.00	1.00	1.00

Table A.6: Empirical levels and powers for testing whether $m = 1$ or $m = 2$ in the autocorrelation function $\rho(t) = \sum_{j=1}^{m} \phi_j \exp(-\theta_j t)$ at level α. The true values are indicated in paranthesis.

When testing whether the autocorrelation function could be the one of a single Ornstein-Uhlenbeck process, we find that the test is very powerful. The overall behavior of the goodness of fit test is good when five lags are considered. For the higher number of lags the level tend to be to high. It seems that the fewer the lags, the better the χ^2-approximation. Surprisingly, the levels are farther off in case of the single Ornstein-Uhlenbeck process. For the sums of two Ornstein-Uhlenbeck processes the level of the test is closer to the formal level than what might be expected. A plausible explanation is the missing parameter

$\alpha = 5\%$										
		$m = 1$ $(m = 1)$			$m = 2$ $(m = 2)$			$m = 1$ $(m = 2)$		
		θ_1			ϕ_1			ϕ_1		
n	k	1	0.1	0.01	0.2	0.5	0.8	0.2	0.5	0.8
1,000	5	0.05	0.05	0.07	0.04	0.05	0.05	0.83	1.00	1.00
1,000	25	0.06	0.13	0.15	0.05	0.07	0.10	0.68	0.99	0.99
1,000	50	0.08	0.19	0.22	0.07	0.12	0.18	0.70	0.99	1.00
5,000	5	0.06	0.05	0.05	0.05	0.05	0.05	1.00	1.00	1.00
5,000	25	0.06	0.07	0.07	0.05	0.06	0.08	1.00	1.00	1.00
5,000	50	0.06	0.11	0.13	0.05	0.08	0.13	1.00	1.00	1.00
10,000	5	0.05	0.05	0.05	0.05	0.05	0.05	1.00	1.00	1.00
10,000	25	0.05	0.06	0.07	0.05	0.05	0.07	1.00	1.00	1.00
10,000	50	0.05	0.08	0.10	0.06	0.06	0.11	1.00	1.00	1.00

Table A.7: Empirical levels and powers for testing whether $m = 1$ or $m = 2$ in the autocorrelation function $\rho(t) = \sum_{j=1}^{m} \phi_j \exp(-\theta_j t)$ at level α. The true values are indicated in paranthesis.

estimates (reported in tables A.3 through A.5), which typically occur when the empirical autocorrelation function diverge the most from the true one.

Appendix: Proofs

Consistency and Uniqueness

Proof of theorem A.2.1:
A1 and **A3** imply that $l_n^W(\theta)$ converge to $l^W(\theta) = \{\boldsymbol{\rho}_k(\theta) - \boldsymbol{\rho}_k\}^T \cdot W_0 \cdot \{\boldsymbol{\rho}_k(\theta) - \boldsymbol{\rho}_k\}$ almost surely, and convergence is uniform on Θ as for all θ

$$|l_n^W(\theta) - l^W(\theta)| \leq k^3 ||W_n - W_0||_{\max}^2 + 5k^3 ||W_0||_{\max} |\mathbf{r}_{n,k} - \boldsymbol{\rho}_k|$$

where $||W||_{\max} = \max_{i,i'=1,\dots,k} |W_{i,i'}|$ defines the matrix norm. Clearly $l^W \geq 0$, and by **A2** and **A3** it holds that $l^W(\theta) = 0$ if and only if $\theta = \theta^\star$. If $B_\epsilon^\star$ denotes the closed ball in Θ centered at $\theta^\star$ with radius ϵ, it suffices to show that minimum of l_n^W on $\Theta^\star$ (Θ in case 2) is attained in $B_\epsilon^\star$ eventually with probability one. Let

$$\delta = \inf\{l^W(\theta) : \theta \in \Theta^\star,\ ||\theta - \theta^\star|| \geq \epsilon\},$$

then $\delta > 0$ by continuity and compactness (by assumption in case 2), and for n so large that $\sup_{\theta \in \Theta} |l_n^W - l^W| < \frac{\delta}{2}$ we find

$$\inf\{l_n^W(\theta) : \theta \in \Theta^\star,\ ||\theta - \theta^\star|| > \epsilon\} > \delta - \tfrac{\delta}{2} = \tfrac{\delta}{2},$$

which together with

$$\inf\{l_n^W(\theta) : \theta \in \Theta^\star,\ ||\theta - \theta^\star|| \leq \epsilon\} \leq l_n^W(\theta^\star) < \tfrac{\delta}{2}$$

imply that l_n^W attains its minimum in $B_\epsilon^\star$. □

Proof of theorem A.2.2:
We consider the derivatives $F_n^W(\theta) = \partial_\theta \partial_{\theta^T} l_n^W(\theta)$ and $F^W(\theta) = \partial_\theta \partial_{\theta^T} l^W(\theta)$. By the above uniqueness will hold for $\hat{\theta}_n$ if F_n^W is positive definite on $B_\epsilon^\star$ for some $\epsilon > 0$. By **A4** this holds true eventually with probability one as $F_n^W \to F^W$ uniformly on compacts, F^W is continuous, and $F^W(\theta^\star)$ is positive definite. □

Asymptotic normality

Proof of theorem A.2.3:
Let $G_n^W(\theta) = \partial_{\theta^T} l_n^W(\theta)$. Apply a Taylor expansion around $\hat{\theta}_n$ to each coordinate of $G_n = G_n^W(\theta^\star)$ the conclusion being that eventually as $G_n(\hat{\theta}_n) = 0$,

$$(G_n)_i = \{F_n^W(\tilde{\theta}_{n,i})\}_i \cdot (\theta^\star - \hat{\theta}_n)$$

for some $\tilde{\theta}_{n,i}$ such that $|\tilde{\theta}_{n,i} - \theta^\star| \leq |\hat{\theta}_n - \theta^\star|$. Let $\widetilde{F}_n$ define the matrix with i'th row equal to the i'th row of $F_n^W(\tilde{\theta}_{n,i})$. Then, as $\tilde{\theta}_{n,i}$ tend to $\theta^\star$ and F_n^W tend to F^W uniformly on compacts, $\widetilde{F}_n \to 2 \cdot V^T W_0 V$ almost surely. In particular, $\widetilde{F}_n$ will be invertible eventually with probability one, implying

$$\sqrt{n} \cdot (\theta^\star - \hat{\theta}_n) = \widetilde{F}_n^{-1} \cdot \sqrt{n} \cdot G_n^T = 2 s_n^{-2} \cdot \widetilde{F}_n^{-1} V^T W_n \cdot \sqrt{n} \cdot \{\boldsymbol{\rho}_k \cdot s_n^2 - \mathbf{c}_{n,k}\}$$

As $2s_n^{-2} \cdot \widetilde{F}_n^{-1} V^T W_n$ converge in probability to $\sigma^{-2} \cdot (V^T W_0 V)^{-1} V^T W_0$, we are done if we can prove that $\sqrt{n} \cdot \{\boldsymbol{\rho}_k \cdot s_n^2 - \mathbf{c}_{n,k}\}$ converge in distribution to a normally distributed random variable with mean zero and variance $\sigma^4 \cdot \Sigma$. From this point on, assume without loss of generality that $\mu = 0$. As the empirical covariances are invariant under translation, we can replace the obeservations Y_i with $Y_i - \mu$. A bit of rearranging yields $\sqrt{n} \cdot (s_n^2 \cdot \rho(t) - c_{n,t})$ equal to

$$\begin{aligned} &\tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n \{\rho(t) \cdot Y_i^2 - Y_i Y_{i+t}\} + \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=n-t+1}^n Y_i Y_{i+t} - \tfrac{t}{\sqrt{n}(n-t)} \textstyle\sum_{i=1}^{n-t} Y_i Y_{i+t} \\ &+ \sqrt{n} \cdot (\tfrac{1}{n-t} \textstyle\sum_{i=1}^{n-t} Y_i) \cdot (\tfrac{1}{n-t} \textstyle\sum_{i=1}^{n-t} Y_{i+t}) - \sqrt{n} \cdot (\rho(t) \tfrac{1}{n} \textstyle\sum_{i=1}^n Y_i)^2. \end{aligned}$$

The four latter terms tend to zero in probability as we now demonstrate. First, note that the process $(\sum_{i=n+t-1}^n Y_i Y_{i+t})_{n \in \mathbb{N}}$ is stationary and thus converges in distribution. It follows that $\frac{1}{\sqrt{n}} \sum_{i=n-t+1}^n Y_i Y_{i+t}$ tends to zero in probability. Secondly, $\frac{t}{\sqrt{n}} \cdot \frac{1}{n-t} \sum_{i=1}^{n-t} Y_i Y_{i+t}$ converge almost surely to zero. Likewise $\frac{1}{n} \sum_{i=1}^n Y_i$ tends to zero almost surely, and $\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$ converges in distribution by Ibragimov's central limit theorem[2]. We conclude that $\sqrt{n} \cdot (\frac{1}{n} \sum_{i=1}^n Y_i)^2$ tends to zero in probability. The same argument finally shows that $\sqrt{n} \cdot (\frac{1}{n-t} \sum_{i=1}^{n-t} Y_i) \cdot (\frac{1}{n-t} \sum_{i=1}^{n-t} Y_{i+t})$ tends to zero in probability.
Define random vectors by $(Z_i)_t = \rho(t) Y_i^2 - Y_i Y_{i+t}$. We have just shown that

$$\sqrt{n} \cdot \{\boldsymbol{\rho}_k \cdot s_n^2 - \mathbf{c}_{n,k}\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i + o_P(1)$$

[2] Nahapetian (1991) theorem 5.1.7

The process $(Z_i)_{i\in\mathbb{N}}$ inherits stationarity as well as mixing properties from $(Y_i)_{i\in\mathbb{N}}$. Hence by the central limit theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i$ converges in distribution to a normal distribution with mean 0 and variance $\widetilde{\Sigma}$ given by

$$\tilde{\sigma}_{s,t} = \lim_{n\to\infty}\frac{1}{n}\operatorname{Cov}(\rho(s)\cdot\textstyle\sum_{i=1}^n Y_i^2 - \sum_{i=1}^n Y_iY_{i+s}\ ,\ \rho(t)\cdot\sum_{i=1}^n Y_i^2 - \sum_{i=1}^n Y_iY_{i+t}).$$

Splitting the covariance into four terms leads us to consider, for $s,t\in\{0,\dots,k\}$,

$$\begin{aligned}&\frac{1}{n}\operatorname{Cov}(\textstyle\sum_{i=1}^n Y_iY_{i+s}, \sum_{i=1}^n Y_iY_{i+t}) = \operatorname{Cov}(Y_1Y_{s+1}, Y_1Y_{t+1})\\ &+ \textstyle\sum_{i=1}^n \frac{n-i}{n}\{\operatorname{Cov}(Y_1Y_{s+1}, Y_{i+1}Y_{i+t+1}) + \operatorname{Cov}(Y_1Y_{t+1}, Y_{i+1}Y_{i+s+1})\}\end{aligned}$$

which by dominated convergence, using **A6** and the covariance inequalities Doukhan (1994) theorem 1.2.3, tends to

$$\operatorname{Cov}(Y_1Y_{s+1}, Y_1Y_{t+1}) + \textstyle\sum_{i=1}^\infty\{\operatorname{Cov}(Y_1Y_{s+1}, Y_{i+1}Y_{i+t+1}) + \operatorname{Cov}(Y_1Y_{t+1}, Y_{i+1}Y_{i+s+1})\}.$$

Under the assumption $\mu = 0$ this quantity is equal to $\sigma^4\cdot S_{s,t}$ as for instance

$$\operatorname{Cov}(Y_1Y_{s+1}, Y_1Y_{t+1}) = E(Y_1Y_{s+1}Y_1Y_{t+1}) - \sigma^4\rho(s)\rho(t) = \sigma^4\{\mu_{1,s+1,1,t+1} - \rho(s)\rho(t)\}$$

All in all we conclude that $\widetilde{\Sigma} = \sigma^4\Sigma$. For a general μ the observations should be replaced by their centralized counterparts. □

Misspecification and goodness of fit

Proof of theorem A.2.6:
Suppose that $\inf_{\theta\in\Theta}|\boldsymbol{\rho}_k(\theta) - \boldsymbol{\rho}_k|^2 > 0$. For all $\theta\in\Theta$ it holds that

$$\{\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}\}^T\cdot W_n\cdot\{\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}\} \geq \lambda_{\min}(W_n)\cdot|\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}|^2$$

where $\lambda_{\min}(W_n)$ denotes the smallest eigenvalue of W_n. From this we conclude that

$$\liminf_{n\to\infty}\inf_{\theta\in\Theta}\{\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}\}^T\cdot W_n\cdot\{\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}\} > 0$$

as **A1** implies that $\inf_{\theta\in\Theta}|\boldsymbol{\rho}_k(\theta) - \mathbf{r}_{n,k}|^2 \to \inf_{\theta\in\Theta}|\boldsymbol{\rho}_k(\theta) - \boldsymbol{\rho}_k|^2$, and **A7** that $\liminf_{n\to\infty}\lambda_{\min}(W_n)$ is greater than zero. Part *1* of the theorem follows readily.
To establish part *2* it suffices to prove that eventually with probability one

$$\{\boldsymbol{\rho}_k - \mathbf{r}_{n,k}\}^T\cdot W_n\cdot\{\boldsymbol{\rho}_k - \mathbf{r}_{n,k}\} \leq \epsilon_n. \tag{A.8}$$

By a diagonalization argument

$$\{\boldsymbol{\rho}_k - \mathbf{r}_{n,k}\}^T\cdot W_n\cdot\{\boldsymbol{\rho}_k - \mathbf{r}_{n,k}\} \leq \lambda_{\max}(W_n)\cdot|\boldsymbol{\rho}_k - \mathbf{r}_{n,k}|^2$$

where $\lambda_{\max}(W_n)$ denotes the largest eigenvalue of W_n. Hence we are led to consider

$$|\boldsymbol{\rho}_k - \mathbf{r}_{n,k}|^2 = s_n^{-4}\cdot\sum_{t=1}^k\{s_n^2\cdot\rho(t) - c_{n,t}\}^2$$

As in the proof of theorem A.2.3 we assume without loss of generality that $\mu = 0$ and rearrange slightly to get

$$\begin{aligned} s_n^2 \cdot \rho(t) - c_{n,t} &= \rho(t) \cdot \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (Y_i^2 - \sigma^2) - \tfrac{1}{n-t} \textstyle\sum_{i=1}^{n-t} (Y_i Y_{i+t} - \rho(t)\sigma^2) \\ & \quad -(\tfrac{1}{n} \textstyle\sum_{i=1}^{n} Y_i)^2 + (\tfrac{1}{n-t} \textstyle\sum_{i=1}^{n-t} Y_i) \cdot (\tfrac{1}{n-t} \textstyle\sum_{i=1}^{n-t} Y_{i+t}). \end{aligned}$$

The law of the iterated logaritm, Nahapetian (1991) theorem 5.4.3, applies to each average. Consequently there exist constants $C_1, \ldots, C_k$, and C such that

$$\begin{aligned} |\boldsymbol{\rho}_k \cdot s_n^2 - \mathbf{c}_{n,k}|^2 &\leq \sum_{t=1}^{k} C_t (\sqrt{\tfrac{\log(\log n)}{n}} + \sqrt{\tfrac{\log\{\log(n-t)\}}{n-t}} + \tfrac{\log(\log n)}{n} + \tfrac{\log\{\log(n-t)\}}{n-t})^2 \\ &\leq C \cdot \tfrac{\log(\log n)}{n} \end{aligned}$$

holds eventually with probability one. It follows that eventually

$$\{\boldsymbol{\rho}_k - \mathbf{r}_{n,k}\}^T \cdot W_n \cdot \{\boldsymbol{\rho}_k - \mathbf{r}_{n,k}\} \leq \lambda_{\max}(W_n) \cdot s_n^{-4} \cdot C \cdot \tfrac{\log(\log n)}{n}$$

By **A7** the sequence $\{\lambda_{\max}(W_n)\}_{n \in \mathbb{N}}$ is almost surely bounded. As $\epsilon_n \cdot n \cdot \{\log(\log n)\}^{-1} \rightarrow \infty$, the proof is hereby completed. □

Examples

Proof of lemma A.3.1:
To prove the first part of the lemma we shall use the following fact:

FACT: *For $m \in \mathbb{N}$, $x_1, \ldots, x_m \in \mathbb{R}$ the $m \times m$-matrix*

$$A(x_1, \ldots, x_m) = \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_m \\ \vdots & & \vdots \\ x_1^{m-1} & \ldots & x_m^{m-1} \end{pmatrix}$$

has non-trivial null space if and only if two or more of $x_1, \ldots, x_m$ are identical.

Clearly, if two or more x's are identical $A(x_1, \ldots, x_m)$ cannot have full rank; Thus, the null space is at least one dimensional. On the other hand, if the null space contains a non-zero vector $(a_1, \ldots, a_m)^T$, then $x_1, \ldots, x_m$ are roots of a polynomial $p(x) = \sum_{i=0}^{m-1} a_i x^i$ of degree at most $m-1$.

To prove *1* we assume without loss of generality that $\widetilde{m} \leq m$ and begin by demonstating that $\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{\widetilde{m}}\} \subseteq \{\lambda_1, \ldots, \lambda_m\}$. By assumption $\boldsymbol{\rho}_{m+\widetilde{m}-1}(m, \lambda, \phi) = \boldsymbol{\rho}_{m+\widetilde{m}-1}(\widetilde{m}, \tilde{\lambda}, \tilde{\phi})$. This we restate as

$$A(\lambda_1, \ldots, \lambda_m, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_{\widetilde{m}}) \cdot (\phi_1, \ldots, \phi_m, -\tilde{\phi}_1, \ldots, -\tilde{\phi}_{\widetilde{m}})^T = 0.$$

The ϕ's are all non-zero. Thus, by the above FACT one of $\lambda_1, \ldots, \lambda_m$ must equal one of $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{\widetilde{m}}$. We assume that $\lambda_1 = \tilde{\lambda}_1$ (changing indices if necessary). Deleting the replicate $\tilde{\lambda}_1$ yields the equation

$$A(\lambda_1, \ldots, \lambda_m, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_{\widetilde{m}}) \cdot (\phi_1 - \tilde{\phi}_1, \phi_2, \ldots, \phi_m, -\tilde{\phi}_2, \ldots, -\tilde{\phi}_{\widetilde{m}})^T = 0.$$

By assumption $\lambda_1 = \tilde{\lambda}_1$ differs from the other λ's and $\tilde{\lambda}$'s. Therefore we conclude that one of $\lambda_2, \ldots, \lambda_m$ equals one of $\tilde{\lambda}_2, \ldots, \tilde{\lambda}_{\widetilde{m}}$, say $\lambda_2 = \tilde{\lambda}_2$ (changing indices again if necessary). Continuing this way leads to the desired conclusion, $\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{\widetilde{m}}\} \subseteq \{\lambda_1, \ldots, \lambda_m\}$. Moreover, if $m > \widetilde{m}$, then

$$A(\lambda_1, \ldots, \lambda_m) \cdot (\phi_1 - \tilde{\phi}_1, \ldots, \phi_{\widetilde{m}} - \tilde{\phi}_{\widetilde{m}}, \phi_{\widetilde{m}+1}, \ldots, \phi_m)^T = 0,$$

which contradics FACT. We conclude that $m = \widetilde{m}$, and $\lambda = \tilde{\lambda}$ follows as $\lambda_1 > \ldots > \lambda_m$ and $\tilde{\lambda}_1 > \ldots > \tilde{\lambda}_m$ by definition. At last, $\phi = \tilde{\phi}$ is deduced from

$$A(\lambda_1, \ldots, \lambda_m) \cdot (\phi_1 - \tilde{\phi}_1, \ldots, \phi_m - \tilde{\phi}_m)^T = 0$$

with a final application of FACT.

To establish part 2 let $\theta \in \Theta_m$ and $V_k(\theta) = \partial_{\theta^T} \boldsymbol{\rho}_k(\theta)$. It suffices to show that $V_{2m-1}(\theta)$ has full rank $2m-1$. To this end we demonstrate that $V_{2m-1}(\theta)^T$ has trivial null space. Assume that $V_{2m-1}(\theta)^T a = 0$ for an $a \in \mathbb{R}^{2m-1}$. That is,

$$0 = \begin{pmatrix} \phi_1 \sum_{t=1}^{2m-1} a_t t \lambda_1^{t-1} \\ \vdots \\ \phi_m \sum_{t=1}^{2m-1} a_t t \lambda_m^{t-1} \\ \sum_{t=1}^{2m-1} a_t \lambda_1^t - \sum_{t=1}^{2m-1} a_t \lambda_m^t \\ \vdots \\ \sum_{t=1}^{2m-1} a_t \lambda_{m-1}^t - \sum_{t=1}^{2m-1} a_t \lambda_m^t \end{pmatrix}$$

implying that the polynomial $p(x) = \sum_{t=1}^{2m-1} a_t x^t - \sum_{t=1}^{2m-1} a_t \lambda_m$ has m distinct double roots, namely $\lambda_1, \ldots, \lambda_m$. This is only possible if $a = 0$.

Finally, to prove part 3 and 4 note that for all k, m it holds that

$$\overline{\{\boldsymbol{\rho}_k(\theta) \colon \theta \in \Theta_m\}} = \{\boldsymbol{\rho}_k(\theta) \colon \theta \in \overline{\Theta_m}\} = \cup_{j=1}^m \{\boldsymbol{\rho}_k(\theta) \colon \theta \in \Theta_j\}.$$

By part 1 if $k \geq 2m-1$, the union is disjoint. Using compactness, continuity (part 2), and identifiablility (part 1) both claims now follow. First, whenever $\theta^\star \in \Theta_m$, $\varepsilon > 0$, and $k \geq 2m-1$ it holds that

$$\begin{aligned} \inf\{|\boldsymbol{\rho}_k(\theta) - \boldsymbol{\rho}_k(\theta^\star)|^2 \colon \theta \in \Theta_m,\ |\theta - \theta^\star| \geq \varepsilon\} &\geq \\ \inf\{|\boldsymbol{\rho}_k(\theta) - \boldsymbol{\rho}_k(\theta^\star)|^2 \colon \theta \in \overline{\Theta_m},\ |\theta - \theta^\star| \geq \varepsilon\} &> 0. \end{aligned}$$

Secondly, for $\theta^\star \in \Theta_{m^\star}$ and $k \geq 2m^\star - 1$.

$$\inf_{m=1,\ldots,m^\star-1} \inf_{\theta \in \Theta_m} |\boldsymbol{\rho}_k(\theta) - \boldsymbol{\rho}_k(\theta^\star)|^2 \geq \inf_{\theta \in \overline{\Theta_{m^\star}}} |\boldsymbol{\rho}_k(\theta) - \boldsymbol{\rho}_k(\theta^\star)|^2 > 0$$

holds true. □

Acknowledgements

I am grateful to Michael Sørensen for many inspiring discussions and useful comments on earlier versions of this paper.

B

The Pearson Diffusions: A Class of Statistically Tractable Diffusion Processes

The Pearson diffusions: A class of statistically tractable diffusion processes

Julie Lyng Forman & Michael Sørensen

Department of Applied Mathematics and Statistics, University of Copenhagen,
Universitetsparken 5, DK-2100 Copenhagen Ø.

Email: julief@math.ku.dk, michael@math.ku.dk

Abstract

The Pearson diffusions is a flexible class of diffusions defined by having linear drift and quadratic squared diffusion coefficient. It is demonstrated that for this class explicit statistical inference is feasible. Explicit optimal martingale estimating functions are found, and the corresponding estimators are shown to be consistent and asymptotically normal. A complete model classification is presented for the ergodic subclass. The class of stationary distributions equals the full Pearson system of distributions. Well-known instances are the Ornstein-Uhlenbeck processes and the square root processes. Also heavy-tailed and skewed marginals are included. Special attention is given to a skewed t-type distribution. Explicit formulae for the conditional moments and the polynomial eigenfunctions are derived. The analytical tractability is inherited by transformed Pearson diffusions, integrated Pearson diffusions, sums of Pearson diffusions, and stochastic volatility models with Pearson volatility process. For the non-Markov models explicit optimal prediction based estimating functions are found and shown to yield consistent and asymptotically normal estimators.

Key words: stochastic differential equation, ergodic diffusion, Pearson system, mixing, martingale estimating function, prediction based estimating function, optimal estimating function, quasi likelihood.

B.1 Introduction

In applications of diffusions the Ornstein-Uhlenbeck process and the square-root process (a.k.a. the CIR process) are often used, more because of their tractability than because they fit the data particularly well. The aim of this paper is to point out that these two diffusion processes belong to a versatile class of tractable diffusion models, which we call the Pearson diffusions. For these diffusion models moments and conditional moments can be calculated explicitly. Moreover, the optimal martingale estimating functions based on eigenfunctions of the generator, introduced by Kessler & Sørensen (1999), can be found explicitly. Thus statistical inference using this method is straightforward. Recently, Sørensen (2007) has proved that optimal martingale estimating functions give estimators that are efficient in a high frequency asymptotics. Parameter estimation is also easy for some diffusion-type models obtained using the Pearson diffusions as building blocks such as transformations and sums of Pearson diffusions, integrated Pearson diffusions, and Pearson stochastic volatility models. Most of these models are non-Markovian processes, for which we derive explicit optimal prediction-based estimating functions, see Sørensen (2000).

We shall use the term Pearson diffusion for any stationary solution of a stochastic differential equation specified by a mean reverting linear drift and a squared diffusion coefficient which is a second order polynomial of the state. The motivation is that when a stationary solution exists, then its invariant density belongs to the Pearson system, Pearson (1895). The class of Pearson diffusions is thus highly flexible and therefore suited for many different applications. Just like the Pearson densities the diffusions can be positive, negative, real valued, or bounded, symmetric or skewed, and heavy- or light-tailed. We give special attention to the Pearson diffusion with type IV marginals (the type IV Pearson distribution is a skewed kind of t-distribution). To our knowledge this process is new to the literature and has a noteworthy potential in, for instance, financial applications.
Most of the Pearson diffusions were derived by Wong (1964) using a different approach and with another aim. In particular, he did not consider the nice statistical properties of the Pearson diffusions on which our paper focuses. Most of the Pearson diffusions are among the diffusion models studied in Bibby, Skovgaard & Sørensen (2005), where no attention was, however, given to statistical inference.
The paper is organized as follows. In Section 2 we give a complete classification of the Pearson diffusions and demonstrate their tractability. We show that all Pearson diffusions have polynomial eigenfunctions that can be found explicitly. It is also demonstrated that estimation is easy for transformations of Pearson diffusions. In Section 3 statistical inference based on martingale estimating functions is investigated, and in Section 4 we explicitly find optimal prediction-based estimating functions for integrated Pearson diffusions, for sums of Pearson diffusions and for stochastic volatility models where the volatility process is a Pearson diffusion or a sum of Pearson diffusions. Also asymptotics for these models are considered.

B.2 The Pearson diffusions

A Pearson diffusion is a stationary solution to a stochastic differential equation of the form

$$dX_t = -\theta(X_t - \mu)dt + \sqrt{2\theta(aX_t^2 + bX_t + c)}dB_t, \qquad \text{(B.1)}$$

where $\theta > 0$, and where a, b and c are such that the square root is well defined when X_t is in the state space. The parameters of (B.1) are referred to as the canonical parameterisation: $\theta > 0$ is a scaling of time that determines how fast the diffusion moves. The parameters μ, a, b, and c determine the state space of the diffusion as well as the shape of the invariant distribution. In particular, μ is the mean of the invariant distribution.

Let us first briefly outline, why the stationary density of the diffusion (B.1) belongs to the Pearson system. The scale and speed densities of the diffusion (B.1) are

$$s(x) = \exp\left(\int_{x_0}^{x} \frac{u-\mu}{au^2+bu+c} du\right) \quad \text{and} \quad m(x) = \frac{1}{2\theta s(x)(ax^2+bx+c)}$$

where x_0 is a fixed point such that $ax_0^2 + bx_0 + c > 0$. Let (l, r) be an interval such that $ax^2 + bx + c > 0$ for all $x \in (l, r)$. A unique ergodic weak solution to (B.1) with values in the interval $(l, r) \ni x_0$ exists if and only if $\int_{x_0}^{r} s(x)dx = \infty$, $\int_{l}^{x_0} s(x)dx = \infty$, and $\int_{l}^{r} m(x)dx < \infty$. Its invariant distribution has density proportional to the speed density, $m(x)$. Since

$$\frac{dm(x)}{dx} = -\frac{(2a+1)x - \mu + b}{ax^2+bx+c} m(x),$$

we see that when a stationary solution to (B.1) exists, the invariant distribution belongs to the Pearson system, which is defined as the class of probability densities obtained by solving a differential equation of this form. If $\int_{x_0}^{r} s(x)dx < \infty$, the boundary l can with positive probability be reached in finite time. In this case a solution for which the invariant distribution has density proportional to the speed density is obtained if the boundary l is made instantaneously reflecting. Similarly for the other boundary, r.

B.2.1 Classification of the stationary solutions

In the following we present a full classification of the ergodic Pearson diffusions. Needless to say, the squared diffusion coefficient must be positive on the state space of the diffusion. We consider six cases according to whether the squared diffusion coefficient is constant, linear, a convex parabola with either zero, one or two roots, or a concave parabola with two roots. The classification problem can be reduced by first noting that the Pearson class of diffusions is closed under translations and scale-transformations. To be specific, if $(X_t)_{t \geq 0}$ is an ergodic Pearson diffusion, then so is $(\tilde{X}_t)_{t \geq 0}$ where $\tilde{X}_t = \gamma X_t + \delta$. The parameters of the stochastic differential equation (B.1) for $(\tilde{X}_t)_{t \geq 0}$ are $\tilde{a} = a$, $\tilde{b} = b\gamma - 2a\delta$, $\tilde{c} = c\gamma^2 - b\gamma\delta + a\delta^2$, $\tilde{\theta} = \theta$, and $\tilde{\mu} = \gamma\mu + \delta$.
Hence, up to translation and transformation of scale the ergodic Pearson diffusions can take the following forms. Note that we consider scale transformations in a general sense where multiplication by -1 is allowed, so that to each case of a diffusion with state space $(0, \infty)$ there corresponds a diffusion with state space $(-\infty, 0)$. Note also that the enumeration of cases does not correspond to the types of the Pearson system.

Case 1: $\sigma^2(x) = 2\theta$.
For all $\mu \in \mathbb{R}$ there exists a unique ergodic solution to (B.1). It is an Ornstein-Uhlenbeck process, and the invariant distribution is the normal distribution with mean μ and variance 1. In the finance literature this model is sometimes referred to as the Vasiček model.

Case 2: $\sigma^2(x) = 2\theta x$.
A unique ergodic solution to (B.1) on the interval $(0, \infty)$ exists if and only if $\mu > 1$. The invariant distribution is the gamma distribution with scale parameter 1 and shape parameter μ. In particular μ is the mean of the invariant distribution. If $0 < \mu \leq 1$, the boundary 0 can with positive probability be reached at a finite time point, but if the boundary is made instantaneously reflecting, we obtain a stationary process for which

the invariant distribution is the gamma distribution with scale parameter 1 and shape parameter μ. The process goes back to Feller (1951), who introduced it as a model of population growth. It is often referred to as the square-root process. In the finance literature it is often refereed to as the CIR-process; Cox, Ingersoll & Ross (1985).

Case 3: $a > 0$ **and** $\sigma^2(x) = 2\theta a(x^2+1)$**.**
The scale and speed densities are given by $s(x) = (x^2+1)^{\frac{1}{2a}} \exp(-\frac{\mu}{a}\tan^{-1} x)$ and $m(x) = (x^2+1)^{-\frac{1}{2a}-1}\exp(\frac{\mu}{a}\tan^{-1}x)$. Hence, for all $a > 0$ and all $\mu \in \mathbb{R}$ a unique ergodic solution to (B.1) exists on the real line. If $\mu = 0$ the invariant distribution is a scaled t-distribution with $\nu = 1 + a^{-1}$ degrees of freedom and scale parameter $\nu^{-\frac{1}{2}}$. If $\mu \neq 0$ the invariant distribution is skew and has tails decaying at the same rate as the t-distribution with $1 + a^{-1}$ degrees of freedom. A fitting name for this distribution is the skew t-distribution. It is also known as Pearson's type IV distribution. In either case the mean is μ and the invariant distribution has moments of order k for $k < 1 + a^{-1}$. The class of diffusions with $\mu \neq 0$ seem to be new. With its skew and heavy tailed marginal distribution it is potentially very useful in e.g. finance. The skew t-distribution with mean zero, ν degrees of freedom, and skewness parameter ρ has (unnormalized) density

$$f(z) \propto \{(z/\sqrt{\nu}+\rho)^2+1\}^{-(\nu+1)/2}\exp\left\{\rho(\nu-1)\tan^{-1}\left(z/\sqrt{\nu}+\rho\right)\right\}, \qquad \text{(B.2)}$$

which is the invariant density of the diffusion $Z_t = \sqrt{\nu}(X_t - \rho)$ with $\nu = 1 + a^{-1}$ and $\rho = \mu$. By the transformation result above, the corresponding stochastic differential equation is

$$dZ_t = -\theta Z_t dt + \sqrt{2\theta(\nu-1)^{-1}\{Z_t^2 + 2\rho\nu^{\frac{1}{2}}Z_t + (1+\rho^2)\nu\}dB_t}. \qquad \text{(B.3)}$$

For $\rho = 0$ the invariant distribution is the t-distribution with ν degrees of freedom. Figure B.1 shows the density for a range of ρ values.

Case 4: $a > 0$ **and** $\sigma^2(x) = 2\theta a x^2$**.**
The scale and speed densities are $s(x) = x^{\frac{1}{a}}\exp(\frac{\mu}{ax})$ and $m(x) = x^{-\frac{1}{a}-2}\exp(-\frac{\mu}{ax})$. The integrability conditions hold if and only if $\mu > 0$. Hence, for all $a > 0$ and all $\mu > 0$ a unique ergodic solution to (B.1) exists on the positive halfline. The invariant distribution is an inverse gamma distribution with shape parameter $1 + \frac{1}{a}$ and scale parameter $\frac{a}{\mu}$. In particular the mean is μ and the invariant distribution has moments of order k for $k < 1 + \frac{1}{a}$.

Case 5: $a > 0$ **and** $\sigma^2(x) = 2\theta a x(x+1)$**.**
The scale and speed densities are $s(x) = (1+x)^{\frac{\mu+1}{a}}x^{-\frac{\mu}{a}}$ and $m(x) = (1+x)^{-\frac{\mu+1}{a}-1}x^{\frac{\mu}{a}-1}$. The integrability conditions hold if and only if $\frac{\mu}{a} \geq 1$. Hence, for all $a > 0$ and all $\mu \geq a$ a unique ergodic solution to (B.1) exists on the positive halfline. The invariant distribution is a scaled F-distribution with $\frac{2\mu}{a}$ and $\frac{2}{a} + 2$ degrees of freedom and scale parameter $\frac{\mu}{1+a}$. In particular the mean is μ and the invariant distribution has moments of order k for $k < 1 + \frac{1}{a}$. If $0 < \mu < 1$, the boundary 0 can with positive probability be reached at a finite time point, but if the boundary is made instantaneously reflecting, a stationary process is obtained for which the invariant distribution is the indicated F-distribution.

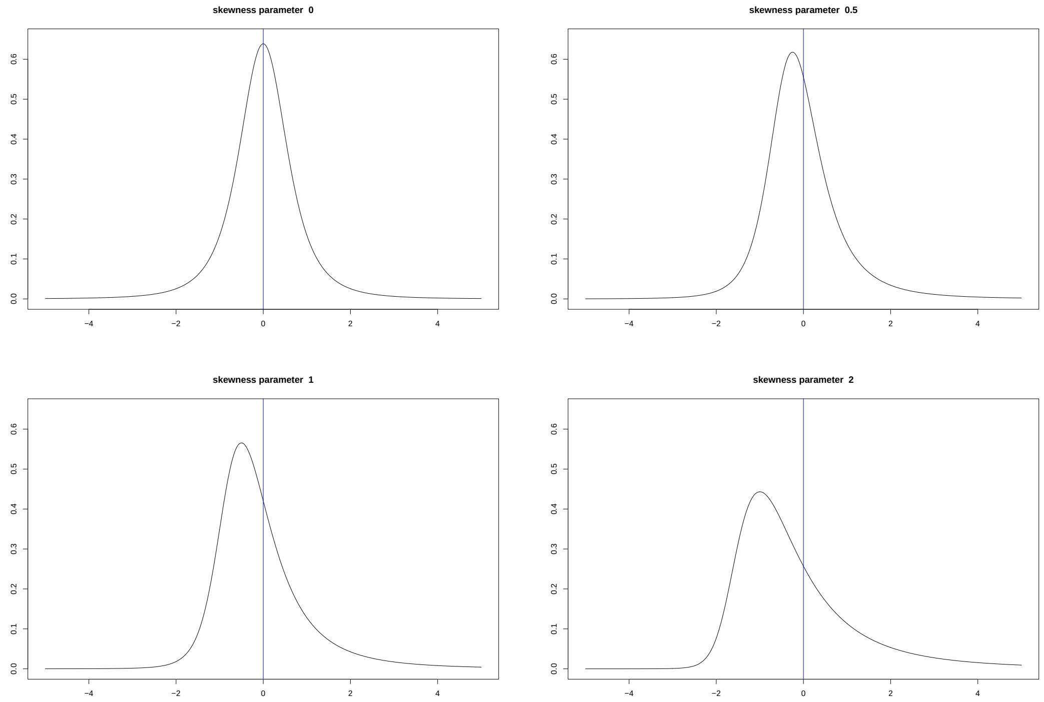

Figure B.1: Densities of skew t-distributions (Pearson type IV distributions) with zero mean for $\rho = 0$, 0.5, 1, and 2 respectively.

Case 6: $a < 0$ **and** $\sigma^2(x) = 2\theta a x(x-1)$**.**
The scale and speed densities are $s(x) = (1-x)^{\frac{1-\mu}{a}} x^{\frac{\mu}{a}}$ and $m(x) = (1-x)^{-\frac{1-\mu}{a}-1} x^{-\frac{\mu}{a}-1}$. The integrability conditions hold if and only if $\frac{\mu}{a} \leq -1$ and $\frac{1-\mu}{a} \leq -1$. Hence, for all $a < 0$ and all $\mu > 0$ such that $\min(\mu, 1-\mu) \geq -a$ a unique ergodic solution to (B.1) exists on the interval $(0,1)$. The invariant distribution is a Beta distribution with shape parameters $\frac{\mu}{-a}$, $\frac{1-\mu}{-a}$. In particular the mean is μ. If $0 < \mu < -a$, the boundary 0 can with positive probability be reached at a finite time point, but if the boundary is made instantaneously reflecting, a stationary process is obtained with the indicated Beta distribution as invariant distribution. Similar remarks apply to the boundary 1 when $0 < 1-\mu < -a$. These diffusions are often referred to as the Jacobi diffusions because the related eigenfunctions are Jacobi polynomials, see below. The model was used (after a position and scale transformation) by De Jong, Drost & Werker (2001) (with $\mu = \frac{1}{2}$) and Larsen & Sørensen (2003) to model the logarithm of exchange rates in a target zone.

B.2.2 Mixing and moments

Common to the stationary solutions of (B.1) is that they are *ergodic and ρ-mixing with exponentially decaying mixing coefficients.* This follows from Genon-Catalot, Jeantheau & Laredo (2000) theorem 2.6 by the fact that the drift is linear, see Hansen, Scheinkman & Touzi (1998), section 5. If the marginal distribution has finite second order moment, the linear drift implies, moreover, that the *autocorrelation function* is given by

$$\rho(t) = \mathrm{Cor}(X_s, X_{s+t}) = e^{-\theta t}$$

see for instance Bibby, Skovgaard & Sørensen (2005). Another important and appealing feature is that explicit expressions of the marginal and conditional moments can be found. We saw in subsection 2.1 that $E(|X_t|^\kappa) < \infty$ if and only if $a < (\kappa - 1)^{-1}$. Thus if $a \leq 0$ all moments exist, while for $a > 0$ only the moments satisfying that $\kappa < a^{-1} + 1$ exist. In particular, the expectation always exists. By Ito's formula

$$dX_t^n = -\theta n X_t^{n-1}(X_t - \mu)dt + \theta n(n-1)X_t^{n-2}(aX_t^2 + bX_t + c)dt + nX_t^{n-1}\sigma(X_t)dB_t, \quad \text{(B.4)}$$

and if $E(X_t^{2n})$ is finite, i.e. if $a < (2n-1)^{-1}$, the integral of the last term is a martingale. Thus, the *moments* of the invariant distribution satisfy

$$E(X_t^n) = a_n^{-1}\{b_n \cdot E(X_t^{n-1}) + c_n \cdot E(X_t^{n-2})\} \quad \text{(B.5)}$$

where $a_n = n\{1-(n-1)a\}\theta$, $b_n = n\{\mu+(n-1)b\}\theta$, and $c_n = n(n-1)c\theta$ for $n = 0, 1, 2, \ldots$. Initial conditions are given by $E(X_t^0) = 1$, and $E(X_t) = \mu$.

Example B.2.1 Equation (B.5) allows us to find the moments of the skewed t-distribution, in spite of the fact that the normalising constant of the density (B.2) is unknown. In particular, for the diffusion (B.3), $E(Z_t^2) = \text{Var}(Z_t) = \frac{(1+\gamma^2)\nu}{\nu-2}$,

$$E(Z_t^3) = \frac{4\gamma(1+\gamma^2)\nu^{\frac{3}{2}}}{(\nu-3)(\nu-2)}, \quad E(Z_t^4) = \frac{24\gamma^2(1+\gamma^2)\nu^2 + 3(\nu-3)(1+\gamma^2)^2\nu^2}{(\nu-4)(\nu-3)(\nu-2)}.$$

Recall that the mean of Z_t is zero. △

The *conditional moments* $q_n(x,t) = E(X_t^n | X_0 = x)$ satisfy the recursive system of first order linear differential equations

$$\frac{d}{dt}q_n(x,t) = -a_n q_n(x,t) + b_n q_{n-1}(x,t) + c_n q_{n-2}(x,t).$$

This follows from (B.4), again under the condition that the $2n$'th moment is finite. Solving for the initial condition $q_n(x,0) = x^n$ yields

$$q_n(x,t) = x^n e^{-a_n t} + b_n I_{n-1}(a_n, x, t) + c_n I_{n-2}(a_n, x, t)$$

where $I_\eta(\alpha, x, t) = \exp(-\alpha t)\int_0^t e^{\alpha s} q_\eta(x,s)ds$. Using once more the recursion, we get

$$I_\eta(\alpha) = \frac{x^\eta\{e^{-a_\eta t} - e^{-\alpha t}\} + b_\eta\{I_{\eta-1}(a_\eta) - I_{\eta-1}(\alpha)\} + c_\eta\{I_{\eta-2}(a_\eta) - I_{\eta-2}(\alpha)\}}{\alpha - a_\eta}.$$

To calculate $I_1(\alpha, x, t)$ we use that $I_0(\alpha, x, t) = \alpha^{-1}\{1 - e^{-\alpha t}\}$ as $q_0(x,t) = 1$ and that $c_1 = 0$. We see that $q_n(x,t)$ is a polynomial of order n in x for any fixed t. A somewhat easier derivation of this result comes by means of the eigenfunctions considered below.

B.2.3 Eigenfunctions

Recall that for a diffusion process

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t$$

the generator is the second order differential operator

$$L = b(x)\frac{d}{dx} + \frac{1}{2}\sigma^2(x)\frac{d^2}{dx^2}.$$

A function h is an eigenfunction if there exist a positive number $\lambda > 0$, an eigenvalue, such that $Lh = -\lambda h$. Under mild regularity conditions, see e.g. Kessler & Sørensen (1999), it follows from Ito's formula that

$$E(h(X_t)|X_0 = x) = e^{-\lambda t}h(x). \tag{B.6}$$

This relationship can be used to construct martingale estimating functions. In case of the Pearson diffusions that have a linear drift and a quadratic squared diffusion coefficient, the generator maps polynomial into polynomials. It is therefore natural to search for eigenfunctions among the polynomials

$$p_n(x) = \sum_{j=0}^{n} p_{n,j}x^j.$$

The polynomial $p_n(x)$ is an eigenfunction if an eigenvalue $\lambda_n > 0$ exist satisfying that $\theta(ax^2 + bx + c)p_n''(x) - \theta(x - \mu)p_n'(x) = -\lambda_n p_n(x)$, i.e.

$$\sum_{j=0}^{n}\{\lambda_n - a_j\}p_{n,j}x^j + \sum_{j=0}^{n-1} b_{j+1}p_{n,j+1}x^j + \sum_{j=0}^{n-2} c_{j+2}p_{n,j+2}x^j = 0.$$

where $a_j = j\{1-(j-1)a\}\theta$, $b_j = j\{\mu + (j-1)b\}\theta$, and $c_j = j(j-1)c\theta$ for $j = 0, 1, 2, \ldots$. Without loss of generality, we assume $p_{n,n} = 1$. Thus, equating the coefficients we find that the eigenvalue is given by $\lambda_n = a_n = n\{1-(n-1)a\}\theta$. If further we define $p_{n,n+1} = 0$, then the coefficients $\{p_{n,j}\}_{j=0,\ldots,n-1}$ solve the linear system

$$(a_j - a_n)p_{n,j} = b_{j+1}p_{n,j+1} + c_{j+2}p_{n,j+2} \tag{B.7}$$

Equation (B.7) is equivalent to a simple recursive formula if $a_n - a_j \neq 0$ for all $j = 0, 1, \ldots, n-1$. Note that $a_n - a_j = 0$ if and only if there exists an integer $n-1 \leq m < 2n-1$ such that $a = m^{-1}$ and $j = m - n + 1$. In particular, $a_n - a_j = 0$ cannot occur if $a < (2n-1)^{-1}$. It is important to notice that λ_n is positive if and only if $a < (n-1)^{-1}$. This is exactly the condition ensuring that $p_n(x)$ is integrable with respect to the invariant distribution. If the stronger condition $a < (2n-1)^{-1}$ is satisfied, the first n eigenfunctions belong to the space of functions that are square integrable with respect to the invariant distribution, and they are orthogonal with respect to the usual inner product in this space. From equation (B.6) the conditional moments can be derived. Sufficient conditions that (B.6) holds is that the drift and diffusion coefficients, b and σ, are of linear growth and that the eigenfunction h is of polynomial growth. These conditions are clearly satisfied here. Thus,

$$E(X_t^n \,|\, X_0 = x) = e^{-a_n t}\sum_{j=0}^{n} p_{n,j}x^j - \sum_{j=0}^{n-1} p_{n,j}E(X_t^j|X_0 = x). \tag{B.8}$$

For any fixed t the conditional expectation is a polynomial of order n in x the coefficients of which are linear combinations of $1, e^{-\lambda_1 t}, \ldots, e^{-\lambda_n t}$. Let $\lambda_0 = 0$ and

$$E(X_t^n \,|\, X_0 = x) = q_n(x,t) = \sum_{k=0}^{n} q_{n,k}(t) x^k = \sum_{k=0}^{n} \sum_{l=0}^{n} q_{n,k,l} \cdot e^{-\lambda_l t} \cdot x^k. \tag{B.9}$$

Initially $q_0(x,t) = 1$. From the above it follows that $q_{n,n}(t) = e^{-a_n t}$ and for $k = 0, \ldots, n-1$,

$$q_{n,k}(t) = p_{n,k} e^{-a_n t} - \sum_{j=k}^{n-1} p_{n,j} q_{j,k}(t). \tag{B.10}$$

In particular, $q_{n,k,n} = p_{n,k}$ and $q_{n,k,l} = -\sum_{j=l}^{n-1} p_{n,j} q_{j,k,l}$ for $l = 0, \ldots, n-1$.
For the diffusions of form (B.1) with $a \leq 0$ there are infinitely many polynomial eigenfunctions. In these cases the eigenfunctions are well-known families of orthogonal polynomials. In case 1, where the marginal distribution is the normal distribution, the eigenfunctions are the Hermite polynomials. In case 2, with gamma marginals, the eigenfunctions are the Laguerre polynomials, and finally in case 6, where the marginals are Beta-distributions, the eigenfunctions are Jacobi polynomials (on the interval $(0,1)$). For these cases all moments of the marginal distribution exists.
In the remaining cases, 3, 4, and 5, $a > 0$ which implies that there is only a finite number of polynomial eigenfunctions. The number is the integer part of $1 + a^{-1}$, which is also the order of the highest finite moment of the marginal distribution. In these cases the marginal distributions are the inverse gamma distributions, the F-distributions, and the skew (and symmetric) t-distributions, respectively.

Example B.2.2 The skew t-diffusion (B.3) has the eigenvalues $\lambda_n = n(\nu - n)(\nu - 1)^{-1}\theta$ for $n < \nu$. The four first eigenfunctions are $p_1(z) = z$,

$$\begin{aligned}
p_2(z) &= z^2 - \frac{4\gamma\nu^{\frac{1}{2}}}{\nu - 3} z - \frac{(1+\gamma^2)\nu}{\nu - 2}, \\
p_3(z) &= z^3 - \frac{12\gamma\nu^{\frac{1}{2}}}{\nu - 5} z^2 + \frac{24\gamma^2\nu + 3(1+\gamma^2)\nu(\nu - 5)}{(\nu - 5)(\nu - 4)} z + \frac{8\gamma(1+\gamma^2)\nu^{\frac{3}{2}}}{(\nu - 5)(\nu - 3)},
\end{aligned}$$

and

$$\begin{aligned}
p_4(z) &= z^4 - \frac{24\gamma\nu^{\frac{1}{2}}}{\nu - 7} z^3 + \frac{144\gamma^2\nu - 6(1+\gamma^2)\nu(\nu - 7)}{(\nu - 7)(\nu - 6)} z^2 \\
&\quad + \frac{8\gamma(1+\gamma^2)\nu^{\frac{3}{2}}(\nu - 7) + 48\gamma(1+\gamma^2)\nu^{\frac{3}{2}}(\nu - 6) - 192\gamma^3\nu^{\frac{3}{2}}}{(\nu - 7)(\nu - 6)(\nu - 5)} z \\
&\quad + \frac{3(1+\gamma^2)^2\nu(\nu - 7) - 72\gamma^2(1+\gamma^2)\nu^2}{(\nu - 7)(\nu - 6)(\nu - 4)},
\end{aligned}$$

provided that $\nu > 4$. Conditional moments are readily obtained from equation (B.8). The most simple cases are $E(Z_t | Z_0 = z) = z e^{-\theta t}$ and

$$E(Z_t^2 | Z_0 = z) = e^{-\frac{2\nu-4}{\nu-1}\theta t} z^2 + \frac{4\gamma\nu^{\frac{1}{2}}}{\nu - 3}\left(e^{-\theta t} - e^{-\frac{2\nu-4}{\nu-1}\theta t}\right) z + \frac{(1+\gamma^2)\nu}{\nu - 2}\left(1 - e^{-\frac{2\nu-4}{\nu-1}\theta t}\right).$$

These formulae are used in Examples B.4.1 and B.4.4 below. △

B.2.4 Transformations

For any diffusion obtained from a solution to (B.1) by a twice differentiable and invertible transformation T, the eigenfunctions of the generator are $p_n\{T^{-1}(x)\}$, which have the same eigenvalues as the original eigenfunctions p_n. Thus the estimation methods discussed below can be used for the much broader class of diffusions obtained by such transformations. Their stochastic differential equations can, of course, be found by Ito's formula. We will just give a couple of examples.

Example B.2.3 For the Jacobi-diffusion (case 6) with $\mu = -a = \frac{1}{2}$, i.e.

$$dX_t = -\theta(X_t - \tfrac{1}{2})dt + \sqrt{\theta X_t(1 - X_t)}dW_t$$

the invariant distribution is the uniform distribution on $(0, 1)$ for any $\theta > 0$. For any strictly increasing and twice differentiable distribution function F we therefore have a class of diffusions given by $Y_t = F^{-1}(X_t)$ or

$$dY_t = -\theta\frac{(F(Y_t) - \frac{1}{2})f(Y_t)^2 + \frac{1}{2}F(Y_t)\{1 - F(Y_t)\}}{f(Y_t)^3}dt + \frac{\theta F(Y_t)\{1 - F(Y_t)\}}{f(Y_t)}dW_t,$$

which has invariant distribution with density $f = F'$. A particular example is the logistic distribution

$$F(x) = \frac{e^x}{1 + e^x} \quad x \in \mathbb{R},$$

for which

$$dY_t = -\theta\left\{\sinh(x) + 8\cosh^4(x/2)\right\}dt + 2\sqrt{\theta}\cosh(x/2)dW_t.$$

If the same transformation $F^{-1}(y) = \log(y/(1 - y))$ is applied to the general Jacoby diffusion (case 6), then we obtain

$$dX_t = -\theta\left\{1 - 2\mu + (1 - \mu)e^x - \mu e^{-1} - 8a\cosh^4(x/2)\right\}dt + 2\sqrt{-a\theta}\cosh(x/2)dW_t,$$

a diffusion for which the invariant distribution is the generalized logistic distribution with density

$$f(x) = \frac{e^{\alpha x}}{(1 + e^x)^{\alpha+\beta}B(\alpha, \beta)}, \quad x \in \mathbb{R},$$

where $\alpha = -(1 - \mu)/a$, $\beta = \mu/a$ and B denotes the Beta-function. This distribution was introduced and studied in Barndorff-Nielsen, Kent & Sørensen (1982). △

Example B.2.4 Let again X be a general Jacobi-diffusion (case 6). If we apply the transformation $T(x) = \sin^{-1}(2x - 1)$ to X_t we obtain the diffusion

$$dY_t = -\rho\frac{\sin(Y_t) - \varphi}{\cos(Y_t)}dt + \sqrt{-a\theta/2}dW_t,$$

where $\rho = \theta(1 + a/4)$ and $\varphi = (2\mu - 1)/(1 + a/4)$. The state space is $(-\pi/2, \pi/2)$. The model was proposed and studied in Kessler & Sørensen (1999) for $\varphi = 0$, where the drift is $-\rho\tan(x)$. The general asymmetric version was proposed in Larsen & Sørensen (2003) as a model for exchange rates in a target zone. △

B.3 Martingale estimating functions

Suppose $\{Y_i\}_{i=0,1,\ldots,n}$ is a sequence of observations from an ergodic Pearson diffusion made at the time points $t_i = i\Delta$ for $i = 0, \ldots, n$. Our goal is to estimate a parameter ψ belonging to the parameter space $\Psi \subset \mathbb{R}^d$. The parameter ψ might be the parameter (θ, μ, a, b, c) of the full class of Pearson diffusions, or it might be a subclass, e.g. a class corresponding to one of the Pearson types. If the diffusion has moments of order N, then the N first eigen-polynomials $p_1(\cdot, \psi), \ldots, p_N(\cdot, \psi)$ are well defined. Thus, we can apply a martingale estimating function of the type introduced by Kessler & Sørensen (1999),

$$G_n(\psi) = \sum_{i=1}^{n} \sum_{j=1}^{N} \alpha_j(Y_{i-1}, \psi)\{p_j(Y_i, \psi) - e^{-\lambda_j(\psi)\Delta} p_j(Y_{i-1}, \psi)\} \tag{B.11}$$

where $\alpha_1, \ldots, \alpha_N$ are weight functions and $\lambda_1(\psi), \ldots, \lambda_N(\psi)$ are the eigenvalues. Written on matrix form the associated estimating equation take the form

$$G_n(\psi) = \sum_{i=1}^{n} \alpha(Y_{i-1}, \psi) h(Y_{i-1}, Y_i, \psi) = 0. \tag{B.12}$$

where α is the $d \times N$ weight matrix and $h_j(x, y, \psi) = p_j(y, \psi) - e^{-\lambda_j(\psi)\Delta} p_j(x, \psi)$. We shall focus on the optimal estimating function in the sense of Godambe & Heyde (1987) where α is chosen to minimize the asymptotic variance of the related estimator. Also the simple estimating function where α is the $d \times d$ identity matrix is briefly considered, although not all parameters can be estimated in this simple way. For other choices of weight functions we refer to the general theory in Bibby, Jacobsen & Sørensen (2004).

It is well known that the transition probabilities of an ergodic diffusion have series expansions in terms of the eigenfunctions, see e.g. Karlin & Taylor (1981). As the expansion mainly depends on the first eigenfunctions the optimally weighted martingale estimating function can be interpreted as an approximation to the score function. In fact the optimal martingale estimating function is the L_2 projection of the score function onto the set of martingale functions given by the various selection of weights as was proved by Kessler (1996), see also Sørensen (1997). The series expansions for some of the Pearson diffusions can be found in Wong (1964).

B.3.1 Moment estimators and the simple estimating function

Very often the most simple estimating equations can be solved explicitly, and although the resulting estimators may not be efficient, they are still useful as input when having to solve more complicated estimating equations numerically or to simplify optimal estimating functions as discussed below. The most simple martingale estimating function of form (B.11) is the one with weight matrix equal to the identity. However, it is of limited use as it can only identify parameters in the invariant distribution. Consider for instance the canonical parameter $\tau = (\theta, \mu, a, b, c)$. The simple estimating function, $G_n(\tau)$, yields no sensible estimate of θ because

$$\frac{G_n(\tau)_j}{n} \to (1 - e^{-j(1-(j-1)a)\theta\Delta}) \cdot E_{\tau_0}(p_j(Y, \tau)),$$

almost surely as $n \to \infty$. The limit equals zero for $(\mu, a, b, c) = (\mu_0, a_0, b_0, c_0)$ regardless of the value of θ. Estimators of the parameters in the marginal distribution might as well be obtained from solving the first four instances of equation (B.5) with empirical moments inserted. Let $M_n(j) = \frac{1}{n}\sum_{i=1}^{n} Y_i^j$, then $\tilde{\mu}_n = M_n(1)$ and

$$\begin{pmatrix} \tilde{a}_n \\ \tilde{b}_n \\ \tilde{c}_n \end{pmatrix} = \begin{pmatrix} M_n(2) & M_n(1) & 1 \\ 2M_n(3) & 2M_n(2) & 2M_n(1) \\ 3M_n(4) & 3M_n(3) & 3M_n(2) \end{pmatrix}^{-1} \cdot \begin{pmatrix} M_n(2) - M_n(1)^2 \\ M_n(3) - M_n(1)M_n(2) \\ M_n(4) - M_n(1)M_n(3) \end{pmatrix}$$

are consistent and asymptotically normal estimators of μ, a, b, and c due to the mixing properties of the Pearson diffusion. In fact, the moment estimators are asymptotically equivalent to the estimators obtained from the simple martingale estimating function with θ (any θ) held fixed. Finally, θ can be estimated by least squares estimation as in Forman (2005), by means of a linear estimating function as in Bibby & Sørensen (1995), or with a modification of Kessler's estimator, Kessler (2000). A simultaneous estimator of τ can also be obtained by replacing the first diagonal element in the identity matrix by Y_{i-1}.

B.3.2 Optimal martingale estimating function

A noteworthy feature of the Pearson diffusions is that the optimal weights in the sense of Godambe & Heyde (1987) are simple and explicit. For the optimal weights the asymptotic variance of the corresponding estimator is minimal. An account of the theory of optimal estimating functions can be found in Heyde (1997).

Assume that the Pearson diffusion is ergodic and has moments of order $2N$. In particular, $a < (2N-1)^{-1}$. Further assume that the mapping $\psi \mapsto \tau = (\theta, \mu, a, b, c)$ is differentiable. Then the optimal weights for the martingale estimating function (B.11) are given by proposition 3.1 of Kessler & Sørensen (1999) as

$$\alpha^\star(x, \psi) = -S(x, \psi)^T \cdot V(x, \psi)^{-1} \tag{B.13}$$

where ^{T} denotes transposition and

$$\begin{aligned} S_{j,k}(x, \psi) &= -E_\psi\{\partial_{\psi_k} p_j(Y_i, \psi) | Y_{i-1} = x\} + \partial_{\psi_k}\{e^{-\lambda_j(\psi)\Delta} p_j(x, \psi)\} \\ V_{j,k}(x, \psi) &= E_\psi\{p_j(Y_i, \psi) p_k(Y_i, \psi) | Y_{i-1} = x\} - e^{-\{\lambda_j(\psi)+\lambda_k(\psi)\}\Delta} p_j(x, \psi) p_k(x, \psi). \end{aligned}$$

Note that the indicated conditions imply that S and V are well defined. The proof that V is invertible is implicitly given as part of the proof of Theorem B.3.1 below. Moreover, the formula defining the optimal weights can be made explicit by means of the recursive formula (B.8) and (B.7) of Section B.2. Note that

$$\begin{aligned} V_{j,k}(x, \psi) &= \sum_{j'=0}^{j}\sum_{k'=0}^{k} p_{j,j'}(\psi) p_{k,k'}(\psi) q_{j'+k'}(x, \Delta, \psi) - e^{-(\lambda_j(\psi)+\lambda_k(\psi))\Delta} p_j(x, \psi) p_k(x, \psi) \\ S_{j,k}(x, \psi) &= p_j(x, \psi) e^{-\lambda_j(\psi)\Delta} \partial_{\psi^T} \lambda_j(\psi) + \sum_{j'=0}^{j} \{q_{j'}(x, \Delta, \psi) - e^{-\lambda_j(\psi)\Delta} x^{j'}\} \partial_{\psi^T} p_{j,j'}(\psi). \end{aligned}$$

where $q_j(x,t,\psi) = E_\psi(X_t^j | X_0 = x)$ is specified by equations (B.8) and (B.9). Hence, the j,k'th element of $V(x,\psi)$ is a polynomial $v_{j,k}(x) = \sum_{l=0}^{j+k} v_{j,k,l} x^l$ with coefficients given by

$$v_{j,k,l} = \sum_{j'=0}^{j} \sum_{k'=0}^{k} p_{j,j'} p_{k,k'} \cdot \left(q_{j'+k',l}(\Delta) - e^{-(\lambda_j+\lambda_k)\Delta} I_{\{j'+k'=l\}}\right),$$

where $I_{\{j'+k'=l\}}$ denotes the indicator function. Similarly, the j,k'th element of $S(x,\psi)$ is the j'th order polynomial $s_{j,k}(x) = \sum_{l=0}^{j} s_{j,k,l} x^l$ the coefficients of which are

$$s_{j,k,l} = e^{-\lambda_j \Delta}(p_{j,l} \partial_{\psi_k} \lambda_j - \partial_{\psi_k} p_{j,l}) + \sum_{j'=0}^{l} \partial_{\psi_k} p_{j,j'} \cdot q_{j',l}(\Delta).$$

It is important to notice that the derivatives $d_{j,l} = \partial_{\psi^T} p_{j,l}$ satisfy the recursion

$$d_{j,l} = \frac{b_{l+1} d_{j,l+1} + c_{l+2} d_{j,l+2} + p_{j,l} \partial_{\psi^T}(a_l - a_j) + p_{j,l+1} \partial_{\psi^T} b_{l+1} + p_{j,l+2} \partial_{\psi^T} c_{l+2}}{a_l - a_j}$$

for $l = j-1, j-2, \ldots, 0$ where initially $d_{j,j} = d_{j,j+1} = 0$.

In practice, it is often a good idea to replace the weight matrix $\alpha^\star(x,\psi)$ by

$$\tilde{\alpha}_n(x) = \alpha^\star(x, \tilde{\psi}_n), \tag{B.14}$$

where $\tilde{\psi}_n$ is a $\sqrt{n}$-consistent estimator of ψ. For instance $\tilde{\psi}_n$ could be a moment estimator as described in Section B.3.1. The resulting estimating equations are much easier to solve numerically because of the simpler dependence on θ and because the weight matrix need only be evaluated once for every observation. Moreover, replacing the weights by estimates does not affect the asymptotic distribution of the estimator so there is no loss of efficiency (see Theorem B.3.1 below).

B.3.3 Asymptotic theory

The optimally weighted martingale estimating function (B.11) provides consistent and asymptotically normal estimators of the parameters of a Pearson diffusion under mild regularity conditions. In what follows ψ_0 denotes the true parameter value.

Theorem B.3.1 *Suppose that the following hold true:*

R0: *The Pearson diffusion is ergodic and has moments of order $2N$ where $N \geq 2$.*

R1: *ψ_0 belongs to the interior of Ψ.*

R2: *The mapping $\psi \mapsto \tau = (\theta, \mu, a, b, c)$ is differentiable and $\partial_\psi \tau(\psi_0)$ has full rank d.*

Then with probability tending to one as $n \to \infty$ there exist a solution $\hat{\psi}_n$ to the estimating equation (B.12) with weights specified by either (B.13) or (B.14) such that $\hat{\psi}_n$ converges to ψ_0 in probability and

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, W(\psi_0)^{-1})$$

where $W(\psi_0) = E_{\psi_0}\{S(Y_i,\psi_0)^T V(Y_i,\psi_0)^{-1} S(Y_i,\psi_0)\}$.

Note: Condition **R0** ensures that the eigenfunctions are well defined and that $h_1, \dots, h_N$ have finite variance so that $G_n(\psi_0)$ is indeed a martingale. In fact, **R0** implies that $G_n(\psi_0)$ is a square integrable martingale. The proof of Theorem B.3.1 is given in the appendix.

Example B.3.1 For the skewed t-diffusion with parameter $\psi = (\theta, \nu, \rho)$ the canonical parameter is

$$(\theta, \mu, a, b, c) = \left(\theta, 0, \frac{1}{\nu - 1}, \frac{2\rho\nu^{\frac{1}{2}}}{\nu - 1}, \frac{(1+\rho^2)\nu}{\nu - 1}\right)$$

and

$$\frac{\partial \tau}{\partial \psi^T} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{(\nu-1)^2} & \frac{\rho}{\nu^{\frac{1}{2}}(\nu-1)} - \frac{2\rho\nu^{\frac{1}{2}}}{(\nu-1)^2} & \frac{1+\rho^2}{(\nu-1)} - \frac{\nu(1+\rho^2)}{(\nu-1)^2} \\ 0 & 0 & 0 & \frac{2\nu^{\frac{1}{2}}}{\nu-1} & \frac{2\nu\rho}{\nu-1} \end{pmatrix}$$

which has full rank three. Hence, consistent and asymptotically normal estimates are obtained by means of the optimally weighted martingale estimating function under the further assumption that $\nu_0 > 2N$. △

B.4 Derived diffusion-type models

The Pearson diffusion processes can be used as building blocks to obtain more general diffusion-type models. In what follows we consider inference for integrated diffusions, sums of diffusions, and stochastic volatility models. These derived processes are not Markovian. Therefore explicit martingale estimating functions are no longer available. In stead we suggest to base the statistical inference on prediction based estimation functions, introduced in Sørensen (2000). We will demonstrate that such estimating functions can be found explicitly for models based on Pearson diffusions. We start by briefly reviewing the method of prediction based estimating functions.

B.4.1 Prediction based estimating functions

Here we focus on estimating functions based on prediction of powers of the observations of the process. Suppose that we have observed the random variables $Y_1, \dots, Y_n$ that form a stationary stochastic process the distribution of which is parametrised by $\Psi \subseteq \mathbb{R}^d$. Assume that $E_\psi(Y_i^{2m}) < \infty$ for all $\psi \in \Psi$ for some $m \in \mathbb{N}$. For each $i = r+1, \dots, n$ and $j = 1, \dots, m$ let the class $\{Z_{jk}^{(i-1)} \mid k = 1, \dots, q_j\}$ be a subset of the random variables $\{Y_{i-\ell}^{\kappa} \mid \ell = 1, \dots, r, \kappa = 0, \dots, j\}$, where $Z_{j1}^{(i-1)}$ is always equal to 1. We wish to predict Y_i^j by means of linear combinations of the $Z_{jk}^{(i-1)}$-s for each of the values of i and j listed above and then to use suitable linear combinations of the prediction errors to estimate the parameter ψ. Let $\mathcal{P}_{i-1,j}$ denote the space of predictors of Y_i^j, i.e. the space of square integrable random variables spanned by $Z_{j1}^{(i-1)}, \dots, Z_{jq_j}^{(i-1)}$. The elements of $\mathcal{P}_{i-1,j}$ are of the form $a^T Z_j^{(i-1)}$, where $a^T = (a_1, \dots, a_{q_j})$ and $Z_j^{(i-1)} = (Z_{j1}^{(i-1)}, \dots, Z_{jq_j}^{(i-1)})^T$ are q_j-dimensional vectors.

We will use estimating functions of the type

$$G_n(\psi) = \sum_{i=r+1}^{n} \sum_{j=1}^{m} \Pi_j^{(i-1)}(\psi) \left[Y_i^j - \hat{\pi}_j^{(i-1)}(\psi) \right] \tag{B.15}$$

where $\Pi_j^{(i-1)}(\psi)$ is a d-dimensional data dependent vector of weights, the coordinates of which belong to $\mathcal{P}_{i-1,j}$, and where $\hat{\pi}_j^{(i-1)}(\psi)$ is the minimum mean square error predictor of Y_i^j in $\mathcal{P}_{i-1,j}$, which is the usual L_2-projection of Y_i^j onto $\mathcal{P}_{i-1,j}$. When ψ is the true parameter value, we define $C_j(\psi)$ as the covariance matrix of $(Z_{j2}^{(r)}, \ldots, Z_{jq_j}^{(r)})^T$ and $b_j(\psi) = (\mathrm{Cov}_\psi(Z_{j2}^{(r)}, Y_{r+1}^j), \ldots, \mathrm{Cov}_\psi(Z_{jq_j}^{(r)}, Y_{r+1}^j))^T$. Then we have

$$\hat{\pi}_j^{(i-1)}(\psi) = \hat{a}_j(\psi)^T Z_j^{(i-1)}$$

where $\hat{a}_j(\psi)^T = (\hat{a}_{j1}(\psi), \hat{a}_{j*}(\psi)^T)$ with $\hat{a}_{j*}(\psi)^T = (\hat{a}_{j2}(\psi), \ldots, \hat{a}_{jq_j}(\psi))$ defined by

$$\hat{a}_{j*}(\psi) = C_j(\psi)^{-1} b_j(\psi) \tag{B.16}$$

and

$$\hat{a}_{j1}(\psi) = E_\psi(Y_1^j) - \sum_{k=2}^{q_j} \hat{a}_{jk}(\psi) E_\psi(Z_{jk}^{(r)}). \tag{B.17}$$

Thus to find $\hat{\pi}_j^{(i-1)}(\psi)$, $j = 1, \ldots, m$, we need to calculate moments of the form

$$E_\psi(Y_1^\kappa Y_k^j), \quad 0 \le \kappa \le j \le m, \quad k = 1, \ldots, r. \tag{B.18}$$

Once we have calculated these moments, the vector of coefficients $\hat{a}_j$ can easily be found by means of the m-dimensional Durbin-Levinson algorithm applied to $\{(Y_i, Y_i^2, \ldots, Y_i^m)\}_{i \in \mathbb{N}}$, see Brockwell & Davis (1991). The non-Markovian diffusion-type models considered in this paper inherit the exponential ρ-mixing property from the Pearson diffusions. Therefore constants $K > 0$ and $\lambda > 0$ exist such that $\left|\mathrm{Cov}_\psi(Y_1^j, Y_k^j)\right| \le K e^{-\lambda k}$ (λ is typically the smallest speed of mean reversion of the involved Pearson diffusions). Therefore r will usually not need to be chosen particularly large. If Y_i^j is restricted to have mean zero, we need not include a constant in the space of predictors, i.e. we need only the space spanned by $Z_{j2}^{(i-1)}, \ldots, Z_{jq_j}^{(i-1)}$.

In many situations $m = 2$ with $Z_{jk}^{(i-1)} = Y_{i-k}$, $k = 1, \ldots, r, j = 1, 2$ and $Z_{2k}^{(i-1)} = Y_{i+r-k}^2$, $k = r+1, \ldots, 2r$, will be a reasonable choice. In this case the minimum mean square error predictor of Y_i can be found using the Durbin-Levinson algorithm for real processes, while the predictor of Y_i^2 can be found by applying the two-dimensional Durbin-Levinson algorithm to the process (Y_i, Y_i^2).

Including predictors in the form of lagged terms $Y_{i-k}Y_{i-k-l}$ for a number of lags l's might also be of relevance. These terms enter into the least squares estimator of Forman (2005), which produces good estimates for a sum of Ornstein-Uhlenbeck processes.

The choice of the weights $\Pi_j^{(i-1)}(\psi)$ in (B.15) for which the asymptotic variance of the estimators is minimized is the Godambe optimal prediction-based estimating function, that was derived in Sørensen (2000). An account of the theory of optimal estimating functions can be found in Heyde (1997). The optimal estimating function of the type (B.15) can be written in the form

$$G_n^*(\psi) = A_n^*(\psi) \sum_{i=r+1}^{n} H^{(i)}(\psi), \tag{B.19}$$

where

$$H^{(i)}(\psi) = Z^{(i-1)} \left(F(Y_i) - \hat{\pi}^{(i-1)}(\psi) \right), \tag{B.20}$$

with $F(x) = (x, x^2, \ldots, x^m)^T$, $\hat{\pi}^{(i-1)}(\psi) = (\hat{\pi}_1^{(i-1)}(\psi), \ldots \hat{\pi}_m^{(i-1)}(\psi))^T$ and

$$Z^{(i-1)} = \begin{pmatrix} Z_1^{(i-1)} & 0_{q_1} & \cdots & 0_{q_1} \\ & & & \\ 0_{q_2} & Z_2^{(i-1)} & \cdots & 0_{q_2} \\ \vdots & \vdots & & \vdots \\ 0_{q_m} & 0_{q_m} & \cdots & Z_m^{(i-1)} \end{pmatrix}. \tag{B.21}$$

Here 0_{q_j} denotes the q_j-dimensional zero-vector. Finally,

$$A_n^*(\psi) = \partial_\psi \hat{a}(\psi)^T \bar{C}(\psi) \bar{M}_n(\psi)^{-1}, \tag{B.22}$$

with

$$\bar{M}_n(\psi) = E_\psi\Big(H^{(r+1)}(\psi) H^{(r+1)}(\psi)^T\Big) + \tag{B.23}$$

$$\sum_{k=1}^{n-r-1} \frac{(n-r-k)}{(n-r)} \Big[E_\psi\Big(H^{(r+1)}(\psi) H^{(r+1+k)}(\psi)^T\Big) + E_\psi\Big(H^{(r+1+k)}(\psi) H^{(r+1)}(\psi)^T\Big) \Big],$$

$$\bar{C}(\psi) = E_\psi \left(Z^{(i-1)} (Z^{(i-1)})^T \right), \tag{B.24}$$

and

$$\hat{a}(\psi)^T = \left(\hat{a}_1(\psi)^T, \ldots, \hat{a}_m(\psi)^T \right), \tag{B.25}$$

where $\hat{a}_j(\psi)$ is given by (B.16) and (B.17). A necessary condition that the moments in (B.23) exist is that $E_\psi(Y_i^{4m}) < \infty$ for all $\psi \in \Psi$. For (B.19) to be optimal we need that the matrix $\partial_\psi \hat{a}(\psi)^T$ has full rank. The matrix $\bar{M}_n(\psi)$ is always invertible.

Because the processes considered below inherit the exponential ρ-mixing property from the Pearson diffusions, there exist constants $K > 0$ and $\lambda > 0$ such that the absolute values of all entries in the expectation matrices in the sum in (B.23) are dominated by $K\mathrm{e}^{-\lambda(k-r-1)}$ when $k > r$. Therefore, the sum in (B.23) can in practice often be truncated so that fewer moments need to be calculated. The matrix $\bar{M}_n(\psi)$ can also be approximated by a truncated version of the limiting matrix

$$\bar{M}(\psi) = E_\psi\Big(H^{(r+1)}(\psi) H^{(r+1)}(\psi)^T\Big) + \tag{B.26}$$

$$\sum_{k=1}^{\infty} \Big[E_\psi\Big(H^{(r+1)}(\psi) H^{(r+1+k)}(\psi)^T\Big) + E_\psi\Big(H^{(r+1+k)}(\psi) H^{(r+1)}(\psi)^T\Big) \Big],$$

obtained for $n \to \infty$. In practice, it is usually also a good idea to replace $A_n^*(\psi)$ by $A_n^*(\bar{\psi}_n)$, where $\bar{\psi}_n$ is a $\sqrt{n}$-consistent estimator of ψ (and similarly for approximations to $A_n^*(\psi)$).

This has the advantages that (B.23) or (B.26) need only be calculated once and that a simpler estimating equation is obtained, while the asymptotic variance of the estimator is unchanged. The estimator $\bar{\psi}_n$ can, for instance, be obtained from an estimating function similar to (B.19), where $A_n^*(\psi)$ has been replaced by a suitable simple matrix independent of ψ, but such that the estimating equation has a solution. Usually it is enough to use the first d coordinates of $H^{(i)}(\psi)$, where d is the dimension of the parameter. In order to calculate (B.23) or (B.26), we need mixed moments of the form

$$E_\psi[Y_1^{k_1} Y_{t_1}^{k_2} Y_{t_2}^{k_3} Y_{t_3}^{k_4}], \quad 1 \leq t_1 \leq t_2 \leq t_3 \quad k_1 + k_2 + k_3 + k_4 \leq 4m \tag{B.27}$$

where k_i, $i = 1, \ldots, 4$ are non-negative integers. In the following subsections we demonstrate that in three diffusion-type models derived from Pearson diffusions, explicit expressions can be found for the necessary moments, (B.18) and (B.27). Thus the optimal prediction-based estimating functions are explicit.

B.4.2 Integrated Pearson diffusions

Let X be a stationary Pearson diffusion, i.e. a solution to (B.1). Suppose that the diffusion cannot be observed directly, but that the data are

$$Y_i = \frac{1}{\Delta} \int_{(i-1)\Delta}^{i\Delta} X_s \, ds, \qquad i = 1, \ldots, n \tag{B.28}$$

for some fixed Δ. Such observations can be obtained if the process X is observed after passage through an electronic filter. Another example is provided by ice-core records. The isotope ratio $^{18}O/^{16}O$ in the ice, measured as an average in pieces of ice, each piece representing a time interval with time increasing as a function of the depth, is a proxy for paleo-temperatures. The variation of the paleo-temperature can be modelled by a stochastic differential equation, and it is natural to model the ice-core data as an integrated diffusion process, see Ditlevsen, Ditlevsen & Andersen (2002). Estimation based on this type of data was considered by Gloter (2001), Ditlevsen & Sørensen (2004) and Gloter (2006). Since X is stationary, the random variables Y_i, $i = 1, \ldots, n$ form a stationary process with the same mixing properties as X, i.e. it is exponentially mixing. However, the observed process is not Markovian, so martingale estimating functions are not available in a tractable form, but explicit prediction-based estimating functions can be found.

Suppose that $4m$'th moment of X_t is finite. The moments (B.18) and (B.27) can be calculated by

$$E\left[Y_1^{k_1} Y_{t_1}^{k_2} Y_{t_2}^{k_3} Y_{t_3}^{k_4}\right] = \frac{\int_A E[X_{v_1} \cdots X_{v_{k_1}} X_{u_1} \cdots X_{u_{k_2}} X_{s_1} \cdots X_{s_{k_3}} X_{r_1} \cdots X_{r_{k_4}}] \, d\mathbf{t}}{\Delta^{k_1+k_2+k_3+k_4}}$$

where $1 \leq t_1 \leq t_2 \leq t_3$, $A = [0\,, \Delta]^{k_1} \times [(t_1 - 1)\Delta\,, t_1\Delta]^{k_2} \times [(t_2 - 1)\Delta\,, t_2\Delta]^{k_3} \times [(t_3 - 1)\Delta\,, t_3\Delta]^{k_4}$, and $d\mathbf{t} = dr_{k_4} \cdots dr_1 \, ds_{k_3} \cdots ds_1 \, du_{k_2} \cdots du_1 \, dv_{k_1} \cdots dv_1$. The domain of integration can be reduced considerably by symmetry arguments, but here the point is that we need to calculate moments of the type $E(X_{t_1}^{\kappa_1} \cdots X_{t_k}^{\kappa_k})$, where $t_1 < \cdots < t_k$. Since $E(X_t^n \mid X_0 = x)$ is a polynomial in x given by (B.9), it follows that we can find the needed moments iteratively

$$E(X_{t_1}^{\kappa_1} \cdots X_{t_k}^{\kappa_k}) = \sum_{j=1}^{\kappa_k} q_{\kappa_k, j}(t_k - t_{k-1}) E(X_{t_1}^{\kappa_1} \cdots X_{t_{k-1}}^{\kappa_{k-1}+j}),$$

where $q_{\kappa_k,j}(t_k - t_{k-1})$ is given by (B.10). The coefficient depends on time through an exponential function, so $E(X_{t_1}^{\kappa_1} \cdots X_{t_k}^{\kappa_k})$ depends on $t_1, \ldots, t_k$ through sums and products of exponential functions. Therefore the integral above can be explicitly calculated.

Example B.4.1 *Integrated skew t-diffusion.* We will now calculate an optimal estimating function for the integrated skew t-diffusion (B.3). To simplify the exposition we consider the simple case where $m = 2$, $Z_{1,1}^{(i-1)} = Y_{i-1}$, $Z_{2,1}^{(i-1)} = 1$, and $Z_{2,2}^{(i-1)} = Y_{i-1}^2$ (i.e. $q_1 = r = 1, q_2 = 2$). The estimating equations take the form

$$G_n(\theta, \rho, \nu) = \sum_{i=2}^{n} \begin{bmatrix} Y_{i-1}Y_i - \beta_1 Y_{i-1}^2 \\ Y_i^2 - \sigma^2(1-\beta_2) - \beta_2 Y_{i-1}^2 \\ Y_{i-1}^2 Y_i^2 - \sigma^2(1-\beta_2)Y_{i-1}^2 - \beta_2 Y_{i-1}^4 \end{bmatrix} = 0, \tag{B.29}$$

with $\sigma^2 = \mathrm{Var}(Y_{i-1})$ and $\beta_j = \mathrm{Cov}(Y_{i-1}^j, Y_i^j) \cdot \mathrm{Var}(Y_{i-1}^j)^{-1}$ for $j = 1, 2$. In particular,

$$\sigma^2 = \frac{2\nu(1+\rho^2)}{\nu - 2} \cdot \left\{ \frac{1}{\theta\Delta} - \frac{1 - e^{-\theta\Delta}}{(\theta\Delta)^2} \right\}, \quad \beta_1 = \frac{(1 - e^{-\theta\Delta})^2}{2(\theta\Delta - 1 + e^{-\theta\Delta})}.$$

In order to get an explicit expression of β_2 let $f_n(x) = x^{-n}(1 - e^{-x})$, then

$$\mathrm{Cov}(Y_{i-1}^2, Y_i^2) = 4\gamma_1(\lambda\Delta - \theta\Delta)^{-2}\{f_1(\lambda\Delta) - f_1(\theta\Delta)\}^2 + 4\gamma_2(\theta\Delta)^{-2}\{1 - (1+\theta\Delta)f_1(\theta\Delta)\}^2$$

where $\lambda = \frac{2\theta(\nu-2)}{\nu-1}$, $\gamma_1 = \frac{(3\nu^3 - 10\nu^2 - 4\nu)\rho^2\sigma^2}{(\nu-4)(\nu-3)^2} + \frac{3\nu\sigma^2}{\nu-4} - \sigma^4$, and $\gamma_2 = \frac{16\nu\rho^2\sigma^2}{(\nu-3)^2}$. Likewise,

$$\begin{aligned} \mathrm{Var}(Y_{i-1}^2) &= 24\gamma_1[(\lambda\Delta - \theta\Delta)^{-2}\{f_2(\theta\Delta) - f_2(\lambda\Delta)\} + (\theta\Delta)^{-2}\{(\lambda\Delta)^{-1} + (\lambda\Delta - \theta\Delta)^{-1}\}] \\ &- 24\gamma_1(\theta\Delta)^{-1}(\theta\Delta - \lambda\Delta)^{-1}(2 + \theta\Delta)f_2(\theta\Delta) \\ &+ 12\gamma_2(\theta\Delta)^{-2}[1 + 6(\theta\Delta)^{-1} - \{(\theta\Delta)^2 + 4\theta\Delta + 6\}f_2(\theta\Delta)] \\ &+ 12\sigma^4(\theta\Delta)^{-2}\{1 - 6(\theta\Delta)^{-1} + (2\theta\Delta + 6)f_2(\theta\Delta)\} - 4\sigma^4(\theta\Delta)^{-2}\{1 - f_1(\theta\Delta)\}^2. \end{aligned}$$

Solving equation (B.29) for β_1, β_2, and σ^2 we get

$$\hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}Y_i}{\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}^2}, \quad \hat{\beta}_2 = \frac{\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}^2 Y_i^2 - (\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}^2)(\frac{1}{n-1}\sum_{i=2}^n Y_i^2)}{\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}^4 - (\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}^2)^2}$$

and

$$\hat{\sigma}^2 = \frac{1}{1-\hat{\beta}_2}\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}^2 + \frac{\hat{\beta}_2}{1-\hat{\beta}_2}\frac{1}{n-1}\sum_{i=2}^n Y_i^2.$$

Hence, if $0 < \hat{\beta}_1 < 1$, which happens eventually with probability one, the expression of β_1 yields a unique estimate $\hat{\theta} > 0$ satisfying

$$2\hat{\beta}_1\{\hat{\theta}\Delta - (1 - e^{-\hat{\theta}\Delta})\} - (1 - e^{-\hat{\theta}\Delta})^2 = 0.$$

The remaining equations are solved by substituting $\hat{\rho}(\nu)^2 = \frac{\nu-2}{2\nu}\{(\hat{\theta}\Delta)^{-1} - f_2(\hat{\theta}\Delta)\}^{-1} - 1$ into the equation $\beta_2(\hat{\theta}, \hat{\rho}(\nu)^2, \nu) = \hat{\beta}_2$, which has to be solved numerically. To estimate the sign of ρ, note that for instance $E(Y_1^3) = \frac{24\sqrt{\nu}\rho\sigma^2}{\nu-3} \cdot (\theta\Delta)^{-2}\{2 - (2 + \theta\Delta)f_1(\theta\Delta)\}$ has the same sign as ρ. △

B.4.3 Sums of diffusions

The simple exponentially decreasing autocorrelation function of the Pearson diffusions is too simple in some applications, but we can obtain a much richer autocorrelation structure by considering sums of Pearson diffusions:

$$\begin{aligned} Y_t &= X_{1,t} + \ldots + X_{M,t} && \text{(B.30)} \\ dX_{i,t} &= -\theta_i(X_{i,t} - \mu_i) + \sigma_i(X_{i,t})dB_{i,t}, \quad i = 1, \ldots, M, && \text{(B.31)} \end{aligned}$$

where $\theta_1, \ldots, \theta_M > 0$ and $B_1, \ldots, B_M$ are independent Brownian motions. The diffusion coefficients $\sigma_1, \ldots, \sigma_M$ are of the form of a Pearson diffusion (B.1). Suppose all $X_{i,t}$ have finite second moment. Then the autocorrelation function of Y is

$$\rho(t) = \phi_1 \exp(-\theta_1 t) + \ldots + \phi_M \exp(-\theta_M t) \tag{B.32}$$

with

$$\phi_i = \frac{\mathrm{Var}(X_{i,t})}{\mathrm{Var}(X_{1,t}) + \cdots + \mathrm{Var}(X_{M,t})}.$$

Thus $\phi_1 + \ldots + \phi_M = 1$. The expectation of Y_t is $\mu_1 + \cdots + \mu_M$. Sums of diffusions with a pre-specified marginal distribution of Y were considered by Bibby & Sørensen (2003), Bibby, Skovgaard & Sørensen (2005) and Forman (2005). Here we specify instead the distributions of the $X_{i,t}$'s, which implies that the models are simpler to handle. Sums of Ornstein-Uhlenbeck processes driven by Lévy processes were introduced and studied in Barndorff-Nielsen, Jensen & Sørensen (1998). An autocorrelation function of the form (B.32) fits turbulence data well, see Barndorff-Nielsen, Jensen & Sørensen (1990) and Bibby, Skovgaard & Sørensen (2005).

Example B.4.2 *Sum of Ornstein-Uhlenbeck processes.* If $\sigma_i^2(x) = 2\theta_i c_i$, the stationary distribution of Y_t is a normal distribution with mean $\mu_1 + \cdots + \mu_M$ and variance $c_1^2 + \cdots + c_M^2$. $\triangle$

Example B.4.3 *Sum of CIR processes.* If $\sigma_i^2(x) = 2\theta_i bx$ and $\mu_i = \alpha_i b$, then the stationary distribution of Y_t is a Gamma-distribution with shape parameter $\alpha_1 + \cdots + \alpha_M$ and scale parameter b. The weights in the autocorrelation function are $\phi_i = \alpha_i/(\alpha_1 + \cdots + \alpha_M)$. $\triangle$

In the other cases of Pearson diffusions, the class of marginal distributions is not closed under convolution, so the stationary distribution of Y_t is not in the Pearson class and is, in fact, not any of the standard distributions. It has recently been proven that the sum of two t-distributions with odd degrees of freedom is a finite mixture (over degrees of freedom) of scaled t-distributions, see Berg & Vignat (2006). In the case of the Jacobi-diffusions it might be preferable to consider Y_t/M to obtain again a process with state space $(0, 1)$.

A sum of diffusions is not a Markov process, so also for this type of model we use prediction-based estimating functions rather than martingale estimating functions. Suppose that the process Y has been observed at the time points $t_i = \Delta i$, $i = 1, \ldots, n$. The necessary moments of the form (B.18) and (B.27) can, provided they exist, be obtained from the mixed moments of the Pearson diffusions because by the multinomial formula we find, for instance

$$E(Y_{t_1}^{\kappa} Y_{t_2}^{\nu}) = \sum\sum \binom{\kappa}{\kappa_1, \ldots, \kappa_M}\binom{\nu}{\nu_1, \ldots, \nu_M} E(X_{1,t_1}^{\kappa_1} X_{1,t_2}^{\nu_1}) \ldots E(X_{M,t_1}^{\kappa_M} X_{M,t_2}^{\nu_M})$$

where

$$\binom{\kappa}{\kappa_1, \ldots, \kappa_M} = \frac{\kappa!}{\kappa_1! \cdots \kappa_M!}$$

is the multinomial coefficient, and where the first summation is over $0 \leq \kappa_1, \ldots, \kappa_M$ such that $\kappa_1 + \ldots \kappa_M = \kappa$ and the second summation is the same just for the ν's. The higher order mixed moments of the form (B.27) can be found by a similar formula with four sums and four multinomial coefficients. Such formulae may appear daunting, but are easy to programme. Mixed moments of the form $E(X_{t_1}^{\kappa_1} \cdots X_{t_k}^{\kappa_k})$ can be calculated iteratively as explained in Subsection B.4.2.

Example B.4.4 *Sum of two skew t-diffusions.* If, for i=1,2, $\sigma_i^2(x) = 2\theta_i(\nu_i - 1)^{-1}\{x^2 + 2\rho\sqrt{\nu_i}x + (1+\rho^2)\nu\}$, the stationary distribution of $X_{i,t}$ is a skew t-diffusion. The distribution of Y_t is a convolution of skew t-diffusions,

$$\mathrm{Var}(Y) = (1+\rho^2)\left(\frac{\nu_1}{\nu_1 - 2} + \frac{\nu_2}{\nu_2 - 2}\right),$$

and $\phi_i = \nu_i(\nu_i - 2)^{-1}/\{\nu_1(\nu_1-2)^{-1} + \nu_2(\nu_2-2)^{-1}\}$. To simplify the exposition we assume that the correlation parameters θ_1, θ_2, ϕ_1, and ϕ_2 are known or have been estimated in advance (the least squares estimator of Forman (2005) applies and so does the predictions based estimating function with $m = 1$, $Z_{1,k}^{(i-1)} = Y_{i-k}$, $k = 1, \ldots, r$). We will find the optimal estimating function in the simple case where predictions of Y_i^2 are made based on $Z_{1,1}^{(i-1)} = 1$ and $Z_{1,2}^{(i-1)} = Y_{i-1}$. The estimating equations take the form

$$G_n(\theta, \rho, \nu) = \sum_{i=2}^{n} \left[\begin{array}{c} Y_i^2 - \sigma^2 - \beta_{21}Y_{i-1} \\ Y_{i-1}Y_i^2 - \sigma^2 Y_{i-1} - \beta_{21}Y_{i-1}^2 \end{array}\right] = 0, \tag{B.33}$$

with $\sigma^2 = \mathrm{Var}(Y_{i-1})$ and $\beta_{21} = \mathrm{Cov}(Y_{i-1}, Y_i^2) \cdot \mathrm{Var}(Y_{i-1})^{-1}$. To be specific

$$\sigma^2 = (1+\rho^2)\left\{\frac{\nu_1}{\nu_1 - 2} + \frac{\nu_2}{\nu_2 - 2}\right\}, \quad \beta_{21} = 4\rho\left\{\frac{\sqrt{\nu_1}}{\nu_1 - 3}\phi_1 e^{-\theta_1\Delta} + \frac{\sqrt{\nu_2}}{\nu_2 - 3}\phi_2 e^{-\theta_2\Delta}\right\}.$$

Solving equation (B.33) for β_{21} and σ^2 we get

$$\begin{aligned} \hat{\beta}_{21} &= \frac{\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}Y_i^2 - (\frac{1}{n-1}\sum_{i=2}^n Y_{i-1})(\frac{1}{n-1}\sum_{i=2}^n Y_i^2)}{\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}^2 - (\frac{1}{n-1}\sum_{i=2}^n Y_{i-1})^2}, \\ \hat{\sigma}^2 &= \tfrac{1}{n-1}\textstyle\sum_{i=2}^n Y_i^2 + \hat{\beta}_{21}\tfrac{1}{n-1}\sum_{i=2}^n Y_{i-1}. \end{aligned}$$

In order to estimate ρ we restate β_{21} as

$$\beta_{21} = \sqrt{32(1+\rho^2)} \cdot \rho \cdot \left\{\frac{\sqrt{9(1+\rho^2) - \phi_1\sigma^2}}{3(1+\rho^2) - \phi_1\sigma^2}\phi_1 e^{-\theta_1\Delta} + \frac{\sqrt{9(1+\rho^2) - \phi_2\sigma^2}}{3(1+\rho^2) - \phi_2\sigma^2}\phi_2 e^{-\theta_2\Delta}\right\}$$

and insert $\hat{\sigma}^2$ for σ^2. Thus, we get a one-dimensional estimating equation, $\beta_{21}(\theta, \phi, \hat{\sigma}^2, \rho) = \hat{\beta}_{21}$, which can be solved numerically. Finally by inverting $\phi_i = \frac{1+\rho^2}{\sigma^2}\frac{\nu_i}{\nu_i - 2}$ we find the estimates $\hat{\nu_i} = \frac{2\phi_i\hat{\sigma}^2}{\phi_i\hat{\sigma}^2 - (1+\hat{\rho}^2)}$, $i = 1, 2$. △

A more complex model is obtained if the observations are integrals of Y in analogy with the previous subsection:

$$Z_i = \frac{1}{\Delta}\int_{(i-1)\Delta}^{i\Delta} Y_s \, ds = \frac{1}{\Delta}\left(\int_{(i-1)\Delta}^{i\Delta} X_{1,t} ds + \cdots + \int_{(i-1)\Delta}^{i\Delta} X_{M,t} ds\right), \tag{B.34}$$

$i = 1, \ldots, n$. Also here the moments of form (B.18) and (B.27) can be found explicitly because each of the observations Z_i is a sum of processes of the type considered in the previous subsection. To calculate $E(Z_1^{k_1} Z_{t_1}^{k_2} Z_{t_2}^{k_3} Z_{t_3}^{k_4})$, first apply the multinomial formula to express this quantity in terms of moments of the form $E(Y_{j,1}^{\ell_1} Y_{j,t_1}^{\ell_2} Y_{j,t_2}^{\ell_3} Y_{j,t_3}^{\ell_4})$, where

$$Y_{j,i} = \frac{1}{\Delta}\int_{(i-1)\Delta}^{i\Delta} X_{j,s} \, ds.$$

Now proceed as in Subsection B.4.2.

B.4.4 Stochastic volatility models

A stochastic volatility model is a generalization of the Black-Scholes model for the logarithm of an asset price $dX_t = (\kappa + \beta\sigma^2)dt + \sigma dW_t$, that takes into account the empirical finding that the volatility σ^2 varies randomly over time:

$$dX_t = (\kappa + \beta v_t)dt + \sqrt{v_t} dW_t. \tag{B.35}$$

Here the volatility v_t is a stochastic process that cannot be observed directly. If the data are observations at the time points Δi, $i = 0, 1, 2, \ldots, n$, then the returns $Y_i = X_{i\Delta} - X_{(i-1)\Delta}$ can be written in the form

$$Y_i = \kappa\Delta + \beta S_i + \sqrt{S_i} A_i, \tag{B.36}$$

where

$$S_i = \int_{(i-1)\Delta}^{i\Delta} v_t dt, \tag{B.37}$$

and where the A_i's are independent, standard normal distributed random variables. Here we consider the case where v is a sum of independent Pearson diffusions with state-space $(0, \infty)$ (the cases 2, 4 and 5). Barndorff-Nielsen & Shephard (2001a) demonstrated that an autocorrelation function of the type (B.32) fits empirical autocorrelation functions of volatility well, while an autocorrelation function like that of a single Pearson diffusion is too simple to obtain a good fit. We assume that v and W are independent, so that the sequences $\{A_i\}$ and $\{S_i\}$ are independent.

By the multinomial formula we find that

$$E\left(Y_1^{k_1} Y_{t_1}^{k_2} Y_{t_2}^{k_3} Y_{t_3}^{k_4}\right) =$$

$$\sum K_{k_{11},\ldots,k_{43}} E(S_1^{k_{12}+k_{13}/2} S_{t_1}^{k_{22}+k_{23}/2} S_{t_2}^{k_{32}+k_{33}/2} S_{t_3}^{k_{42}+k_{43}/2}) E(A_1^{k_{13}}) E(A_{t_1}^{k_{23}}) E(A_{t_2}^{k_{33}}) E(A_{t_3}^{k_{43}}),$$

where the sum is over all non-negative integers k_{ij}, $i = 1, 2, 3, 4$, $j = 1, 2, 3$ such that $k_{i1} + k_{i2} + k_{i3} = k_i$ $(i = 1, 2, 3, 4)$, and where

$$K_{k_{11},\ldots,k_{43}} = \binom{k_1}{k_{11}, k_{12}, k_{13}}\binom{k_2}{k_{21}, k_{22}, k_{23}}\binom{k_3}{k_{31}, k_{32}, k_{33}}\binom{k_4}{k_{41}, k_{42}, k_{43}}(\kappa\Delta)^{k_{\cdot 1}}\beta^{k_{\cdot 2}}$$

with $k_{\cdot j} = k_{1j} + k_{2j} + k_{3j} + k_{4j}$. The moments $E(A_i^{k_{i3}})$ are the well-known moments of the standard normal distribution. When k_{i3} is odd, these moments are zero. Thus we only need to calculate the mixed moments of the form $E(S_1^{\ell_1} S_{t_1}^{\ell_2} S_{t_2}^{\ell_3} S_{t_3}^{\ell_4})$, where $\ell_1, \ldots, \ell_4$ are integers. However, when the volatility process is a sum of independent Pearson diffusions, S_i of the same form as Z_i in (B.34) (apart from $1/\Delta$), so we can proceed as in the previous section. Thus also for the stochastic volatility models defined in terms of Pearson diffusions we can explicitly find the optimal estimating function based on prediction of powers of returns.

B.4.5 Asymptotics

In this subsection we will briefly discuss the asymptotic distribution of the estimators obtained from prediction based estimating functions for the model types discussed above. We assume that the estimating function has the form

$$G_n(\psi) = A(\psi) \sum_{i=r+1}^{n} H^{(i)}(\psi) \tag{B.38}$$

with $H^{(i)}(\psi)$ given by (B.20), and that it is based on predicting powers up to m of the observations. The observations are either Y_i given by (B.28), (B.30) or (B.36) or Z_i given by (B.34). We denote the true value of ψ by ψ_0.

Theorem B.4.1 *Assume that the underlying Pearson diffusions have finite $(4m + \epsilon)$'th moment ($a < (4m - 1 + \epsilon)^{-1}$) for some $\epsilon > 0$. Suppose, moreover, that $A(\psi)$ is twice continuously differentiable, and that the matrices $A(\psi)$, $A(\psi)\bar{M}(\psi)A(\psi)^T$ and $\partial_{\psi^T}\hat{a}$ have full rank, d, where $\hat{a}(\psi)$ is given by (B.25) and $\bar{M}(\psi)$ by (B.26). Then with probability tending to one as $n \to \infty$ there exists a solution $\hat{\psi}_n$ to the estimating equation $G_n(\psi) = 0$ such that $\hat{\psi}_n$ converges to ψ_0 in probability and*

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, W^{-1}(\psi_0)V(\psi_0)(W^{-1}(\psi_0))^T\right),$$

where $W(\psi_0) = A(\psi_0)\bar{C}(\psi_0)\partial_{\psi^T}\hat{a}(\psi_0)$ and $V(\psi_0) = A(\psi_0)\bar{M}(\psi_0)A(\psi_0)^T$ with $\bar{C}(\psi)$ given by (B.24). For the optimal matrix $A^*(\psi) = \partial_\psi \hat{a}(\psi)^T \bar{C}(\psi)\bar{M}(\psi)^{-1}$, the asymptotic covariance matrix of $\hat{\psi}_n$ simplifies to

$$\left[\partial_\psi \hat{a}(\psi_0)^T \bar{C}(\psi_0)\bar{M}(\psi_0)^{-1}\bar{C}(\psi_0)\partial_{\psi^T}\hat{a}(\psi_0)\right]^{-1}.$$

Proof: The result follows from Theorem 6.2 in Sørensen (2000). We just need to check the conditions of that theorem. First, we note that the observations are exponentially α-mixing. In the cases of integrated Pearson diffusions and sums of Pearson diffusions, this follows immediately from the exponential α-mixing of the Pearson diffusions. That the sequence of observations of a stochastic volatility model with exponentially α-mixing volatility process is exponentially α-mixing, was proven in Sørensen (2000) in the case $\kappa = \beta = 0$. The proof holds in the more general case too, see also the more general result in Genon-Catalot, Jeantheau & Laredo (2000). Secondly, we need that $H^{(i)}(\psi)$, given by (B.20), has finite $(2+\delta)$'th moment for some $\delta > 0$. In the case of an integrated Pearson diffusion this follows from Jensen's inequality and Fubini's theorem:

$$E_\psi \left(\left| \int_0^\Delta X_s ds \right|^{2m(2+\delta)} \right) \leq \int_0^\Delta E_\psi \left(|X_s|^{2m(2+\delta)} \right) ds = \Delta E_\psi \left(|X_0|^{2m(2+\delta)} \right) < \infty.$$

In the case of a sum of Pearson diffusions, it follows from Minkowski's inequality that the $(2+\delta)$'th moment of $H^{(i)}(\psi)$ is finite. For Pearson stochastic volatility models, Minkowski's inequality shows that it is sufficient that the integrated volatility process has finite $(4m+\epsilon)$'th moment, and integrated Pearson diffusion were considered above. Finally, it follows from (B.5) that the (finite) moments of a Pearson diffusion are twice continuously differentiable, so that $\hat{a}$ is twice continuously differentiable, cf. (B.16) and (B.17). Now all conditions of Theorem 6.2 in Sørensen (2000) have been shown to hold. □

Appendix: Proofs and general asymptotics

Theorem B.3.1 can be established using standard asymptotic techniques. Regularity conditions to ensure the existence of a consistent and asymptotically normal sequence of solutions to a general martingale estimating equation of form (B.12) can be found in Sørensen (1999). In case estimates are inserted for the parameter in the weights, as in (B.14), we need somewhat stronger conditions. The following result is taken from Jacod & Sørensen (2007).

Theorem B.4.2 *Suppose that $\{Y_i\}_{i\in\mathbb{N}_0}$ is a stationary ergodic process with state-space D and that*

$$G_n(\psi) = \sum_{i=1}^n \alpha(Y_{i-1},\psi) h(Y_{i-1}, Y_i, \psi)$$

is a martingale estimating function such that the following holds.

A1: *The true parameter ψ_0 belongs to the interior of Ψ.*

A2: *For all ψ in a neighborhood of ψ_0 each of the variables $\alpha(Y_{i-1},\psi_0)h(Y_{i-1},Y_i,\psi)$ is P_{ψ_0}-integrable and $\alpha(Y_{i-1},\psi_0)h(Y_i,Y_{i-1},\psi_0)$ is square integrable.*

A3: *The mappings $\psi \mapsto \alpha(x,\psi)$ and $\psi \mapsto h(x,y,\psi)$ are continuously differentiable in a neighborhood of ψ_0 for all $x, y \in D$.*

A4: *For all ψ, ψ' in a neighborhood of ψ_0 each of the entries of $\partial_{\psi_k}\alpha(Y_{i-1},\psi)h(Y_{i-1},Y_i,\psi_0)$, $\alpha(Y_{i-1},\psi_0)\partial_{\psi_k}h(Y_{i-1},Y_i,\psi)$, and $\partial_{\psi_k}\alpha(Y_{i-1},\psi)\partial_{\psi_{k'}}h(Y_{i-1},Y_i,\psi')$ $(k,k' = 1,\ldots,d)$ is dominated by a P_{ψ_0}-integrable function.*

A5: *The $d \times d$ matrix $W(\psi_0) = E_{\psi_0}\left\{\partial_{\psi^T}\left[\alpha(Y_{i-1},\psi)h(Y_{i-1},Y_i,\psi)\right]\right\}$ is invertible.*

Then with probability tending to one as $n \to \infty$ the estimating equation $G_n(\psi) = 0$ has a solution, $\hat{\psi}_n$, satisfying that $\hat{\psi}_n \to \psi_0$ in probability and

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, W(\psi_0)^{-1}V(\psi_0)W(\psi_0)^{-1})$$

where $V(\psi_0) = E_{\psi_0}\{\alpha(Y_{i-1},\psi_0)h(Y_{i-1},Y_i,\psi_0)h(Y_{i-1},Y_i,\psi_0)^T\alpha(Y_{i-1},\psi_0)^T\}$. The same result holds for the estimating function

$$\tilde{G}_n(\psi) = \sum_{i=1}^{n} \alpha(Y_{i-1},\tilde{\psi}_n)h(Y_{i-1},Y_i,\psi),$$

where $\tilde{\psi}_n$ is a $\sqrt{n}$-consistent estimator of ψ.

Proof of Theorem B.3.1:
Preliminarily we demonstrate that $V(x,\psi)$ is positive definite for all (x,ψ) and that the smallest eigenvalue is bounded away from zero uniformly in (x,ψ) when ψ belongs to a compact subset $\Psi_0 \subset \Psi$. Clearly $V(x,\psi)$ is positive semidefinite for all (x,ψ). Moreover, for $z \in \mathbb{R}^N$ it holds that $z^TV(x,\psi)z = 0$ if and only if

$$\sum_{j=1}^{N} z_j\{p_j(y,\psi) - e^{-\lambda_j(\psi)\Delta}p_j(x,\psi)\} = 0$$

for almost every y with respect to the conditional distribution of Y_i given $Y_{i-1} = x$ under ψ. However, the above is a polynomial in y and thus cannot equal zero almost surely unless the order is zero. As $p_j(y,\psi)$ is a j'th order polynomial with leading coefficient $p_{j,j} = 1$ we deduce that that $z^TV(x,\psi)z = 0$ if and only if $z = 0$. Hence, $V(x,\psi)$ is positive definite. By continuity the smallest eigenvalue

$$\varepsilon_1\{V(x,\psi)\} = \inf\{z^TV(x,\psi)z \;:\; |z| = 1\}$$

is bounded away from zero on compact subsets of $\Psi \times \mathcal{X}$ where $\mathcal{X}$ is the state space. To make the bound valid for all $x \in \mathcal{X}$ we need only check that it holds as $|x| \to \infty$. To this end note that $z^TV(x,\psi)z$ is a non-zero polynomial in x of order at most $2N$ the coefficient of which are given as continuous functions of z and ψ. If a sequence (x_n, z_n, ψ_n) were to exist such that $z_n^TV(x_n,\psi_n)z_n \to 0$, $|x_n| \to \infty$, $|z_n| = 1$, and $\{\psi_n\} \subset \Psi_0$, then we would find an accumulation point (z_0,ψ_0) such that $z_0^TV(x_n,\psi_0)z_0 \to 0$ although $z_0^TV(x,\psi_0)z_0$ defines a non-zero polynomial in x. By contradiction we conclude that $\inf\{\varepsilon_1\{V(x,\psi)\} \;:\; x \in \mathcal{X}, \psi \in \Psi_0\} > 0$.

As to the regularity conditions, **A1** holds true by assumption, and **A3** follows from **R2** as $\alpha^\star(x,\cdot)$ and $h(y,x,\cdot)$ are continuously differentiable with respect to the canonical parameter.
In order to check the integrability condition **A2** let Ψ_0 be a compact neighbourhood of ψ_0 and denote by $||B|| = \max_{j,k}|B_{j,k}|$ the max-norm of a matrix. A diagonalization argument shows that

$$||V(x,\psi)^{-1}|| \leq \frac{N^2}{C_1(\Psi_0)}$$

for all x and all $\psi \in \Psi_0$ where $C_1(\Psi_0)$ is the lower bound on the smallest eigenvalue of $V(x,\psi)$ on $\mathcal{X} \times \Psi_0$. Thus, by continuity of the coefficients a constant $C_2(\Psi_0)$ exist such that

$$\begin{aligned} |\alpha^\star(Y_{i-1},\psi_0)h(Y_i,Y_{i-1},\psi)| &\leq d \cdot N \cdot ||S(Y_{i-1},\psi_0)|| \cdot ||V(Y_{i-1},\psi_0)^{-1}|| \cdot |h(Y_i,Y_{i-1},\psi)| \\ &\leq C_2(\Psi_0)(1 + Y_{i-1}^{2N} + Y_i^{2N}) \end{aligned}$$

for all $\psi \in \Psi_0$. The latter is integrable by **R0**. Further we note that

$$\begin{aligned} E_{\psi_0}\{\alpha^\star(Y_{i-1},\psi_0)h(Y_i,Y_{i-1},\psi_0)h(Y_i,Y_{i-1},\psi_0)^T\alpha^\star(Y_{i-1},\psi_0)^T\} \\ = E_{\psi_0}\{S(Y_{i-1},\psi_0)^T V(Y_{i-1},\psi_0)^{-1} S(Y_{i-1},\psi_0)\}, \end{aligned}$$

which by **R0** is finite because

$$||S(x,\psi_0)^T V(x,\psi_0)^{-1} S(x,\psi_0)|| \leq N^2 \cdot ||S(x,\psi_0)||^2 \cdot ||V(x,\psi_0)^{-1}|| \leq C_3(\psi_0)(1 + x^{2N})$$

for some constant $C_3(\psi_0)$.
Similar bounds can be established for the derivatives of **A4**.
Finally, let us check that **A5** holds true. Clearly, $W(\psi_0)$ is negative semidefinite as

$$W(\psi_0) = -E_{\psi_0}\{S(Y_{i-1},\psi_0)^T V(Y_{i-1},\psi_0)^{-1} S(Y_{i-1},\psi_0)\}.$$

Let $z \in \mathbb{R}^d$ be such that $z^T W(\psi_0) z = 0$. The task is to demonstrate that $z = 0$. As $V(x,\psi_0)$ is positive definite for all x the assumption is that $S(x,\psi_0)z = 0$ for almost every x. We assume without loss of generality that $\psi = \tau = (\theta,\mu,a,b,c)$ is the canonical parameter. The general case follows readily as

$$S(x,\psi_0) = S(x,\tau_0) \cdot \partial_{\psi^T}\tau(\psi_0)$$

where by **R3** $\partial_{\psi^T}\tau(\psi_0)$ has full rank d. Hence, the assumption is

$$E_{\tau_0}(\partial_{\tau^T}\{p_j(Y_i,\tau_0) - e^{-\lambda_j(\tau_0)\Delta}p_j(Y_{i-1},\tau_0)\} \cdot z | Y_{i-1} = x) = 0$$

for $j = 1,\ldots,N$ and almost every x. The first equation reads

$$z_1(x - \mu_0)\Delta e^{-\theta_0\Delta} + z_2(e^{-\theta_0\Delta} - 1) \overset{\text{a.e.x}}{=} 0$$

which only holds true if $z_1 = z_2 = 0$. As $N \geq 2$ at least one more equation is available, namely

$$z_3 S_{2,3}(x,\tau_0) + z_4 S_{2,4}(x,\tau_0) + z_5 S_{2,5}(x,\tau_0) = 0$$

where

$$\begin{aligned} S_{2,3}(x,\tau_0) &= -2\theta\Delta e^{-2(1-a)\theta\Delta}p_2(x,\tau_0) + \frac{4(\mu+b)}{(2a-1)^2}(e^{-2(1-a)\theta\Delta} - e^{-\theta\Delta})x \\ &\quad -\frac{4\mu(\mu+b)}{(2a-1)^2}(1 - e^{-\theta\Delta}) + \left\{\frac{\mu(\mu+b)(4a-3)}{(2a-1)^2(a-1)^2} + \frac{c}{(a-1)^2}\right\}(e^{-2(1-a)\theta\Delta} - 1), \\ S_{2,4}(x,\tau_0) &= \frac{2}{2a-1}(e^{-\theta\Delta} - e^{-2(1-a)\theta\Delta})x + \frac{2\mu}{2a-1}(1 - e^{-\theta\Delta}) + \frac{\mu(1 - e^{-2(1-a)\theta\Delta})}{(2a-1)(a-1)}, \\ S_{2,5}(x,\tau_0) &= \frac{1}{a-1}(1 - e^{-2(1-a)\theta\Delta}) \end{aligned}$$

from which we deduce that $z_3 = z_4 = z_5$. □

C

Goodness of Fit Based on Downsampling with Applications to Diffusion Type Models

Goodness of Fit based on Downsampling with Applications to Continuous-Time Models

Julie Lyng Forman[1], Bo Markussen[2], Helle Sørensen[2]

[1]Department of Mathematical Sciences, University of Copenhagen,
Universitetsparken 5, DK-2100 Copenhagen Ø.

[2] Department of Natural Sciences, University of Copenhagen (LIFE),
Thorvaldsensvej 40, DK-1871 Frederiksberg C

Email: `julief@math.ku.dk`, `bomar@life.ku.dk`, `helle@dina.kvl.dk`

Abstract

In this paper we develop a goodness of fit test based on comparison of distributions for different sampling frequencies. More specifically the test compares parameter estimates for downsamples of the data. We prove asymptotic results and apply the test to various diffusion models. In particular we develop a test for a linear drift hypothesis. Simulations indicate that the finite sample properties are satisfactory and that the test indeed is able to detect certain deviations from the hypothesis.

Key words: continuous time model, goodness of fit, generalized estimating equation, generalized method of moments, linear drift hypothesis, martingale estimating function.

C.1 Introduction

Continuous-time models based on diffusions have a wide range of applications. In biology, chemistry and physics the models are used to represent phenomena that evolve continuously and randomly in time. In finance diffusions and stochastic volatility models are used to model various price processes. The analysis of these models however is complicated since the functional form of the likelihood is rarely explicitly known. Through the last decade the estimating problem has received much attention, see for instance Bibby, Jacobsen & Sørensen (2004), Gallant & Tauchen (2004), Aït-Sahalia, Hansen & Scheinkman (2003), and Sørensen (2004) for reviews. On the other hand the literature on goodness of fit testing is limited. Nevertheless goodness of fit is a matter of importance; in case a model is misspecified the related estimators may be inconsistent and the conclusions of the statistical analysis may be invalid.

In this paper we develop a goodness of fit test applicable for diffusion-type models as well as other continuous-time models. The basic idea is to check if the distributions for different sampling frequencies are consistent with another. More specifically we compare parameter estimates computed from the original sample to those obtained from downsampled data. If the model is true, then one would expect the estimates of the parameters to be alike, whereas if the model is not true, then one would expect estimates to differ for parameters related to the misspecification of the model.

Aït-Sahalia (1996b) was perhaps the first to consider goodness of fit for stationary diffusion models. He proposed to compare a kernel density estimator of the invariant density to the density implied by the parametric model. The test extends naturally to other stationary process models with a parameterized marginal density such as for instance the summed diffusion models of Bibby, Skovgaard & Sørensen (2005). However, the test aims solely at the marginal distribution of the data and, hence, is not suited for detecting misspecification in the dependence structure of the model. Moreover, numerical studies have shown that the test has a poor finite sample performance in case of high persistence in the data, see Pritsker (1998) and Chapman & Pearson (2000).

Fan & Zhang (2003) suggest applying the generalized likelihood ratio test of Fan, Zhang & Zhang (2001) to perform goodness of fit in plain diffusion models. It is not clear how the generalized likelihood ratio test can be extended to other diffusion-type processes. The test is based on parametric and nonparametric estimates of the drift and diffusion coefficient derived from a discretization scheme. Hence, in a low frequency asymptotics the estimators are inconsistent. Whereas the bias is negligible in the examples of Fan & Zhang (2003) this need not always be the case, see the discussion following Fan (2005) for an example. We suspect that the discretization bias may in some cases have a damaging effect to the test.

Recently Hong & Li (2005) launched a test based on the uniform residuals, that is, the observations transformed with the conditional distribution function given the past observations. The uniform residuals is a highly useful diagnostics, and the idea is to test if the residuals are independent and uniformly distributed. However, it may be computationally demanding to compute the residuals when the conditional distributions are not explicitly known and, in particular, for non-Markovian models. The test of Hong & Li (2005) is omnibus in a large class of univariate stationary processes and due to the approximate independence of the residuals it has a good small sample behavior. An application to interest rate data suggests that the test is almost too powerful in the sense that it firmly rejects all of the preceding models. A more adequate test for multivariate data based on non-parametric estimation estimation of the conditional characteristic function is discussed by Chen & Hong (2005).

An overall concern with omnibus tests such as the ones proposed by Hong & Li (2005) and Chen & Hong (2005) is that it may be hard to tell what kind of deviation the test detects. In both papers additional diagnostics are proposed to gauche possible sources of misspecification. Still the conditional transformations blur the relation to the original data and it is an open question whether or not the reported discrepancy seriously affect the model application.

In practice we are often content applying a model which may in some regards be misspecified as long as the estimators are robust. In this paper we propose a test which compares estimates based on various sampling frequencies in order to detect misspecification in the dependence structure of the model. The test is thus likely to detect any misspecification leading the estimates to vary systematically with the sampling frequency. On the other hand, if the estimator is robust against a certain alternative, the model most likely passes the test.

In practice the sampling frequency can be varied by downsampling the data, picking out for instance every second, third or fourth datum.

The proposed test is within the framework of general estimating functions. The theory of estimating functions covers virtually any estimating scheme but we shall be mostly interested in fairly simple estimating equations. This makes the test quite simple from a computational point of view as it only requires evaluation of quantities which are already used for estimation and inference.

The test is closely related to the overidentifying restrictions test of the generalized method of moments, see Hansen (1982), as downsampling can be viewed as a generic way of constructing excess moment conditions. The idea of recycling the moment conditions for varying sampling frequencies is particularly useful in models where simple explicit moment conditions are hard to come by. The test resembles the tests of Hausman (1978) and Newey (1985) as it compares different estimators for the same parameter(s).

In the following we apply the test to various diffusion models, in particular we use it to test the adequacy of linear drift diffusions. The basic idea, however, of comparing the distributions for downsamples of the original data is certainly more generally applicable.

The structure of the paper is as follows: In Section C.2 we review some important properties of estimating functions, define the test and discuss some of its properties. In Section C.3 we apply the test to diffusion models, in particular we propose a test checking for linear drift. The diffusion applications are illustrated by simulation studies in Section C.4, and finally conclusions are drawn in Section C.5.

C.2 Inference from downsampled estimating functions

In what follows we briefly review the statistical inference of continuous time models based on a general estimating function. Throughout the chapter we assume that $\{X_t\}_{t\geq 0}$ is a stationary stochastic process which we observe at discrete time-points $t_i = i\Delta$. That is, our observations are $Y_i = X_{i\Delta}$, $i = 1, \dots, n$ where Δ^{-1} is the sampling frequency. For the unknown distribution of $\{X_t\}_{t\geq 0}$ we assume a (semi) parametric model parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$. By convention we treat all vectors as rows, for instance θ is a $1 \times d$ matrix. The true parameter is denoted by θ_0. In order to check that the model is correctly specified we suggest comparing the estimates based on varying sampling frequencies $\Delta^{-1}, (2\Delta)^{-1}$, $(3\Delta)^{-1}$, etc.

C.2.1 General estimating functions

Ideally we would base inference on the likelihood function. However, many continuous time models such as diffusion type models do not admit an explicit likelihood function. Thus, we consider instead a general estimating function, i.e. a function of the parameter and the data

$$F(\theta) = F(Y_1, \dots, Y_n, \theta) \in \mathbb{R}^d.$$

An estimate is obtained by solving the estimating equation $F(\theta) = 0$. The prime example of an estimating function is the score function from which the maximum likelihood estimator is obtained. Some examples of simple explicit estimating functions are given in Section C.3 below: see also Bibby, Jacobsen & Sørensen (2004) for a recent review. In general any estimator obtained from minimizing or maximizing a differential criterion function is the solution to the estimating equation where the estimating function is the derivative of the criterion. The generalized method of moments estimators of Hansen (1982) is an important class of such estimators in relation to what follows, see also Hall (2005) for a thorough account of the method.

To simplify matters we consider only estimating functions of the form (note that the dependence on Δ is made explicit)

$$F_1(\theta) = \sum_{i=1}^{n-r} f(Y_i, \dots, Y_{i+r}, \theta, \Delta)$$

satisfying

A0: $Ef(X_\Delta, \dots, X_{r\Delta}, \theta, \Delta) = 0$ if and only if $\theta = \theta_0$.

Denote by $\hat{\theta}_1$ a solution to the estimating equation $F_1(\theta) = 0$. Theorem C.2.1 below states that the estimator $\hat{\theta}_1$ is consistent and asymptotically normal provided that the following regularity conditions are valid:

A1: The parameter θ_0 belongs to the interior of Θ.

A2: The process $\{Y_i\}_{i\in\mathbb{N}}$ is stationary and ergodic.

A3: $f(y_1, \dots, y_{r+1}, \theta, \Delta)$ is twice continuously differentiable w.r.t. θ for all $(y_1, \dots, y_{r+1})$.

A4: There exists a neighborhood $B_\varepsilon(\theta_0)$ of θ_0 in which the variables $f(Y_1, \dots, Y_{r+1}, \theta, \Delta)$, $\partial_{\theta_i} f(Y_1, \dots, Y_{r+1}, \theta, \Delta)$, and $\partial_{\theta_i}\partial_{\theta_j} f(Y_1, \dots, Y_{r+1}, \theta, \Delta)$ indexed by $i, j = 1, \dots, d$ and $\theta \in B_\varepsilon(\theta_0)$ are dominated by an integrable function.

A5: The matrix $S_1(\theta_0) = E\{\partial_{\theta^T} f(Y_1, \dots, Y_r, \theta_0, \Delta)\}$ is invertible.

A6: $n^{-1/2} F_1(\theta_0) \to \mathcal{N}(0, \Gamma_1)$ where $\Gamma_1 = \lim_{n\to\infty} n^{-1} E\{F_1(\theta_0)^T F_1(\theta_0)\}$.

Theorem C.2.1 *If* **A0** - **A6** *are satisfied then, with probability tending to one as $n \to \infty$, a unique solution $\hat{\theta}_1$ to the estimating equations exist, and it furthermore holds that $\hat{\theta}_1 \to \theta_0$ in probability and*

$$n^{1/2}(\hat{\theta}_1 - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \{S_1(\theta_0)^{-1}\}^T \Gamma_1 S_1(\theta_0)^{-1}).$$

Proof: Theorem C.2.1 is proven in Jacod & Sørensen (2007) in the case $r = 1$, and the general case is a straight-forward extension. □

C.2.2 The goodness of fit test

It is important to notice that the continuous-time model specifies the distribution of $(X_\Delta, \dots, X_{n\Delta})$ not just for the given sampling frequency but for any $\Delta > 0$. In particular, the moment conditions **A0** should hold for all $\Delta > 0$ if the Δ-dependence is specified correctly according to the true data generating model. On the other hand if the model is misspecified, it is likely that the moment conditions fail to hold simultaneously for all Δ causing the parameter estimates to vary with the choice of Δ. Define

$$\mu(\theta, \Delta) = Ef(X_\Delta, \dots, X_{r\Delta}, \theta, \Delta).$$

Ideally we would test the hypothesis that $\mu(\theta_0, \Delta) = 0$ for all $\Delta > 0$. However, as data is only available at sampling frequencies Δ^{-1}, $(2\Delta)^{-1}$, etc., we derive a statistic for testing

$$H_0: \text{ there exists a } \theta_0 \text{ such that } \mu(\theta_0, \Delta) = \mu(\theta_0, k\Delta) = 0$$

against the alternative

$$H_A: \text{ for all } \theta \in \Theta \;\; \mu(\theta, \Delta) \neq 0 \text{ or } \mu(\theta, k\Delta) \neq 0$$

where k is a pre-specified integer. To this end we consider the "downsampled" estimating functions F_k given by

$$F_k(\theta) = \sum_{i=1}^{n-rk} f(Y_i, Y_{k+i}, \dots, Y_{rk+i}, \theta, k\Delta).$$

Note that F_k is based on all observations, not only every k'th. In other words, F_k is not based on a single downsample, but is the sum of k estimating functions based on different downsamples. For example, F_2 is the sum two estimating functions, one based on all the even-indexed observations and another based on all the odd-indexed observations. This seems beneficial as no observations are wasted, but one should be aware that certain properties, for example the martingale property, of the original estimating function F_1 are not inherited by its downsampled version. In particular the asymptotic variance for $n^{-1/2}F_k(\theta_0)$ is usually far more complicated than the one for $n^{-1/2}F_1(\theta_0)$.

Denote by $\hat\theta_k$ the estimator associated to F_k. In case the null hypothesis hold true $\hat\theta_1$ and $\hat\theta_k$ both are consistent estimators of θ_0 and we would thus expect that $\hat\theta_1 \approx \hat\theta_k$ for n sufficiently large. Hence, we would reject the hypothesis when observing a large value of

$$\tau(k) = n^{-1} \cdot (\hat\theta_1 - \hat\theta_k) W_n (\hat\theta_1 - \hat\theta_k)^T \tag{C.1}$$

where $\{W_n\}_{n\in\mathbb{N}}$ is an appropriate sequence of positive semi-definite weight matrices. Hausman (1978) and Newey (1985) also based their goodness of fit tests on the difference of two estimators, the one assumed to be consistent under the null and inconsistent under the alternative and the other assumed to be consistent under both the null and the alternative. In comparison, the above estimators $\hat\theta_1$ and $\hat\theta_k$ usually are either both consistent or both inconsistent under the alternative as they rely on essentially the same moment conditions.

Theorem C.2.2 *Suppose that* **A0** *through* **A5** *hold true and that the analogous conditions hold for the downsampled data (replace Δ with $k\Delta$ etc). Further assume that*

A7: $n^{-1/2}\{F_1(\theta_0), F_k(\theta_0)\} \to \mathcal{N}(0, \Gamma_0)$*, where* $\Gamma_0 = \lim_{n\to\infty} n^{-1}\operatorname{Var}(\{F_1(\theta_0), F_k(\theta_0)\})$*.*

Define matrices S_0 *and* Σ_0 *by*

$$S_0 = \begin{pmatrix} S_1(\theta_0)^{-1} \\ -S_k(\theta_0)^{-1} \end{pmatrix}$$

and $\Sigma_0 = S_0^T \Gamma_0 S_0$*. If* Σ_0 *is non-singular and* $W_n \to \Sigma_0^{-1}$ *in probability, then*

$$\tau(k) \overset{\mathcal{D}}{\to} \chi^2_d.$$

Proof: The regularity conditions ensure the consistency and asymptotic normality of the estimators $\hat{\theta}_k$ and $\hat{\theta}_1$. The proof relies on the same standard Taylor expansions as the proof of Theorem C.2.1, see for instance Jacod & Sørensen (2007). For $j = 1, k$ we get

$$0 = F_j(\hat{\theta}_j) = F_j(\theta_0) + \{\hat{\theta}_j - \theta_0\}\partial_{\theta^T} F_j(\theta_0) + o_P(n^{1/2}).$$

Combining these by **A5** yields

$$n^{1/2}\{\hat{\theta}_k - \theta_1\} = n^{-1/2}F_1(\theta_0)S_1(\theta_0)^{-1} - n^{-1/2}F_k(\theta_0)S_k(\theta_0)^{-1} + o_P(1).$$

Hence, by **A7** and the assumption that $W_n \to \Sigma_0^{-1}$ the desired limit distribution for $\tau(k)$ is hereby established. □

Note that the theorem demands that W_n is a consistent estimator of Σ_0^{-1}. When it comes to estimating Σ_0^{-1} several different strategies are available. If an explicit expression of $\Sigma_0 = \Sigma(\theta_0)$ is known, then the "plug-in estimator" $W_n = \Sigma(\hat{\theta}_1)^{-1}$ is an obvious choice. As an alternative sample analogues of the terms in $S(\hat{\theta}_1)$ and $\Gamma(\hat{\theta}_1)$ can be used to estimate Σ_0. Please note that

$$\Gamma_0 = EH_1(\theta_0)^T H_1(\theta_0) + \sum_{i=2}^{\infty}\{EH_1(\theta_0)^T H_i(\theta_0) + EH_i(\theta_0)^T H_1(\theta_0)\} \tag{C.2}$$

with $E\{H_1(\theta_0)^T H_{j+1}(\theta_0)\}$ the $j'th$ auto-covariance of $\{H_i(\theta_0)\}_{i\in\mathbb{N}}$ where

$$H_i(\theta) = \{f(Y_i, Y_{i+1}, \ldots, Y_{i+r}, \theta, \Delta) - \mu(\theta, \Delta), f(Y_i, Y_{i+k}, \ldots, Y_{i+rk}, \theta, k\Delta) - \mu(\theta, k\Delta)\}$$

The so-called *heteroscedasticity and auto-correlation consistent covariance estimator* of Newey & West (1987b) is a kernel-type estimator which combines the empirical auto-covariances into a consistent and positive semi-definite estimate. We refer to Andrews (1991), Andrews & Monahan (1992), and Newey & West (1994) for details. Notice that the sample-type estimates are less sensitive to model misspecification than the plug-in estimator as Σ usually depends on additional model features besides the moment conditions **A0**. See Hall (2000) and Hall & Inoue (2003) for results on covariance estimation under misspecification.

In practice some of the estimating equations may have either multiple solutions or no solution at all. When faced with several potential estimates one can choose the one giving rise to the smallest $\tau(k)$-value or the one closest to another consistent estimator. If one or more estimates are missing, a test can still be based on the existing coordinates of $\hat{\theta}_k - \hat{\theta}_1$ reducing the degrees of freedom of the limit χ^2 distribution accordingly. However, it should be noted that the problems of computing the estimates could indicate misspecification or an unfortunate choice of estimating function. A similar approach can be adopted if Σ_0 is singular.

C.2.3 Misspecification and consistency

Under misspecification the limit distribution of $\tau(k)$ depends in part on the moment conditions, in part on the behavior of the weights $\{W_n\}_{n\in\mathbb{N}}$ as demonstrated in Theorem C.2.3 below. See Hall (2005) for a similar result on the behavior of the generalized method of moments estimator under misspecification.

If the model is misspecified and the moment conditions still hold for some $\theta_0(\Delta) \in \Theta$ possibly depending on Δ, the test makes a comparison of the estimates of $\overline{\theta}_1 = \theta_0(\Delta)$ and $\overline{\theta}_k = \theta_0(k\Delta)$ satisfying

A8: $\mu(\theta, \Delta) = 0$ if and only if $\theta = \overline{\theta}_1$ and $\mu(\theta, k\Delta) = 0$ if and only if $\theta = \overline{\theta}_k$.

If $\theta_0(\Delta) = \overline{\theta}$ is constant, then the model is not badly misspecified in the sense that H_0 in fact holds true and the parameter estimates are robust. However, the test typically does not attain the correct size unless the covariance estimate is also robust. Hence, if our interest lies in testing whether a specific parameter estimate is robust, we should choose the weights independent of nuisance parameters. On the other hand if the alternative H_A holds true, then the test is consistent provided that the weights are not too misbehaved.

Theorem C.2.3 *Suppose* **A8** *holds, and* **A1** *through* **A5** *hold with θ_0 replaced by $\overline{\theta}_1$ and that the analogous conditions hold for the downsampled data with θ_0 replaced by $\overline{\theta}_k$. Further assume that $W_n \to W$ in probability where W is a deterministic positive definite matrix and that*

A9: $n^{-1/2}\{F_1(\overline{\theta}_1), F_k(\overline{\theta}_k)\} \to \mathcal{N}(0, \overline{\Gamma})$, *where* $\overline{\Gamma} = n^{-1}\lim_{n\to\infty} \mathrm{Var}[\{F_1(\overline{\theta}_1), F_k(\overline{\theta}_k)\}]$.

Define the matrices $\overline{S}$ and $\overline{\Sigma}$ in analogy with Theorem C.2.2, then

$$n^{1/2}\{\hat{\theta}_k - \hat{\theta}_1 - (\overline{\theta}_k - \overline{\theta}_1)\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \overline{S}^T \overline{\Gamma}\, \overline{S}) \tag{C.3}$$

and the limit distribution of $\tau(k)$ is given by one of the following:

1. *If $\overline{\theta}_k = \overline{\theta}_1$, then*
$$\tau(k) \xrightarrow{\mathcal{D}} \lambda_1\chi_1^2 \star \ldots \star \lambda_d\chi_1^2$$
where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $W^{1/2}\overline{\Sigma}W^{1/2}$ and $\star$ denotes convolution. In particular, if $W = \overline{\Sigma}^{-1}$ the limit distribution is χ_d^2.

2. *If $\overline{\theta}_k \neq \overline{\theta}_1$, then the test is consistent as*
$$n^{-1}\tau(k) \xrightarrow{P} (\overline{\theta}_k - \overline{\theta}_1)W(\overline{\theta}_k - \overline{\theta}_1)^T > 0.$$

Proof: Mimicking the proof of Theorem C.2.2 we get that

$$n^{1/2}\{\hat{\theta}_k - \hat{\theta}_1 - (\overline{\theta}_k - \overline{\theta}_1)\} = n^{-1/2}F_1(\overline{\theta}_1)S_1(\overline{\theta}_1)^{-1} - n^{-1/2}F_k(\overline{\theta}_k)S_k(\overline{\theta}_k)^{-1} + o_P(1).$$

as well as the convergence in (C.3) follows from **A9**. Moreover, noting that $\tau(k)$ equals

$$\begin{aligned}&n\{\hat{\theta}_k - \hat{\theta}_1 - (\overline{\theta}_k - \overline{\theta}_1)\}W\{\hat{\theta}_k - \hat{\theta}_1 - (\overline{\theta}_k - \overline{\theta}_1)\}^T\\ &+ n(\overline{\theta}_k - \overline{\theta}_1)W(\overline{\theta}_k - \overline{\theta}_1)^T + |\overline{\theta}_k - \overline{\theta}_1|O_P(n^{1/2}) + o_P(1),\end{aligned}$$

the assertions made about the limit distribution follow readily from (C.3). □

With Theorem C.2.3 in mind we can aim the goodness of fit test at a specific alternative by choosing the estimating function such that $\overline{\theta}_1$ differs from $\overline{\theta}_k$ as much as possible under this particular kind of misspecification. If the estimating function is explicit, then the regularity conditions can easily be checked for a specific alternative as $\mu(\theta, \Delta)$ can be computed either directly or by simulations. Hall (2000) and Hall & Inoue (2003) consider the behavior of covariance estimators under misspecification. Note that the above conditions assumed to derive the distribution of $\tau(k)$ are stronger than needed for the mere consistency of the test. The condition in the following theorem are sufficient:

Theorem C.2.4 *Assume that*

C1: *There exists sets $\mathcal{Z}_1, \mathcal{Z}_k \subset \Theta$ such that* $\mathrm{dist}(\mathcal{Z}_1, \mathcal{Z}_k) > 0$ *and $P(\hat{\theta}_j \in \mathcal{Z}_j) \to 1$ as $n \to \infty$ for $j = 1, k$.*

C2: *The smallest eigenvalue $\lambda_1(W_n)$ of W_n satisfy that $n\lambda_1(W_n) \to \infty$ in probability as $n \to \infty$.*

Then $\tau(k) \to \infty$ in probability as $n \to \infty$ and the test is consistent.

Proof: Choose δ such that $\delta < \mathrm{dist}(\mathcal{Z}_1, \mathcal{Z}_k)$. Condition **C1** implies that $P(|\hat{\theta}_k - \hat{\theta}_1| > \delta) \to 1$ as $n \to \infty$. Further note that

$$\tau(k) \geq |\hat{\theta}_k - \hat{\theta}_1|^2 \cdot n\lambda_1(W_n).$$

It follows that $P(\tau(k) > \delta^2 \cdot n\lambda_1(W_n)) \to 1$ as $n \to \infty$. Thus, $\tau(k)$ diverges to infinity by **C2**. In particular, $P(\tau(k) > \chi^2_{d,1-\alpha}) \to 1$ for any $0 < \alpha < 1$ with $\chi^2_{d,1-\alpha}$ denoting the $1 - \alpha$ quantile of the χ^2 distribution. □

Note that condition **C1** constrains the estimate $\hat{\theta}_k$ to the set $\mathcal{Z}_k$ where F_1 eventually cannot have zero points and vice versa. If $\{Y_i\}_{i \in \mathbb{N}}$ is ergodic, Θ is compact, and $f(y_1, \ldots, y_{r+1}, \theta, j\Delta)$ is continuous in θ for all $(y_1, \ldots, y_{r+1})$ and $j = 1, k$, then a straightforward extension of the law of large number on Banach spaces, see Ledoux & Talagrand (1991), implies that $n^{-1}F_j(\theta) \to \mu(\theta, j\Delta)$ uniformly on Θ for $j = 1, k$ whenever $\mu(\theta, j\Delta) = E\{f(Y_{j+1}, \ldots, Y_{jr+1}, \theta, j\Delta)\}$ is well defined in $[-\infty; \infty]^d$ for all $\theta \in \Theta$ and $j = 1, k$. Hence, **C1** follows if the zero points of $\mu(\theta, \Delta)$ and $\mu(\theta, k\Delta)$ form disjoint sets. Condition **C2** is a mild condition on the weights. Loosely speaking it ensures that they cannot tend to singularity at any rate faster than or equal to $o(n)$.

C.2.4 The related over-identifying restrictions test

For those familiar with the generalized method of moments, Hansen (1982), it is natural to view downsampling as a way to generate additional moment conditions and to perform goodness of fit by use of the over-identifying restrictions test. Of course, the estimators $\hat{\theta}_1$ and $\hat{\theta}_k$ are identical to the generalized method of moments estimators, obtained by minimizing the criteria $|F_1(\theta)|^2$ and $|F_k(\theta)|^2$, respectively. The pooled moment conditions for Δ and $k\Delta$ yield the optimal generalized method of moments estimator

$$\hat{\theta}_{\text{gmm}} = \arg\min_{\theta \in \Theta} \{F_1(\theta), F_k(\theta)\} V_n \{F_1(\theta), F_k(\theta)\}^T,$$

where $\{V_n\}$ is an optimal sequence of weights, i.e. V_n converges to Γ_0^{-1} in probability, with Γ_0 as defined in Theorem C.2.2.

The associated over-identifying restrictions test statistic is given by the minimum value of the criterion,

$$\tau_{\text{gmm}}(k) = \{F_1(\hat{\theta}_{\text{gmm}}), F_k(\hat{\theta}_{\text{gmm}})\}V_n\{F_1(\hat{\theta}_{\text{gmm}}), F_k(\hat{\theta}_{\text{gmm}})\}^T, \tag{C.4}$$

which converges to a χ^2 distribution with d degrees of freedom under the null. The over-identifying restrictions test is consistent under the conditions of Theorem C.2.3, see Hall (2000) and Hall & Inoue (2003). We would expect the behavior of our test to be similar to that of the over-identifying restrictions test. This is confirmed by simulation studies, see Section C.4.1.

C.3 Goodness of fit for diffusion models

In this section we present some examples that illustrate the theory of the previous section. To simplify matters we consider only diffusion models, but the test is applicable to a broader range of continuous time model.

Let $\{X_t\}_{t\geq 0}$ be a stationary diffusion, the solution of the stochastic differential equation

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t.$$

Regularity conditions on μ and σ that ensure the existence, uniqueness, stationarity, and ergodicity of the solution can be found in Genon-Catalot, Jeantheau & Laredo (2000), for example.

Often the parameters in the drift and/or the diffusion coefficients can be estimated by means of a simple martingale estimating function of the form

$$G_1(\theta) = \sum_{i=1}^{n-1} \omega(X_{i\Delta})[g(X_{(i+1)\Delta}, \theta) - E\{g(X_{(i+1)\Delta}, \theta)|X_{i\Delta}\}]$$

where ω is a weight function. We refer to Bibby, Jacobsen & Sørensen (2004) for an introduction to martingale estimating functions for diffusion models, including generic examples, and to Kessler & Sørensen (1999) for some particularly nice explicit examples. See Sørensen (1999) for asymptotic theory for estimating function on this particular form. It is important to notice that the limit behavior of the corresponding test statistic depends on the weight function under misspecification.

Note that the down-sampled estimating function F_k is not a martingale estimating functions because the observations are mixed together in the different terms in a non-standard way. Still, the martingale property of the original estimating function proves useful: due to it, the matrix Γ_0 defined in Theorem C.2.2 takes the form (C.2) with $E\{H_1(\theta_0)H_j(\theta_0)^T\} = 0$ for $j > k$.

C.3.1 Testing for a linear drift

Due to their simple structure the linear drift diffusions are among the most tractable diffusions. The linear drift family include some popular interest rate models such as the ones suggested by Vasicek (1977), Cox, Ingersoll & Ross (1985), and Chan et al. (1992),

$$\begin{aligned} \text{VAS} &: \ dX_t = -\theta(X_t-\mu)dt + \sigma dB_t & \text{(C.5)} \\ \text{CIR} &: \ dX_t = -\theta(X_t-\mu)dt + \sigma X_t^{1/2} dB_t & \text{(C.6)} \\ \text{CKLS} &: \ dX_t = -\theta(X_t-\mu)dt + \sigma X_t^{\gamma} dB_t & \text{(C.7)} \end{aligned}$$

In order to check the adequacy of these models we would have to asses whether the linear drift is reasonable. Assume for simplicity that we observe $Y_i = X_{i\Delta} - \mu$ (in practice, replace μ by the sample mean). By linearity of drift the conditional means are given by $E(Y_{i+1}|Y_i = y) = ye^{-\theta\Delta}$, provided that $E\{\sigma^2(X_t)\}$ is finite. Hence, for $j = 1, k$

$$G_j(\theta) = \textstyle\sum_{i=1}^{n-j}(Y_{i+j} - e^{-j\theta\Delta}Y_i)Y_i \quad \text{with} \quad \hat{\theta}_j = -(j\Delta)^{-1}\log\{(\sum_{i=1}^{n-j} Y_{i+j}Y_i)/(\sum_{i=1}^{n-j} Y_i^2)\}$$

is a (downsampled) estimating function which clearly satisfies **A0**. The estimator $\hat{\theta}_j$ is well defined and unique provided that $\sum_{i=1}^{n-j} Y_{i+j}Y_i > 0$ and $\sum_{i=1}^{n-j} Y_i^2 > 0$. The test statistic is thus given by

$$\tau(k) = \widehat{\Sigma}_n^{-1} \cdot \Delta^{-2} \cdot [\log\{\textstyle\sum_{i=1}^{n-1} Y_{i+1}Y_i / \sum_{i=1}^{n-1} Y_i^2\} - \log(\{(\sum_{i=1}^{n-k} Y_{i+k}Y_i)/(\sum_{i=1}^{n-k} Y_i^2)\}^{1/k})\,]^2,$$

where $\widehat{\Sigma}_n$ is an estimate of Σ_0.

In order to estimate Σ_0 for one of the specific models, let $\lambda_0 = e^{-\theta_0\Delta}$, $\nu_j = E(Y_1^2 Y_{j+1}^2) - \lambda_0^{2j} E(Y_1^4)$, and $\sigma_{\text{inv}}^2 = E(Y_i^2)$ then

$$S_0 = \begin{pmatrix} (\Delta\lambda_0\sigma_{\text{inv}}^2)^{-1} \\ -(k\Delta\lambda_0^k\sigma_{\text{inv}}^2)^{-1} \end{pmatrix} \text{ and } \Gamma_0 = \begin{pmatrix} \nu_1 & k\lambda_0^{k-1}\nu_1 \\ k\lambda_0^{k-1}\nu_1 & \nu_k + 2\sum_{j=1}^{k-1}\lambda_0^{2j}\nu_{k-j} \end{pmatrix}.$$

It follows that $\Sigma_0 = \sigma_{\text{inv}}^{-4}\Delta^{-2}[(k\lambda_0^k)^{-2}\{\nu_k + 2\sum_{j=1}^{k-1}\lambda_0^{2j}\nu_{k-j}\} - \lambda_0^{-2}\nu_1]$. Obviously λ_0 can be estimated by $\hat{\lambda}_k = e^{-\hat{\theta}_k\Delta}$. For the Vasiček model (C.5) we find the explicit expression $\nu_j = (1-\lambda_0^{2j})\sigma_{\text{inv}}^4$ where $\sigma_{\text{inv}}^2 = \sigma^2/(2\theta)$ is the variance of the invariant normal distribution of the diffusion (C.5). Consequently,

$$\text{VAS}: \ \Sigma_0 = \Delta^{-2} \cdot [(k\lambda_0^k)^{-2}\{2(1-\lambda_0^{2k})(1-\lambda_0^2)^{-1} + (1-2k)\lambda_0^{2k} - 1\} - (1-\lambda_0^2)\lambda_0^{-2}] \quad \text{(C.8)}$$

For the CIR model (C.6), let $\alpha = 2\mu\theta/\sigma^2$ and $\beta = \sigma^2/(2\theta)$ be the shape and scale parameter of the invariant Γ-distribution, then $\nu_j = \alpha^2\beta^4(1-\lambda_0^{2j}) + 4\alpha\beta^4\lambda_0^j(1-\lambda_0^j)$ yields an explicit expression,

$$\begin{aligned} \text{CIR}: \ \Sigma_0 &= \Delta^{-2}[(k\lambda_0^k)^{-2}\{2(1-\lambda_0^{2k})(1-\lambda_0^2)^{-1} + (1-2k)\lambda_0^{2k} - 1\} - (1-\lambda_0^2)\lambda_0^{-2}] \\ &+ 4\alpha^{-1}\Delta^{-2}[(k\lambda_0^k)^{-2}\{2\lambda_0^k(1-\lambda_0^k)(1-\lambda_0)^{-1} + (1-2k)\lambda_0^{2k} - \lambda_0^k\} - (1-\lambda_0)\lambda_0^{-1}]. \end{aligned}$$

Usually we will not be apt to use any of the model specific estimators of Σ_0, though, when we test for linearity rather than for the one of the specific models. Rather, we will use a sample-based estimator like the heteroscedasticity and auto-correlation consistent (HAC) estimator of Newey & West (1987b) for Γ_0 and plug in $\hat{\lambda}$ for λ in S_0.

Due to the specific estimating function G_j we essentially test if the auto-correlation function has the correct form, that is, $\text{Cor}(X_0, X_{k\Delta}) = \text{Cor}(X_0, X_\Delta)^k$. If we were interested in other features of the distribution, we should choose an estimating functions expressing these. Note that the test statistic depends on the exact parameterization of the model, which is of course somewhat unfortunate, but by the Delta-method the test is asymptotically invariant under continuously differentiable re-parameterizations.

Consistency against a sum of linear drift diffusions alternative

In order to better match the auto-correlation function found in measurements of the wind velocity Bibby, Skovgaard & Sørensen (2005) considered sums of independent linear drift diffusions. Our test is consistent against the sum of linear drift diffusions alternative and thus well suited for judging whether an apparent deviation from linearity is significant. Suppose for instance that the true data generating process is the sum of $m \geq 2$ independent stationary linear drift diffusions,

$$X_t = X_{1,t} + \ldots + X_{m,t}, \quad dX_{j,t} = -\theta_j (X_{j,t} - \mu_j)dt + \sigma_j(X_{j,t})dB_{j,t}$$

where $E\{\sigma_j^2(X_{j,t})\} < \infty$. Let $\mu = \mu_1 + \ldots + \mu_m$, $\sigma^2_{\text{inv}} = \sigma^2_{1,\text{inv}} + \ldots + \sigma^2_{m,\text{inv}}$, and let $\phi_j = \sigma^2_{j,\text{inv}}/\sigma^2_{\text{inv}}$ for $i = 1, \ldots, m$. Based on the observations $Y_i = X_{i\Delta} - \mu$ the goodness of fit test is consistent unless $\theta_1 = \ldots = \theta_m$. It is easily verified that for $j = 1, k$

$$\mu(\theta, j\Delta) = \sigma^2_{\text{inv}}\{\phi_1 e^{-j\theta_1\Delta} + \ldots + \phi_m e^{-j\theta_m\Delta} - e^{-j\theta\Delta}\},$$

and that $\hat{\theta}_j$ tends to $\overline{\theta}_j = -(j\Delta)^{-1}\log\{\phi_1 e^{-j\theta_1\Delta} + \ldots + \phi_m e^{-j\theta_m\Delta}\}$ which is the unique solution of $\mu(\theta, j\Delta) = 0$. Note that $\phi_1, \ldots, \phi_m > 0$ and $\phi_1 + \ldots + \phi_m = 1$. Hence, by convexity

$$\phi_1 e^{-\Delta\theta_1} + \ldots + \phi_m e^{-\Delta\theta_m} < \{\phi_1 e^{-\Delta\theta_1} + \ldots + \phi_m e^{-\Delta\theta_m}\}^{1/k}$$

unless $\theta_1 = \ldots = \theta_m$. Thus, $\overline{\theta}_1 \neq \overline{\theta}_k$ and the conditions of Theorem C.2.3 hold true.

Consistency against a nonlinear drift-alternative

Aït-Sahalia (1996b) suggests the following non-linear diffusion model as an alternative to linear drift models:

$$dX_t = (\alpha_{-1}X_t^{-1} + \alpha_0 + \alpha_1 X_t + \alpha_2 X_t^2)dt + \sqrt{\beta_0 + \beta_1 X_t + \beta_2 X_t^\rho}dB_t. \qquad \text{(C.9)}$$

The auto-correlations of the general non-linear drift model are not explicitly known. Hence we cannot check the consistency condition by computing $\mu(\theta, j\Delta)$ and $\overline{\theta}_j$ explicitly. However, as $\hat{\theta}_j$ is explicit and unique we can easily simulate $\hat{\theta}_k - \hat{\theta}_1$ which will approximate $\overline{\theta}_k - \overline{\theta}_1$ for a large sample size n.

Figure C.1 depicts $\overline{\theta}_k - \overline{\theta}_1$ as a function of the parameter α_2 in the nonlinear drift model. The remaining parameters are $\alpha_{-1} = 0$, $\alpha_0 = 0.2$, $\alpha_1 = 0.1$, $\beta_0 = \beta_1 = 0$, $\beta_2 = 0.25$, and $\rho = 1$. For each value of α_2 a time series of length $201,000$ was simulated using the Milstein scheme with five steps for each observation. The one thousand first observations served as burn-in. It is clear that $\overline{\theta}_k - \overline{\theta}_1$ decreases with α_2 and the test is thus consistent for $\alpha_2 < 0$. It is important to notice that for $\alpha_2 = 0$ we recover the CIR model for which the drift is of course linear, whence $\overline{\theta}_k - \overline{\theta}_1 = 0$. In small samples, however, the test does not have much power as the difference in $\overline{\theta}_k$ and $\overline{\theta}_1$ is tiny and cannot easily be detected as described in the power study in section C.4.2 below.

C.3.2 A specification test for the Vasiček model

Of course, the above test for linear drift cannot distinguish the various linear drift models as they have the same correlation structure. Hence, we need to consider other estimating

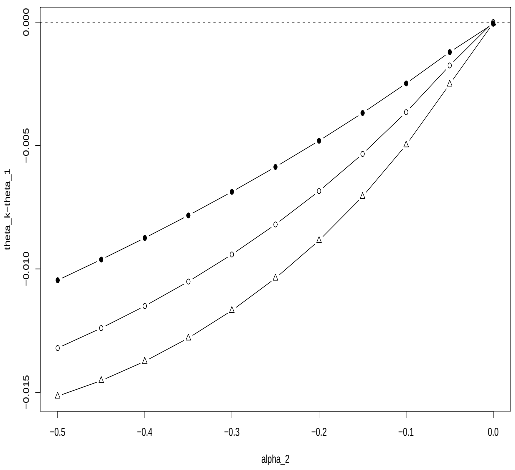

Figure C.1: Simulated values of $\overline{\theta}_k - \overline{\theta}_1$ based on time series of length 200,000 for the nonlinear drift model $dX_t = (0.2 - 0.1X_t + \alpha_2 X_t^2)dt + \sqrt{0.25X_t}dB_t$ with α_2 varying. The values of k are 2 (solid) circles, 3 (open circles), and 4 (triangles).

functions in order to separate these models.

For the Vasiček model (C.5) it is known that $E(Y_{i+1}^2|Y_i = y) = y^2 e^{-2\theta\Delta} + \sigma_{\text{inv}}^2(1 - e^{-2\theta\Delta})$.

$$G_j(\theta) = \textstyle\sum_{i=1}^{n-j}\{Y_{i+j}^2 - \sigma_{\text{inv}}^2 - e^{-2j\theta\Delta}(Y_i^2 - \sigma_{\text{inv}}^2)\}(Y_i^2 - \sigma_{\text{inv}}^2), \quad j = 1, k$$

is a (downsampled) estimating function which is unbiased under the null. The unique zero point is

$$\hat{\theta}_j = -(2j\Delta)^{-1}\log\{(\textstyle\sum_{i=1}^{n-j}(Y_{i+j}^2 - \sigma_{\text{inv}}^2)(Y_i^2 - \sigma_{\text{inv}}^2))/(\sum_{i=1}^{n-j}(Y_i^2 - \sigma_{\text{inv}}^2)^2)\}, \quad j = 1, k$$

assuming that $\sum_{i=1}^{n-j}(Y_{i+j}^2 - \sigma_{\text{inv}}^2)(Y_i^2 - \sigma_{\text{inv}}^2) > 0$ and $\sum_{i=1}^{n-j}(Y_i^2 - \sigma_{\text{inv}}^2)^2 > 0$, and the corresponding test statistic is given by

$$\begin{aligned}\tau(k) \;&=\; \widehat{\Sigma}_n^{-1} \cdot (2\Delta)^{-2} \cdot [\log\{\textstyle\sum_{i=1}^{n-1}(Y_{i+1}^2 - \sigma_{\text{inv}}^2)(Y_i^2 - \sigma_{\text{inv}}^2)/\sum_{i=1}^{n-1}(Y_i^2 - \sigma_{\text{inv}}^2)^2\} \\ &\qquad - \log(\{(\textstyle\sum_{i=1}^{n-k}(Y_{i+k}^2 - \sigma_{\text{inv}}^2)(Y_i^2 - \sigma_{\text{inv}}^2))/(\sum_{i=1}^{n-k}(Y_i^2 - \sigma_{\text{inv}}^2)^2)\}^{1/k})\,]^2,\end{aligned}$$

where $\widehat{\Sigma}_n$ is an estimate of Σ_0. In order to find such an estimate note that the centered and squared Vasiček process is a CIR process as it satisfies a stochastic differential equation of form (C.6),

$$d(X_t - \mu)^2 = -2\theta\{(X_t - \mu)^2 - \sigma_{\text{inv}}^2\}dt + 2\sigma\{(X_t - \mu)^2\}^{1/2}dB_t.$$

Hence, we test for linear drift of the transformed process. In this case, since we are testing for a specific model, is is natural to use the CIR-specific estimator of Σ_0. With the above parameterization we get

$$\begin{aligned}\Sigma_0 \;&=\; 4\Delta^{-2}[(k\lambda_0^{2k})^{-2}\{2(1 - \lambda_0^{4k})(1 - \lambda_0^4)^{-1} + (1 - 2k)\lambda_0^{4k} - 1\} - (1 - \lambda_0^4)\lambda_0^{-4}] \\ &\;+\; 32\Delta^{-2}[(k\lambda_0^{2k})^{-2}\{2\lambda_0^{2k}(1 - \lambda_0^{2k})(1 - \lambda_0^2)^{-1} + (1 - 2k)\lambda_0^{4k} - \lambda_0^{2k}\} - (1 - \lambda_0^2)\lambda_0^{-2}]\end{aligned}$$

where $\lambda_0 = e^{-\theta_0\Delta} = \mathrm{Cor}(X_0, X_\Delta$ is the correlation between two consecutive observations. Summarizing, the procedure is the following: Consider the centered squared process, $\tilde{X} = (X - \mu)^2$, and proceed as in Section C.3.1, using the CIR-specific estimator Σ_0, suitably accommodated to the new situation.

Consistency against the CIR alternative

By Theorem C.2.3 the test is consistent against the CIR alternative (C.6): Denote by θ_0 the drift parameter of the CIR process and by α and β the parameters of the invariant Γ-distribution, then for $j = 1, k$

$$\mu(\theta, j\Delta) = 2\alpha\beta^4\{(\alpha+1)e^{-2j\theta_0\Delta} + 2e^{-j\theta_0\Delta} - (\alpha+3)e^{-j\theta\Delta}\}.$$

and $\hat{\theta}_j$ converges to $\overline{\theta}_j = -(2j\Delta)^{-1}\log\{\psi e^{-2j\theta_0\Delta} + (1-\psi)e^{-j\theta_0\Delta}\}$, the unique solution of $\mu(\theta, j\Delta) = 0$, where $\psi = (\alpha+1)(\alpha+3)^{-1}$. As

$$\psi e^{-2\theta_0\Delta} + (1-\psi)e^{-\theta_0\Delta} < \{\psi e^{-2k\theta_0\Delta} + (1-\psi)e^{-k\theta_0\Delta}\}^{1/k}$$

it follows that $\overline{\theta}_k \neq \overline{\theta}_1$.

C.3.3 A specification test for the CIR model

For the CIR model (C.6) the conditional second order moments are well known, too:

$$E(Y_{i+1}^2|Y_i = y) = y^2e^{-2\theta\Delta} + 2\beta y e^{-\theta\Delta}(1-e^{-\theta\Delta}) + \alpha\beta^2(1-e^{-2\theta\Delta})$$

where α and β are the parameters of the invariant Γ distribution. A (downsampled) estimating function is thus given by

$$G_j(\theta) = \textstyle\sum_{i=1}^{n-j}\{Y_{i+j}^2 - \alpha\beta^2 - e^{-2j\theta\Delta}(Y_i^2 - \alpha\beta^2) - 2\beta e^{-j\theta\Delta}(1-e^{-j\theta\Delta})Y_i\}Y_i, \quad j = 1, k.$$

which may be used as the basis for a test. The estimating equations can be solved explicitly, and using the moment formulas from Forman & Sørensen (2006) one can even compute an explicit expression for Σ_0 under the null. The formulas are quite complicated, though, and therefore left out.

C.4 Simulation studies

In order to investigate the performance of the test we have carried out a number of simulation studies for the diffusion case, cf. Section C.3. In Sections C.4.1 and C.4.2 focus is on the test for linear drift, investigating properties under the null and the power against certain alternatives, respectively, whereas in Section C.4.3 a minor study is performed on the specification test for the Vasiček model.

First a few preliminary comments: As illustrated in Figure C.2, we find that $\tau(k)$ and $\tau_{\text{gmm}}(k)$ defined by (C.1) and (C.4), respectively are almost indistinguishable even for moderate sample sizes. Therefore, for the remaining figures and tables, we have used $\tau(k)$ only.

For estimation of the variance matrix Γ_0 we have systematically used the heteroscedasticity and auto-correlation consistent covariance estimator with the data-generated (automatic) weight- and lag selection from Newey & West (1994) and the pre-whitening technique from Andrews & Monahan (1992). This yields, per construction, a positive semi-definite variance matrix, which from now on will be referred to as the HAC estimator (or sample based) of Γ_0. In practice we used (slightly modified) versions of functions

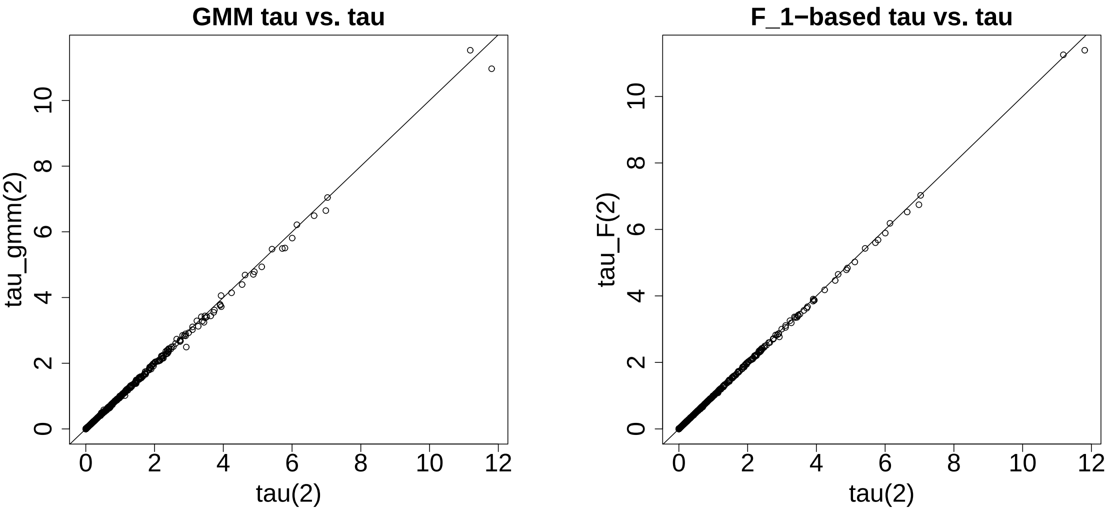

Figure C.2: Comparison of three test statistics, $\tau(2)$, $\tau_{\text{gmm}}(2)$ and $\tau_F(2)$, for 500 samples of length 500 from the Vasiček model.

from the R-package `sandwich`, see Zeileis (2004). Initially, we also tried a simpler version, with weights one for the first k lags and zero otherwise (which is the correct form under the null), but for misspecified models, small sample sizes and extreme parameter values this often produced estimated variance matrices that were not positive definite. In some studies we used the Vasiček model specific estimator of Σ_0 given by (C.8). This is essentially only meaningful if one is really interested in testing for the Vasiček model (and not just for linear drift) and it is mainly included for comparison with the HAC estimator.

C.4.1 Properties of the test statistics under the null

In this section we investigate the properties of the goodness of fit test under the null hypothesis of a linear drift. First, let us compare the asymptotically equivalent test statistics $\tau(2)$ and $\tau_{\text{gmm}}(2)$ for a relatively small sample size. We simulated 500 processes of length $n = 500$ from the Vasiček process (C.5) with parameters $(\mu, \theta, \sigma) = (0, 0.1, 1)$ and computed, for each sample, $\tau(2)$ and $\tau_{\text{gmm}}(2)$ with the HAC estimator of Γ_0. The left part of Figure C.2 where $\tau_{\text{gmm}}(2)$ is plotted against $\tau(2)$, shows that the two test statistics are almost indistinguishable. We got similar results for $k > 2$ and when we simulated from misspecified models. Yet another asymptotically equivalent test statistic is based on $F_1(\hat{\theta}_2)$, squared and correctly scaled in order to be asymptotically $\chi^2(1)$-distributed. This test statistic, $\tau_F(2)$, is plotted against $\tau(2)$ in the right part of Figure C.2. Again, the test statistics are essentially identical.

Second, Figure C.3 compares $\tau(2)$ and $\tau(3)$ for the same simulated datasets as above, still using the HAC estimator of Γ_0. We see that the two test statistics differ to some degree; the Pearson correlation is 0.80. We got similar results when we simulated sums of two Vasiček processes. We will return to the comparison of different values of k in the following.

Third, we check the distribution of $\tau(k)$ under the null. Again we simulated Vasiček

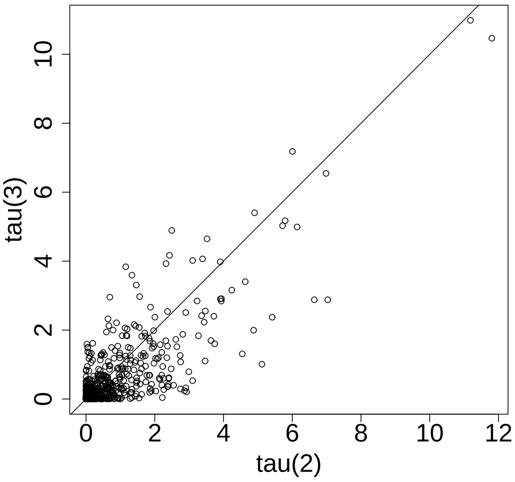

Figure C.3: Comparison of the test statistics $\tau(2)$ and $\tau(3)$ for 500 samples of length 500 from the Vasiček model.

processes with parameters $(\mu, \theta, \sigma) = (0, 0.1, 1)$, this time 1,000 processes of length $n = 1,000$. We computed $\frac{1}{\sqrt{n}}(\hat{\theta}_1 - \hat{\theta}_k)/\sqrt{W_n}$ which is asymptotically $N(0,1)$-distributed and equal to $\tau(k)$ while squared. We used $k = 2$ and $k = 3$ and the HAC estimator (sample based) as well as the Vasiček model specific estimator of Γ_0. Recall that the choice of estimation technique for Γ_0 solely has to do with the normalization of the test statistic. Normal probability plots (QQ-plots) are shown in Figure C.4 and compared to the $N(0,1)$-distribution (the straight line). The plots are quite nice for $k = 2$, not quite as nice for $k = 3$, but they of course improve with increasing sample size. An AR(1)-process is fitted for the pre-whitening technique, and we believe that perhaps fitting an AR(p)-process of order $p > 1$ might improve the finite sample behavior of the HAC based $\tau(k)$ for $k > 2$, see Andrews & Monahan (1992). Comparing the upper and lower left panels, the different normalizations of the test statistic (sample based vs. model based) do not seem to give different distributions for $k = 2$.

Fourth, we investigate the actual size of the test for varying sample size. For n ranging from 200 to 6,000 we simulated 5,000 samples from the same Vasiček model as above and carried out the goodness of fit test for each sample on the 5% significance level. That is, we rejected the hypothesis of a linear drift when $\tau(k)$ was larger than 3.84. Figure C.5 shows the rejection rates plotted against sample size: we used the HAC (sample based) estimator for Γ_0 (solid lines) as well as the Vasiček model specific estimator (dashed lines) for $k = 2$ (solid circles), $k = 3$ (open circles) and $k = 4$ (triangles). For all sample sizes the size of the HAC test is close to 5% for $k = 2$, but too small for k equal to 3 and 4. For $n = 50,000$ we got 0.051 ($k = 2$), 0.049 ($k = 3$), and 0.044 ($k = 4$) for 1,000 samples so the size does approach the correct size as n increases, but the convergence is quite slow. The test procedure using the model specific estimator for Γ_0 hits the level satisfactorily for all three values of k.

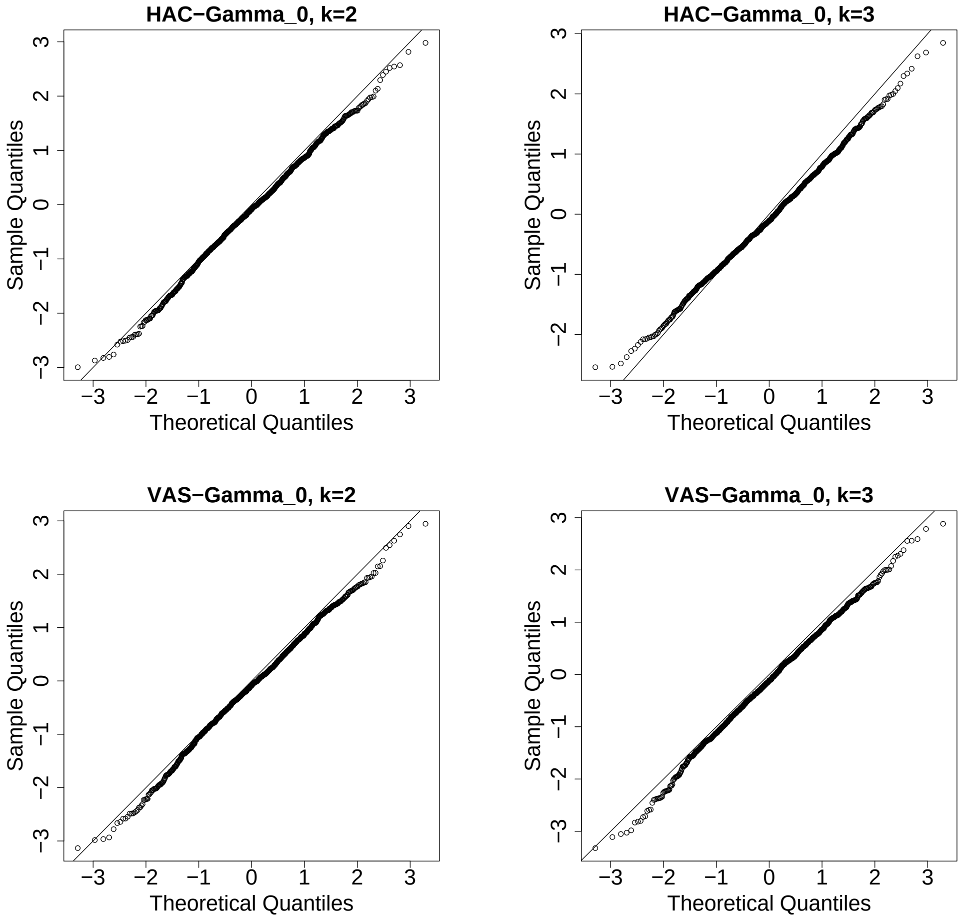

Figure C.4: Normal probability plots for $\frac{1}{\sqrt{n}}(\hat{\theta}_1 - \hat{\theta}_k)\sqrt{W_n}$ (equal to $\tau(k)$ while squared), based on 1,000 Vasiček processes of length 1,000. The values of k are 2 (left) and 3 (right). The upper and lower panels differ by the estimation procedure for Γ_0 (sample based in the top, model based in the bottom).

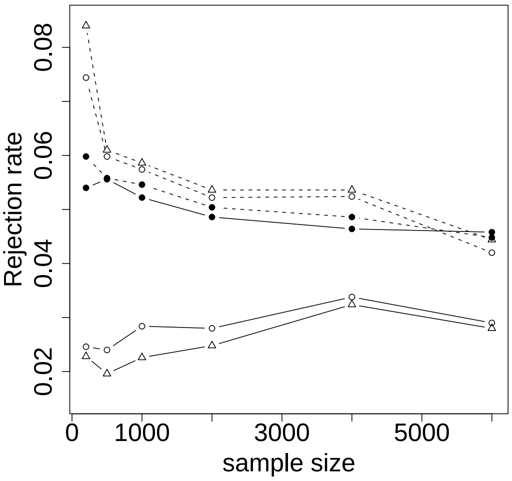

Figure C.5: Simulated size of the goodness of fit test for varying sample size. For the solid lines we have used the HAC estimator of Γ_0 for normalization; for the dashed lines we have used the Vasiček model specific normalization. The values of k are 2 (solid circles), 3 (open circles) and 4 (triangles). Each point is based on 5,000 samples.

C.4.2 Power investigations

We proceed to investigate the power of the linear drift goodness of fit test under various alternatives. All tests are carried out at the 5% significance level. As the first alternative we assume that the true process is a sum of two Vasicik processes. We simulated processes $X_1 + X_2$ where X_1 has parameters $(\mu_1, \theta_1, \sigma_1) = (0, 0.1, \sqrt{0.5})$ and X_2 has $\mu_2 = 0$ and varying (θ_2, σ_2) such that the variance of X is the same in all cases. In Figure C.6 the rejection rates are plotted against the true value of θ_2. In the left panel we have used the HAC estimator of Γ_0, in the right we have used the Vasiček model specific estimator. As before solid circles, open circles and triangles correspond to k equal to 2, 3 and 4, respectively. The sample size is $n = 1,000$ (solid lines) and $n = 2,000$ (dashed lines), and each point is based on 5,000 samples. We see that the test procedure actually detects the discrepancy from the hypothesis. Note that for $\theta_2 = \theta_1 = 0.1$ the sum itself is a Vasiček process and hence we find the size for the test once again. Naturally the power increases as the sample size increases (this is just the consistency of the test). As for the different values of k we see that the power increases with k. Larger values of k turned out not to improve the power any further, however. Recall that the size of the HAC test was not quite as precise for $k = 3$ and $k = 4$ as for $k = 2$ (cf. Figure C.5) so it is not obvious that $k > 2$ is to be preferred. Comparing the left and the right panel we see surprisingly little difference between the two different normalization procedures.

As a second alternative we have used transformations of the CIR-process (C.6) with parameters $(\mu, \theta, \sigma) = (1, 0.1, \sqrt{0.2})$; then the stationary distribution is the Gamma distribution with shape parameter and scale parameter both equal to one. We simulated 1,000 CIR processes and transformed each process with four different functions: $\log(x)$, $\log((x-1)^2)$, $1/x$, and $(x-1)^2$. We also used the integrated process, that is, $Y_i = \int_{(i-1)\Delta}^{i\Delta} X_s \, ds$. The goodness of fit test for linear drift was applied to the transformed processes, and the

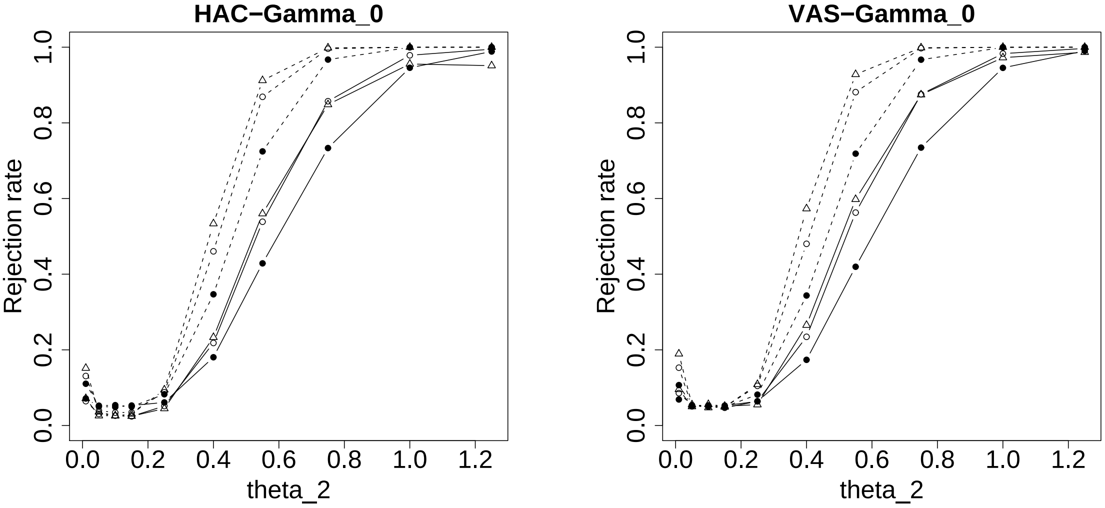

Figure C.6: Power against the alternative of a sum of two Vasiček processes. Sample sizes are 1,000 (solid lines) and 2,000 (dashes lines). See the main text for further details.

rejection rates are listed in Table C.1. Note that the reported sample sizes vary quite a lot. The results are quite different for the various transformations: The test procedure very easily detects that the hypothesis is not true for the integrated CIR process: this happens for roughly 90% of the simulated datasets for a sample size as small as 300. The discrepancy from a linear drift process is detected with a probability of roughly 80% for a sample size of 2,000, 1,000, and 5,000 for $\log(X)$, $\log((X-1)^2)$ and $1/X$, respectively. On the other hand, the test is useless for $(X-1)^2$: the hypothesis is almost never rejected even for a sample size of 8,000. Note that for $\log(X)$ the power increases with k, whereas it decreases for the other transformations.

As a third alternative we used the non-linear diffusion process (C.9) from Aït-Sahalia (1996b). The results from four different sets of parameter values are listed in Table C.2 with the parameter values themselves in the top eight lines. The parameter values in case I are the same as in Figure C.1 so the test is indeed consistent in this case. The parameter values in case II are almost the same; only the coefficient α_{-1} has been changed. The parameter values in case III are the estimated parameters from Aït-Sahalia (1996b) for a dataset on 5505 daily spot rates from 1973 to 1995, and finally case IV is a control case corresponding to a CIR process where the linear drift hypothesis is indeed true. The means and variances are quite different in the four settings and are listed in lines 9–10 of the table (computed by simulation in cases I–III). In each of the four cases we simulated 1,000 processes with 10,000 observations, for case III also 1,000 samples with 5,500 observations corresponding to the spot rate data from Aït-Sahalia (1996b). The last lines list the rejection rates for the test where we have used the sample based HAC estimator for Γ_0. We not not try all combinations of parameter values and k: a star ($\star$) indicates a combination we have not tries; a horizontal line (—) indicates numerical problems implying unreliable results (not listed).

The results are not too impressive. Although, from Figure C.1, we know that the test is consistent in case I, the power against this alternative turns out to be extremely

Transformation	n	$k=2$	$k=3$	$k=4$
$\int Xds$	100	0.466	0.248	—
$\int Xds$	300	0.879	0.596	0.358
$\int Xds$	500	0.973	0.808	0.569
$\int Xds$	1,000	0.999	0.973	0.894
$\log(X)$	1,000	0.454	0.535	0.561
$\log(X)$	2,000	0.814	0.874	0.898
$\log(X)$	3,000	0.928	0.968	0.981
$\log((X-1)^2)$	500	0.351	0.122	—
$\log((X-1)^2)$	1,000	0.835	0.615	0.095
$\log((X-1)^2)$	2,000	0.955	0.988	0.0561
$1/X$	3,000	0.341	0.039	0
$1/X$	5,000	0.762	0.207	0.001
$1/X$	8,000	0.982	0.511	0.005
$(X-1)^2$	8,000	0.039	0.044	0.048

Table C.1: Power under various alternatives where the true process is a transformed CIR process with parameters $(1, 0.1, \sqrt{0.2})$. Each rejection rate is based in 1,000 samples. For the entries marked with a horizontal line (—) we encountered numerical problems which we did not yet pursue any further.

low: for a sample size of 10,000 the linear drift is only rejected with a probability of 10%. The reason is, of course, that the drift is close to linear in the part of the state space where X actually takes its values. The results are even worse for case II where $\alpha_{-1} > 0$. What happens is that the process is strongly forced away from zero, so there are few observations in the area where the drift is seriously non-linear. For the parameters from Aït-Sahalia (1996b), case III, we find that the power maximal for $k = 6$ (we also tried larger values but 6 was optimal in this sense). The hypothesis of a linear drift is rejected for 32% of the samples with 5,000 observations and for 47% of the samples with 10,000 observations. For the control case IV, the results are as expected.

As a fourth alternative we use the discrete-time AR(2)-process. Indeed, this is not a continuous-process but it appears as the Euler approximation to certain delay SDE's. We simulated from an AR(2)-process

$$X_t = \rho_1 X_{t-1} + \rho_2 X_{t-2} + e_t$$

with the e_t's are iid. $N(0, \sigma^2)$, $\rho_1 = 0.8$ and (ρ_2, σ) varies such that the stationary variance is constantly equal to 0.6944, the variance for $(\rho_2, \sigma) = (0, 0.5)$. Figure C.7 shows the results for sample size $n = 1,000$. We have used $k = 2, 3, 4$ (solid circles, open circles, triangles as above) and the HAC estimator (solid lines) as well as the Vasiček model specific estimator of Γ_0 (dashed lines). Each point is based on 5,000 samples. The test indeed detects that this is not a linear drift process, and the power decreases with k. Note that numerical problems occurred for $k = 4$ and large negative values of ρ_2, and that the two different estimation strategies for Γ_0 produce almost identical power for $k = 2$ but not for $k = 3, 4$.

Case	I		II	III		IV
α_{-1}	0		0.25	$1.304 \cdot 10^{-4}$		0
α_0	0.2		0.2	$-4.643 \cdot 10^{-3}$		0.2
α_1	-0.1		-0.1	0.04333		-0.1
α_2	-0.3		-0.3	-0.1143		0
β_0	0		0	$1.108 \cdot 10^{-4}$		0
β_1	0		0	$-1.883 \cdot 10^{-3}$		0
β_2	0.25		0.25	$9.681 \cdot 10^{-3}$		0.25
ρ	1		1	2.073		1
$\mathrm{E}\,X$	0.58		1.04	0.097		2.00
$\mathrm{Var}\,X$	0.14		0.14	0.0030		2.50
n	10,000	50,000	10,000	5,000	10,000	10.000
$k = 2$	0.095	0.359	0.05	0.171	0.241	0.054
$k = 3$	0.072	0.299	—	0.154	0.237	0.026
$k = 4$	0.060	0.203	—	0.211	0.318	0.032
$k = 5$	⋆	⋆	⋆	0.237	0.344	⋆
$k = 6$	⋆	⋆	⋆	0.319	0.465	⋆

Table C.2: Power against the non-linear diffusion model (C.9) from Aït-Sahalia (1996b). Each rejection rate is based on 1,000 samples. Stars (⋆) indicate that we have not performed the corresponding simulations; horizontal lines (—) indicate that we encountered numerical problems and hence do not have reliable results.

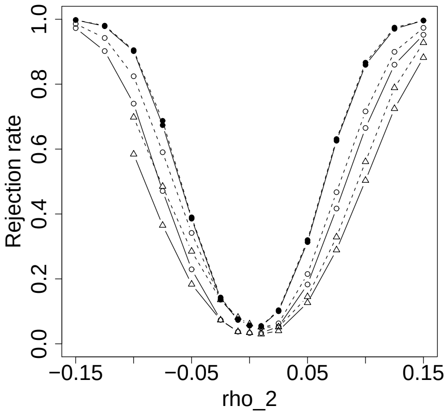

Figure C.7: Power against the discrete-time AR(2)-alternative, where $\rho_1 = 0.8$ and $\sigma^2 = \mathrm{Var}\, e_t$ is such that the stationary variance is 0.6944. The sample size is 1,000 and each point is based on 5,000 samples.

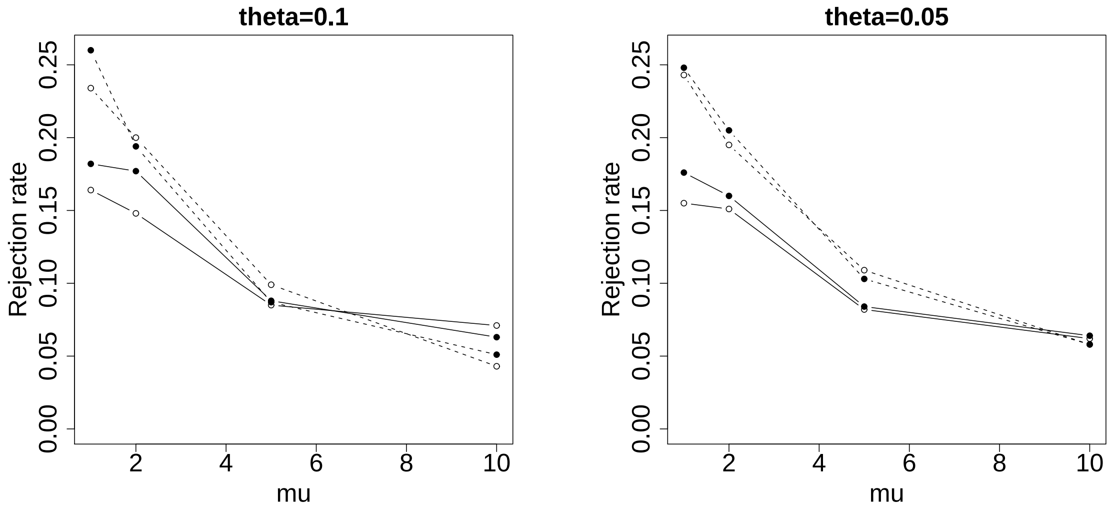

Figure C.8: Power of the Vasiček model specification test against the CIR model with parameters (μ, θ_1, σ) where θ_1 is equal to 0.1 (left panel) and 0.05 (right panel), and $\sigma = \sqrt{2\theta/\mu}$. The sample size is 1,000 (solid lines) and 5,000 (dashes lines), and k equals 2 (solid circles) and 3 (open circles). Each point is based on 5,000 observations.

C.4.3 Specification test for the VAS-model

As a final application we have applied the specification test for the Vasiček model from Section C.3.2. Since the drift is linear for the Vasiček as well as for the CIR process, say, the test for linear drift applied to the original observations would not (and should not) be able to distinguish between these processes. The idea from Section C.3.2 was to instead apply the test to the centered and squared observations in order check is the transformed process is a linear drift diffusion. This is the case for the Vasiček process but not for the CIR process.

We simulated 1,000 CIR processes with two different values of θ ($\theta = 0.1$ and $\theta = 0.05$) and four different values of $\mu \geq 1$. In all cases we used $\sigma = \sqrt{2\theta/\mu}$, such that the stationary variance is constantly one. For μ approaching infinity the drift for the squared and centered process approaches a linear drift, so the hypothesis is true.

In Figure C.8 the rejection rates are shown for $\theta = 0.1$ to the left and $\theta = 0.05$ to the right. The solid lines are for sample size 1,000, the dashed lines for sample size 5,000, and As usual the solid circles corresponds to $k = 2$ and the open circles to $k = 3$. Each point is based on 5,000 samples. The results are not impressive. The best rejection rate, obtained for $\mu = 1$ as expected, is only around 0.25 for a sample size of 5,000 and below 0.20 for a sample size of 1,000. In other words, very large samples are necessary in order to use the goodness of fit test to distinguish between these two processes with a satisfactory power. (Note that for certain parameter values a simple analysis of the marginal distributions easily detects the difference between the two models.)

C.5 Conclusion

In this paper we have developed a goodness of fit test based on comparison of distributions for different sampling frequencies. More specifically the test compares parameter estimates for downsamples of the data. We have proved results on the asymptotic distribution under the null as well as consistency results, and the properties of the test have been investigated by simulation in various diffusion models. In this paper we have applied the test to diffusion models only, but the it applies to (continuous-time) processes in general.

Of course, the simulations illustrate only partly the properties of the test in the sense that other circumstances (other processes, parameter values, sample sizes etc.) could have been considered. Nevertheless, for the test for linear drift we believe to have illustrated (i) that the size of the test is satisfactory even for moderate sample sizes, at least for $k = 2$; and (ii) that the test is indeed able to detect certain discrepancies from the hypothesis of a linear drift diffusion. In particular the test was quite strong against three non-Markovian alternatives: a sum of two VAS-processes, an integrated CIR-process and a discrete-time AR(2)-process. On the other hand, there were other alternatives where the performance was less than good, even in situations where it could be proved theoretically that the test is consistent. This was also the case for the specification test for the Vasiček model.

Comparing the results for varying values of k we got different results for the various alternatives. Intuitively, a discrepancy between two models will be magnified for increasing k. On the other hand, the estimates get more imprecise (for example, elements of Γ_0 increases dramatically in certain cases), so there is a trade-off. Our advice is to try out a few different values of k, but one should be aware that the tests are far from independent.

As for all other goodness of fit tests it is necessary to understand which aspects of the model that are actually tested for, that is, which alternatives the test is capable of detecting. In our setting this is determined by the estimating function. For example, the test for linear drift essentially tests whether the correlation over time lag 2Δ is the square of the correlation over time lag Δ. If we are interested in other features of the distribution, then we should choose estimating functions matching those.

It would be interesting to compare our test to other goodness of fit tests. In particular we have the generalized likelihood ratio test Fan, Zhang & Zhang (2001; Fan & Zhang (2003) in mind, where local linear estimators of the drift are compared to their parametric counterparts. Moreover, the specification test for the Vasiček model could be compared to the test based on generalized uniform residuals Hong & Li (2005).

Bibliography

Aguilar, O. & West, M. (2000). “Bayesian dynamic factor models and portfolio allocation”. *Journal of Business and Economic Statistics*, 18:338–357.

Aït-Sahalia, Y. (1996a). “Nonparametric Pricing of Interest Rate Derivative Securities”. *Econometrica*, 64:527–560.

Aït-Sahalia, Y. (1996b). “Testing Continuous-Time Models of the Spot Interest Rate”. *The Review of Financial Studies*, 9:385–426.

Aït-Sahalia, Y. (2002). “Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approximation Approach”. *Econometrica*, 70:223–262.

Aït-Sahalia, Y. (2003). “Closed-Form Likelihood Expansions for Multivariate Diffusions”. NBER working paper No. w8956.

Aït-Sahalia, Y.; Hansen, L. P. & Scheinkman, J. A. (2003). “Operator methods for continuous-time Markov models”. In Ait-Sahalia, Y. & Hansen, L. P., editors, *Handbook of Financial Econometrics*. North-Holland, Amsterdam. forthcoming.

Aït-Sahalia, Y. & Kimmel, R. (2004). “Maximum Likelihood Estimation of Stochastic Volatility Models”. NBER working paper No. w10579.

Andersen, T. & Sørensen, B. (1996). “GMM estimation of a stochastic volatility model: A Monte Carlo study”. *Journal of Business and Economic Statistics*, 14:329–352.

Andersen, T. G. & Bollerslev, T. (1998). “Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts”. *International Economic Review*, 39:885–905.

Andersen, T. G.; Bollerslev, T.; Diebold, F. & Labys, P. (2001). “The Distribution of Realized Exchange Rate Volatility”. *Journal of the American Statistical Association*, 96:42–55.

Andrews, D. W. K. (1991). “Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation”. *Econometrica*, 59:817–858.

Andrews, D. W. K. & Monahan, C. (1992). “An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator”. *Econometrica*, 60:953–966.

Bandi, F. M. & Phillips, P. C. B. (2003). "Fully Nonparametric estimation of Scalar Diffusion Models". *Econometrica*, 71:241–283.

Barndorff-Nielsen, O. E. & Cox, D. (1989). *Asymptotic Techniques for Use in Statistics.* Chapman and Hall.

Barndorff-Nielsen, O. E.; Jensen, J. L. & Sørensen, M. (1990). "Parametric Modelling of Turbulence". *Phil. Trans. R. Soc. Lond. A*, 332:439–455.

Barndorff-Nielsen, O. E.; Jensen, J. L. & Sørensen, M. (1998). "Some stationary processes in discrete and continuous time". *Advances in Applied Probability*, 30:989–1007.

Barndorff-Nielsen, O. E.; Kent, J. & Sørensen, M. (1982). "Normal variance-mean mixtures and z-distributions". *International Statistical Review*, 50:145–159.

Barndorff-Nielsen, O. E. & Shephard, N. (2001a). "Non-Gaussian Ornstein-Uhlenbeck-Based Models and some of their uses in Financial Econometrics (with discussion)". *Journal of the Royal Statistical Society* **B**, 63:167–241.

Barndorff-Nielsen, O. E. & Shephard, N. (2001b). "Superposition of Ornstein-Uhlenbeck type processes". *Theory of Probability and its Applications*, 45:175–194.

Barndorff-Nielsen, O. E. & Shephard, N. (2002). "Econometric analysis of realized volatility and its use in estimating stochastic volatility models". *Journal of the Royal Statistical Society* **B**, 64:253–280.

Barndorff-Nielsen, O. E. & Stelzer, R. (2006). "Positive-Definite Matrix Processes of Finite Variation". Research Report no. 11, Thiele Centre, University of Aarhus.

Berg, C. & Vignat, C. (2006). "Linearization coefficients of Bessel polynomials and properties of Student t-distributions". To appear in *Constr. Approx.*, published online DOI:10.1007/s00365-006-0643-6.

Beskos, A.; Papaspiliopoulos, O.; Roberts, G. & Fearnhead, P. (2006). "Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion)". *Journal of the Royal Statistical Society* **B**, 68:333–382.

Beskos, A. & Roberts, G. (2005). "Exact simulation of diffusions". *Annals of Applied Probability*, 15:2422–2444.

Bibby, B. M.; Jacobsen, M. & Sørensen, M. (2004). "Estimating functions for Discretely Sampled Diffusion-type Models". In Ait-Sahalia, Y. & Hansen, L. P., editors, *Handbook of Financial Econometrics.* North-Holland, Amsterdam. forthcoming.

Bibby, B. M.; Skovgaard, I. M. & Sørensen, M. (2005). "Diffusion-type Models with given Marginals and Autocorrelation Function". *Bernoulli*, 1:191–220.

Bibby, B. M. & Sørensen, M. (2003). "Hyperbolic processes in finance". In Rachev, S., editor, *Handbook of Heavy Tailed Distributions in Finance*, pages 211–248. Elsevier Science.

Bibby, B. M. & Sørensen, M. (1995). "Martingale Estimation Functions for Discretely Observed Diffusion Processes". *Bernoulli*, 1:17–39.

Billingsly, P. (1961). *Statistical Inference for Markov Processes.* The University of Chicago Press.

Bollerslev, T. & Zhou, H. (2002). "Estimating stochastic volatility diffusion using conditional moments of integrated volatility". *Journal of Econometrics*, 109:33–65.

Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes - Estimation and Prediction.* Springer-Verlag.

Brandt, M. W. & Santa-Clara, P. (2002). "Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets". *Journal of Financial Economics*, 63:161–210.

Brockwell, P. J. & Davis, R. A. (1991). *Time Series: Theory and Methods.* Springer Verlag, 2. ed. edition.

Chan, K. C.; Karolyi, A. G.; Longstaff, F. A. & Sanders, A. B. (1992). "Empirical Comparison of Alternative Models of the Short-Term Interest Rate". *Journal of Finance*, 47:1209–1227.

Chapman, D. A. & Pearson, N. D. (2000). "Is the Short Rate Drift Actually Nonlinear?". *The Journal of Finance*, 55:355–388.

Chen, B. & Hong, Y. (2005). "Diagnostics of Multivariate Continuous-Time Models with Application to Affine Term Structure Models". Working paper, Cornell and Xiamen University, submitted to Econometrica.

Chib, S. & Jeliazkov, I. (2001). "Marginal likelihood from the Metropolis-Hastings output". *Journal of the American Statistical Association*, 96:270–281.

Comte, F. & Genon-Catalot, V. (2006). "Penalized Projection Estimator for Volatility Density". *Scandinavian Journal of Statistics*, 33:875–893.

Cox, Jr., J. C.; Ingersoll, J. E. & Ross, S. A. (1985). "A Theory of the Term Structure of Interest Rates". *Econometrica*, 53:385–407.

Dacunha-Castelle, D. & Florens-Zmirou, D. (1986). "Estimation of the coefficient of a diffusion from discrete observations". *Stochastics*, 19:263–284.

Danielsson, J. (1998). "Multivariate stochastic volatility models: Estimation and comparison with VGARCH models". *Journal of Empirical Finance*, 5:155–173.

De Jong, F.; Drost, F. C. & Werker, B. J. M. (2001). "A jump-diffusion model for exchange rates in a target zone". *Statistica Neerlandica*, 55:270–300.

Dembo, A. & Peres, Y. (1994). "A topological criterion for hypothesis testing". *The Annals of Statistics*, 22(1):106–117.

Diggle, P. J.; Heagerty, P.; K.-Y., L. & Zeger, S. L. (2002). *Analysis of Longitudinal Data.* Oxford University Press, second edition.

Ditlevsen, P. D.; Ditlevsen, S. & Andersen, K. K. (2002). "The fast climate fluctuations during the stadial and interstadial climate states". *Annals of Glaciology*, 35:457–462.

Ditlevsen, S. & Sørensen, M. (2004). "Inference for observations of integrated diffusion processes". *Scandinavian Journal of Statistics*, 31:417–429.

Doukhan, P. (1994). *Mixing.* Springer-Verlag.

Durham, G. B. (2003). "Likelihood-Based Specification Analysis of Continuous-Time Models of the Shorth-Term Interest Rate". *Journal of Financial Economics*, 70.

Durham, G. B. & Gallant, R. A. (2002). "Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes (with discussion)". *Journal of Business and Economic Statistics*, 20.

Elerian; Chib & Shephard (2001). "Likelihood Inference for Discretely Observed Nonlinear Diffusions". *Econometrica*, 69:959–993.

Eraker, B. (2001). "Markov chain Monte Carlo analysis of diffusion models with application to finance". *Journal of Business and Economic Statistics*, 19:177–191.

Fan, J. (2005). "A Selective Overview of Nonparametric Methods in Financial Econometrics (with discussion)". *Statistical Science*, 20:317–357.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman and Hall.

Fan, J. & Yao, Q. (1998). "Efficient estimation of conditional variance functions in stochastic regression". *Biometrika*, 85:645–660.

Fan, J.; Yao, Q. & Tong, H. (1998). "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems". *Biometrika*, 83:189–206.

Fan, J. & Zhang, C. (2003). "A Reexamination of Diffusion Estimators With Applications to Financial Model Validation". *Journal of the American Statistical Association*, 98:118–134.

Fan, J.; Zhang, C. & Zhang, J. (2001). "Generalized likelihood ratio statistics and Wilks phenomenon". *The Annals of Statistics*, 29:153–193.

Feller, W. (1951). "Diffusion processes in genetics". In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statisitcs and Probability*, pages 227–246. University of California Press.

Florens-Zmirou, D. (1993). "On Estimating the Diffusion Coefficient from Discrete Observations". *Journal of Applied Probability*, 30:790–804.

Forman, J. L. (2005). "Least squares estimation for correlation parameters with applications to sums of Ornstein-Uhlenbeck type processes". AMS preprint 3.

Forman, J. L.; Markusen, B. & Sørensen, H. (2007). "Goodness of fit based on downsampling with applications to continuous-time processes". working paper, Department of Mathematical Sciences.

Forman, J. L. & Sørensen, M. (2006). "The Pearson Diffusions: A class of Statistically Tractable Diffusion Processes". AMS preprint 7.

Gallant, R. A. & Tauchen, G. (1996). "Which Moments to Match?". *Econometric Theory*, 12:657–681.

Gallant, R. A. & Tauchen, G. (2004). "Simulated score methods and indirect inference for continuous-time models". In Ait-Sahalia, Y. & Hansen, L. P., editors, *Handbook of Financial Econometrics*. North-Holland, Amsterdam. forthcoming.

Genon-Catalot, V.; Jeantheau, T. & Laredo, C. (1999). "Parameter Estimation for Discretely Observed Stochastic Volatility Models". *Bernoulli*, 5:855–872.

Genon-Catalot, V.; Jeantheau, T. & Laredo, C. (2000). "Stochastic volatility models as hidden Markov models and statistical applications". *Bernoulli*, 6:1051–1079.

Gloter, A. (2001). "Parameter estimation for a discrete sampling of an integrated Ornstein-Uhlenbeck process". *Statistics*, 35:225–243.

Gloter, A. (2006). "Parameter estimation for a discretely observed integrated diffusion process". *Scandinavian Journal of Statistics*, 33:83–104.

Gobet, E.; Hoffmann, M. & Reiß, M. (2004). "Nonparametric estimation of scalar diffusions based on low frequency data". *The Annals of Statistic*, 32:2223–2253.

Godambe, V. P. & Heyde, C. C. (1987). "Quasi likelihood and optimal estimation". *International Statistical Review*, 55:231–244.

Gourieroux, C.; Jasiak, J. & Sufana, R. (2004). "The Wishart Autoregressive Process of Multivariate Stochastic Volatility". forthcoming, Journal of Econometrics.

Gourieroux, C. & Monfort, A. (1994). "Testing non-nested hypotheses.". In Engle, R. F. & McFadden, D., editors, *The Handbook of Econometrics*, pages 2583–2637. North-Holland, Amsterdam.

Gourieroux, C.; Monfort, A. & Renault, E. (1993). "Indirect Inference". *Journal of Applied Econometrics*, 8:85–118.

Hall, A. R. (2000). "Covariance matrix estimation and the power of the overidentifying restrictions test". *Econometrica*, 68:1517–1527.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press.

Hall, A. R. & Inoue, A. (2003). "The large sample behaviour of the Generalized Method of Moments estimator in misspecified models". *Journal of Econometrics*, 114:361–394.

Hall, P. & Heyde, C. C. (1980). *Martingale Limit Theory and its Applications*. Academic Press, New York.

Hansen, B. E. (1992). "Consistent covariance matrix estimation for dependent heterogeneous processes". *Econometrica*, 60:967–972.

Hansen, L. P. (1982). "Large sample properties of generalized method of moments estimator". *Econometrica*, 50:1029–1054.

Hansen, L. P. & Scheinkman, J. A. (1995). "Back to the future: Generating moment implications for continuous-time markov processes". *Econometrica*, 63:767–804.

Hansen, L. P.; Scheinkman, J. A. & Touzi, N. (1998). "Spectral methods for identifying scalar diffusions". *Journal of Econometrics*, 86:1–32.

Harvey, A.; Ruiz, E. & Shephard, N. (1994). "Multivariate stochastic variance models". *Review of Economic Studies*, 61:247–264.

Hausman, J. (1978). "Specification tests in econometrics". *Econometrica*, 46:1251–1271.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application.* Springer-Verlag, New York.

Hoffmann, M. (1999). "L_p Estimation of the Diffusion coefficient". *Bernoulli*, 5:447–481.

Hoffmann, M. (2002). "Rate of convergence for parametric estimation in a stochastic volatility model". *Stochastic Processes and their Applications*, 97:147–170.

Hong, Y. & Li, H. (2005). "Nonparametric Specification Testing for Continuous-Time Models with Applications to Term Structure of Interest Rates". *The Review of Financial Studies*, 18.

Hull, J. & White, A. (1987). "The pricing of options on assts with stochastic volatilities". *The Journal of Finance*, 42.

Jacobsen, M. (2001). "Discretely Observed Diffusions; Classes of Estimating Functions and Small Δ-optimality". *Scandinavian Journal of Statistics*, 28:123–150.

Jacobsen, M. (2002). "Optimality and small Δ-optimality of martingale estimating functions". *Bernoulli*, 8:643–668.

Jacod, J. & Sørensen, M. (2007). "Asymptotic statistical theory for stochastic processes: A review.". Preprint, Department of Applied Mathematics and Statistics, University of Copenhagen. In preparation.

Jacquier, E.; Polson, N. G. & Rossi, P. E. (1994). "Bayesian analysis of stochastic volatility models (with discussion)". *Journal of Busimess and Economic Statistics*, 12:371–417.

Jansson, M. (2002). "Consistent covariance matrix estimation for linear processes". *Econometric Theory*, 18:1449–1459.

Jensen, B. & Poulsen, R. (2002). "Transition densities of diffusion processes: Numerical comparison of approximation techniques". *Journal of Derivatives*, 9:18–32.

Karatzas, I. & Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus.* Springer-Verlag, second edition.

Karlin, S. & Taylor, H. M. (1981). *A Second Course in Stochastic Processes.* Academic Press.

Kauermann, G. & Carroll, R. J. (2004). "A note on the efficiency of sandwich covariance matrix estimation". *Journal of the American Statistical Association*, 96(46):1387–1396.

Kessler, M. (1996). *Estimation paramétrique des coefficients d'une diffusion ergodique à partir d'observations discrètes.* Ph.d. thesis, Laboratoire de Probabilités, Université Paris VI.

Kessler, M. (2000). "Simple and Explicit Estimating Functions for a Discretely Observed Diffusion Process". *Scandinavian Journal of Statistics*, 27:65–82.

Kessler, M. & Sørensen, M. (1999). "Estimating equations based on eigenfunctions for a discretely observed diffusion". *Bernoulli*, 5:299–314.

Kloeden, P. E. & Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations.* Springer-Verlag, New York, 3rd revised printing edition.

Larsen, K. S. & Sørensen, M. (2003). "A diffusion model for exchange rates in a target zone". Preprint No. 6, Department of Applied Mathematics and Statistics, University of Copenhagen. To appear in *Mathematical Finance.*

Ledoux, M. & Talagrand, M. (1991). *Probability on Banach Spaces.* Springer-Verlag, Berlin Heidelberg.

Liang, K. Y.; Zeger, S. L. & Qaqish, B. (1992). "Multivariate Regression Analysis for Categorical Data (with discussion)". *Journal of the Royal Statistical Society* **B**, 54:3–40.

Liesenfeld, R. & Richard, J. (2003). "Univariate and multivariate stochastic volatility models: estimation and diagnostics". *Journal of Empirical Finance*, 10:505–531.

Lo, A. W. (1988). "Maximum Likelihood Estimation of Generalized Ito Processes with Discretely Sampled Data". *Econometric Theory*, 4:231–247.

Masuda, H. (2004). "On multidimensional Ornstein-Uhlenbeck processes driven by a genral Levy process". *Bernoulli*, 10(1):97–120.

Melino, A. & Turnbull, S. M. (1990). "Pricing foreign currency options with stochastic volatility". *Journal of Econometrics*, 45:239–265.

Nagahara, Y. (1996). "Non-Gaussian Distribution for Stock Returns and related Stochastic Differential Equation". *Financial Engineering and the Japanese Markets*, 3:121–149.

Nahapetian, B. (1991). *Limit Theorems and Some Applications in Statistical Physics.* Teubner-Texte zur Mathematik.

Nelson, D. B. (1990). "ARCH models as diffusion approximations". *Journal of Econometrics*, 45:7–38.

Newey, W. K. (1985). "Generalized Method of Moments specification testing". *Journal of Econometrics*, 29:229–256.

Newey, W. K. & West, K. D. (1987a). "Hypothesis testing with efficient Method of Moments testing". *International Economic Review*, 28:777–787.

Newey, W. K. & West, K. D. (1987b). "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix". *Econometrica*, 55:703–708.

Newey, W. K. & West, K. D. (1994). "Automatic Lag-selection in Covariance Matrix Estimation". *Review of Economic Studies*, 61:631–653.

Pearson, K. (1895). "Contributions to the Mathematical Theory of Evolution II. Skew Variation in Homogeneous Material". *Philosophical Transactions of the Royal Society of London. A*, 186:343–414.

Pedersen, A. R. (1994). "Uniform residuals for discretely observed diffusion processes". Research Report No. 295, Department of Theoretical Statistics, University of Aarhus.

Pedersen, A. R. (1995a). "Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes". *Bernoulli*, 1:257–279.

Pedersen, A. R. (1995b). "A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations". *Scandinavian Journal of Statistics*, 22:55–71.

Philipov, A. & Glickman, M. E. (2006). "Multivariate Stochastic Volatility via Wishart Processes". forthcoming, Journal of Busines and Economic Statistics.

Pitt, M.; Chib, S. & Shephard, N. (2006). "Likelihood based inference for diffusion driven state space models". working paper, Nuffield College, Oxford.

Pitt, M. & Shephard, N. (1999a). "Filtering via simulation: auxiliary particle filter.". *Journal of the American Statistical Association*, 94:590–599.

Pitt, M. & Shephard, N. (1999b). "Time varying covariances: A factor stochastic volatility approach". In Bernardo; Berger; David & Smith, editors, *Bayesian Statistics*, volume 6, pages 547–570. Oxford University Press.

Pritsker, M. (1998). "Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models". *The Review of Financial Studies*, 11:449–487.

Richard, J.-F. & Zhang, W. (1998). "Efficient High-Dimensional Monte Carlo Importance Sampling". Unpublished paper, University of Pittsburgh, Dept. of Economics.

Ritz, C. (2000). "Markov processes and their infinitesimal operators with an application to estimating functions". Masters thesis, Department of Theoretical Statistics, University of Copenhagen.

Roberts, G. & Stramer, O. (2001). "On inference for non-linear diffusion models using the Metropolis-Hastings algorithms". *Biometrika*, 88:603–621.

Rogers, L. & Williams, D. (1987). *Diffusions, Markov Processes, and Martingales*, volume 2: *Itô Calculus*. John Wiley and Sons.

Rosenblatt, M. (1952). "Remarks on a Mulivariate Transformation". *Annals of Mathematical Statistics*, 23:470–472.

Shephard, N. (2005). *Stochastic Volatility: Selected Readings.* Oxford University Press.

Sørensen, H. (2003). "Simulated Likelihood Approximations for Stochastic Volatility Models". *Scandinavian Journal of Statistics*, 30:257–276.

Sørensen, H. (2004). "Parametric inference for diffusion processes observed at discrete points in time: a survey". *International Statistical Review*, 72:337–354.

Sørensen, M. (1997). "Estimating Functions for Discretely Observed Diffusions: A Review". In Basawa, I. V.; Godambe, V. P. & Taylor, R. L., editors, *Selected Proceedings of the Symposium on Estimating Functions*, pages 305–325. Hayward: Institute of Mathematical Statistics. IMS Lecture Notes – Monograph Series, Vol. 32.

Sørensen, M. (1999). "On asymptotics of Estimating Functions". *Brazilian Journal of Probability and Statistics*, 13:111–136.

Sørensen, M. (2000). "Prediction-Based Estimating Functions". *Econometrics Journal*, 3:123–147.

Sørensen, M. (2007). "Efficient martingale estimating functions for discretely sampled diffusions". In preparation.

Stanton, R. (1997). "A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk". *The Journal of Finance*, 52:1973–2002.

van Es, B.; Spreij, P. & van Zanten, H. (2003). "Nonparametric volatility density estimation". *Bernoulli*, 9:451–465.

Vasicek, O. A. (1977). "An Equilibrium Characterization of the Term Structure". *Journal of Financial Economics*, 5:177–188.

White, H. (1984). *Asymptotic Theory for Econometricians.* Academic press.

Wong, E. (1964). "The Construction of a Class of Stationary Markoff Processes". In Bellman, R., editor, *Stochastic Processes in Mathematical Physics and Engineering*, pages 264–276. American Mathematical Society, Rhode Island.

Yu, J. & Meyer, R. (2004). "Multivariate stochastic volatility models: Bayesian estimation and model comparison". Working paper, Singapore Management University.

Zeileis, A. (2004). "Econometric Computing with HC and HAC Covariance Matrix Estimators". *Journal of Statistical Software*, 11.