Inference for Diffusion Processes and Stochastic Volatility Models

Ph.D. thesis

Helle Sørensen

University of Copenhagen
September 2000

Inference for Diffusion Processes and Stochastic Volatility Models

Ph.D. thesis

Helle Sørensen
Department of Statistics and Operations Research
Institute for Mathematical Sciences
Faculty of Science
University of Copenhagen

Thesis advisor:	Martin Jacobsen, University of Copenhagen
Thesis committee:	Michael Sørensen, University of Copenhagen
	Uwe Küchler, Humboldt University of Berlin
	Bo Martin Bibby, KVL, Copenhagen

Helle Sørensen
Department of Statistics and Operations Research
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen East
Denmark
hsoeren@math.ku.dk
http://www.math.ku.dk/~hsoeren

Preface

This thesis has been prepared in partial fulfillment of the requirements for the Ph.D. degree at the Department of Statistics and Operations Research, Institute for Mathematical Sciences at the University of Copenhagen. The work has been carried out in the period from May 1997 to July 2000 with Martin Jacobsen as thesis advisor.

The thesis contains a brief overall introduction, two introductory chapters and three papers. The introductory chapters have been prepared for this thesis exclusively whereas the papers have been (or will shortly be) submitted for publication. Each chapter and paper is self-contained and can be read independently from the rest. The first page of each of the three papers contain an abstract and details on publication. Page numbers *within* the papers are given in parentheses at the bottom of each page, underlining that the papers have been prepared and written separately. To emphasize the unity of the thesis, pages are also numbered consecutively (at the top of each page) and the lists of references are collected in *one* bibliography placed at the end of the thesis.

The present version differs from the original one which was submitted for the Ph.D. degree on July 20, 2000, by this preface and in that a minor number of typos and misprints have been corrected.

Acknowledgements

I would like to thank my supervisor Martin Jacobsen for his encouragement and for numerous valuable suggestions and discussions during the last three years. Thanks are also due to Martin for careful reading of earlier versions of the chapters and papers. Also, I would like to thank Søren Feodor Nielsen for his ideas and help on empirical processes. Thanks are also due to everyone at the department for making everyday life enjoyable.

Part of the work was done while I was visiting Department of Statistics at University of California, Berkeley. I thank everyone there for making it such a pleasant stay.

Jens Lund, Bo Markussen, Søren Feodor Nielsen, Henning Niss and Martin Richter have all read various parts of the manuscript. I am grateful for their comments and constructive critics.

Copenhagen, September 2000

Helle Sørensen

Summary

Diffusion processes have a wide range of applications. In physics and biology they are used for modeling phenomena assumed to evolve randomly and continuously in time. In mathematical finance they are used for modeling various price processes. Data are essentially always sampled at discrete points in time only. This leaves the statistician in a dilemma because the few models that are easy to handle statistically, in general do not describe data adequately. For example, it is well known that stock price data usually violate the assumptions of the geometric Brownian motion (or in finance terms, the Black-Scholes model) classically used for stock price modeling. For more complicated models maximum likelihood estimation is usually not possible because the discrete-time transitions implicitly defined by the continuous-time model are not known analytically. Consequently, there is a need for alternative statistical methods.

The first part of this thesis (Chapter 2 and Papers I and II) is about *parametric inference for stationary and ergodic diffusion processes* with general, often non-linear, specifications of the drift and diffusion functions. Chapter 2 provides an overview of existing techniques with emphasis on estimating functions. Furthermore, new results on identification for martingale estimating functions are presented. In Paper I a simple, explicit approximation of the continuous-time score function is derived in terms of the infinitesimal generator and the invariant density. As opposed to the usual Riemann-Itô approximation, it is unbiased and provides consistent estimators. Paper II presents a method suitable for estimation of parameters in the diffusion term. It is based on a functional relationship between the drift, the diffusion function and the invariant density, and provides satisfactory estimates in the difficult CKLS model. The usual limit theory does not apply; instead empirical process theory is employed in order to prove asymptotic properties of the estimator.

The second part of the thesis (Chapter 3 and Paper III) is about *parametric inference for stochastic volatility models*, that is, two-dimensional diffusion models with a special structure and one of the coordinates unobservable. The introduction of a latent process makes it possible to retain a simple (linear) structure of the model and still create the complex data structures known from empirical studies. However, it also complicates the statistical analysis because the model is only partially observed. Chapter 3 provides an introduction to stochastic volatility models with special emphasis on four particular models and on statistical analysis. A comparison of different models shows that the increments of the observable process can have almost identical distributions although the underlying latent processes

are specified quite differently. Still, the models differ in their ability to create highly leptokurtic distributions. The overview of estimation methods covers a wide range of techniques from simple moment-based methods to quite complicated techniques relying on very intensive computations. In Paper III a new approximate maximum likelihood method is presented. The idea is to pretend that the increments of the observable process form a k'th order Markov chain for some relatively small k. The corresponding approximate score function is unbiased, and the estimators therefore consistent, for each fixed k because the *true* conditional distributions given the k previous observations are used. These conditional densities are not known analytically but can be computed by simulation. The method makes it thereby possible to compute quite natural approximations to the likelihood function.

Dansk resumé

Diffusionsprocesser har anvendelsesmuligheder indenfor adskillige fagområder. De benyttes til beskrivelse af fænomener der varierer kontinuert og stokastisk over tid, for eksempel i fysik og biologi. De benyttes også intensivt i matematisk finansiering til beskrivelse af prisfluktuationer på forskellige finansielle aktiver. Uanset antagelsen om kontinuert variation er observationer af processerne dog altid diskrete af natur idet målinger foretages på endeligt mange, adskilte tidspunkter. Dette komplicerer den statistiske analyse betydeligt fordi overgangssandsynlighederne, implicit defineret af modellen, kun er kendt analytisk for ganske få modeller. Disse modeller er som regel for simple til at beskrive strukturen i de observerede data tilfredsstillende. For eksempel er det velkendt at faktisk observerede aktiekurser er i klar modstrid med den geometriske brownske bevægelse (eller med terminologi fra finansiering: Black-Scholes modellen) som ellers klassisk set er blevet brugt som model for aktiekurser. Det er med andre ord sjældent muligt udføre maksimaliseringsestimation, og der er således behov for alternative estimationsmetoder.

Afhandlingens første del (kapitel 2 og artikel I og II) handler om *parametrisk inferens for generelle stationære og ergodiske diffusionsprocesser*. Kapitel 2 giver en oversigt over eksisterende estimationsmetoder med hovedvægt på teorien for estimationsfunktioner. Udover en redegørelse for velkendte metoder og resultater præsenteres også et nyt resultat om identifikation for martingalestimationsfunktioner. I artikel I udledes en simpel, eksplicit approksimation af scorefunktionen hørende til en observation i kontinuert tid. Approksimationen er en central estimationfunktion og giver derfor, til forskel fra den sædvanlige Riemann-Itô approksimation, konsistente estimatorer. I artikel II beskrives en metode til estimation af parametre i diffusionsfunktionen. Metoden er baseret på en punktvis sammenhæng mellem driftfunktionen, diffusionsfunktionen og tætheden for den stationære begyndelsesfordeling, og den giver fornuftige estimater i den ellers vanskelige CKLS model. De klassiske grænsesætniger kan ikke anvendes; i stedet benyttes teorien om empiriske processer til at bevise asymptotiske egenskaber for estimatorerne.

Afhandlingens anden del (kapitel 3 og artikel III) handler om *parametrisk inferens for stokastiske volatilitetsmodeller*, dvs. todimensionale diffusionsmodeller der har en speciel form og hvor kun den ene af koordinaterne er observerbar. Indførelsen af den ekstra proces gør det muligt at frembringe fænomenerne kendt fra empiriske analyser ved hjælp af relativt simple (lineære) modeller, men den statistiske analyse kompliceres fordi modellen kun observeres partielt. Kapitel 3

er en introduktion til stokastiske volatilitetsmodeller med særligt henblik på fire specifikke modeller og på statistisk analyse. En sammenligning viser at forskellige modeller for den ikke-observerbare process kan frembringe næsten identiske fordelinger for tilvæksterne af den observerbare process, men at modellerne adskiller sig fra hinanden ved deres evne til at skabe tilvækster med meget tunge haler. Oversigten over estimationsmetoder for stokastiske volatilitetsmodeller spænder fra enkle momentbaserede metoder til ganske komplicerede og meget beregningskrævende metoder. I artikel III præsenteres en ny approksimativ maximumlikelihoodmetode. Ideen er at opføre sig som om tilvæksterne for den observerbare process udgør en markovkæde af orden k for et relativt lille k. Centraliteten af den tilsvarende scorefunktion bibeholdes såfremt de *sande* betingede tætheder givet de k foregående observationer benyttes. Således bliver estimatoren konsistent og asymptotisk normalfordelt for ethvert fast k. De betingede tætheder er ikke kendt analytisk men kan beregnes ved simulation. Metoden gør det dermed muligt at beregne naturlige approksimationer til likelihoodfunktionen.

Table of Contents

1 Introduction

Diffusion models have a large range of applications. They have been used for a long time to model phenomena evolving randomly and continuously in time, *e.g.* in physics and biology. During the last thirty years or so the models have also been applied intensively in mathematical finance for describing stock prices, exchange rates, interest rates, *etc.* (although it is well-known that such quantities do not really change continuously in time).

Data are essentially always recorded at discrete points in time only (*e.g.* weekly, daily or each minute) and can thus be interpreted as time series data. Still, continuous-time models are often preferred to classical time series models. There are (at least) two reasons for this. First, if data are sampled at irregularly spaced time-points, then an appropriate discrete-time model should incorporate this explicitly. As opposed to this, continuous-time models implicitly define transitions over time intervals of any length in a consistent way. For example, missing data in a sample where time-points for observations are otherwise regularly spaced, do not give rise to serious problems in the continuous-time setting as they are treated just like the values not observed due to discrete-time sampling. Second, all the machinery from stochastic calculus is at our disposal when we use diffusion models. This has proved important in finance theory where derivation of various price formulas usually relies heavily on this theory.

Thus convinced that diffusion models are important and useful alternatives to classical time series models I turn to the statistical analysis. I shall be concerned with parametric inference exclusively. For a few models, estimation is straightforward because the corresponding stochastic differential equation can be solved explicitly. This is the case for the geometric Brownian motion, the Ornstein-Uhlenbeck process and the square-root process which have log-normal, normal and non-central chi-square transition probabilities respectively. However, "nature" (or "the market") most often generates data not adequately described by such simple models. For example, empirical studies clearly reveal that increments of logarithmic stock prices are not independent and Gaussian as implied by the geometric Brownian motion classically used for stock price modeling. Rather, they exhibit temporal dependence and leptokurtosis. Consequently, more complex models are needed in order to obtain reasonable agreement with data. This complicates the statistical analysis considerably because the discrete-time transitions (implicitly defined by the model) are no longer known analytically. Specifically, *the likelihood function is usually not tractable*. In other words, one has to use models for which likelihood analysis is not possible, and there is consequently a need for alternative

methods.

In this thesis I am concerned with parametric inference for two types of generalizations of the above simple models, namely (one-dimensional) diffusion models with more general, typically non-linear, specifications of drift and diffusion functions, and continuous-time stochastic volatility models. By the latter I mean two-dimensional diffusion processes with a special structure and one of the coordinates unobservable. The introduction of an extra, latent process makes it possible to retain a simple (linear) structure of the stochastic differential equation for the observable process and still create the characteristic features known from empirical studies. However, the extra process also complicates the statistical analysis because the model is only partially observed.

Further introductory comments on the two model types and the corresponding estimation problems are given in the beginning of Chapters 2 and 3.

Structure of the thesis

My main contributions in this thesis are contained in three papers: Papers I and II on (pure) diffusion models and Paper III on stochastic volatility models. In addition I provide two introductory chapters: Chapter 2 on diffusions and Chapter 3 on stochastic volatility models. The aim of the two introductory chapters is mainly to provide overviews of existing estimation methods, but they also contain a few new results. I do not know of any review papers with quite the same focus. The chapters and papers may be read independently. This has the unfortunate consequence that models, notation, *etc.* are defined several times. Attempts have been made in order to customize notation; still, there may be slight differences which should cause no confusion. The lists of references have been collected to *one* bibliography placed at the end of the thesis.

Estimation in (pure) diffusion models. **Chapter 2** provides an overview of existing estimation techniques for stationary and ergodic diffusion processes. Main emphasis is on estimating functions, in particular on martingale estimating functions and so-called simple estimating functions. Well-known properties and results are reviewed, and some some new results concerning identification for martingale estimating functions are presented: one of the regularity conditions needed in order for the estimator to be asymptotically well-behaved is explained in terms of reparametrizations. In addition to estimating functions, the chapter covers three approximate maximum likelihood methods, Bayesian analysis and methods based on auxiliary models.

Papers I and II contain my main contributions in the area of estimation in diffusion models. Brief reviews are given in Sections 2.3.2 and 2.7. In **Paper I** (*Discretely Observed Diffusions: Approximation of the Continuous-time Score Function*) I study how the structure of the continuous-time score function can be used when only discrete-time observations are available. The usual Riemann-Itô approximation is biased; I derive an alternative, unbiased approximation in terms of

the infinitesimal generator and the invariant density. The approximation is an explicit, so-called simple estimating function; is is invariant to data transformations; and it provides consistent and asymptotically normal estimators as the number of observations increases (for any fixed time interval between observations). The approach carries over to multi-dimensional diffusions (to some extent at least), and I study a few examples where the method works very well.

In **Paper II** (*Estimation of Diffusion Parameters for Discretely Observed Diffusion Processes*) I discuss a method suitable for estimation of parameters in the diffusion term when the drift is known. It is based on a functional relationship between the drift, the diffusion function and the invariant density. I apply the method to simulated data from the relatively difficult CKLS model and get satisfactory estimates. The estimators are probably not efficient, though. From a theoretical point of view the derivation of asymptotic results is perhaps most interesting. The usual limit theory does not apply; instead I employ empirical process theory. I am not aware of other applications of empirical process theory to problems related to discretely observed diffusions.

Stochastic volatility models. **Chapter 3** is an introduction to stochastic volatility models in continuous time. I study four particular models in detail and conclude that they mainly differ in their ability to create processes for which the increments are highly leptokurtic. If parameter values are chosen appropriately, then the models are hard to distinguish. I do not know of any similar comparisons in the literature. Chapter 3 also provides an overview of existing estimation methods, some of which are developed very recently. The overview covers moment methods, approximations to the marginal distribution of the increments, prediction-based estimating functions, Bayesian analysis, indirect inference and EMM, and a filtering-based method. Strikingly, most methods are extremely computationally intensive.

My main contribution consists of a new approximate maximum likelihood method, developed in **Paper III** (*Simulated Likelihood Approximations for Stochastic Volatility Models*) and reviewed in Section 3.4.7. The method provides a sequence of approximations to the likelihood function. For the k'th approximation, the idea is to pretend that the increments of the observable process form a k'th order Markov chain. The corresponding approximate score function is unbiased because the *true* conditional distributions given the k previous observations are used. For any fixed k the estimator is invariant to transformations of data, consistent and asymptotically normal (for any fixed time interval between observations). There is no closed-form expression for the approximate likelihood function (just as for the true likelihood function) but it can be computed by simulation. I apply the method to simulated data in Paper III and to Microsoft stock price data in Section 3.4.7.

Finally, let me stress that although diffusion-type models are perhaps most widely applied in finance these days, and although the applications mentioned

originate from finance, *the focus of this thesis is purely statistical!* My main interest in the models lies in their statistical properties rather than their financial applications.

2 Inference for diffusion processes

Statistical inference for diffusion processes has been an active research area during the last two or three decades. The work has developed from estimation of linear systems from continuous-time observations (see Le Breton (1974) and the references therein) to estimation of non-linear systems (parametric or non-parametric) from discrete-time observations. In this chapter, as well as in Papers I and II, we shall be concerned with *parametric inference for discrete-time observations* exclusively. The models may be linear or non-linear.

This branch of research commenced in the mid eighties (with the paper by Dacunha-Castelle & Florens-Zmirou (1986) on the loss of information due to discretization as an important reference) and accelerated in the nineties. Important references from the mid of the decade are Bibby & Sørensen (1995) on martingale estimating functions, Gourieroux, Monfort & Renault (1993) on indirect inference, and Pedersen (1995*b*) on approximate maximum likelihood methods, among others. Later work includes Bayesian analysis (Elerian, Chib & Shephard 2000) and further approximate likelihood methods (Aït-Sahalia 1998, Poulsen 1999).

Ideally, the parameter should be estimated by maximum likelihood but, except for a few models, the likelihood function is not available analytically. In this chapter we review some of the alternatives proposed in the literature. There exist review papers on estimation via estimating functions (Bibby & Sørensen 1996, Sørensen 1997), but we do not know of any surveys covering all the techniques discussed in this chapter.

Papers I and II contain my main contributions in this area. Furthermore, there are some new results on identification for martingale estimating functions in Section 2.3.1. In Paper I we discuss a particular estimating function derived as an approximation to the continuous-time score function. The estimating function is of the so-called simple type, it is unbiased and invariant to data transformations and provides consistent and asymptotically normal estimators. In Paper II we discuss a method suitable for estimation of parameters in the diffusion term when the drift is known. It is based on a functional relationship between the drift, the diffusion function and the invariant density, and provides asymptotically well-behaved estimators. The asymptotic results are proved using empirical process theory.

In the following we focus on fundamental ideas and refer to the literature for rigorous treatments. In particular, we consider one-dimensional diffusions only, although most methods apply in the multi-dimensional case as well. Also, we do not account for technical assumptions, regularity conditions *etc.* An exception is

Section 2.3.1, though, where the new identification results are presented.

The chapter is organized as follows. The model is defined in Section 2.1, and Section 2.2 contains preliminary comments on the estimation problem. Section 2.3 is about estimating functions with special emphasis on martingale estimating functions and so-called simple estimating functions, including the one from Paper I. In Sections 2.4 we discuss three approximations of the likelihood which can in principle be made arbitrarily accurate, and Section 2.5 is about Bayesian analysis. In Section 2.6 we discuss indirect inference and EMM which both introduce auxiliary (but wrong) models and correct for the implied bias by simulation. The method from Paper II is reviewed in Section 2.7 and conclusions are finally drawn in Section 2.8.

2.1 Model, assumptions and notation

In this section we present the model and the basic assumptions, and introduce notation that will be used throughout the chapter. We consider a one-dimensional, time-homogeneous stochastic differential equation

$$dX_t = b(X_t,\theta)\,dt + \sigma(X_t,\theta)\,dW_t \tag{2.1}$$

defined on a filtered probability space $(\Omega,\mathscr{F},\mathscr{F}_t,Pr)$. Here, W is a one-dimensional Brownian motion and θ is an unknown p-dimensional parameter from the parameter space $\Theta\subseteq\mathbb{R}^p$. The true parameter value is denoted θ_0. The functions $b:\mathbb{R}\times\Theta\to\mathbb{R}$ and $\sigma:\mathbb{R}\times\Theta\to(0,\infty)$ are known and assumed to be suitably smooth.

The state space is denoted $I=(l,r)$ for $-\infty\le l<r\le+\infty$ (implicitly assuming that it is open and the same for all θ). We shall assume that for any $\theta\in\Theta$ and any $\mathscr{F}_0$-measurable initial condition U with state space I, equation (2.1) has a unique strong solution X with $X_0=U$. Assume furthermore that there exists an *invariant distribution* $\mu_\theta=\mu(x,\theta)dx$ such that the solution to (2.1) with $X_0\sim\mu_\theta$ is strictly stationary and ergodic. It is well-known that sufficient conditions for this can be expressed in terms of the scale function and the speed measure (see Section II.2, or the textbook by Karatzas & Shreve (1991)), and that $\mu(x,\theta)$ is given by

$$\mu(x,\theta)=\big(M(\theta)\sigma^2(x,\theta)s(x,\theta)\big)^{-1} \tag{2.2}$$

where $\log s(x,\theta)=-2\int_{x_0}^x b(y,\theta)/\sigma^2(y,\theta)\,dy$ for some $x_0\in I$ and $M(\theta)$ is a normalizing constant.

For all $\theta\in\Theta$ the distribution of X with $X_0\sim\mu_\theta$ is denoted by P_θ. Under P_θ all $X_t\sim\mu_\theta$. Further, let for $t\ge0$ and $x\in I$, $p_\theta(t,x,\cdot)$ denote the conditional density (transition density) of X_t given $X_0=x$. Since X is time-homogeneous $p_\theta(t,x,\cdot)$ is actually the density of X_{s+t} conditional on $X_s=x$ for *all* $s\ge0$. Note that the transition probabilities are most often analytically intractable whereas the invariant density is easy to find (at least up the normalizing constant).

We are going to need some matrix notation: Vectors in $\mathbb{R}^p$ are considered as $p\times1$ matrices and A^T is the transpose of A. For a function $f=(f_1,\ldots,f_q)^T$:

$\mathbb{R} \times \Theta \to \mathbb{R}^q$ we let $f'(x,\theta)$ and f'' denote the matrices of first and second order partial derivatives with respect to x, and $\dot{f}(x,\theta) = \partial_\theta f(x,\theta)$ denote the $q \times p$ matrix of partial derivatives with respect to θ, *i.e.* $\dot{f}_{jk} = \partial f_j / \partial \theta_k$, assuming that the derivatives exist.

Finally, introduce the differential operator $\mathscr{A}_\theta$ given by

$$\mathscr{A}_\theta f(x,\theta) = b(x,\theta) f'(x,\theta) + \tfrac{1}{2}\sigma^2(x,\theta) f''(x,\theta) \tag{2.3}$$

for twice continuously differentiable functions $f : \mathbb{R} \times \Theta \to \mathbb{R}$. When restricted to a suitable subspace, $\mathscr{A}_\theta$ is the *infinitesimal generator* of X (see Rogers & Williams (1987), for example).

2.2 Preliminary comments on estimation

The objective of this chapter is estimation of the parameter θ. First note that if X is observed *continuously* from time zero to time T then parameters from the diffusion coefficient can be determined (rather than estimated) from the quadratic variation process of X, and the remaining part can be estimated by maximum likelihood: if the diffusion function is completely known, that is $\sigma(x,\theta) = \sigma(x)$, then the likelihood function for $X_{0 \le t \le T}$ is given by

$$L_T^c(\theta) = \exp\left(\int_0^T \frac{b(X_s,\theta)}{\sigma^2(X_s)}\, dX_s - \frac{1}{2}\int_0^T \frac{b^2(X_s,\theta)}{\sigma^2(X_s)}\, ds \right). \tag{2.4}$$

An informal argument for this formula is given below; for a proper proof see Lipster & Shiryayev (1977, Chapter 7).

From now on we shall consider the situation where X is observed at discrete time-points only. For convenience we consider equidistant time-points $\Delta, 2\Delta, \ldots, n\Delta$ for some $\Delta > 0$. Conditional on the initial value X_0, the likelihood function is given as the product

$$L_n(\theta) = \prod_{i=1}^n p_\theta(\Delta, X_{(i-1)\Delta}, X_{i\Delta})$$

because X is Markov. Ideally, θ should be estimated by the value maximizing $L_n(\theta)$, but since the transition probabilities are not analytically known, neither is the likelihood function.

There are a couple of obvious, very simple alternatives which unfortunately are not satisfactory. First, one could ignore the dependence structure and simply approximate the conditional densities by the marginal density. Then all information due to the time evolution of X is lost, and it is usually not possible to estimate the full parameter vector. See Section 2.3.2 for further details.

As a second alternative, one could use the *Euler scheme* (or some higher-order scheme) given by the approximation

$$X_{i\Delta} \approx X_{(i-1)\Delta} + b(X_{(i-1)\Delta},\theta)\Delta + \sigma(X_{(i-1)\Delta},\theta)\sqrt{\Delta}\varepsilon_i$$

where ε_i, $i = 1, \ldots, n$ are independent, identically $N(0,1)$-distributed. This approximation is good for small values of Δ but may be bad for larger values. The approximation is two-fold: the moments are not the true conditional moments, and the true conditional distribution need not be Gaussian. The moment approximation introduces bias implying that the corresponding estimator is inconsistent as $n \to \infty$ for any fixed Δ (Florens-Zmirou 1989). The Gaussian approximation introduces no bias per se, but usually implies inefficiency: if the conditional mean and variance are replaced by the true ones, but the Gaussian approximation is maintained, then the corresponding approximation to the score function is a non-optimal martingale estimating function, see Section 2.3.1.

Note that the Euler approximation provides an informal explanation of formula (2.4): if σ does not depend on θ, then the Euler approximation to the discrete-time likelihood function is given by (except for a constant)

$$\exp\left\{\sum_{i=1}^{n} \frac{b(X_{(i-1)\Delta}, \theta)}{\sigma^2(X_{(i-1)\Delta})}\left(X_{i\Delta} - X_{(i-1)\Delta}\right) - \frac{1}{2}\Delta\sum_{i=1}^{n} \frac{b^2(X_{(i-1)\Delta}, \theta)}{\sigma^2(X_{(i-1)\Delta})}\right\} \tag{2.5}$$

which is the Riemann-Itô approximation of (2.4).

2.3 Estimating functions

Estimating functions provide estimators in very general settings where an unknown p-dimensional parameter θ is to be estimated from data X^{obs} of size n. Basically, an estimating function F_n is simply a $\mathbb{R}^p$-valued function which takes the data as well as the unknown parameter as arguments. An estimator is obtained by solving $F_n(X^{\text{obs}}, \theta) = 0$ for the unknown parameter θ. General theory for estimating functions may be found in Heyde (1997) or Sørensen (1998*b*).

The prime example of an estimating function is of course the score function, yielding the maximum likelihood estimator. When the score function is not available an alternative estimating function should of course be chosen with care. In order for the corresponding estimator to behave (asymptotically) "nicely" it is crucial that the estimating function is unbiased and is able to distinguish the true parameter value from other values of θ:

$$\mathrm{E}_{\theta_0} F_n(X^{\text{obs}}, \theta) = 0 \quad \text{if and only if} \quad \theta = \theta_0. \tag{2.6}$$

Now, let us turn to the case of discretely observed diffusions again. The score function

$$S_n(\theta) = \partial_\theta \log L_n(\theta) = \sum_{i=1}^{n} \partial_\theta \log p_\theta(\Delta, X_{(i-1)\Delta}, X_{i\Delta})$$

is a sum of n terms where the i'th term depends on data through $(X_{(i-1)\Delta}, X_{i\Delta})$ only. As we are trying to mimic the behaviour of the score function, it is natural

to look for estimating functions with the same structure. Hence, we shall consider estimating functions of the form

$$F_n(\theta) = \sum_{i=1}^{n} f(X_{(i-1)\Delta}, X_{i\Delta}, \theta) \tag{2.7}$$

where we have omitted the dependence of data on F_n from the notation. Condition (2.6) simplifies to: $\mathrm{E}_{\theta_0} f(X_0, X_\Delta, \theta) = 0$ if and only if $\theta = \theta_0$.

Sørensen (1997) and Jacobsen (1998) provide overviews of estimating functions in the diffusion case. In the following we shall concentrate on two special types, namely *martingale estimating functions* ($F_n(\theta)$ being a P_θ-martingale) and *simple estimating functions* (each term in F_n depending on one observation only).

2.3.1 Martingale estimating functions

There are (at least) two good reasons for looking at estimating functions that are martingales: (*i*) the score function which we are basically trying to imitate is a martingale; and (*ii*) we have all the machinery from martingale theory (*e.g.* limit theorems) at our disposal. Also, martingale estimating functions are important as any asymptotically well-behaved estimating function is asymptotically equivalent to a martingale estimating function (Jacobsen 1998).

Definition, asymptotic results and optimality

Consider the conditional moment condition

$$\mathrm{E}_\theta\big(\tilde{h}(X_0, X_\Delta, \theta) | X_0 = x\big) = \int_I \tilde{h}(x, y, \theta) p_\theta(\Delta, x, y)\, dy = 0, \quad x \in I, \theta \in \Theta \tag{2.8}$$

for a function $\tilde{h} : I^2 \times \Theta \to \mathbb{R}$. If all coordinates of f from (2.7) satisfy this condition, and $(\mathscr{G}_i)$ is the discrete-time filtration generated by the observations, then

$$\mathrm{E}_\theta\big(F_n(\theta) | \mathscr{G}_{n-1}\big) = F_{n-1}(\theta) + \mathrm{E}_\theta\big(f(X_{(n-1)\Delta}, X_{n\Delta}, \theta) | X_{(n-1)\Delta}\big) = F_{n-1}(\theta),$$

so $F_n(\theta)$ is a P_θ-martingale with respect to $(\mathscr{G}_i)$. Usually, when $p_\theta(\Delta, x, \cdot)$ is not known, functions satisfying (2.8) cannot be found explicitly but should be calculated numerically.

Suppose that $h_1, \ldots, h_N : I^2 \times \Theta \to \mathbb{R}$ all satisfy (2.8) and let $\alpha_1, \ldots, \alpha_N : I \times \Theta \to \mathbb{R}^p$ be *arbitrary weight functions*. Then each coordinate of f defined by

$$f(x, y, \theta) = \sum_{j=1}^{N} \alpha_j(x, \theta) h_j(x, y, \theta) = \alpha(x, \theta) h(x, y, \theta)$$

satisfies (2.8) as well. Here we have used the notation α for the $\mathbb{R}^{p \times N}$-valued function with (k, j)'th element equal to the k'th element of α_j and h for $(h_1, \ldots, h_N)^T$. Note that the score function is obtained as a special case: for $N = p$, $h(x, y, \theta) = (\partial_\theta \log p_\theta(\Delta, x, y))^T$ and $\alpha(x, \theta)$ equal to the $p \times p$ unit matrix.

Classical limit theory for stationary martingales (Billingsley 1961) is employed for asymptotic results of F_n with f as above. Under differentiability and integrability conditions $\dot{F}_n(\theta)/n \to A(\theta)$ in P_{θ_0}-probability for all θ and $F_n(\theta_0)/\sqrt{n} \to N(0, V_0)$ in distribution wrt. P_{θ_0}. Here,

$$A(\theta) = \mathrm{E}_{\theta_0} \dot{f}(X_0, X_\Delta, \theta) = \sum_{j=1}^{N} \mathrm{E}_{\theta_0} \alpha_j(X_0, \theta) \dot{h}_j(X_0, X_\Delta, \theta) = \mathrm{E}_{\theta_0} \alpha(X_0, \theta) \dot{h}(X_0, X_\Delta, \theta)$$

$$V_0 = \mathrm{E}_{\theta_0} f(X_0, X_\Delta, \theta_0) f(X_0, X_\Delta, \theta_0)^T = \mathrm{E}_{\theta_0} \alpha(X_0, \theta_0) \tau_h(X_0, \theta_0) \alpha^T(X_0, \theta_0),$$

where $\tau_h(x, \theta) = \mathrm{Var}_\theta(h(X_0, X_\Delta, \theta) | X_0 = x)$. If the convergence $\dot{F}_n(\theta)/n \to A(\theta)$ is suitably uniform in θ and $A_0 = A(\theta_0)$ is non-singular then a solution $\tilde{\theta}_n$ to $F_n(\theta) = 0$ exists with a probability tending to 1, $\tilde{\theta}_n \to \theta_0$ in probability, and $\sqrt{n}(\tilde{\theta}_n - \theta_0) \to N(0, A_0^{-1} V_0 A_0^{-1^T})$ in distribution wrt. P_{θ_0} (Sørensen 1998*b*). The condition that A_0 is non-singular is discussed below.

For $h_1 \dots, h_N$ given it is easy to find optimal weights $\alpha^\star$ in the sense that the corresponding estimator has the smallest asymptotic variance, where $V \le V'$ as usual means that $V' - V$ is positive semi-definite (Sørensen 1997):

$$\alpha^\star(x, \theta) = \Big(\tau_h(x, \theta)^{-1} \mathrm{E}_\theta \big(\dot{h}(X_0, X_\Delta, \theta) | X_0 = x \big) \Big)^T.$$

How to construct martingale estimating functions in practice

The question on how to choose $h_1, \dots, h_N$ (and N) is far more subtle (when the score function is not known), and the optimal $h_1, \dots, h_N$ within some class (typically) change with Δ. Jacobsen (1998) investigates optimality as $\Delta \to 0$, and it is clear that the score for the invariant measure is optimal as $\Delta \to \infty$. Not much work has been done for fixed values of Δ in between. Here we mention two particular ways of constructing martingale estimating functions.

First, consider functions of the form

$$h_j(x, y, \theta) = g_j(y) - \mathrm{E}_\theta(g_j(X_\Delta) | X_0 = x) \tag{2.9}$$

for some (simple) functions $g_j : I \to \mathbb{R}$ in $L^1(\mu_\theta)$, $j = 1, \dots, N$. Obvious choices are polynomials $g_j(y) = y^{k_j}$ for some (small) integers k_j (Bibby & Sørensen 1995, Bibby & Sørensen 1996). In some models low-order conditional moments are known analytically although the transition probabilities are not. But even if this is not the case, the conditional moments are easy to calculate by simulation. Kessler & Paredes (1999) investigates the influence of simulations on the asymptotic properties of the estimator.

Second, let $g_j(\cdot, \theta) : I \to \mathbb{R}$, $j = 1, \dots, N$ be eigenfunctions for $\mathscr{A}_\theta$ with eigenvalues $\lambda_j(\theta)$. Under mild conditions (Kessler & Sørensen 1999) $\mathrm{E}_\theta(g_j(X_\Delta, \theta) | X_0 = x) = \exp(-\lambda_j(\theta)\Delta) g_j(x, \theta)$ so

$$h_j(x, y, \theta) = g_j(y, \theta) - e^{-\lambda_j(\theta)\Delta} g_j(x, \theta)$$

satisfies (2.8). Note that this h_j has the same form as (2.9) except that g_j depends on θ. The estimating functions based on eigenfunctions have two advantages: they are invariant to twice continuously differentiable transformations of data and the optimal weights are easy to simulate (Sørensen 1997). However, the applicability is rather limited as the eigenfunctions are known only for a few models; see Kessler & Sørensen (1999) for some non-trivial examples, though.

Considerations on identification

In order for the estimator to behave asymptotically nicely, the matrix A_0 should be regular. Below we shall see how this condition may be explained in terms of reparametrizations. For simplicity we assume that $N = 1$ such that $f(x,y,\theta) = \alpha(x,\theta)h(x,y,\theta)$ for an $\alpha : I \times \Theta \to \mathbb{R}^p$ and an $h : I^2 \times \Theta \to \mathbb{R}$ satisfying (2.8). Note that $\tau_h(x,\theta) = \mathrm{E}_\theta(h(X_0,X_\Delta,\theta)^2|X_0 = x)$ is a real number. From now on we let $\alpha_j : I \times \Theta \to \mathbb{R}$, $j = 1\ldots,p$, denote the coordinate functions of α and λ the Lebesgue measure on I.

Obviously, $\tau_h(x,\theta)$ should be positive; otherwise the conditional distribution of $h(X_0,X_\Delta,\theta)$ given $X_0 = x$ is degenerate at zero and provides no information. It is also obvious that the coordinates of α should be linearly independent; otherwise there are essentially fewer than p equations for estimation of p parameters. The following proposition shows that linear independence of the coordinates of $\alpha(\cdot,\theta_0)$ is equivalent to regularity of the variance matrix V_0 of $f(X_0,X_\Delta,\theta_0)$ and that regularity of A_0 implies regularity of V_0.

Proposition 2.1 *If $\tau_h(x,\theta_0) > 0$ for all $x \in \mathbb{R}$, then (i) V_0 is singular if and only if there exists $\beta \in \mathbb{R}^p\backslash\{0\}$ such that $\beta^T\alpha(x,\theta_0) = 0$ for λ-almost all $x \in \mathbb{R}$; and (ii) V_0 is positive definite if A_0 is regular.*

Proof Since

$$\begin{aligned} V_0 &= \mathrm{E}_{\theta_0}\,\alpha(X_0,\theta_0)\tau_h(X_0,\theta_0)\alpha(X_0,\theta_0)^T \\ &= \mathrm{E}_{\theta_0}\left(\tau_h(X_0,\theta_0)^{1/2}\alpha(X_0,\theta_0)\right)\left(\tau_h(X_0,\theta_0)^{1/2}\alpha(X_0,\theta_0)\right)^T, \end{aligned}$$

it holds that V_0 is singular if and only if there exists a linear combination of the coordinates of $\tau_h(X_0,\theta_0)^{1/2}\alpha(X_0,\theta_0)$ that is zero μ_{θ_0}-a.s. *i.e.* if and only if $\beta \in \mathbb{R}^p\backslash\{0\}$ exists such that $\beta^T\alpha(X_0,\theta_0) = 0$ μ_{θ_0}-a.s. (since $\tau_h(x,\theta) > 0$). The first assertion now follows as μ_{θ_0} has strictly positive density wrt. λ.

For the second assertion we show that singularity of V_0 implies singularity of A_0. Assume that V_0 is singular and find β as above. Then

$$\beta^T A_0 = \beta^T\,\mathrm{E}_{\theta_0}\,\alpha(X_0,\theta_0)\dot{h}(X_0,\theta_0) = \mathrm{E}_{\theta_0}\,\beta^T\alpha(X_0,\theta_0)\dot{h}(X_0,\theta_0) = 0,$$

and $A(\theta_0)$ is singular as claimed. □

In the following we shall only consider h of the form $h(x,y,\theta) = g(y) - G(x,\theta)$ where $G(x,\theta) = \mathrm{E}_\theta(g(X_\Delta)|X_0 = x)$, see (2.9). Since α is nothing but a weight function, a natural requirement is that G determines the full parameter vector uniquely. In essence, the proposition below claims that this is also sufficient in order for the matrix $A_0^\star$ corresponding to the optimal weight function $\alpha^\star = -\dot{G}/\tau_h$ to be regular.

Below we write A_0^α to stress the dependence of α on A_0. In particular, $A_0^\star = A_0^{\alpha^\star}$. We need some further terminology: say that a bijective transformation γ from a neighbourhood Θ_0 of θ_0 to a set $\Gamma_0 \subseteq \mathbb{R}^p$ is a *reparametrization around* θ_0. The inverse of γ is denoted by γ^{-1} or θ, and $\gamma_0 = \gamma(\theta_0)$. The function $G_\gamma : I \times \Gamma_0$ is defined by $G_\gamma(x,\gamma) = G(x,\theta(\gamma))$; hence $G(x,\theta) = G_\gamma(x,\gamma(\theta))$.

Proposition 2.2 *If there exist $j_1,\dots,j_q \subseteq \{1,\dots,p\}$ with $j_k \neq j_{k'}$ for $k \neq k'$ and a reparametrization around θ_0 such that for $j = j_1,\dots,j_q$*

$$\partial G_\gamma(x,\gamma_0)/\partial\gamma_j = 0, \quad \lambda - \text{a.s.}, \tag{2.10}$$

then A_0^α has rank at most q for any α. Conversely, if $A_0^\star = A_0^{\alpha^\star}$ corresponding to the optimal $\alpha^\star$ has rank $q < p$ and $\tau_h(x,\theta) > 0$ for all $x \in I$ then there exists a reparametrization γ around θ_0 such that (2.10) holds for all $j = q+1,\dots,p$.

Proof By the chain rule it holds for any α that

$$\begin{aligned} A_0^\alpha &= -\mathrm{E}_{\theta_0}\,\alpha(X_0,\theta_0)\dot{G}(X_0,\theta_0) \\ &= -\mathrm{E}_{\theta_0}\,\alpha(X_0,\theta_0)\dot{G}_\gamma(X_0,\gamma_0)\dot{\gamma}(\theta_0) \\ &= -\big(\mathrm{E}_{\theta_0}\,\alpha(X_0,\theta_0)\dot{G}_\gamma(X_0,\gamma_0)\big)\dot{\gamma}(\theta_0) \end{aligned}$$

where $\dot{G}_\gamma$ is the matrix of derivatives wrt. γ of G_γ and $\dot{\gamma}$ is the matrix of derivatives of γ wrt. θ. By assumption the j_k'th column of $\dot{G}_\gamma(X_0,\gamma_0)$ has all elements equal to zero almost surely, $k = 1,\dots,q$, so A_0^α has rank at most q as claimed.

For the second assertion, assume that

$$\begin{aligned} A_0^\star &= \mathrm{E}_{\theta_0}\,\dot{G}(X_0,\theta_0)^T\dot{G}(X_0,\theta_0)/\tau_h(X_0,\theta_0) \\ &= \mathrm{E}_{\theta_0}\big(\dot{G}(X_0,\theta_0)\tau_h(X_0,\theta_0)^{-1/2}\big)^T\big(\dot{G}(X_0,\theta_0)\tau_h(X_0,\theta_0)^{-1/2}\big) \end{aligned}$$

has rank $q < p$ and assume without loss of generality that the upper left $q \times q$ submatrix is positive definite (possibly after the coordinates of θ have been renumbered).

According to Lemma 2.3 below, $x_1,\dots,x_q$ exist such that

$$\begin{pmatrix} \partial G(x_1,\theta_0)/\partial\theta_1 & \cdots & \partial G(x_1,\theta_0)/\partial\theta_q \\ \vdots & & \vdots \\ \partial G(x_q,\theta_0)/\partial\theta_1 & \cdots & \partial G(x_q,\theta_0)/\partial\theta_q \end{pmatrix}$$

is regular. Hence, there is a neighbourhood Θ_0 of θ_0 such that $\gamma\colon \Theta_0 \to \mathbb{R}^p$ defined by

$$\gamma(\theta) = \big(G(x_1,\theta),\dots,G(x_q,\theta),\theta_{q+1},\dots,\theta_p\big)$$

is injective. Let $\Gamma_0 = \gamma(\Theta_0)$ and $\gamma_0 = \gamma(\theta_0)$. The first q rows of $\dot{\gamma}(\theta_0)$ are given by

$$\begin{pmatrix} \partial G(x_1,\theta_0)/\partial\theta_1 & \cdots & \partial G(x_1,\theta_0)/\partial\theta_p \\ \vdots & & \vdots \\ \partial G(x_q,\theta_0)/\partial\theta_1 & \cdots & \partial G(x_q,\theta_0)/\partial\theta_p \end{pmatrix}$$

and the last $p-q$ rows are $(0_{p-q\times q}, I_{(p-q)\times(p-q)})$.

Next, let $\dot{G}^j = (\dot{G}_1,\ldots,\dot{G}_q,\dot{G}_j)$ be the $1\times(q+1)$ matrix of derivatives wrt. $\theta_1,\ldots,\theta_q,\theta_j$ for $j = q+1,\ldots,p$. Since $A_0^\star$ has rank q, the matrix

$$\mathrm{E}_{\theta_0}\big(\dot{G}^j(X_0,\theta_0)\tau_h(X_0,\theta_0)^{-1/2}\big)^T\big(\dot{G}^j(X_0,\theta_0)\tau_h(X_0,\theta_0)^{-1/2}\big)$$

is singular implying that $\tilde{\beta}^j \in \mathbb{R}^{q+1}\backslash\{0\}$ exists such that $\dot{G}^j(X_0,\theta_0)\tilde{\beta}^j = 0$ almost surely wrt. μ_{θ_0}. Here, $\tilde{\beta}^j_{q+1} \neq 0$ because the upper left $q\times q$ sub-matrix of $A_0^\star$ is regular. If $\beta^j \in \mathbb{R}^p\backslash\{0\}$ is defined by

$$\beta^j_k = \begin{cases} \tilde{\beta}^j_k/\tilde{\beta}^j_{q+1}, & k = 1,\ldots,q \\ 1, & k = j \\ 0, & \textit{otherwise} \end{cases}$$

it follows that

$$\dot{G}(X_0,\theta_0)\beta^j = 0 \quad \mu_{\theta_0}-\text{a.s.} \tag{2.11}$$

for all $j = q+1,\ldots,p$ and hence $\dot{G}(x,\theta_0)\beta^j = 0$ λ-a.s. for all $j = q+1,\ldots,p$.

From the expression for the derivative $\dot{\gamma}(\theta_0)$ it now follows that $\dot{\gamma}(\theta_0)\beta^j$ equals the j'th unit column. Hence, since the inverse θ of γ has derivative $\dot{\theta}(\gamma) = \dot{\gamma}(\theta(\gamma))^{-1}$ it holds that

$$\beta^j = \Big(\frac{\partial\theta_1(\gamma(\theta_0))}{\partial\gamma_j},\ldots,\frac{\partial\theta_p(\gamma(\theta_0))}{\partial\gamma_j}\Big)^T, \quad j = q+1,\ldots,p.$$

Finally, by the chain rule

$$\frac{\partial G_\gamma(x,\gamma_0)}{\partial\gamma_j} = \dot{G}(x,\theta_0)\big(\partial\theta_1(\gamma_0)/\partial\gamma_j,\ldots,\partial\theta_p(\gamma_0)/\partial\gamma_j\big)^T = \dot{G}(x,\theta_0)\beta^j = 0$$

almost surely wrt. the Lebesgue measure λ for all $j = q+1,\ldots,p$ as claimed. □

Note that (2.11) implies that the coordinates of $\alpha^\star(\cdot,\theta_0)$ are linearly dependent λ-a.s., compare with Proposition 2.1. Also note that the reparametrization around θ_0 is not necessarily a global one as it may not be injective on all of Θ. In the proof we used the following lemma.

Lemma 2.3 *Let Y be a real random variable and $d:\mathbb{R}\to\mathbb{R}^q$ be a function such that $\mathrm{E}\,d(Y)d(Y)^T$ is positive definite. Then $y_1,\ldots,y_q$ exist such that the $q\times q$ matrix $D^{(q)}(y_1,\ldots,y_q)$ defined coordinate-wise by $D^{(q)}_{ij}(y_1,\ldots,y_q) = d_j(y_i)$ is regular.*

Proof By assumption it holds for all $\beta \in \mathbb{R}^q \setminus \{0\}$ that

$$0 < \beta^T \big(\mathrm{E}\, d(Y) d(Y)^T\big)\beta = \mathrm{E}\big(\beta^T d(Y) d(Y)^T \beta\big) = \mathrm{E}\big(\beta^T d(Y)\big)^2$$

so $\beta^T d(Y)$ is not zero almost surely and y_β exists with $\beta^T d(y_\beta) \neq 0$.

The points $y_1, \ldots, y_q$ are chosen recursively as follows. First, let β_1 be the first unit vector and choose y_1 such that $\beta_1^T d(y_1) = d_1(y_1) \neq 0$. Next, let $\beta_2 = (-d_2(y_1), d_1(y_1), 0, \ldots, 0)^T$ and choose y_2 such that

$$\beta_2^T d(y_2) = d_1(y_1) d_2(y_2) - d_2(y_1) d_1(y_2) = \det D^{(2)}(y_1, y_2),$$

i.e. such that $D^{(2)}(y_1, y_2)$ is regular. Continue in the same manner: for y_r, assume that $y_1, \ldots, y_{r-1}$ are chosen such that $D^{(r-1)}(y_1, \ldots, y_{r-1})$ is regular, and note that the determinant of $D^{(r)}(y_1, \ldots, y_{r-1}, Y)$ is a linear combination $\beta_r^T d(Y)$ with coefficients β_r depending on $d_j(y_i)$, $j = 1, \ldots, r$ and $i = 1, \ldots, r-1$. Consequently, we can find y_r such that $\beta_r^T d(y_r) = \det D^{(r)}(y_1, \ldots, y_r) \neq 0$. The assertion now follows for $r = q$. □

2.3.2 Simple estimating functions

An estimating function is called *simple* if it has the form $F_n(\theta) = \sum_{i=1}^n f(X_{i\Delta}, \theta)$ where $f : I \times \Theta \to \mathbb{R}^p$ takes only one state variable as argument (Kessler 2000). Condition (2.6) simplifies to: $\mathrm{E}_{\theta_0} f(X_0, \theta) = 0$ if and only if $\theta = \theta_0$. It involves the marginal distribution only which has two important consequences: First, since the invariant distribution is known explicitly, it is easy to find functionals f analytically with $\mathrm{E}_{\theta_0} f(X_0, \theta_0) = 0$. Second, simple estimating functions completely ignore the dependence structure of X and can only be used for estimation of (parameters in) the marginal distribution. This is of course a very serious objection.

Kessler (2000) shows asymptotic results for the corresponding estimators and is also concerned with optimality. This work was continued by Jacobsen (1998). However, it is usually not possible to find f optimally so f is chosen somewhat ad hoc. An obvious possibility is the score corresponding to the invariant distribution, $f = \partial_\theta \log \mu$. Another is moment generated functions $f_j(x, \theta) = x^{k_j} - \mathrm{E}_\theta X_0^{k_j}$, $j = 1, \ldots, p$. Also, functions could be generated by the infinitesimal generator $\mathscr{A}_\theta$ defined by (2.3): let $h_j : I \times \Theta \to \mathbb{R}$, $j = 1, \ldots, p$, be such that the martingale part of $h_j(X, \theta)$ is a true martingale wrt. P_θ. Then $f = (\mathscr{A}_\theta h_1, \ldots, \mathscr{A}_\theta h_p)^T$ gives rise to an unbiased, simple estimating function. Kessler (2000) suggests to use low-order polynomials for $h_1, \ldots, h_p$ — regardless of the model.

In Paper I we study the *model-dependent* choice $(h_1, \ldots, h_p) = \partial_\theta \log \mu$. We show that the corresponding estimating function based on $f_j = \mathscr{A}_\theta(\partial_{\theta_j} \log \mu)$, $j = 1, \ldots, p$, may be interpreted as an approximation to minus twice the continuous-time score function when σ does not depend on θ (Proposition I.1). Intuitively, we would thus expect it to work well for small values of Δ, and it is indeed small Δ-optimal in the sense of Jacobsen (1998); still if σ does not depend on θ.

There are two important differences from the usual Riemann-Itô approximation of the continuous-time score, that is, the logarithmic derivative wrt. θ of (2.5): the above approximation is unbiased which the Riemann-Itô approximation is not; and each term in the Riemann-Itô approximation depends on *pairs* of observations whereas each term in the above approximation depends on a single observation only.

Also note that the estimating function from Paper I is invariant to bijective and twice differentiable transformations of the data if σ does not depend on θ (Proposition I.2); this is not the case for the simple estimating functions discussed earlier. The ideas carry over (to some extent at least) to multi-dimensional diffusions, and the estimating function works quite well in simulation studies.

Finally, a remark connecting a simple estimating function $F_n(\theta) = \sum_{i=1}^n f(X_{i\Delta}, \theta)$ to a class of martingale estimating functions. Define

$$h_f(x,y,\theta) = U_\theta f(y,\theta) - \big(U_\theta f(x,\theta) - f(x,\theta)\big)$$

where U_θ is the potential operator given by $U_\theta f(x,\theta) = \sum_{k=0}^\infty \mathrm{E}_\theta(f(X_{k\Delta},\theta)|X_0 = x)$. Then h_f satisfies condition (2.8), and the martingale estimating functions $\sum_{i=1}^n h_f(X_{(i-1)\Delta}, X_{i\Delta}, \theta)$ and $F_n(\theta)$ are asymptotically equivalent (Jacobsen 1998).

However, the martingale estimating function may be improved by introducing weights α (unless of course the optimal weight $\alpha^\star(\cdot,\theta)$ is constant). In this sense martingale estimating functions are always better (or at least as good) as simple estimating functions. In practice it is not very helpful, though, as the potential operator in general is not known! Also, the improvement may be very small as we shall see in the following example.

Example *(The Ornstein-Uhlenbeck process)* Consider the solution to $dX_t = \theta X_t\,dt + dW_t$ where $\theta < 0$. Kessler (2000) shows that the *optimal* simple estimating function is obtained for $f(x,\theta) = 2\theta x^2 + 1$. It is easy to see that $h_f(x,y,\theta) \propto f(y,\theta) - \psi f(x,\theta)$ where $\psi = \psi(\theta,\Delta) = \exp(2\theta\Delta)$ and that the optimal weight function is given by

$$\alpha^\star(x,\theta) = \frac{\mathrm{E}_\theta\big(\dot h_f(X_0,X_\Delta)|X_0 = x\big)}{\tau_{h_f}(x,\theta)} = \frac{-4\theta\Delta\psi x^2 - (1-\psi+2\theta\Delta\psi)/\theta}{-8\theta\psi(1-\psi)x^2 + 2(1-\psi)^2}.$$

Since $\alpha^\star(\cdot,\theta)$ is not constant, improvement is indeed possible. It turns out, however, that the asymptotic variance is only reduced by about 1% (for $\theta_0 = -1$). □

It is well-known that the optimal simple estimating function is nearly (globally) efficient in the Ornstein-Uhlenbeck model, and the example does not rule out the possibility that the improvement could be considerable for other models (and other simple estimating functions).

2.3.3 Comments

Obviously, there are lots of unbiased estimating functions that are neither martingales nor simple. For example,

$$f(x,y,\theta) = h_2(y,\theta)\mathscr{A}_\theta h_1(x,\theta) - h_1(x,\theta)\mathscr{A}_\theta h_2(y,\theta)$$

generates a class of estimating functions which are transition dependent and yet explicit (Hansen & Scheinkman 1995, Jacobsen 1998).

Estimating functions of different kinds may of course be combined. For example, one could firstly estimate parameters from the invariant distribution by solving a simple estimating equation and secondly estimate parameters from the conditional distribution one step ahead. See Bibby & Sørensen (1998) for a succesful application.

Also, estimating functions may be used as building blocks for the *generalized method of moments* (GMM), the much favored estimation method in the econometric literature (Hansen 1982). Estimation via GMM is essentially performed by choosing an estimating function F_n of dimension $p' > p$ and minimizing the quadratic form $F_n(\theta)^T \Omega F_n(\theta)$ for some weight matrix Ω.

2.4 Approximate maximum likelihood estimation

We now describe three approximate maximum likelihood methods. They all supply approximations, analytical or numerical, of $p_\theta(\Delta, x, \cdot)$ for fixed x and θ. In particular they supply approximations of $p_\theta(\Delta, X_{(i-1)\Delta}, X_{i\Delta})$, $i = 1, \ldots, n$, and therefore of $L_n(\theta)$. The approximate likelihood is finally maximized over $\theta \in \Theta$.

2.4.1 An analytical approximation

A naive, explicit approximation of the conditional distribution of X_Δ given $X_0 = x$ is provided by the Euler approximation. The Gaussian approximation may be poor even if the conditional moments are replaced by accurate approximations (or perhaps even the true moments). A sequence of *explicit, non-Gaussian approximations* of $p_\theta(\Delta, x, \cdot)$ is suggested by Aït-Sahalia (1998). For fixed x and θ the idea is to (*i*) transform X to a process Z which, conditional on $X_0 = x$, has $Z_0 = 0$ and Z_Δ "close" to standard normal; (*ii*) define a truncated Hermite series expansion of the density of Z_Δ around the standard normal density; and (*iii*) invert the Hermite approximation in order to obtain an approximation of $p_\theta(\Delta, x, \cdot)$.

For step (*i*) define $Z = g_{x,\theta}(X)$ where

$$g_{x,\theta}(y) = \frac{1}{\sqrt{\Delta}} \int_x^y \frac{1}{\sigma(u, \theta)} du.$$

Then Z solves $dZ_t = b_Z(Z_t, \theta)\, dt + 1/\sqrt{\Delta}\, dW_t$ with drift function given by Itô's formula and $Z_0 = 0$ (given $X_0 = x$). Note that $g'_{x,\theta}(y) = (\Delta\sigma^2(y, \theta))^{-1/2} > 0$ for all $y \in I$ so that $g_{x,\theta}$ is injective.

For step (*ii*) note that $N(0,1)$ is a natural approximation of the conditional distribution of Z_Δ given $Z_0 = 0$, as increments of Z over time intervals of length Δ has approximately unit variance. Let $p^Z_\theta(\Delta, 0, \cdot)$ denote the true conditional density of Z_Δ given $Z_0 = 0$ and let $p^{Z,J}_\theta(\Delta, 0, \cdot)$ be the *Hermite series expansion truncated after J terms* of $p^Z_\theta(\Delta, 0, \cdot)$ around the standard normal density.

For step (*iii*) note that the true densities $p_\theta(\Delta,x,\cdot)$ and $p^Z_\theta(\Delta,0,\cdot)$ are related by

$$p_\theta(\Delta,x,y) = \frac{1}{\sqrt{\Delta}\sigma(x,\theta)} p^Z_\theta\big(\Delta,0,g_{x,\theta}(y)\big), \quad y \in I$$

and apply this formula to invert the approximation $p^{Z,J}_\theta(\Delta,0,\cdot)$ of $p^Z_\theta(\Delta,0,\cdot)$ into an approximation $p^J_\theta(\Delta,x,\cdot)$ of $p_\theta(\Delta,x,\cdot)$ in the natural way:

$$p^J_\theta(\Delta,x,y) = \frac{1}{\sqrt{\Delta}\sigma(x,\theta)} p^{Z,J}_\theta\big(\Delta,0,g_{x,\theta}(y)\big), \quad y \in I.$$

Then $p^J_\theta(\Delta,x,y)$ converges to $p_\theta(\Delta,x,y)$ as $J\to\infty$, suitably uniformly in y and θ. Furthermore, if $J = J(n)$ tends to infinity fast enough as $n\to\infty$ then the estimator maximizing $\prod_{i=1}^n p^{J(n)}_\theta(\Delta, X_{(i-1)\Delta}, X_{i\Delta})$ is asymptotically equivalent to the maximum likelihood estimator (Aït-Sahalia 1998, Theorems 1 and 2).

Note that the coefficients of the Hermite series expansion cannot be computed explicitly but could be replaced by analytical approximations in terms of the infinitesimal generator. Hence, the technique provides explicit, though *very complex*, approximations to $p_\theta(\Delta,x,\cdot)$. Aït-Sahalia (1998) performs numerical experiments that indicate that the error $p^J_\theta(\Delta,x,y) - p_\theta(\Delta,x,y)$ decreases quickly; roughly with a factor 10 for each extra term included in the expansion of $p^Z_\theta(\Delta,0,\cdot)$.

2.4.2 Numerical solutions of the Kolmogorov forward equation

A classical result from stochastic calculus states that the transition densities under certain regularity conditions are characterized as solutions to the *Kolmogorov forward equations*. Lo (1988) employs a similar result and finds explicit expressions for the likelihood function for a log-normal diffusion with jumps and a Brownian motion with zero as an absorbing state. Poulsen (1999) seems to be the first to employ numerical procedures for non-trivial diffusion models.

For x and θ fixed the forward equation for $p_\theta(\cdot,x,\cdot)$ is a partial differential equation: for $(t,y)\in(0,\infty)\times I$,

$$\frac{\partial}{\partial t}p_\theta(t,x,y) = -\frac{\partial}{\partial y}\big(b(y,\theta)p_\theta(t,x,y)\big) + \frac{1}{2}\frac{\partial^2}{\partial(y)^2}\big(\sigma^2(y,\theta)p_\theta(t,x,y)\big),$$

with initial condition $p_\theta(0,x,y) = \delta(x-y)$ where δ is the Dirac delta function. In order to calculate the likelihood $L_n(\theta)$ one has to solve n of the above forward equations, one for each $X_{(i-1)\Delta}$, $i = 1,\ldots,n$. Note that the forward equation for $X_{(i-1)\Delta}$ determines $p_\theta(t,X_{(i-1)\Delta},y)$ for *all* values of (t,y), but that we only need it at a single point, namely $(\Delta,X_{i\Delta})$.

Poulsen (1999) employs the so-called Crank-Nicholson finite difference method for each of the n forward equations. For fixed θ he obtains a second order approximation of $\log L_n(\theta)$ in the sense that the numerical approximation $\log L^h_n(\theta)$ satisfies

$$\log L^h_n(\theta) = \log L_n(\theta) + h^2 f^\theta_n(X_0,X_\Delta,\ldots,X_{n\Delta}) + o(h^2)g^\theta_n(X_0,X_\Delta,\ldots,X_{n\Delta})$$

for suitable functions f_n^θ and g_n^θ. The parameter h determines how fine-grained a (t,y)-grid used in the numerical procedure is (and thus the accuracy of approximation). If $h = h(n)$ tends to zero faster than $n^{-1/4}$ as $n \to \infty$ then the estimator maximizing $\log L_n^h(\theta)$ is asymptotically equivalent to the maximum likelihood estimator (Poulsen 1999, Theorem 3).

Poulsen (1999) fits the CKLS model to a dataset of 655 observations (in a revised version, even a six-parameter extension is fitted) and is able to do it in quite reasonable time. Although n partial differential equations must be solved the method seems to be much faster than the simulation based method below.

2.4.3 Approximation via simulation

Pedersen (1995*b*) defines a sequence of approximations to $p_\theta(\Delta,x,\cdot)$ via a missing data approach. The basic idea is to (*i*) split the time interval from 0 to Δ into pieces short enough that the Euler approximation holds reasonably well; (*ii*) consider the joint Euler likelihood for the augmented data consisting of the observation X_Δ and the values of X at the endpoints of the subintervals; (*iii*) integrate the unobserved variable out of the joint Euler density; and (*iv*) calculate the resulting expectation by simulation. The method has been applied successfully to the CKLS model (Honoré 1997).

To be precise, let x and θ be fixed, consider an integer $N \geq 0$, and split the interval $[0,\Delta]$ into $N+1$ subintervals of length $\Delta_N = \Delta/(N+1)$. Use the notation $X_{0,k}$ for the (unobserved) value of X at time $k/(N+1)$, $k = 1,\ldots,N$. Then (with $x_{0,0} = x$ and $x_{0,N+1} = y$),

$$\begin{aligned} p_\theta(\Delta,x,y) &= \int_{I^N} \prod_{i=1}^{N+1} p_\theta\left(\Delta_N, x_{0,i-1}, x_{0,i}\right) d(x_{0,1},\ldots,x_{0,N}) \\ &= \int_I p_\theta\left(N\Delta_N, x, x_{0,N}\right) p_\theta\left(\Delta_N, x_{0,N}, y\right) dx_{0,N} \\ &= \mathrm{E}_\theta\Big(p_\theta\left(\Delta_N, X_{0,N}, y\right) \Big| X_0 = x \Big), \quad y \in I \end{aligned} \tag{2.12}$$

where we have used the Chapman-Kolmogorov equations.

Now, for Δ_N small (N large), $p_\theta(\Delta_N, x_{0,N}, \cdot)$ is well approximated by the normal density with mean $x_{0,N} + b(x_{0,N},\theta)\Delta_N$ and variance $\sigma^2(x_{0,N},\theta)\Delta_N$. Let $\tilde{p}_\theta^N(\Delta_N, x_{0,N}, \cdot)$ denote this density. Following (2.12),

$$p_\theta^N(\Delta,x,y) = \mathrm{E}_\theta\Big(\tilde{p}_\theta^N\left(\Delta_N, X_{0,N}, y\right) \Big| X_0 = x \Big)$$

is a natural approximation of $p_\theta(\Delta,x,y)$, $y \in I$. Note that $N = 0$ corresponds to the simple Euler approximation.

The approximate likelihood functions $L_n^N(\theta) = \prod_{i=1}^n p_\theta^N(\Delta, X_{(i-1)\Delta}, X_{i\Delta})$ converge in probability to $L_n(\theta)$ as $N \to \infty$ (Pedersen 1995*b*, Theorems 3 and 4). Furthermore, there exists a sequence $N(n)$ such that the estimator maximizing $L_n^{N(n)}(\theta)$

is asymptotically equivalent (as $n \to \infty$) to the maximum likelihood estimator (Pedersen 1995*a*, Theorem 3).

In practice we could calculate $p_\theta^N(\Delta, x, y)$ as the average of a large number of values $\{\tilde{p}_\theta^N(\Delta, X_{0,N}^r, y)\}_r$ where $X_{0,N}^r$ is the last element of a simulated discrete-time path $X_0, X_{0,1}^r, \ldots, X_{0,N}^r$ started at x. Note that the paths are simulated conditional on $X_0 = x$ only which implies that the simulated values $X_{0,N}^r$ at time $N\Delta_N$ may be far from the observed value at time Δ. This is not very appealing as the continuity of X makes a large jump over a small time interval unlikely to occur in practice. Also, it has the unfortunate numerical implication that a very large number of simulations are needed in order to obtain convergence of the average. Elerian *et al.* (2000, Section 3.1) suggest an importance sampling technique which utilizes the observation at time Δ as well, but is far more difficult to perform than the above (see also Section 2.5 below).

2.5 Bayesian analysis

Bayesian analysis of discretely observed diffusions has been discussed by Eraker (1998) and Elerian *et al.* (2000). The unknown model parameter is treated as a missing data point, and Markov Chain Monte Carlo (MCMC) methods are used for simulation of the posterior distribution of the parameter with density

$$f(\theta | X_0, X_\Delta, \ldots, X_{n\Delta}) \propto f(X_0, X_\Delta, \ldots, X_{n\Delta} | \theta) f(\theta). \tag{2.13}$$

The Bayesian estimator of θ is simply the mean (say) of this posterior. Note that we use f generically for densities. In particular, $f(\theta)$ denotes the prior density of the parameter and $f(X_0, \ldots, X_{n\Delta}|\theta)$ denotes the likelihood function evaluated at θ.

The Bayesian approach deals with the intractability of $f(X_0, \ldots, X_{n\Delta}|\theta)$ in a way very similar to that of Pedersen (1995*b*), namely by introducing auxiliary data and employing the Euler approximation over small time intervals. However, the auxiliary data are generated and used quite differently in the two approaches.

As in Section 2.4.3 each interval $[(i-1)\Delta, i\Delta]$ is split into $N+1$ subintervals of length $\Delta_N = \Delta/(N+1)$. We use the notation $X_{i\Delta,k}$ for the value of X at time $i\Delta + k/(N+1)$, $i = 0, \ldots, n-1$ and $k = 0, \ldots, N+1$. The value is observed for $k = 0$ and $k = N$, and $X_{(i-1)\Delta,N+1} = X_{i\Delta,0}$. Further, let $\tilde{X}_{i\Delta}$ be the collection of latent variables $X_{i\Delta,1}, \ldots, X_{i\Delta,N}$ between $i\Delta$ and $(i+1)\Delta$, let $\tilde{X} = (\tilde{X}_0, \ldots, \tilde{X}_{(n-1)\Delta})$ be the nN-vector of all auxiliary variables, and let X^{obs} be short for the vector of observations $X_0, X_\Delta, \ldots, X_{n\Delta}$.

For N large enough the Euler approximation is quite good and the density of $(X^{\text{obs}}, \tilde{X})$, conditional on θ (and X_0), is roughly

$$f^N(X^{\text{obs}}, \tilde{X}|\theta) = \prod_{i=0}^{n-1} \prod_{k=1}^{N+1} \varphi\Big(X_{i\Delta,k}, X_{i\Delta,k-1} + b(X_{i\Delta,k-1}, \theta)\Delta_N, \sigma^2(X_{i\Delta,k-1}, \theta)\Delta_N\Big) \tag{2.14}$$

where $\varphi(\cdot, m, v)$ is the density of $N(m, v)$. The idea is now to generate a Markov chain $\{\tilde{X}^j, \theta^j\}_j$ with invariant (and limiting) density equal to the approximate

posterior density

$$f^N(\tilde{X},\theta|X^{\text{obs}}) = \frac{f^N(X^{\text{obs}},\tilde{X}|\theta)f(\theta)}{f(X^{\text{obs}})} \propto f^N(X^{\text{obs}},\tilde{X}|\theta)f(\theta). \tag{2.15}$$

Then $\{\theta^j\}_j$ has invariant density equal to the marginal of $f^N(\tilde{X},\theta|X^{\text{obs}})$. This is interpreted as an approximation of the posterior (2.13) of θ and the Bayes estimator of θ is simply the average of the simulated values $\{\theta^j\}_j$ (after some burn-in time).

In order to start off the Markov chain, θ^0 is drawn according to the prior density $f(\theta)$, and $\tilde{X}^0$ is defined by linear interpolation between the observed values of X, say. The j'th iteration in the Markov chain is conducted in two steps: first, $\tilde{X}^j = (\tilde{X}^j_0,\ldots,\tilde{X}^j_{(n-1)\Delta})$ is updated from $f(\tilde{X}|X^{\text{obs}},\theta^{j-1})$, and second, θ^j is updated from $f(\theta|X^{\text{obs}},\tilde{X}^j)$.

For the first step, note that the Markov property of X implies that the conditional distribution of $\tilde{X}_{i\Delta}$ given (X^{obs},θ) depends on $(X_{i\Delta},X_{(i+1)\Delta},\theta)$ only so the vectors $\tilde{X}^j_{i\Delta}$, $i=0,\ldots,n-1$ may be drawn one at a time. We focus on how to draw $\tilde{X}_0 = (X_{0,1},\ldots,X_{0,N})$ conditional on $(X_0,X_\Delta,\theta^{j-1})$; the target density being proportional to

$$\prod_{k=1}^{N+1} \varphi\Big(X_{0,k}, X_{0,k-1} + b(X_{0,k-1},\theta^{j-1})\Delta_N, \sigma^2(X_{0,k-1},\theta^{j-1})\Delta_N\Big),$$

cf. (2.14). It is (usually) not possible to find the normalizing constant so direct sampling from the density is not feasible. However, *the Metropolis-Hastings algorithm* may be employed; for example with suitable Gaussian proposals. Eraker (1998) suggests to sample only one element of $\tilde{X}_0$ at a time whereas Elerian *et al.* (2000) suggests to sample block-wise, with random block size. The latter is supposed to increase the rate of convergence of the Markov chain (of course, all the usual problems with convergence of the chain should be investigated). Note the crucial difference from the simulation approach in Section 2.4.3 where $\tilde{X}_{i\Delta}$ was simulated conditional on $X_{i\Delta}$ only: here $\tilde{X}_{i\Delta}$ is simulated conditional on both $X_{i\Delta}$ and $X_{(i+1)\Delta}$.

For the second step it is sometimes possible to find the posterior of θ explicitly from (2.15) in which case θ is updated by direct sampling from the density. Otherwise the Metropolis-Hastings algorithm is imposed again.

The method is easily extended to cover the multi-dimensional case. Also, it applies to models that are only partially observed (*e.g.* stochastic volatility models) in which case the values of the unobserved coordinates are simulated like $\tilde{X}$ above (Eraker 1998). Eraker (1998) analyses US interest rate data and simulated data, using the CKLS model $dX_t = \alpha(\beta - X_t)\,dt + \sigma X_t^\gamma$ as well as a stochastic volatility model (see Section 3.4.4). Elerian *et al.* (2000) apply the method on simulated Cox-Ingersoll-Ross data and on interest rate data using a non-standard eight-parameter model.

2.6 Estimation based on auxiliary models

We now discuss *indirect inference* (Gourieroux *et al.* 1993) and the so-called *efficient method of moments*, or EMM for short (Gallant & Tauchen 1996). The methods are essentially applicable whenever simulation from the model is possible and there exists a suitable auxiliary model. This flexibility must be the reason why the methods are fairly often applied by econometricians in empirical studies. However, we find the methods somewhat artificial and awkward and believe that the term "efficient" in EMM is misleading.

The idea is most easily described in a relatively general set-up: let $(Y_1, \dots, Y_n)$ be data from a (complicated) time series model Q_θ, indexed by the parameter of interest θ. Estimation is performed in two steps: First, the model Q_θ is approximated by a simpler one $\tilde{Q}_\rho$ — *the auxiliary model*, indexed by ρ — and the auxiliary parameter ρ is estimated. Second, the two parameters ρ and θ are linked in order to obtain an estimate of θ. This is done via a GMM procedure, and the first step may simply be viewed as a way of finding moment functionals for the GMM procedure.

Let us be more specific. Assume that $(Y_1, \dots, Y_n)$ has density $\tilde{q}_n$ wrt. $\tilde{Q}_\rho$ and let $\hat{\rho}_n$ be the maximum likelihood estimator of ρ, that is,

$$\hat{\rho}_n = \operatorname{argmax}_\rho \log \tilde{q}_n(Y_1, \dots, Y_n, \rho),$$

with first-order condition

$$\tfrac{\partial}{\partial \rho} \log \tilde{q}_n(Y_1, \dots, Y_n, \hat{\rho}_n) = 0.$$

Loosely speaking, $\hat{\theta}_n$ is now defined such that simulated data drawn from $Q_{\hat{\theta}_n}$ resembles data drawn from $\tilde{Q}_{\hat{\rho}_n}$.

For $\theta \in \Theta$ let $Y_1^\theta, \dots, Y_R^\theta$ be a long trajectory simulated from Q_θ and let $\hat{\rho}_R(\theta)$ be the maximum likelihood estimator of ρ based on the simulated data. The indirect inference estimator of θ is the value minimizing the quadratic form

$$\left[\hat{\rho}_n - \hat{\rho}_R(\theta)\right] \Omega \left[\hat{\rho}_n - \hat{\rho}_R(\theta)\right]^T$$

where Ω is some positive semidefinite matrix of size $\dim(\rho) \times \dim(\rho)$. In EMM computation of $\hat{\rho}_R(\theta)$ is avoided as

$$\left[\tfrac{\partial}{\partial \rho} \log \tilde{q}_R(Y_1^\theta, \dots, Y_n^\theta, \hat{\rho}_n)\right] \tilde{\Omega} \left[\tfrac{\partial}{\partial \rho} \log \tilde{q}_R(Y_1^\theta, \dots, Y_R^\theta, \hat{\rho}_n)\right]^T$$

with $\tilde{\Omega}$ like Ω above, is minimized.

Both estimators of θ are consistent and asymptotically normal, and they are asymptotically equivalent (if Ω and $\tilde{\Omega}$ are chosen appropriately). If θ and ρ have same dimension, then the two estimators coincide and simply solve $\hat{\rho}_R(\hat{\theta}_n) = \hat{\rho}_n$. However, as the auxiliary model should be both easy to handle statistically and flexible enough to resemble the original model, it is often necessary to use one with higher dimension than the original model.

Now, how should we choose the auxiliary model? For the diffusion models considered in this chapter the discrete-time Euler scheme

$$X_{i\Delta} = X_{(i-1)\Delta} + b(X_{(i-1)\Delta}, \rho)\Delta + \sigma(X_{(i-1)\Delta}, \rho)\sqrt{\Delta}U_i$$

with $U_1, \ldots, U_n$ independent and identically $N(0,1)$-distributed, is a natural suggestion (Gourieroux *et al.* 1993). The second step in the estimation procedure corrects for the discrepancy between the true conditional distributions and those suggested by the Euler scheme. In a small simulation study for the Ornstein Uhlenbeck process (solving $dX_t = \theta X_t\,dt + \sigma dW_t$) the indirect inference estimator was highly inefficient (compared to the maximum likelihood estimator). In the EMM literature it is generally suggested to use auxiliary densities based on expansions of a non-parametric density (Gallant & Long 1997). Under certain (strong) conditions EMM performed with these auxiliary models is claimed to be as efficient as maximum likelihood.

However, we are convinced that EMM is by no means efficient in practice. The choice of auxiliary model is still quite arbitrary (and fairly incomprehensible), and the whole idea seems slightly artificial. We believe that for many models it is possible to do some kind of (simulated) likelihood approximation that is as fast and efficient — and far more comprehensible. This has already been done for the diffusion models (Section 2.4) and Paper III provides ideas for stochastic volatility models in continuous time.

2.7 Estimation of parameters in the diffusion term

In Paper II we discuss a method for estimation of parameters in the diffusion function which does not fit into any of the previous sections. We briefly sketch the idea here and refer to Paper II for details.

Assume that the drift is known, $b(x,\theta) = b(x)$ (or has been estimated by some other method). Recall that $\mu(\cdot,\theta)$ is the invariant density and define $f = \sigma^2\mu : I \times \Theta \to (0,\infty)$. By equation (2.2) it is easy to verify that $f' = 2b\mu$. Aït-Sahalia (1996) uses this relation for non-parametric estimation of σ^2 via kernel estimation methods. In Paper II the relation is used for parametric estimation. The idea is to define a pointwise consistent estimator of $f(\cdot,\theta)$ and estimate θ by the value that makes the uniform distance between the "true" function $f(\cdot,\theta)$ and the estimated version minimal.

It is crucial that f converges to zero at at least one of the endpoints, l and r, of the state space. If $f(x,\theta) \to 0$ as $x \searrow l$, then $f(x,\theta) = 2\int_l^x b(u)\mu(u,\theta)\,du$ for all $x \in I$ and

$$\hat{f}_{1,n}(x) = \frac{2}{n}\sum_{i=1}^{n}\left(b(X_{i\Delta})1_{\{X_{i\Delta}\le x\}}\right)$$

is consistent for $f(x,\theta)$, $x \in I$. The uniform distance $\sup_{x\in I}\left|f(x,\theta) - \hat{f}_{1,n}(x)\right|$ is minimized in order to obtain an estimator of θ. Similarly, if $f(x,\theta) \to 0$ as $x \nearrow r$,

then

$$\hat{f}_{2,n}(x) = -\frac{2}{n}\sum_{i=1}^{n}\left(b(X_{i\Delta})1_{\{X_{i\Delta}>x\}}\right)$$

is consistent for $f(x,\theta)$, $x \in I$, and $\sup_{x\in I}\left|f(x,\theta) - \hat{f}_{2,n}(x)\right|$ is minimized. If $f(x,\theta) \to 0$ at both l and r then both $\hat{f}_{1,n}$ and $\hat{f}_{2,n}$ provide pointwise consistent estimators of $f(\cdot,\theta)$, and we may use a weighted average $\hat{f}_n$ of the two in order to reduce variance.

The estimators are $\sqrt{n}$-consistent and in certain cases weakly convergent (Theorems II.7 and II.9) but the limit distribution need not be Gaussian. Note that the observations are mixed in a quite complex way in the uniform distance so the usual limit theorems do not apply. Instead, the asymptotic results are proved using *empirical process theory*. We are not aware of any other applications of empirical process theory to problems related to inference for diffusion processes.

In Paper II we apply the method to simulated data from the CKLS model, $dX_t = (\alpha + \beta X_t)\,dt + \sigma X_t^{\gamma} dW_t$, and get reasonable estimators for both γ and σ. The drift parameters are estimated beforehand using martingale estimating functions. Note that this model is relatively hard to identify as different values of the pair (γ,σ) may yield very similar diffusion functions.

There are two objections to the method. First, it provides estimators of the parameters in the diffusion function only; the drift needs to be estimated beforehand. This is possible via martingale estimating functions if the drift is linear (as in many popular models, *e.g.* the CKLS model above), but is otherwise difficult. Second, the approach is perhaps somewhat ad hoc and the estimators need not be efficient.

2.8 Conclusion

Maximum likelihood estimation is typically not possible for diffusion processes that have been observed at discrete time-points only. In this chapter we have reviewed a number of alternatives from the literature.

From a classical point of view, the most appealing methods are those based on approximations of the true likelihood that in principle can be made arbitrarily accurate. We reviewed three types above: One provides analytical approximations to the likelihood function and is therefore in principle the easiest one to use. The expressions are quite complicated, though, even for low-order approximations. The other two rely on numerical techniques, one on numerical solutions to partial differential equations and one on simulations. Even with today's efficient computers both methods are quite computationally demanding so faster procedures are often valuable.

Estimation via estimating functions is generally much faster. So-called simple estimating functions are available in explicit form but provide only estimators for parameters from the marginal distribution. Still, they may be useful for preliminary analysis. Paper I investigates a special simple estimating function which can

be interpreted as an approximation of the continuous-time score function. The corresponding estimator is invariant to transformations of data. Martingale estimating functions are analytically available for a few models but must in general be calculated by simulated. This basically amounts to simulating conditional expectations, which is faster than calculating conditional densities as required by the direct likelihood approximations above. Under regularity conditions, estimators obtained by martingale estimating functions are consistent and asymptotically normal. We studied one of the regularity conditions in some detail and showed how it may be explained in terms of reparametrizations.

The Bayesian approach is to consider the parameter as random and make simulations from its (posterior) distribution. This is quite hard and requires simulation, conditional on the observations, of the diffusion process at a number of time-points in between those where it was observed. The posterior distribution depends on the prior distribution which is chosen more or less arbitrarily. Indirect inference and EMM remove bias due to the discrete-time auxiliary model by simulation methods. The quality of the estimators is bound to depend on the auxiliary model which is chosen somewhat arbitrarily, and we believe that more direct approaches are preferable. The procedure from Section 2.7 (and Paper II) for estimation of the diffusion parameters (when the drift is known) provides satisfactory estimates in the difficult CKLS model. The estimators are probably not efficient, though. The application of empirical process theory for proving asymptotic results is interesting from a theoretical point of view.

3 Stochastic volatility models

In this chapter we discuss continuous-time stochastic volatility models. By this we mean two-dimensional diffusion models where only one of the coordinates is observable and where the stochastic differential equation has a special form. The models were introduced in the mathematical finance literature in the late eighties as modifications of the classical Black-Scholes model. However, only very recently satisfactory estimation methods have been developed.

This chapter provides an overview of existing estimation techniques and a comparison of four specific models. There exist review papers on stochastic volatility models (Ghysels, Harvey & Renault 1996, Shephard 1996), but they are mainly concerned with models defined in discrete time. The continuous-time case is somewhat more delicate because not even the *distribution* of the latent process is known. Hence, not all discrete-time methods can be applied, and those that can are in general more troublesome for continuous-time models.

My main contribution is the development of a new estimation technique relying on simulated approximations to the likelihood. The estimation method is discussed in detail in Paper III and reviewed in Section 3.4.7 where it is also applied to Microsoft stock prices. Furthermore, I have compared four particular models that have all been used in the literature (Section 3.3).

The chapter is organized as follows. We give a motivation from finance in Section 3.1 and discuss the models and their probabilistic properties in Section 3.2. In Section 3.3 we compare specific models. Section 3.4 provides reviews of existing methods as well as of the new estimation technique from Paper III. Finally, related models are briefly discussed in Section 3.5 and conclusions are drawn in Section 3.6.

3.1 A modification of the Black-Scholes model

Consider the classical *Black-Scholes model* (or geometric Brownian motion)

$$dP_t = \alpha P_t\,dt + \tau P_t\,dW_t \tag{3.1}$$

where $\alpha \in \mathbb{R}$ and $\tau > 0$ are constants and W is a standard Brownian motion. The famous Black-Scholes formula (Black & Scholes 1973) for option prices was derived in a set-up with the price of the underlying stock governed by (3.1), and in this section we shall indeed think of the model as a model of stock prices.

If the stock price P solves (3.1) then the process $\log P$ has independent, Gaussian increments: if stock prices are sampled at discrete time-points $i\Delta$, $i = 0, \ldots, n$, for some $\Delta > 0$, then the returns $Z_i = \log P_{i\Delta} - \log P_{(i-1)\Delta}$ are independent and identically $N((\alpha - \tau^2/2)\Delta, \tau^2\Delta)$-distributed. However, it is well-known that these properties are inconsistent with empirical findings: typically stock returns (*i*) are heavy-tailed; (*ii*) are uncorrelated but *not* independent; and (*iii*) have variance that varies (randomly) over time.

Of course, it is possible to generate such features by allowing for more complicated (non-linear) drift and diffusion functions for P; thereby staying in the class of one-dimensional diffusion models. In the stochastic volatility approach, however, the linearity of the drift and diffusion for P is retained, but an additional source of noise is introduced as the constant τ in (3.1) is replaced by the value of a diffusion process $\sqrt{V}$. The process V is latent and is interpreted as the random variance, or *volatility*, at the market. To be specific, the modified model is given by the two-dimensional stochastic differential equation

$$dP_t = \alpha P_t\, dt + \sqrt{V_t} P_t\, dW_t \tag{3.2}$$

$$dV_t = b(V_t, \theta)\, dt + \sigma(V_t, \theta)\, d\tilde{W}_t \tag{3.3}$$

where only P is observable at certain time-points. This kind of model is indeed able to generate data with the above properties. In this chapter, as well as in Paper III, we shall consider models where the drift function for P may depend on V as well.

Stochastic volatility models of the above type (and slight generalizations) were introduced in the finance literature in the late eighties and early nineties (for references, see Section 3.3). Focus was on option pricing which is not a simple issue for stochastic volatility models; essentially because volatility is not a traded asset. The pricing problem was investigated for fixed, known value of the parameter θ determining the distribution of V. The majority of the papers paid no, or very little, attention to estimation of this parameter.

3.2 The class of models

Consider the pair of stochastic differential equations

$$dX_t = \xi(V_t)\, dt + \sqrt{V_t}\, dW_t \tag{3.4}$$

$$dV_t = b(V_t, \theta)\, dt + \sigma(V_t, \theta)\, d\tilde{W}_t \tag{3.5}$$

defined on a filtered probability space $(\Omega, \mathscr{F}, \mathscr{F}_t, Pr)$. The drift and diffusion for V are parameter dependent, and in Section 3.4 we shall be concerned with estimation of θ from discrete-time observations $X_0, \ldots, X_{n\Delta}$ of X. The parameter θ is p-dimensional and varies in a set $\Theta \subset \mathbb{R}^p$. Note that, by Itô's formula, $P = e^X$ solves $dP_t = (\xi(V_t) + V_t/2)\, dt + \sqrt{V_t} P_t\, dW_t$ which simplifies to (3.2) if $\xi(v) = \alpha - v/2$.

The functions ξ, b and σ are assumed to be such that for all $\theta \in \Theta$ there is a unique, strong solution (X, V) with V positive almost surely. The Brownian

motions W and $\tilde{W}$ are assumed to be independent, and the drift and diffusion for X do not depend on X itself. Both assumptions are fundamental for the distributional result below and for the approximate maximum likelihood method described in Section 3.4.7 and Paper III. Although the method could easily be modified to work for models where the drift and diffusion for X are parameter dependent, we shall for simplicity assume that this is not the case.

Now, let us briefly mention some probabilistic properties of the model. We refer to Section III.2 for proofs and further details. For fixed $\Delta > 0$ define increments Z_i and integrals M_i and S_i for $i \in \mathbb{N}$ as

$$Z_i = X_{i\Delta} - X_{(i-1)\Delta}; \quad M_i = \int_{(i-1)\Delta}^{i\Delta} \xi(V_s)\, ds; \quad S_i = \int_{(i-1)\Delta}^{i\Delta} V_s\, ds,$$

and let $Z = (Z_1, Z_2, \dots)$ be the sequence of increments.

It is not possible to characterize the distribution of Z explicitly, but we have the following well-known result on the conditional distribution of Z given V (Proposition III.2): *Conditional on the process V, the increments $Z_1, Z_2, \dots$ are independent and Z_i is Gaussian with mean M_i and variance S_i. Furthermore, if V is strictly stationary, then so is Z.*

In the following we shall always assume that V is stationary. Let P_θ be the distribution of Z (on $\mathbb{R}^\infty$) when the parameter is θ and V is started according to its stationary distribution. It is easy to write moments of Z in terms of moments of S and M because of the conditional independence and normality given V. For example, if the relevant moments exist,

$$\mathrm{E}_\theta Z_i = \mathrm{E}_\theta M_1 \tag{3.6}$$

$$\mathrm{Var}_\theta Z_i = \mathrm{E}_\theta S_1 + \mathrm{Var}_\theta M_1 \tag{3.7}$$

$$\mathrm{E}_\theta Z_i^4 = 3\,\mathrm{E}_\theta S_1^2 + \mathrm{E}_\theta M_1^4 + 6\,\mathrm{E} M_1^2 S_1 \tag{3.8}$$

$$\mathrm{Cov}_\theta(Z_i, Z_j) = \mathrm{Cov}_\theta(M_1, M_{j-i+1}) \tag{3.9}$$

$$\mathrm{Cov}_\theta(Z_i^2, Z_j^2) = \mathrm{Cov}_\theta(S_1 + M_1^2, S_{j-i+1} + M_{j-1+1}^2) \tag{3.10}$$

for all $i, j \in \mathbb{N}$ with $j > i$.

Note that $\xi \equiv 0$ implies that for all $i \neq j$ (*i*) $\mathrm{E}_\theta Z_i^l = 0$ and $\mathrm{E}_\theta Z_i^l Z_j^l = 0$ if l is odd; (*ii*) $\mathrm{E}_\theta Z_i^4 / (\mathrm{E}_\theta Z_i^2)^2 > 3$; and (*iii*) $\mathrm{Corr}_\theta(Z_i^2, Z_j^2) < 1/3$. In particular, the stationary distribution of Z always has heavier tails than the normal distribution and the Z's are uncorrelated — but not independent — if $\xi \equiv 0$.

The two-dimensional diffusion process (X, V) is Markov, but the Markov property of X is spoiled by the latency of V, and neither $(X_0, X_\Delta, \dots)$ nor $(Z_1, Z_2, \dots)$ is Markov. Note however that the model is a *hidden Markov model* with hidden chain $\tilde{H}$ where $\tilde{H}_i = (V_{i\Delta}, M_i, S_i)$, see Genon-Catalot, Jeantheau & Laredo (1998*b*), and that the hidden chain has continuous state space.

3.3 Four particular models

In this section we study four particular stochastic volatility models. All have $\xi \equiv 0$ so the increments $Z_1, \ldots, Z_n$ are uncorrelated and have mean zero. As models for V we consider two mean-reverting models and two transformations of the Ornstein-Uhlenbeck process.

3.3.1 Mean-reverting models

Consider models of the type

$$dV_t = \alpha(\beta - V_t)\,dt + \sigma(V_t)\,d\tilde{W}_t$$

where α and β are positive parameters and σ is such that V is positive and stationary with finite second order moment. Furthermore, assume that the martingale part of V is a genuine martingale (not only local). The function σ may be parameter-dependent, and we write θ for the full parameter.

Many of the following moment calculations were carried out by Genon-Catalot *et al.* (1998*b*) but they are repeated here for completeness. First we compute moments of V. By the above assumptions, the conditional expectation of V_t given V_0 is given by

$$\mathrm{E}_\theta(V_t|V_0 = v) = e^{-\alpha t}(v-\beta) + \beta = e^{-\alpha t}v + \beta\big(1 - e^{-\alpha t}\big).$$

Hence, by stationarity, $\mathrm{E}_\theta V_0 = \beta$ and

$$\mathrm{E}_\theta V_0 V_t = \mathrm{E}_\theta V_0\, \mathrm{E}_\theta(V_t|V_0) = e^{-\alpha t}\,\mathrm{Var}_\theta V_0 + \beta^2.$$

In other words, β is the level of the volatility process and α controls the degree of temporal dependence in V. For α small the mean-reversion is weak and V has a tendency to stay above (or below) the mean level β for longer periods. In other words: there will be periods with large variability in Z and periods with small variability in Z. In finance this is referred to as *volatility clustering*.

Next we calculate moments of S: $\mathrm{E}_\theta S_1 = \int_0^\Delta \mathrm{E}_\theta V_s\,ds = \beta\Delta$ and for $j \in \mathbb{N}$ it holds that

$$\mathrm{E}_\theta S_1 S_j = \int_0^\Delta \int_{(j-1)\Delta}^{j\Delta} \mathrm{E}_\theta V_s V_u\,du\,ds = \beta^2\Delta^2 + \mathrm{Var}_\theta V_0 \int_0^\Delta \int_{(j-1)\Delta}^{j\Delta} e^{-\alpha|u-s|}\,du\,ds.$$

By direct computations and subtraction of $(\mathrm{E}_\theta S_1)^2 = (\mathrm{E}_\theta S_2)^2 = \beta^2\Delta^2$, it follows that

$$\mathrm{Var}_\theta S_1 = \frac{2\big(\alpha\Delta - 1 + e^{-\alpha\Delta}\big)}{\alpha^2}\,\mathrm{Var}_\theta V_0; \quad \mathrm{Cov}_\theta(S_1, S_2) = \frac{\big(1 - e^{-\alpha\Delta}\big)^2}{\alpha^2}\,\mathrm{Var}_\theta V_0.$$

It finally follows from (3.6)–(3.10) that $\mathrm{E}_\theta Z_1 = 0$, $\mathrm{Var}_\theta Z_1 = \beta\Delta$ and that

$$\mathrm{Var}_\theta Z_1^2 = 2\beta^2\Delta^2 + \frac{6\big(\alpha\Delta - 1 + e^{-\alpha\Delta}\big)}{\alpha^2}\,\mathrm{Var}_\theta V_0 \tag{3.11}$$

$$\mathrm{Cov}_\theta(Z_1^2, Z_2^2) = \frac{\big(1 - e^{-\alpha\Delta}\big)^2}{\alpha^2}\,\mathrm{Var}_\theta V_0. \tag{3.12}$$

Note that the latter two expressions only depend on β through the variance of V_0.

For (financial) applications it is important that the models are able to generate highly leptokurtic distributions. The (excess) kurtosis κ_θ of the stationary distribution of Z is given by

$$\kappa_\theta(Z_1) = \frac{\mathrm{E}_\theta Z_1^4}{(\mathrm{E}_\theta Z_1^2)^2} - 3 = \frac{6\big(\alpha\Delta - 1 + e^{-\alpha\Delta}\big)}{\alpha^2\Delta^2}\frac{\mathrm{Var}_\theta V_0}{\beta^2}$$

which is positive (as we knew) and less than $3\,\mathrm{Var}_\theta V_0/\beta^2$ — use the inequality $(\alpha\Delta - 1 + e^{-\alpha\Delta})/(\alpha\Delta)^2 < 1/2$. Similarly, by taking the reciprocal and using the inequalities $(\alpha\Delta)^2/(1-e^{-\alpha\Delta})^2 > 1$ and $(\alpha\Delta - 1 + e^{-\alpha\Delta})/(1-e^{-\alpha\Delta})^2 > 1/2$, we find that

$$\mathrm{Corr}_\theta(Z_1^2, Z_2^2) = \frac{\big(1-e^{-\alpha\Delta}\big)^2\,\mathrm{Var}_\theta V_0}{2\alpha^2\beta^2\Delta^2 + 6\big(\alpha\Delta - 1 + e^{-\alpha\Delta}\big)\,\mathrm{Var}_\theta V_0} < \frac{\mathrm{Var}_\theta V_0}{2\beta^2 + 3\,\mathrm{Var}_\theta V_0}.$$

Hence, if $\mathrm{Var}_\theta V_0/(\mathrm{E}_\theta V_0)^2 = \mathrm{Var}_\theta V_0/\beta^2$ is bounded by a constant K_θ, then the excess kurtosis is bounded by $3K_\theta$ (and positive), and the correlation is bounded by $K_\theta/(2+3K_\theta)$ (and positive).

In the following we shall consider two particular choices of the diffusion function σ.

The Cox-Ingersoll-Ross model

Let $\sigma(v) = \sigma\sqrt{v}$ for a constant σ and consider the equation

$$dV_t = \alpha(\beta - V_t)\,dt + \sigma\sqrt{V_t}\,d\tilde{W}_t.$$

The solution V is called a *Cox-Ingersoll-Ross process* (or square-root process) and was used by Hull & White (1988) and Heston (1993) in a stochastic volatility set-up.

Let $\theta = (\alpha, \beta, \sigma^2)$. If $\sigma^2 \le 2\alpha\beta$ then V is positive and stationary, and the stationary distribution is $\Gamma(2\alpha\beta/\sigma^2, \sigma^2/(2\alpha))$ so V, and therefore also Z, have moments of any order. In particular, $\mathrm{Var}_\theta V_0 = \beta\sigma^2/(2\alpha)$ which can be plugged into (3.11) and (3.12). For a given value of β, $\mathrm{Var}_\theta V_0 \le \beta^2$ since $\sigma^2 \le 2\alpha\beta$. Hence, it follows from the above that the excess kurtosis of Z is at most 3 and that the correlation between Z_1^2 and Z_2^2 is a most 1/5.

To get a better understanding of the model we have simulated 10.000 observations from it. We have used $\Delta = 1$ and parameter values $\alpha = 0.075$, $\beta = 1$ and $\sigma^2 = 0.12$. With these values of the parameters,

$$\mathrm{E}_\theta V_0 = 1, \quad \mathrm{Var}_\theta V_0 = 0.8, \quad \mathrm{Corr}_\theta(V_0, V_\Delta) = 0.928 \tag{3.13}$$

$$\mathrm{E}_\theta Z_1^2 = 1, \quad \mathrm{Var}_\theta Z_1^2 = 4.341, \quad \mathrm{Corr}_\theta(Z_1^2, Z_2^2) = 0.171. \tag{3.14}$$

Note that we have chosen α small in order to create longer periods with high volatility.

In practice the simulations were generated as follows: the Millstein scheme was used for simulation of V on the interval from 0 to $10.000\Delta = 10.000$, dividing each Δ-interval into 1000 subintervals; the integrals S were approximated by simple Riemann sums; and finally the Z's were drawn independently, Z_i from $N(0, S_i)$.

The top of Figure 3.1 shows the last 1000 simulated values of Z. The bottom shows the corresponding values $V_{i\Delta}$, $i = 9001, \ldots, 10.000$, of the volatility process (which would not be observable in applications). Clearly, the Z's are more volatile in periods with large values of V than in periods with low values of V.

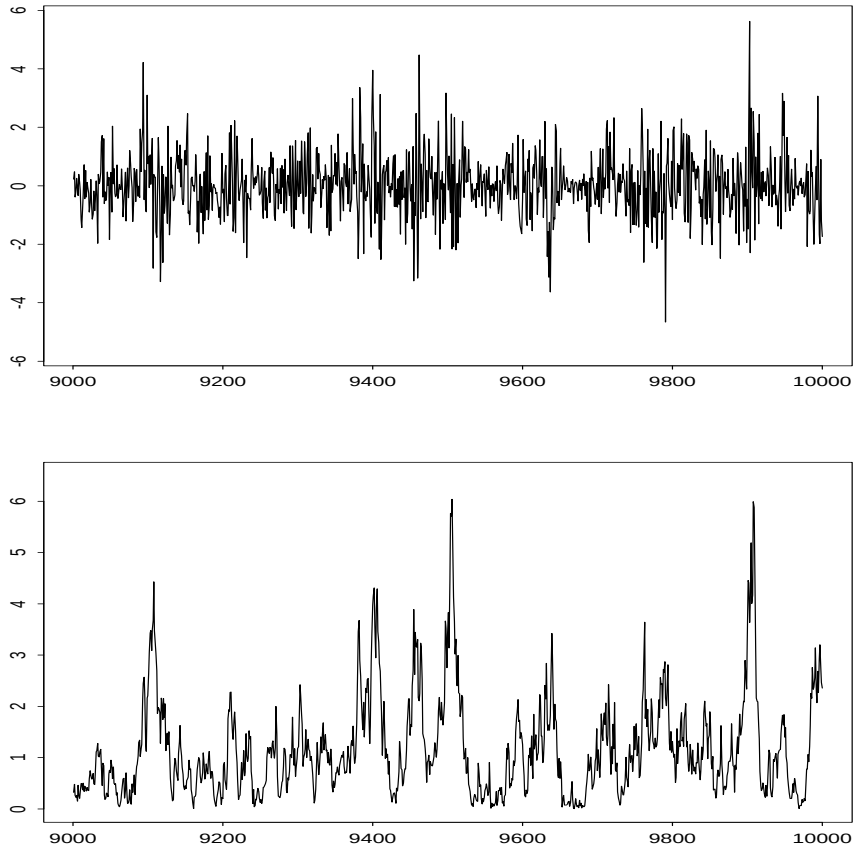

Figure 3.1: Simulated values of Z_i (top figure) and $V_{i\Delta}$ (bottom figure), $i = 9001, \ldots, 10.000$, for the Cox-Ingersoll-Ross model. The model parameters are $\alpha = 0.075$, $\beta = 1$ and $\sigma^2 = 0.12$, and $\Delta = 1$.

As expected, a correlogram of Z shows absolutely no activity and is hence not shown here. Correlograms for Z^2 and $|Z|$ (based on all 10.000 observations) are shown in Figure 3.2. The two correlograms are very similar, but there is a tendency that correlations between absolute values are slightly larger than correlations between squared values. It takes about 25 lags for the correlations to die out.

Figure 3.3 is a QQ-plot of Z (based on all 10.000 simulations): the empirical quantiles of the marginal distribution of Z are plotted against the quantiles of the normal distribution with mean zero and the same variance as Z. The dashed line

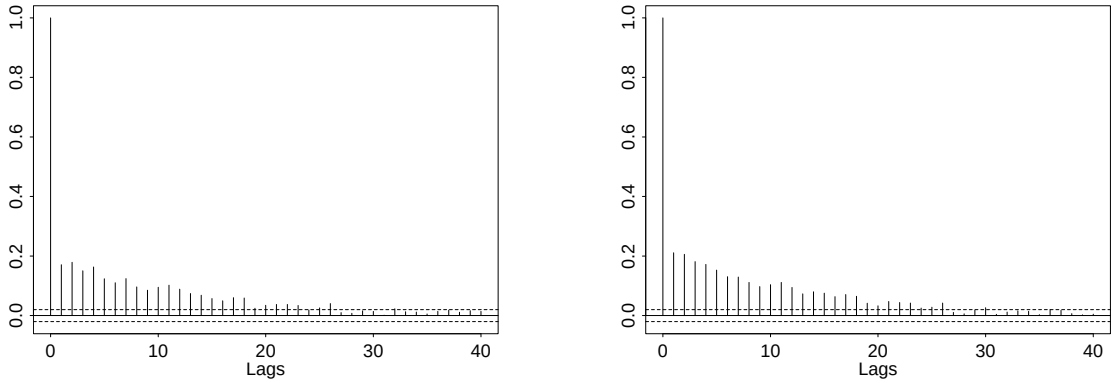

Figure 3.2: Correlograms of Z^2 (to the left) and $|Z|$ (to the right) from the Cox-Ingersoll-Ross model. The dashed lines are approximate confidence intervals.

goes through (0,0) and has slope 1. As expected, the distribution of Z has heavier tails than the normal distribution. The solid line shows the quantiles of a scaled t-distribution, $\rho\, t(f)$. The parameters ρ and f are estimated by requiring that the second and fourth order moment equal those of the empirical distribution of Z. This holds for $\rho = 0.87$ and $f = 8.30$. The scaled t-distribution fits quite well.

Inverse Gamma model

Consider $\sigma(v) = \sigma v$ and the corresponding equation

$$dV_t = \alpha(\beta - V_t)\, dt + \sigma V_t\, d\tilde{W}_t.$$

This model is the continuous-time limit (in a suitable sense) of the GARCH(1,1)-model in discrete time (Nelson 1990). The solution V is positive and stationary. The stationary distribution is the inverse Gamma distribution with parameters $1+2\alpha/\sigma^2$ and $\sigma^2/(2\alpha\beta)$, that is, the stationary distribution of $1/V$ is $\Gamma(1+2\alpha/\sigma^2, \sigma^2/(2\alpha\beta))$. Again let $\theta = (\alpha, \beta, \sigma^2)$. In the following we shall simply refer to the model as *the inverse Gamma model*, with parameter $\theta = (\alpha, \beta, \sigma^2)$.

The inverse Gamma distribution with parameters $1+2\alpha/\sigma^2$ and $\sigma^2/(2\alpha\beta)$ has finite moment of order r if and only if $r < 1+2\alpha/\sigma^2$. Note the difference from the Cox-Ingersoll-Ross model which has finite moments of any order. The mean of V is β and if $\sigma^2 < 2\alpha$, then V has variance $\beta^2\sigma^2/(2\alpha - \sigma^2)$. For any fixed value of β it is thus possible to get the fraction $\mathrm{Var}_\theta V_0/\beta^2$ arbitrarily large by choosing 2α and σ^2 close. In particular, the kurtosis is unbounded and the correlation between Z_1^2 and Z_2^2 can be arbitrarily close to the upper limit 1/3.

We have simulated 10.000 observations from the model using the same random numbers as for the Cox-Ingersoll-Ross model. We have used the same values of α and β (0.075 and 1) as above, but σ^2 is chosen differently, equal to 0.0667, in order to make the variance and correlation structure the same for the two models, that is, in order for (3.13)–(3.14) to hold.

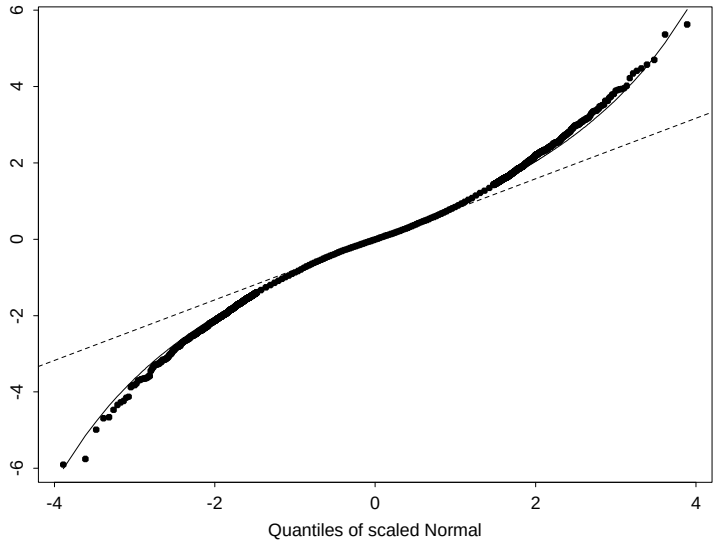

Figure 3.3: The empirical quantiles of Z from the Cox-Ingersoll-Ross model (on the y-axis) plotted against the quantiles of the normal distribution with mean zero and variance equal to the empirical variance of Z (on the x-axis). The dashed line has slope 1 and goes through (0,0). The solid line shows the quantiles of the scaled t-distribution, $\rho\, t(f)$ which has same second and fourth order moment as the empirical distribution of Z. Here, $f = 8.30$ and $\rho = 0.87$.

The simulated volatility process is similar to that of the Cox-Ingersoll-Ross model in the sense that the two processes take large (small) values at the same time. However, the inverse Gamma model produces larger spikes (implying a heavy right tail), whereas the Cox-Ingersoll-Ross model produces many very small values (implying a heavy left tail). This is of course completely in line with their marginal distributions: with the above parameter values $\mathrm{E}_\theta V_0^r$ is finite for $r > -1.25$ in the Cox-Ingersoll-Ross model and for $r < 3.25$ in the inverse Gamma model.

The distribution of Z depends on V through the distribution of the smoothed (integrated) variables S_i only. This smoothing, and the extra Brownian noise W in the equation for X, seem to almost quell the differences between the two models. In Figure 3.4 the quantiles of the two sets of Z's are plotted against each other. They are almost indistinguishable. Also, correlograms from the inverse Gamma model are indistinguishable from those of the Cox-Ingersoll-Ross model and are omitted. Altogether, this suggests that there is not much difference between the two distributions of Z as long as parameters are chosen such that the low order moments of Z are the same.

As mentioned above, one important objection to the Cox-Ingersoll-Ross model is that the Gamma distribution of Z can only be moderately heavy-tailed. The same objection does not apply to the inverse Gamma model: heavier tails are

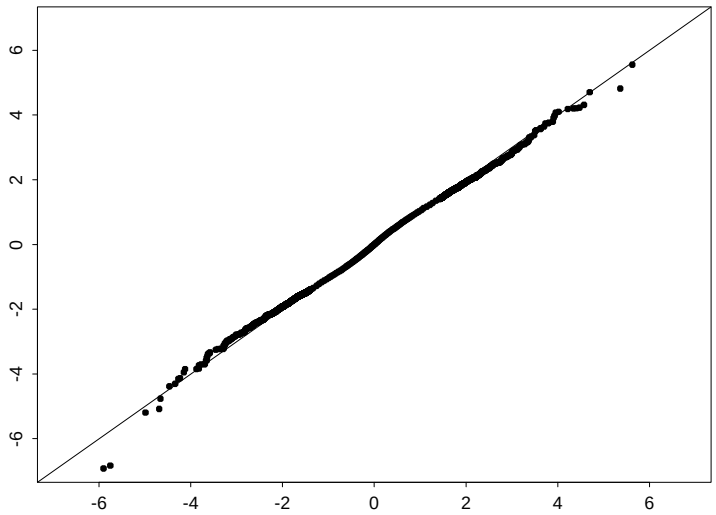

Figure 3.4: The quantiles of Z in the inverse Gamma model (on the y-axis) plotted against the quantiles of Z in the Cox-Ingersoll-Ross model (on the x-axis). The parameter values are $\alpha = 0.075$ and $\beta = 1$ in both models whereas $\sigma^2 = 0.12$ in the Cox-Ingersoll-Ross model and $\sigma^2 = 0.0667$ in the inverse Gamma model.

obtained by choosing σ^2 closer to 2α. To illustrate this we have simulated 10.000 observations from the model with $\alpha = 0.075$, $\beta = 1$ (as above) and $\sigma^2 = 0.12$ (as for the Cox-Ingersoll-Ross model). For these values the distribution of V has finite second, but not third, order moment, so Z has finite fourth, but not sixth, order moment. Figure 3.5 shows a QQ-plot of Z. The dashed line corresponds to a normal distribution with mean zero and variance equal to the empirical variance of Z, and the solid line corresponds to the scaled t-distribution $\rho\, t(f)$ with $\rho = 0.882$ and $f = 9.68$ which has same second and fourth moments as Z. Clearly, the distribution of Z is far more leptokurtic than the scaled t-distribution.

3.3.2 Ornstein-Uhlenback based models

In many respects the Ornstein-Uhlenbeck process is the simplest diffusion process apart from the Brownian motion and the geometric Brownian motion. It solves the equation $d\tilde{V}_t = \alpha(\beta - \tilde{V}_t)\,dt + \sigma\, d\tilde{W}_t$ which can be solved explicitly. For $\alpha \neq 0$ the solution $\tilde{V}$ has Gaussian transition probabilites, $\tilde{V}_t|\tilde{V}_0 \sim N(\lambda_1(t)\tilde{V}_0 + \lambda_2(t), \tau^2(t))$ where $\lambda_1(t) = e^{-\alpha t}$, $\lambda_2(t) = \beta(1 - e^{-\alpha t})$ and $\tau^2(t) = \sigma^2(1 - e^{-2\alpha t})/(2\alpha)$. For $\alpha > 0$, $\tilde{V}$ is stationary with $N(\beta, \sigma^2/(2\alpha))$ as its stationary distribution. The normality implies that the model cannot directly be used as a model of the positive volatility process, but we may transform it and still be able to utilize its nice properties.

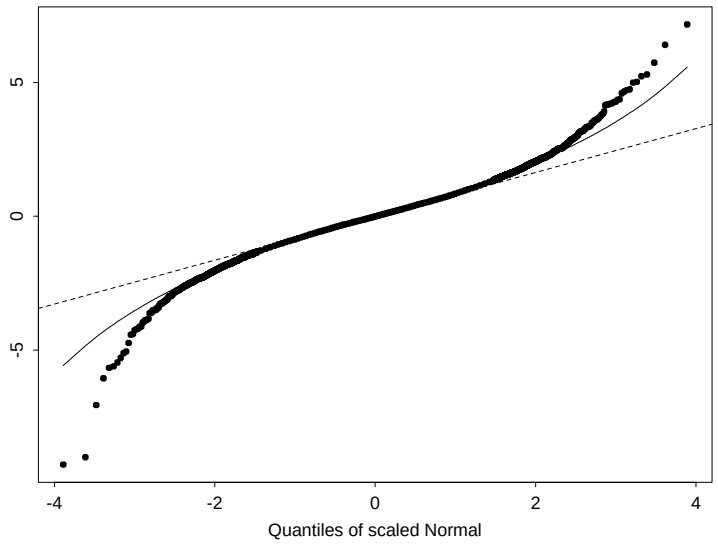

Figure 3.5: The empirical quantiles of Z from the inverse Gamma model with parameters $\alpha = 0.075$, $\beta = 1$ and $\sigma = 0.12$ plotted against the quantiles of the normal distribution with mean zero and variance equal to the empirical variance of Z. The dashed line has slope 1 and goes through (0,0). The solid line shows the quantiles of the scaled t-distribution, $\rho\, t(f)$ which has same second and fourth order moment as the empirical distribution of Z. Here, $f = 9.68$ and $\rho = 0.88$.

The geometric Ornstein-Uhlenbeck process

The specification $V = \exp(\tilde{V})$ was suggested by Wiggins (1987) and Chesney & Scott (1989). We shall refer to V as the *geometric Ornstein-Uhlenbeck process*. Both the stationary distribution and the transition probabilities are log-normal, and V and Z have moments of any order. We easily find

$$\mathrm{E}_\theta V_0 V_t = \exp\bigl(2\beta + \sigma^2/(2\alpha) + e^{-\alpha t}\sigma^2/(2\alpha)\bigr),$$

but it is not easy (if possible at all) to find $\mathrm{E}_\theta S_1^2 = \int_0^\Delta \int_0^\Delta \mathrm{E}_\theta V_u V_s \, du\, ds$ or $\mathrm{E}_\theta S_1 S_2 = \int_0^\Delta \int_\Delta^{2\Delta} \mathrm{E}_\theta V_u V_s \, du\, ds$ explicitly so we have no explicit expression for the moments of Z (except those that are zero, of course).

Note that the approximation $S_1 \approx \Delta V_0$ leads to the approximation

$$\kappa_\theta(Z_1) = \frac{\mathrm{E}_\theta Z_1^4}{\bigl(\mathrm{E}_\theta Z_1^2\bigr)^2} - 3 = 3\frac{\mathrm{E}_\theta S_1^2}{\bigl(\mathrm{E}_\theta S_1\bigr)^2} - 3 \approx 3\frac{\mathrm{Var}_\theta V_0}{\bigl(\mathrm{E}_\theta V_0\bigr)^2} \tag{3.15}$$

of the excess kurtosis of Z. The fraction $\mathrm{Var}_\theta V_0/(\mathrm{E}_\theta V_0)^2$ is unbounded so there is presumably no bound on the kurtosis of Z in the geometric Ornstein-Uhlenbeck model.

We are not able to determine parameter values such that the distribution of Z has certain values. However, we can easily determine values of α, β and σ^2

such that (3.13) holds: $\alpha = 0.0571$, $\beta = -0.294$ and $\sigma^2 = 0.0672$. Note that the value of σ^2 is close to the corresponding value for the inverse Gamma model (0.0667). This is not very surprising since V, by Itô's formula, solves the stochastic differential equation

$$dV_t = \Big(\alpha\big(\beta - \log(V_t)\big)V_t + \sigma^2 V_t/2\Big)\,dt + \sigma V_t\,d\tilde{W}_t$$

with same diffusion function as the inverse Gamma model. Simulations of the geometric Ornstein-Uhlenbeck process with the above parameters are almost indistinguishable from those of the two mean-reverting models.

The squared Ornstein-Uhlenbeck process

The *squared Ornstein-Uhlenbeck process* $V = \tilde{V}^2$ was used by Scott (1987) and Stein & Stein (1991). Under this model, V and thus Z have moments of any order. It is easily verified that $\mathrm{Var}_\theta V_0/(\mathrm{E}_\theta V_0)^2 \leq 2$ so the kurtosis of Z is presumably bounded by a value around 6, *cf.* (3.15). The model has several disadvantages compared to the previous models: (*i*) V is not strictly positive but hits zero; (*ii*) V is not a diffusion unless $\beta = 0$ (the drift term in the stochastic differential equation for V cannot be written in terms of V but involves $\tilde{V}$ as well); and (*iii*) there are no explicit expressions for covariances between V_0 and V_t, say.

The covariances may be calculated by simulation from the invariant distribution of V, though. For this, note that $\mathrm{E}_\theta V_0 V_t$ equals

$$\big(\tau^2(t) + \lambda_2^2(t)\big)\,\mathrm{E}_\theta V_0 + \lambda_1^2(t)\,\mathrm{E}_\theta V_0^2 + 2\lambda_1(t)\lambda_2(t)\,\mathrm{E}_\theta\big\{V_0^{3/2} g\big(V_0, \beta, \sigma^2/(2\alpha)\big)\big\}$$

where the function g is given by

$$g(v, m, s^2) = \frac{\exp(m\sqrt{v}/s^2) - \exp(-m\sqrt{v}/s^2)}{\exp(m\sqrt{v}/s^2) + \exp(-m\sqrt{v}/s^2)}.$$

The formula is derived via repeated expectations

$$\mathrm{E}_\theta\big(V_t | V_0\big) = \mathrm{E}_\theta\big(\tilde{V}_t^2 | \tilde{V}_0^2\big) = \mathrm{E}_\theta\Big(\mathrm{E}_\theta\big(\tilde{V}_t^2 \big| \tilde{V}_0\big)\Big|\tilde{V}_0^2\Big),$$

where the inner expectation is computable as the transitions of $\tilde{V}$ are normal.

For $\alpha = 0.0661$, $\beta = 0.880$ and $\sigma^2 = 0.0298$ the values of $\mathrm{E}_\theta V_0$, $\mathrm{Var}_\theta V_0$ and $\mathrm{Corr}_\theta(V_0, V_\Delta)$ are (roughly) as in (3.13). Simulations of Z from the model with these parameter values are very similar to the simulations of the three previous models.

3.3.3 Concluding remarks

The above investigation indicates that the four models produce very similar distributions of Z — as long as the model parameters are chosen such that the low-order moments of V are the same. The fact that the eighth order moment of Z

is infinite for the inverse Gamma model and finite for the other models is hardly recognizable from the simulations. The reason is that differences in the volatility distributions are suppressed by smoothing (when V is transformed to integrals S) and by the extra noise in the equation for X. Still, there are important differences between the four models in their ability to create strong leptokurtosis and large auto-correlations: only two models (the inverse Gamma and the geometric Ornstein-Uhlenbeck) allow for arbitrarily large kurtosis and maximal autocorrelation (which is 1/3).

Of course one could think of many other models for the volatility process. Presumably, most of them would generate distributions of Z very similar to those above as long as parameters are chosen appropriately. In conclusion: If data are heavy-tailed to an extent that cannot be modeled by the two restrictive models, then one should use the inverse Gamma model or the geometric Ornstein-Uhlenbeck model. On the other hand, if there are no stylized facts contradicting any of the models and if there are no prior (economic) reasons to prefer one specification to another, then it seems less important which one of the models is applied. Note however that the squared Ornstein-Uhlenbeck model has some disadvantages which makes it the least attractive of the four models.

3.4 Estimation methods

In the majority of the early finance papers on stochastic volatility models (referenced in Section 3.3) no or little attention is paid to estimation problems. The possibility of doing parameter estimation via historical option prices and reversed versions of the option pricing formulas is mentioned but not carried out in practice. Anyway, this is a relatively indirect estimation approach. Only Scott (1987) and Chesney & Scott (1989) address the estimation problem seriously and derive moment-like estimators based on historical stock prices.

Recently there has been some progress in the statistics literature concerning stochastic volatility models, and the aim of this section is to give an overview of existing methods. Furthermore, a new method based on simulated approximations to the likelihood is reviewed in Section 3.4.7; a detailed discussion is provided in Paper III. There exist review papers on statistical analysis of stochastic volatility models defined in *discrete time* (Ghysels *et al.* 1996, Shephard 1996), but as mentioned in the beginning of this chapter the continuous-time models are in general more difficult to handle.

Now, the situation is the following. Consider the model given by (3.4)–(3.5), and assume that observations $X_0, X_\Delta, \ldots, X_{n\Delta}$ of X are available for some fixed Δ while the volatility process is unobserved. Because of the nice conditional distribution (given V) of the increments $Z_i = X_{i\Delta} - X_{(i-1)\Delta}$, $i = 1, \ldots, n$, it is natural to base estimation on the increments. The estimation problem is inherently difficult: apart from the usual problems due to discrete-time observations of a continuous-time system, we are faced with yet another missing data problem due to the latency of V. As a consequence, most of the estimation procedures below are very

computationally intensive.

Some remarks before reviewing the methods. First note that exact maximum likelihood estimation is not really an option: For an observation $(z_1, \ldots, z_n)$ it follows by conditional independence and normality given V that the likelihood function is given by

$$L_n(\theta) = \int \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s_i}} \exp\Big(-\frac{(z_i - m_i)^2}{2s_i}\Big)\, d\pi_\theta^n(h^n) = \mathrm{E}_{\pi_\theta^n} \prod_{i=1}^{n} \varphi(z_i, M_i, S_i) \tag{3.16}$$

where we have used the notation $\varphi(\cdot, m, s)$ for the density of $N(m,s)$ and π_θ^n for the distribution of $H^n = ((M_1, S_1), \ldots, (M_n, S_n))$. There is no closed-form expression for the likelihood, not even for very simple models of V. In principle, values of the likelihood function could be computed by simulation as follows: (*i*) simulate a large number of paths V up to time $n\Delta$ according to (3.5); (*ii*) calculate for each simulation (an approximation to) the integrals M_i and S_i, $i = 1, \ldots, n$, and the above product; and (*iii*) calculate the average of the simulated product values. However, this is not feasible in practice as a huge number of simulations would be required in order for the average to converge, that is, in order to compute the likelihood for just a *single* value of the parameter. Our approach in Section 3.4.7 and Paper III will be to use the simulation approach on *approximations* to the likelihood function.

Second, as we are faced with discrete-time observations of a continuous-time model, a natural approach would be to perform estimation in a discrete-time approximation to the original model. If we use the Euler scheme or some stochastic volatility model in discrete time, we are still left with an unobserved component which makes estimation difficult. However, Nelson (1990) showed that some diffusion processes can be approximated by ARCH type processes. For example, the limit of a GARCH(1,1) model is the stochastic volatility model with inverse Gamma-distributed volatility discussed in Section 3.3.1. Approximation by ARCH type models is advantageous as estimation is relatively simple. The problem is of course that the approximation is only good if the time between observations, Δ, is small so consistent estimators are obtained only if $\Delta \to 0$. The methods based on auxiliary models (Section 3.4.5) correct for this bias by simulation.

Third, apart from estimation of the model parameters, one could also be interested in filtering, that is, estimation of the unobserved volatility process V. Nelson (1992) suggests a filtering method based on ARCH approximations. Given estimates, $\hat{V}_0, \hat{V}_\Delta, \ldots, \hat{V}_n$, one could estimate θ by one of the methods in Chapter 2 as if $\hat{V}_0, \hat{V}_\Delta, \ldots, \hat{V}_{n\Delta}$ were actual observations of V. Nielsen, Vestergaard & Madsen (2000) use another technique that simultaneously delivers parameter estimates and estimates of the volatility process.

Except from Section 3.4.6 where the latter filter method is reviewed, the rest of this chapter is exclusively about parameter estimation. We discuss simple moment estimators in Section 3.4.1, estimation based on a simple approximation to the marginal distribution in Section 3.4.2, prediction-based estimating functions in Section 3.4.3, Bayesian analysis in Section 3.4.4, and methods based on auxiliary

models in Section 3.4.5. The latter two sections are very brief as the methods have been discussed in Sections 2.5 and 2.6 for pure diffusion models. The new simulated, approximate maximum likelihood method is discussed in Section 3.4.7 (and Paper III). Conclusions are drawn in Section 3.4.8.

3.4.1 Moment estimation

Moment estimators are obtained by matching empirical and theoretical moments. As already noted, moments of Z are easily expressed in terms of moments of the integrals M and S. These can be calculated explicitly in simple mean-reverting models (Section 3.3.1) and by simulation in more complex models.

Recall that p is the dimension of the parameter and choose p functionals $g_1, \ldots, g_p$, for example of type $z_1 \to z_1^j$ and $(z_1, z_2) \to z_1^j z_2^j$ for suitable (small) values of j. The parameter should be uniquely determined by the theoretical moments of $g_1, \ldots, g_p$. With sloppy notation, the requirement is

$$\big(\mathrm{E}_\theta\, g_1, \ldots, \mathrm{E}_\theta\, g_p\big) \neq \big(\mathrm{E}_{\theta'}\, g_1, \ldots, \mathrm{E}_{\theta'}\, g_p\big), \quad \theta \neq \theta'.$$

Then a natural estimate of θ is the value that makes the theoretical moments of $g_1, \ldots, g_p$ match their empirical counterparts. Also, moments may be used as building blocks for *the generalized method of moments*, GMM (Hansen 1982) which is very popular in the econometric literature. Here $q > p$ moment functionals are selected, and θ is estimated such that certain linear combinations of the theoretical moments are close to their empirical counterparts.

Genon-Catalot *et al.* (1998*b*) showed consistency and asymptotic normality of the empirical moments (in the model with $\xi \equiv 0$). By transformation, the properties carry over to the moment estimators. Moment matching is quick compared to other methods, in particular for models where moments are known analytically. However, a simulation study in Section III.7 indicates that a solution to the estimating equation fairly often does *not* exist and that moments estimators can be very poor for a sample size of 500.

3.4.2 Approximation of the marginal density

In Section 3.4.7 we study approximations to the likelihood. The simplest of these approximations, denoted L_n^0, corresponds to pretending that $Z_1, \ldots, Z_n$ are independent, identically distributed according to the stationary (marginal) distribution, that is $L_n^0(\theta) = \prod_{i=1}^n p_\theta^1(z_i)$ where

$$p_\theta^1(z) = \int \varphi(z, m, s)\, d\pi_\theta^1(m, s) = \mathrm{E}_{\pi_\theta^1}\, \varphi(z, M, S)$$

is the stationary density of Z and π_θ^1 is the distribution of (M_1, S_1).

The distribution of (M_1, S_1) is not known so we have no explicit expression for the above density. In Section 3.4.7 we suggest to determine it by simulation.

Alternatively, the density could be approximated as suggested by Genon-Catalot, Jeantheau & Laredo (1999): use the approximations

$$M_1 = \int_0^\Delta \xi(V_s)\,ds \approx \Delta\xi(V_0), \quad S_1 = \int_0^\Delta V_s\,ds \approx \Delta V_0 \tag{3.17}$$

and the corresponding approximation

$$\tilde{p}_\theta^1(z) = \int \varphi\big(z, \Delta\xi(v), \Delta v\big)\,d\mu_\theta(v) = \mathrm{E}_{\mu_\theta}\,\varphi\big(z, \Delta\xi(V_0), \Delta V_0\big)$$

of the stationary density where μ_θ is the invariant distribution of V (which is known analytically). Genon-Catalot *et al.* (1999) calculate the density $\tilde{p}_\theta^1$ explicitly for the two mean-reverting models from Section 3.3.1; more generally it could be calculated by simulation.

The approximations (3.17) are good for "small" values of Δ, and Genon-Catalot *et al.* (1999) indeed show that the estimator obtained by maximizing $\tilde{L}_n^0(\theta) = \prod_{i=1}^n \tilde{p}_\theta^1(z_i)$ is consistent as $n \to \infty$ if $\Delta = \Delta_n \to 0$ and $n\Delta = n\Delta_n \to \infty$. If furthermore $n\Delta_n^2 \to 0$ then the estimator is asymptotically normal. The proofs are based on limit theorems proved in an earlier paper (Genon-Catalot, Jeantheau & Laredo 1998*a*). If Δ is fixed then the estimator is inconsistent.

The method has two severe disadvantages: (*i*) the bias can be considerable if Δ is not small; and (*ii*) only parameters from the marginal distribution of V can be estimated. Both disadvantages are resolved if the *true* marginal density p_θ^1 is used instead of $\tilde{p}_\theta^1$: the corresponding estimator is consistent as $n \to \infty$ for any fixed Δ, and the marginal distribution of Z typically determines all parameters (at least theoretically). The drawback is of course that the density p_θ^1 is not analytically tractable and must be simulated. See Section 3.4.7 and Paper III for further details.

3.4.3 Prediction-based estimating functions

Martingale estimating functions are important tools for estimation in the (pure) diffusion models where the Markov structure gives rise to natural martingales based on conditional expectations one step ahead. For non-markovian models there are no such simple martingales, and so-called *prediction-based estimating functions* may be useful (Sørensen 1999).

To keep things simple we consider a somewhat simpler set-up than Sørensen (1999). In particular, he considers estimating functions that are sums of N terms of type (3.18) below. This is probably more powerful for high-dimensional parameters. Also, note that it is nowhere important that Z stems from a stochastic volatility model; indeed the estimating functions are applicable for a large range of models.

We need some notation. Let $\mathscr{F}_i^Z$ denote the σ-algebra generated by $Z_1, \ldots, Z_i$. Let $\mathscr{H}_i^\theta$ be the L^2-space of $\mathscr{F}_i^Z$-measurable random variables that have finite second order moment wrt. P_θ and let $\mathscr{P}_i^\theta$ be a closed, linear subspace of $\mathscr{H}_i^\theta$. Furthermore, let $f : \mathbb{R} \to \mathbb{R}$ be a function with $\mathrm{E}_\theta\, f^2(Z_i) < \infty$.

Now, consider the estimating function

$$F_n(\theta) = \sum_{i=1}^{n} w_i(\theta)\Big(f(Z_i) - \hat{\pi}_i(\theta)\Big) \tag{3.18}$$

where $\hat{\pi}_i(\theta)$ is the orthogonal projection of $f(Z_i)$ on $\mathscr{P}_{i-1}^{\theta}$ and $w_i(\theta)$ is a p-dimensional vector with coordinates belonging to $\mathscr{P}_{i-1}^{\theta}$. It is well-known that $\hat{\pi}_i(\theta)$ is the minimum mean square error predictor of $f(Z_i)$ in $\mathscr{P}_{i-1}^{\theta}$. The properties of the orthogonal projection ensures that F_n is an unbiased estimating function, *i.e.* $\mathrm{E}_\theta F_n(\theta) = 0$ for all $\theta \in \Theta$.

Note that F_n is a martingale if $\mathscr{P}_i^{\theta} = \mathscr{H}_i^{\theta}$ (or more generally if and only if the conditional expectation $\mathrm{E}_\theta(f(Z_i)|Z_1,\ldots,Z_{i-1})$ of $f(Z_i)$ given all the past belongs to $\mathscr{P}_{i-1}^{\theta}$ for all i). The sets $\mathscr{P}_i^{\theta}$ are called sets of predictors. As an example, $\mathscr{P}_i^{\theta}$ could be spanned by $1, h(Z_i), \ldots, h(Z_{i-k+1})$ for some function $h : \mathbb{R} \to \mathbb{R}$ and some $k \geq 0$. The constant is included in order to ensure unbiasedness.

Sørensen (1999) shows consistency and asymptotic normality of the estimator obtained as solution to the equation $F_n(\theta) = 0$. Also, given f and a finite-dimensional set of predictors he finds optimal weights, yielding estimators with the least possible asymptotic variance. There is no theory on how to select the basis function f optimally. In practice one would probably use low order polynomials or other simple functions, regardless of the model. This need not be efficient, though, and indicates some amount of built-in arbitrariness. In particular the method is not invariant to transformations of data.

The method seems more promising than the previous ones as it (*i*) does not introduce bias and (*ii*) is able to take into account more features of the distribution than the simple moment estimators. The drawback is of course the need for relatively time-consuming numerical procedures: the projection (and the optimal weights) must typically be computed by simulation so the method is far slower than the previous ones.

3.4.4 Bayesian analysis

Bayesian analysis has been applied to stochastic volatility models by Eraker (1998), see also Elerian *et al.* (2000). The parameter is considered as random, and a sequence $(\theta_j)_j$ is simulated that has the posterior of θ as the limiting distribution. For each j this involves simulation of X and V as well: values of X are simulated at a number of time-points in between those where X is observed; V also at the time-points $i\Delta$, $i = 1,\ldots,n$. See Section 2.5 (or the above quoted papers, of course) for details. The method is computationally quite demanding. However, it is also extremely flexible as it does not rely on probabilistic properties of the model, but purely on simulation (an on a prior distribution for θ which is chosen somewhat arbitrarily).

The flexibility is indicated by an application by Eraker (1998). He analyses US interest rate data as well as simulated data using the CKLS-inspired model

$$dX_t = \theta_1(\theta_2 - X_t)\,dt + X_t^{\theta_3}\sqrt{V_t}\,dW_t \tag{3.19}$$

where $\log V_t$ is an Ornstein-Uhlenbeck process. The model does not belong to the class of models discussed so far as equation (3.19) for X is not completely determined by the volatility process. The conditional independence and normality of the Z's (given V) do not hold, and the methods discussed so far do not easily apply to the model. Neither does the approximate maximum likelihood method from Section 3.4.7.

3.4.5 Estimation based on auxiliary models

In the empirical econometric literature stochastic volatility models have been estimated by *indirect inference* (Gourieroux *et al.* 1993) and the so-called *efficient methods of moments*, EMM (Gallant & Tauchen 1996). Essentially the idea is to find the parameter value for which simulated data resembles the observed data the most in the sense that the simulated data and the observations yield the same estimator in some auxiliary model. See Section 2.6 for more details.

Gourieroux *et al.* (1993) apply indirect inference to the modified Black-Scholes model with $\log V$ being an Ornstein-Uhlenbeck process (and the two Brownian motions possibly correlated). As auxiliary model they use a discrete-time stochastic volatility model which is estimated via the Kalman filter. In our opinion an ARCH type model would be a more obvious choice as it would be easier to handle statistically; a relatively simple one like GARCH(1,1) would probably suffice for this application.

Andersen & Lund (1997) apply EMM on interest rate data using the model (3.19) where V is again the exponential of an Ornstein-Uhlenbeck process. They try a few different, though similar, ARCH type auxiliary models with up to 26 parameters whereas the model of interest has six parameters only (three from (3.19) and three from the Ornstein-Uhlenbeck process)!

As mentioned above, most of the methods in this section do not apply to the model (3.19), and EMM may thus be helpful. However, we are in general critical to the methods based on auxiliary models, see Section 2.6. For simple models where more direct estimation techniques are possible we believe that these should indeed be preferred.

3.4.6 Estimation based on non-linear filters

In some applications it might be of interest to use the observations of X for *filtering*, *i.e.* estimation of the unobserved volatility process V. Nielsen *et al.* (2000) discuss a method that simultaneously provides estimates of the unknown parameter as well as of the values $V_{i\Delta}$ of V at the time-points where X is observed. Their set-up is somewhat more general than ours as the Brownian motions W and $\tilde{W}$ may be correlated and the observable process is observed with noise. Also, the method applies (at least in principle) to systems of higher dimensions and of a more complicated nature. Here we only consider the model given by (3.4) and (3.5), with W and $\tilde{W}$ independent and X observed without noise.

We need some notation. For $i = 1, \ldots, n$ let $\mathscr{F}_i^X$ denote the σ-algebra generated by $\{X_{j\Delta}\}_{j=0,\ldots,i}$ and let

$$\left(\hat{X}^\theta_{t|i-1}, \hat{V}^\theta_{t|i-1}\right) = \mathrm{E}_\theta\left((X_t, V_t)\middle|\mathscr{F}^X_{i-1}\right), \quad t \in [(i-1)\Delta, i\Delta]$$

be the prediction of (X_t, V_t) at time $(i-1)\Delta$ and $P^\theta_{t|i-1}$ the corresponding prediction variance. For $t = i\Delta$ we also write $(\hat{X}^\theta_{i|i-1}, \hat{V}^\theta_{i|i-1})$ and $P^\theta_{i|i-1}$.

For fixed θ the filter consists of two sets of equations: the prediction equations and the updating equations. The *prediction equations* determine equations for the time derivatives $\partial \hat{X}^\theta_{t|i-1}/\partial t$, $\partial \hat{V}^\theta_{t|i-1}/\partial t$ and $\partial P^\theta_{t|i-1}/\partial t$. The equations are only approximate. They are derived via Taylor expansions of the drift and diffusion functions truncated after the second order term. Third and fourth order conditional moments are approximated by simple expressions in terms of $\hat{X}^\theta_{t|i-1}$, $\hat{V}^\theta_{t|i-1}$ and $P^\theta_{t|i-1}$, corresponding to normality of the predictions. For the above model, this amounts to

$$\begin{pmatrix} \partial \hat{X}^\theta_{t|i-1}/\partial t \\ \partial \hat{V}^\theta_{t|i-1}/\partial t \end{pmatrix} = \begin{pmatrix} \hat{\xi}_t + \frac{1}{2}\hat{\xi}_t'' P_t^{22} \\ \hat{b}_t + \frac{1}{2}\hat{b}_t'' P_t^{22} \end{pmatrix} \tag{3.20}$$

$$\frac{\partial P^\theta_{t|i-1}}{\partial t} = \begin{pmatrix} \hat{V}^\theta_{t|i-1} + 2\hat{\xi}_t' P_t^{12} & \hat{\xi}_t' P_t^{22} + \hat{b}_t' P_t^{12} \\ \hat{\xi}_t' P_t^{22} + \hat{b}_t' P_t^{12} & \hat{\sigma}_t^2 + P_t^{22}\left(2\hat{b}_t' + (\hat{\sigma}_t')^2 + \hat{\sigma}_t \hat{\sigma}_t'\right), \end{pmatrix} \tag{3.21}$$

where $\hat{b}_t' = b'(\hat{V}_{t|i-1}, \theta)$, for example, and P_t^{jk} is short for the (j,k)'th element of $P^\theta_{t|i-1}$.

The *updating equations* express how the estimate of $V_{i\Delta}$ and its variance are modified as a new observation $X_{i\Delta}$ becomes available:

$$\begin{aligned} \hat{V}^\theta_{i|i} &= \mathrm{E}_\theta\left(V_{i\Delta}|\mathscr{F}_i^X\right) = \hat{V}^\theta_{i|i-1} + \left(X_{i\Delta} - \hat{X}^\theta_{i|i-1}\right) P^{\theta,12}_{i|i-1}/P^{\theta,11}_{i|i-1} \\ R^\theta_{i|i} &= \mathrm{Var}_\theta\left(V_{i\Delta}|\mathscr{F}_i^X\right) = P^{\theta,22}_{i|i-1} - \left(P^{\theta,12}_{i|i-1}\right)^2/P^{\theta,11}_{i|i-1}. \end{aligned} \tag{3.22}$$

The factor $P^{\theta,12}_{i|i-1}/P^{\theta,11}_{i|i-1}$ is called the Kalman gain and determines how important the new observation of X is for the updated estimate of $V_{i\Delta}$: the new observation is ascribed large weight if the correlation between $\hat{X}_{i|i-1}$ and $\hat{V}_{i|i-1}$ is large.

The prediction and updating equations together constitute *the Gaussian truncated second order filter* which for a fixed θ and initial guesses $\hat{V}^\theta_{0|0}$ and $R^\theta_{0|0}$ is solved recursively as follows: First, solve the prediction equations (3.20)–(3.21) for $i = 1$ with initial conditions X_0 (which has been observed), $\hat{V}^\theta_{0|0}$ and $\tilde{R}^\theta_{0|0}$ where $\tilde{R}^\theta_{0|0}$ is the 2×2 matrix with lower right element equal to $R^\theta_{0|0}$ and all other elements equal to zero (at time zero X_0 is observed without noise). This yields predictors $\hat{X}^\theta_{1|0}$ and $\hat{V}^\theta_{1|0}$ and a prediction variance matrix $P^\theta_{1|0}$. Next, an updated estimate $\hat{V}^\theta_{1|1}$ and its variance $R^\theta_{1|1}$ are calculated according to the updating equations (3.22). The updated estimates and X_Δ are then used as initial values in the prediction equations for $i = 2$ and so forth.

All the above was for a fixed value of θ. Estimation of θ is possible via the one-step predictions $\hat{X}^{\theta}_{i|i-1}$ of the observed values $X_{i\Delta}$, $i = 1,\ldots,n$: Let $r_i(\theta) = X_{i\Delta} - \hat{X}^{\theta}_{i|i-1}$, $i = 1,\ldots,n$, be the prediction errors and assume that they are independent with $r_i(\theta) \sim N(0, P^{\theta,11}_{i|i-1})$. Then the joint density is proportional to

$$\prod_{i=1}^{n} \left(P^{\theta,11}_{i|i-1}\right)^{-1/2} \exp\left(-r_i^2 / \left(2P^{\theta,11}_{i|i-1}\right)\right)$$

which is maximized in order to obtain an estimate of θ.

A few important remarks: First, the predictions and their variances are only approximations to the true ones. This implies bias of the estimator if Δ is not "small". Second, the prediction errors need not be Gaussian. This should cause no bias but affect efficiency only. Third, the method requires very intensive computations as n (the number of observations) differential equations of dimension five must be solved for *every evaluation* of the above density.

Nielsen *et al.* (2000) apply the method to simulated data from the model where the Black-Scholes specification is used together with the inverse Gamma specification for $\sqrt{V}$ (and also to a slightly more complicated model). The method produces volatility estimates that are similar, but noticeably less variable, than the actual simulated values (this is not surprising as some smoothing has taken place). Also, reasonable estimates are obtained for some, but not all, parameters. For example, the estimator of the mean-reverting parameter in the inverse Gamma model is strongly biased.

In conclusion, if we are interested in estimates of the volatility process, the method is indeed fine (though slow). However, if we are only interested in parameter estimation, then more direct — and unbiased — approaches are preferable, as there is no reason to spend time and energy simultaneously estimating the volatility process.

3.4.7 Approximate maximum likelihood estimation

Values of the likelihood may as mentioned in principle — but not in practice — be computed by simulation. In Paper III we consider a sequence of approximations $L^k_n(\theta)$, $k = 0,\ldots,n-1$, to $L_n(\theta)$ which for low values of k are computable in practice. In this section the method is reviewed and applied to Microsoft stock prices.

In the following, let for $i \in \mathbb{N}$, p^i_θ be the density of $(Z_1,\ldots,Z_i)$ and $p^{c,i}_\theta(\cdot|\tilde{z}_1,\ldots,\tilde{z}_i)$ be the conditional density of Z_{i+1} given $Z_j = \tilde{z}_j$, $j = 1,\ldots,i$. Also, write $p^{c,0}_\theta = p^1_\theta$ for the marginal density.

Basic idea and results

Recall that Z is not Markov of any order. Yet, the idea of Paper III is to *pretend that Z is k'th order Markov for some relatively small $k \geq 0$* and simplify the likelihood

function accordingly. Since Z is stationary this amounts to

$$L_n^k(\theta) = p_\theta^k(z_1, \ldots, z_k) \prod_{i=k}^{n-1} p_\theta^{c,k}(z_{i+1}|z_{i-k+1}, \ldots, z_i),$$

where Z_i for $i \geq k+2$ contributes with the conditional density given the k previous observations, rather than given *all* the past. Of course L_n^k is maximized in order to obtain an estimator $\hat{\theta}_n^k$ of θ.

No approximation is made for $k = n-1$, but the idea is to use a small k. In particular, $k = 0$ corresponds to pretending that the observations are independent and identically distributed according to the invariant distribution of Z. Note the crucial difference from the method described in Section 3.4.2: we use the true invariant density $p_\theta^{c,0} = p_\theta^1$ whereas Genon-Catalot *et al.* (1999) use an approximation to it which is only good for small values of Δ.

Generally, we use the true k-lag conditional density rather than some approximation. Consequently, the estimator $\hat{\theta}_n^k$ is invariant to bijective transformations of the data and furthermore consistent and asymptotically normal for *any fixed* Δ *and any fixed* $k \geq 0$ (under regularity conditions of course, see Theorems III.7 and III.9). The size of k is thus a question of efficiency rather than bias (see below for some further comments). The identifiability condition that the k-lag conditional distribution uniquely determines θ, is usually satisfied even for $k = 0$ (at least theoretically) because the invariant distribution of Z involves the distribution of $(V_t)_{t \leq \Delta}$.

There are no explicit expressions for the densities p_θ^k and $p_\theta^{c,k}$ but they can be calculated by simulation: replace n in (3.16) by $k+1$ in order to express p_θ^{k+1} as an expectation with respect to the distribution of $(M_i, S_i)_{i \leq k+1}$. Similarly for p_θ^k. Finally, $p_\theta^{c,k}$ is computed as the quotient between p_θ^{k+1} and p_θ^k. In other words we compute $L_n(\theta)$, or in practice rather its logarithm, by simulation of V on the interval from $[0, (k+1)\Delta]$. See Section III.4 for details.

Some comments on the applicability of the method: It is easily modified to cover models where the drift ξ for X is parameter dependent. However, it is crucial that the conditional distribution of Z given V is analytically known (and preferably simple). Hence, the method cannot easily be applied to models where (*i*) the drift or diffusion for X depends on X itself or (*ii*) the Brownian motions W and $\tilde{W}$ are correlated. On the other hand the method applies immediately to hidden Markov models, and the basic idea of using k-lag conditional densities generally provides quite natural approximations to the likelihood for models with a complicated dependence structure.

A few efficiency considerations

Intuitively we would expect the approximations $L_n^k(\theta)$ of $L_n(\theta)$ to improve as k increases since, loosely speaking, more features of the dependence structure in data are taken into account. Slightly more rigorously, it is easy to see that $\mathrm{E}_{\theta_0} \log L_n^k(\theta_0)$

increases with k (Proposition III.10). However, it is not clear whether the estimators $\hat{\theta}_n^k$ improve. With the asymptotic normality in hand for each k it is natural to compare $\hat{\theta}_n^k$ for different k's by their asymptotic variances. This project is not feasible, though, because the expressions for the asymptotic variances are very complicated and not computable, even for a one-dimensional parameter. In fact, we have worked quite hard on the efficiency question, also from other points of views, but we have not been able to come up with final answers. The reflections below are only indicative.

In Paper III we try the method on simulated data from the model where $\xi \equiv 0$ and V is a Cox-Ingersoll-Ross model. The study is small and thus only suggestive for the true behaviour. When only one parameter is considered unknown there does not seem to be any substantial differences among different values of k, and even $k = 0$ yields quite satisfactory estimates. On the other hand, when all three parameters are unknown estimation is almost impossible for $k = 0$ and $k = 1$. The problems seem to diminish as k increases, suggesting that estimation actually improves with k.

Now, let us consider a much simpler situation. It is not at all related to the stochastic volatility set-up, but it illustrates the method and a simulation study in large scale is easily carried out.

Example *(Autoregressive process of order 4)* Let $(\varepsilon_i)_{i\geq 5}$ be independent, standard normal and consider the AR(4) process $Y = (Y_i)_{i\geq 1}$ given by

$$Y_i = \beta_1 Y_{i-1} + \beta_2 Y_{i-2} + \beta_3 Y_{i-3} + \beta_4 Y_{i-4} + \sigma\varepsilon_i, \quad i \geq 5$$

where $\beta_1, \ldots, \beta_4$ are such that Y is stationary and $(Y_1, \ldots, Y_4)$ is distributed as to obtain strict stationarity of Y.

The marginal distribution is normal with mean zero and variance denoted $\omega_0^2\sigma^2$. The k-lag conditional distributions, $k = 1, \ldots, 4$, are Gaussian with

$$\begin{aligned} \mathrm{E}\big(Y_i | Y_{i-1}, \ldots, Y_{i-k}\big) &= \varphi_{k,1} Y_{i-1} + \ldots + \varphi_{k,k} Y_{i-k} \\ \mathrm{Var}\big(Y_i | Y_{i-1}, \ldots, Y_{i-k}\big) &= \omega_k^2 \sigma^2, \end{aligned}$$

for $i \geq k+1$. Of course, $\varphi_{4j} = \beta_j$, $j = 1, \ldots, 4$, and $\omega_4^2 = 1$. For $k = 1, 2, 3$, the parameters $\varphi_{k,j}$, $j = 1, \ldots, k$, are functions of $\beta_1, \ldots, \beta_4$ only and they are easily determined recursively. The variance parameters ω_k^2 are given recursively by $\omega_k^2 = \omega_{k+1}^2/(1 - \varphi_{k+1,k+1})$, $k = 0, \ldots, 3$.

Our concern is estimation of σ^2 from data $(Y_1, \ldots, Y_n)$. The regression parameters, and therefore also the φ's and the ω's, are assumed to be known. The above conditional distributions give rise to five natural estimators:

$$\hat{\sigma}_n^{2,0} = \frac{1}{n-4} \sum_{i=5}^{n} Y_i^2/\omega_0^2$$

$$\hat{\sigma}_n^{2,k} = \frac{1}{n-4} \sum_{i=5}^{n} \big(Y_i - \varphi_{k,1} Y_{i-1} - \ldots - \varphi_{k,k} Y_{i-k}\big)^2/\omega_k^2, \quad k = 1, \ldots, 4.$$

In particular $k = 4$ yields the maximum likelihood estimator.

Figure 3.6 shows box-plots for 4000 simulated values of the five estimators, each based on 1000 observations. The true value of the unknown σ^2 is 1 whereas the known regression parameter is $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.6, -0.5, 0.4, -0.4)$. As expected, the maximum likelihood estimator has the least spread. For $k \leq 3$ it seems that the spread reduces slightly as k increases, but the improvement is not substantial. □

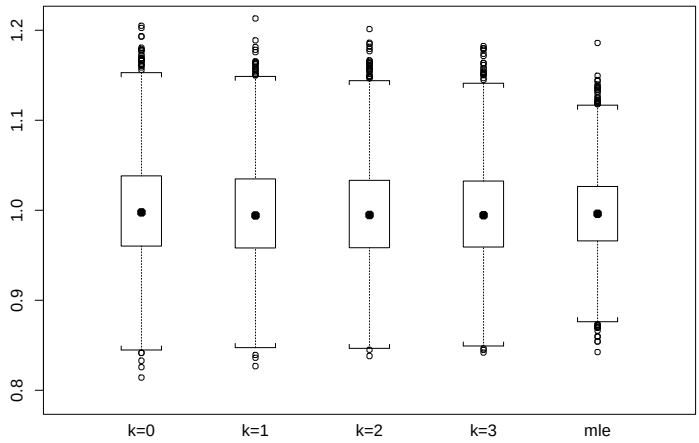

Figure 3.6: Box-plots for 4000 simulated values of $\hat{\sigma}_n^{2,k}$ for $k = 0, \ldots, 4$. The maximum likelihood estimator corresponds to $k = 4$. The dots denote medians; the boxes lower and upper quartiles; the horizontal lines the so-called lower and upper adjacent values; and the circles observations outside the adjacent interval. The upper adjacent value is the largest observations less than the upper quartile plus 1.5 times the interquartile range; the lower adjacent value is defined similarly.

Open problems and future work

Now some ideas to possible future work related to the approximate maximum likelihood method. First, in order for the method to be really useful in practice, one should be able to calculate or estimate the variance of the estimator. The expression for the asymptotic variance is not worth much in practice as it is (a quite complicated expression) given in terms of the k-lag conditional densities which are not known explicitly. It is not obvious how to estimate the variance either. In principle it could be done via simulation of a large number of processes, calculating the corresponding estimators but since estimation for each simulated dataset is relatively complicated and time consuming this is not feasible in practice.

Second, there are possibilities of model control built into the method: An estimator of the same parameter is obtained for all values of k. Consequently, significantly different estimators are indications of misspecification of the model. Again,

in order for this to be applicable (and formalized properly) we need knowledge of the distribution of the estimators.

Third, note that there are no results on the asymptotic behaviour of the true maximum likelihood estimator. There is no reason to believe that it is not well-behaved but the usual limit theorems do not apply. This is because the score function cannot be written as a sum of *one* function, evaluated at consecutive observations. Rather, different terms originate from different functions, the $i+1$'st from $\partial_\theta \log p_\theta^{c,i}(z_{i+1}|z_1,\dots,z_i)$. However, since L_n^k is an approximation of the likelihood function and since the usual limit theorems apply to $\partial_\theta \log L_n^k$, one could hope that properties of L_n^k might be applied in order to derive asymptotic results for the maximum likelihood estimator.

Fourth, when proving asymptotic properties for $\hat{\theta}_n^k$, it is implicitly assumed that the approximate likelihood function can be computed accurately. It would be interesting to see how computation of L_n^k via simulation influence the estimators. Similar work was done for martingale estimating functions (Kessler & Paredes 1999).

Fifth, it would be interesting to see how approximate maximum likelihood estimation performs compared to other methods. Also, in relation to the discussion in Section 3.3, we could estimate several models to the same data and see if they have roughly the same implications for the volatility process, for examples in terms of low-order moments.

Application to Microsoft stock prices

We now apply the approximate maximum likelihood method to a dataset consisting of 1838 observations of Microsoft stock prices on NASDAQ from May 1991 to August 1998. The logarithm of the prices and the returns are plotted in Figure 3.7. Figure 3.8 shows correlograms for the returns (to the left) and the squared returns (to the right), and Figure 3.9 is a QQ-plot of the returns. The returns seem to be uncorrelated but not independent. The auto-correlations of the squared returns die out relatively quickly and are below 0.2 at all lags. The marginal distribution of the returns have moderately heavy tails compared to the normal distribution. The excess kurtosis of the returns is 1.23.

This indicates that all four models from Section 3.3 should fit well with the data. Here we use the Cox-Ingersoll-Ross specification for the latent stochastic volatility process V and let $\xi \equiv 0$. If X denotes the logarithmic stock prices, time is measured in days, and we ignore weekends and holidays, then the model for the returns is specified by $Z_i = X_{i\Delta} - X_{(i-1)\Delta}$ for $\Delta = 1$, where

$$dX_t = \sqrt{V_t}\, dW_t \tag{3.23}$$

$$dV_t = \alpha(\beta - V_t)\, dt + \sigma\sqrt{V_t}\, d\tilde{W}_t. \tag{3.24}$$

In order to avoid numerical inaccuracy due to values of Z very close to zero we multiply the observed returns by a factor 100. Now, $d(100X_t) = \sqrt{10^4 V_t}\, dW_t$ and

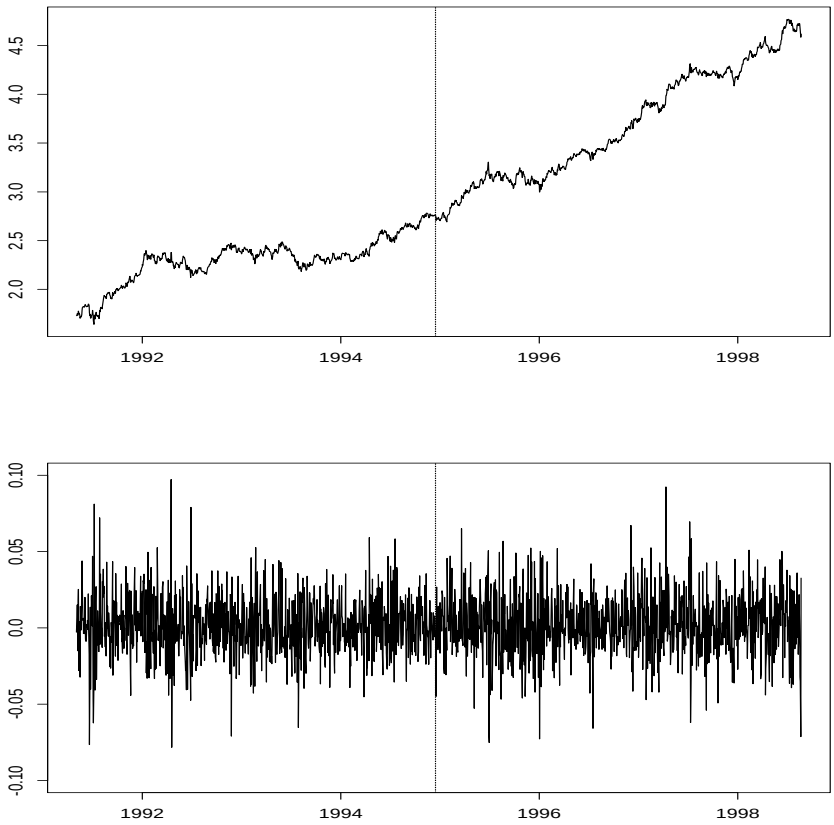

Figure 3.7: The logarithm of Microsoft stock prices (the top plot) and their increments (the bottom plot) from May 1991 to August 1998. The dashed line divides the period into two halves.

$10^4 V$ is a Cox-Ingersoll-Ross process with parameters α, $10^4\beta$ and $10^4\sigma^2$. Hence, these are the parameters estimated in the following.

We have computed estimators based on all 1837 observed returns and for comparison also those based on only the first and second half of the data, respectively. The parameters are estimated by the approximate maximum likelihood method with $k = 2, 3, 4$ (for $k = 0$ and $k = 1$ we did not find well-defined maxima).[1] Furthermore, we have estimated the parameters by matching the empirical and theoretical values of $\mathrm{E}_\theta Z_1^2$, $\mathrm{E}_\theta Z_1^4$ and $\mathrm{E}_\theta Z_1^2 Z_2^2$ (see Sections 3.3.1 and 3.4.1). All estimates are listed in Table 3.1; the estimates based on all observations are listed in the upper third, those based on only half the data in the lower two thirds.

The estimates of β do not differ much for different estimation methods. This is not surprising as β is simply the variance of Z which is easily estimated. The variance is larger for the first half of data than for the second. The estimates of

[1] For each evaluation of $\log L_n^k(\theta)$ 10.000 paths of V on the interval $[0,(k+1)\Delta]$ were simulated (via the Millstein scheme, splitting each Δ-interval into ten pieces) and used as described in Section III.4. As initial points for the numerical maximization routine we used the maximum point on a curve in $\mathbb{R}^3$ determined by estimates of the invariant distribution of V, see Section III.7 for details.

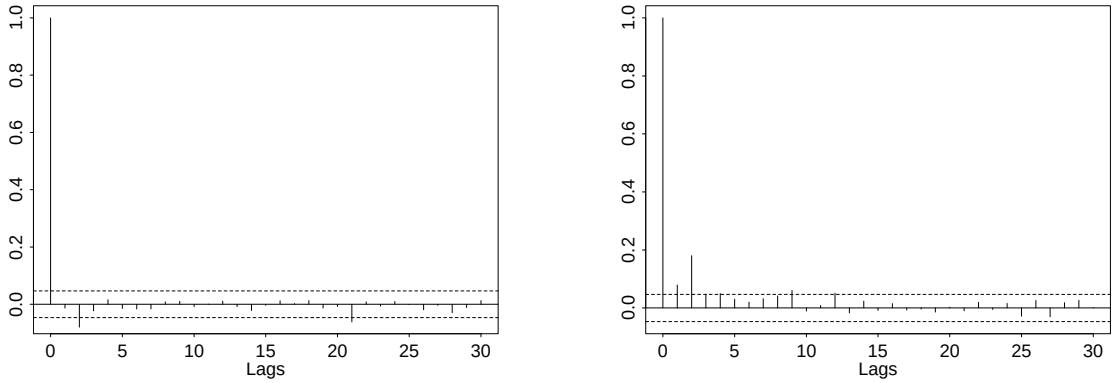

Figure 3.8: Correlogram for the Microsoft returns (to the left) and the squared returns (to the right). The dashed lines are approximate 95% confidence intervals.

α and σ^2 are less stable across methods. Most notably, the moment estimates are *very* different from the approximate maximum likelihood estimates. We are not too concerned about this, however, because a small simulation study in Section III.7 indicates that moment estimators are extremely imprecise! Consequently we are more confident in the likelihood estimates. For the two halfs of the data the approximate maximum likelihood estimates differ relatively much for different values of k whereas they seem more stable when all data are used. It would be interesting to see how much they would stabilize for larger values of k.

The above considerations are very loose and at most indicative as we have no variance estimates of the parameter estimates. However, the application indeed demonstrates that it is practically feasible to perform the necessary computations.

3.4.8 Concluding remarks

Above we have reviewed estimation methods for stochastic volatility models. The most striking characteristic is perhaps the need for extremely time consuming numerical techniques, most often simulation based. The only exceptions are the methods from Sections 3.4.1 and 3.4.2. Neither are very appealing, though: Moment estimation indeed provides consistent estimators for any fixed Δ as $n \to \infty$ but seems to work poorly in practice. The simple approximation to the marginal density introduces bias and furthermore implies that important information on the dependence structure is lost. However, the methods may prove valuable in a preliminary analysis of the data.

The filtering method introduces bias as well but is of course useful if estimates of the volatility process are of interest. Also, the method is quite flexible. So are Bayesian analysis and EMM as they are completely simulation-based. Both approaches require simulation of both the observable process and the volatility process from time zero to $n\Delta$ (the time for the last observation). Bayesian analysis furthermore requires simulation of the parameter which is considered as random,

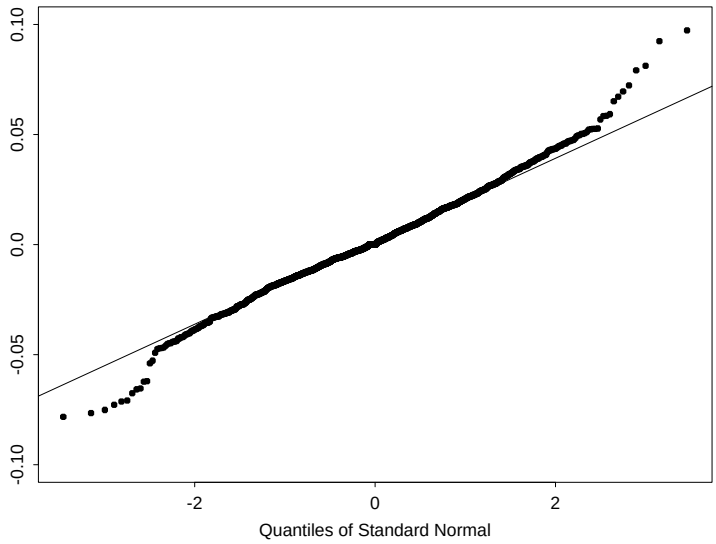

Figure 3.9: QQ-plot for the Microsoft returns with the quantiles of the normal distribution on the x-axis and quantiles of the returns on the y-axis.

and simulation is performed conditional on the observations. In both cases the simulations correct for bias but the estimators are bound to depend on the prior distribution of θ (in the Bayesian analysis) or the auxiliary model (in EMM) which are both selected quite arbitrarily.

Prediction-based estimating functions and the approximate maximum likelihood method provide consistent estimators as well. It is often natural to use predictions based on k lags of the data for some k. In that sense prediction-based estimation is in line with the approximate maximum likelihood method. For fixed k, the functional generating the prediction-based estimating function is chosen slightly arbitrarily (as low order polynomials, say). As opposed to this, the approximate maximum likelihood method suggests always to use the score corresponding to the k-lag conditional density. This makes the method invariant to data transformations but, admittedly, it need not provide efficient estimators. The k'th approximation to the likelihood is computed by simulation, but only of the volatility process and only at the interval from zero to $(k+1)\Delta$. Hence, the computational effort needed is presumably considerably smaller than for the Bayesian and auxiliary-based approaches.

3.5 Related models

So far we have been concerned with continuous-time models driven by Brownian motions. We now discuss related models. First we discuss continuous-time models driven by general Lévy processes, next models defined in discrete time.

Data	Method	$\hat{\alpha}_n$	$10^4 \cdot \hat{\beta}_n$	$10^4 \cdot \hat{\sigma}_n^2$
All	MOM	0.76	4.14	1.81
	$k=2$	0.19	4.13	0.53
	$k=3$	0.23	4.12	0.66
	$k=4$	0.21	4.13	0.59
Part 1	MOM	1.35	3.90	2.88
	$k=2$	0.11	3.88	0.28
	$k=3$	0.20	3.89	0.58
	$k=4$	0.24	3.88	0.75
Part 2	MOM	0.42	4.36	1.14
	$k=2$	0.17	4.36	0.42
	$k=3$	0.27	4.36	0.73
	$k=4$	0.19	4.39	0.52

Table 3.1: Parameter estimates for Microsoft stock prices in the Cox-Ingersoll-Ross model given by (3.23)–(3.24). "Part 1" refers to the first half of the data, "Part 2" to the second. "MOM" refers to moment estimation (method of moments) where the empirical and theoretical values of $\mathrm{E}_\theta Z_1^2$, $\mathrm{E}_\theta Z_1^4$ and $\mathrm{E}_\theta Z_1^2 Z_2^2$ are matched.

3.5.1 Continuous-time models driven by Lévy processes

As an alternative to Brownian motions one could use general Lévy processes (processes that are continuous in probability and have independent increments) as building blocks for the volatility process. This is the approach taken by Barndorff-Nielsen & Shephard (1999) who discuss models on the form

$$dX_t = \left(\xi_1 + \xi_2 V_t\right) dt + \sqrt{V_t}\, dW_t \tag{3.25}$$

$$dV_t = -\lambda V_t\, dt + dz(\lambda t) \tag{3.26}$$

(and slightly more general models). Here, $\lambda > 0$ is a parameter and z is a Lévy process with *positive* increments implying positivity of V. Models for V of the above type are referred to as *Ornstein-Uhlenbeck type processes*. Note that X exhibits jumps (if $\xi_2 \neq 0$) since V does.

The class of Levy processes is large enough that any selfdecomposable distribution on $(0,\infty)$ may be generated as the stationary distribution of an Ornstein-Uhlenbeck type process — retaining the linear drift and the unit diffusion and thus some amount of analytical tractability (Barndorff-Nielsen, Jensen & Sørensen 1998, Barndorff-Nielsen & Shephard 1999). The selfdecomposability condition is not very restrictive. For example, the generalized inverse Gaussian distributions are selfdecomposable; the Gamma, inverse Gamma, inverse Gaussian, and the positive hyperbolic distributions all occur as special cases.

Despite the linear formulation of the volatility process, estimation (and filtering) is not easy, though. Sørensen (1999) uses prediction-based estimating functions on discrete-time observations of V. If only X is observed, we are basically left with the same estimation problems as in the Brownian case. Barndorff-Nielsen & Shephard (1999) mainly discuss estimation for a related and simplified discrete-

time model and focus on two filtering approaches and on Bayesian analysis. The two filtering methods carry over to a simplified version of the above continuous-time model (with $\xi_1 = \xi_2 = 0$). The possibility of parameter estimation via spectral analysis is also mentioned.

3.5.2 Related discrete-time models

So far we have been concerned with models that are defined in continuous time but observed at discrete time-points only. Another possibility is of course to directly define models in discrete time. Such models are often easier to interpret as they usually specify movements from observation to observation in a relatively direct way. On the other hand, continuous-time modeling has advantages: First, the theory of derivative pricing, for example, most often relies on stochastic calculus. Second, it is easier to handle irregularly sampled data as a continuous-time model implicitly defines transitions over time intervals of any length, whereas in discrete time one would have to specify separate (though coherent) models for different time intervals.

Apart from being important models in their own right the discrete-time versions may serve as approximations to the continuous-time models (for example, if estimation is performed by indirect inference or EMM). Or the other way around: the continuous-time versions may be interpreted as limits of discrete-time models as the time interval between observations gets smaller (Nelson 1990).

Essentially, discrete-time models of changing variance are given by an equation for the observations

$$Y_i = \mu_i + \sigma_i \varepsilon_i, \quad i = 1, \ldots, n,$$

together with models for the mean μ_i and variance σ_i. We let $\mu_i \equiv 0$ as we shall mainly be interested in the variance structure. The innovations $(\varepsilon_i)_i$ are assumed to be white noise (*e.g.* Gaussian) with unit variance. The models can roughly be divided into two groups: ARCH type models and stochastic volatility models. We refer to survey papers for a thorough treatment of similarities and differences between the two discrete-time type models (Shephard 1996) and between the continuous-time and discrete-time versions of the stochastic volatility models (Ghysels *et al.* 1996).

ARCH type models

In ARCH type models $(\varepsilon_i)_i$ is the only source of noise and σ_i is assumed to be a (non-random) function of lagged values of Y and σ^2. Consequently, the conditional distribution of Y_i given the past is directly specified and it is straightforward, at least in principle, to do maximum likelihood estimation. Note that although the ARCH type models are driven by one source of noise only, their continuous-time limits (defined in a certain sense, see Nelson (1990)) can be stochastic volatility models of the type from this chapter. For example, the GARCH(1,1) model converges to the inverse Gamma model from Section 3.3.1.

There is a vast literature on ARCH type models, their applications and related statistical issues, and it is beyond the scope of this thesis to go into this. See Bera & Higgins (1993), for example, for a survey of the ARCH literature.

Stochastic volatility models

In the stochastic volatility set-up $(\sigma_i)_i$ is assumed to evolve independently from — or at least not to be perfectly correlated with — $(\varepsilon_i)_i$. For simplicity we shall consider a particular model that by far is the most popular model in the literature: assume that $\sigma_i^2 = \exp(H_i)$ where $(H_i)_i$ is an auto-regressive process of order one,

$$\begin{aligned} Y_i &= \varepsilon_i \exp(H_i/2) \\ H_i &= \gamma_0 + \gamma_1 H_{i-1} + \eta_i, \end{aligned}$$

where the sequences $(\varepsilon_i)_i$ and $(\eta_i)_i$ are Gaussian white noise with variances 1 and σ_η^2 respectively, independent of each other. This model is the natural discrete-time counterpart of the geometric Ornstein-Uhlenbeck model from Section 3.3.2.

We are interested in estimation of $\theta = (\gamma_0, \gamma_1, \sigma_\eta^2)$ from observations $(y_1, \ldots, y_n)$ of $(Y_1, \ldots, Y_n)$. The methods from Section 3.4 are all applicable (when suitably modified) — and most of them have actually been applied. Not surprisingly moment estimation (generalized method of moments and simulated versions like EMM) has been popular: Andersen & Sørensen (1996), among many others, apply GMM, and Gallant, Hsieh & Tauchen (1997) apply EMM to the above model. Markov Chain Monte Carlo methods for stochastic volatilities were developed and applied by Jacquier, Polson & Rossi (1994) and later refined by Kim, Shephard & Chib (1998). The so-called quasi maximum likelihood estimator (Harvey, Ruiz & Shephard 1994) relies on the Kalman filter which is applied to the linear state space model for $\log Y_i^2$: $\log Y_i^2 = H_i + \log \varepsilon_i^2$. This yields consistent (but inefficient) parameter estimates although $\log \varepsilon_i^2$ is not Gaussian. The above list of applications is only a small selection; we refer to the survey papers by Shephard (1996) and Ghysels *et al.* (1996) for many more references.

Maximum likelihood estimation is not possible — for the exact same reasons as in continuous time: the likelihood is given only in integral form

$$L_n(\theta) = \int_{\mathbb{R}^n} f(y|h)\, f_\theta(h)\, dh = \mathrm{E}_\theta\, f(y|H)$$

where we have written y for the vector $(y_1, \ldots, y_n)$ and similarly for h and H and used f generically for densities. Note that the density $f_\theta(h)$ is actually explicitly known. As in continuous time the likelihood could in principle be computed as the average of simulated values of $f(y|H)$ where H are simulated from $f_\theta(h)$. Again, this is not feasible in practice as a *huge* number of simulations would be necessary.

The new approximate maximum likelihood method from Section 3.4.7 and Paper III is one way to circumvent the problem. The methodology immediately carries over to the discrete-time set-up, and simulation of the k-lag conditional densities is much easier than in the continuous-time case because the distribution

of the conditional variances is known explicitly. Note that the method for $k = 0$ (pretending independence) only provides estimates of the parameters $\gamma_0/(1-\gamma_1)$ and $\sigma_\eta^2/(1-\gamma_1^2)$ determining the invariant distribution of H.

Another answer is importance sampling. Basically, the idea is to choose a function $g_{y,\theta}$ that satisfies $\int g_{y,\theta}(h)\,dh = 1$ for all θ and rewrite the likelihood to

$$L_n(\theta) = \int_{\mathbb{R}^n} \frac{f(y|h)\,f_\theta(h)}{g_{y,\theta}(h)} g_{y,\theta}(h)\,dh = \mathrm{E}_{y,\theta}\Big(\frac{f(y|h)\,f_\theta(h)}{g_{y,\theta}(h)}\Big) \tag{3.27}$$

where $\mathrm{E}_{y,\theta}$ is the expectation corresponding to the density $g_{y,\theta}$. Then the likelihood may be calculated as the average of simulated values of $f(y|H)\,f_\theta(H)/g_{y,\theta}(H)$ where H is drawn from $g_{y,\theta}$.

The question is of course how to choose the density $g_{y,\theta}$ cleverly, *i.e.* such that relatively few simulations are necessary. Danielsson & Richard (1993) and Danielsson (1994) suggest a product of univariate Gaussian densities. In each term the Gaussian mean and variance depend on some auxiliary parameters which are estimated beforehand in order to obtain the largest possible variance reduction. The technique reduces the number of required simulations of H impressively. Danielsson (1994) applies the technique to a dataset of roughly 2000 observations (and a somewhat more complicated model than the above) and obtains convergence using only 5000 simulations. However, the method requires heavy computations for estimation of the auxiliary parameters and is still very computer intensive.

Finally, note that it is absolutely crucial that the density of H is known explicitly. Otherwise, the integrand in (3.27) is not known. Hence, the importance sampling approach *cannot* immediately be modified to cover the continuous-time models where the distribution of the conditional variances $(S_1, \ldots, S_n)$ is not known.

3.6 Conclusion

In this chapter (and in Paper III) we have studied a class of continuous-time stochastic volatility models, mainly from a statistical point of view. The main conclusion are the following (see Section 3.3.3 and Section 3.4.8 for more detailed conclusions). An investigation of four particular models showed differences in their ability to generate data with highly leptokurtic distributions. In other respects the models were hard to distinguish. A new estimation method based on simulated approximations to the likelihood function was derived. The method provides consistent and asymptotically normal estimators for any time distance between observations. There are other methods with the same properties, some of which are more widely applicable. However, for the models from this chapter the new technique provides very natural approximations to the likelihood and is thus quite appealing.

I

Discretely Observed Diffusions: Approximation of the Continuous-time Score Function

Abstract

We discuss parameter estimation for discretely observed, ergodic diffusion processes where the diffusion coefficient does not depend on the parameter. We propose using an approximation of the continuous-time score function as an estimating function. The estimating function can be expressed in simple terms through the drift and the diffusion coefficient and is thus easy to calculate. Simulation studies show that the method performs well.

Key words

Continuous-time score function; diffusion process; discrete observations; estimating function.

Publication details

This paper has been accepted for publication in *Scandinavian Journal of Statistics*. An earlier version was printed as Preprint no. 8, 1998 at Department of Theoretical Statistics, University of Copenhagen (Sørensen 1998*a*). In this version a few notational changes have been made.

I.1 Introduction

This paper is about parameter estimation for discretely observed diffusion models with known diffusion function. The idea is to use an approximation of the continuous-time score function as estimating function.

This idea is very much in the spirit of the early work by Le Breton (1976) and Florens-Zmirou (1989). They both studied the usual Riemann-Itô discretization of the continuous-time log-likelihood function, and Florens-Zmirou (1989) showed that the corresponding estimator is inconsistent when the length of the time interval between observations is constant.

More recently, various methods providing consistent estimators have been developed, *e.g.* methods based on approximations of the true, discrete-time likelihood function (Pedersen 1995*b*, Aït-Sahalia 1998); methods based on auxiliary models (Gallant & Tauchen 1996, Gourieroux *et al.* 1993); and methods based on estimating functions (Bibby & Sørensen 1995, Hansen & Scheinkman 1995, Kessler 2000, Jacobsen 1998).

The estimating function discussed in this paper is of the simple, explicit type discussed by Hansen & Scheinkman (1995) and Kessler (2000), that is, on the form $\sum_{i=1}^{n} \mathscr{A}_\theta h(X_{t_{i-1}}, \theta)$ where $\mathscr{A}_\theta$ is the diffusion generator. Hansen & Scheinkman (1995) focus on identifiability and asymptotic behaviour of the estimating function whereas Kessler (2000) focuses on asymptotic behaviour and efficiency of the estimator.

The main contribution of this paper is to recognize that, with a special choice of h, the corresponding estimating function can be interpreted as an approximation to the continuous-time score function. The approximating estimating function is unbiased, it is invariant to data transformations, it provides consistent and asymptotically normal estimators, and it can be explicitly expressed in terms of the drift and diffusion coefficient. The estimating function is also — at least in some cases — available for multi-dimensional processes.

The main objection against the method is the need for a completely known diffusion function. In case of a parameter dependent diffusion function the suggested estimating function is still unbiased and can thus in principle be used, but there is no longer justification for using it since the continuous-time likelihood function does not exist.

We present the model and the basic assumptions in Section I.2, and the estimating function is derived in Section I.3. We give the asymptotic results in Section I.4, and examples and simulation studies in Section I.5. Sections I.2–I.5 discuss one-dimensional diffusion processes exclusively; we study the multi-dimensional case in Section I.6.

I.2 Model and notation

In this section we present the diffusion model, state the assumptions and introduce some notation.

We consider the one-dimensional, time-homogeneous stochastic differential equation

$$dX_t = b(X_t, \theta)\,dt + \sigma(X_t)\,dW_t, \quad X_0 = x_0 \tag{I.1}$$

where θ is an unknown p-dimensional parameter from the parameter space $\Theta \subseteq \mathbb{R}^p$ and W is a one-dimensional Brownian motion. The functions $b : \mathbb{R} \times \Theta \to \mathbb{R}$ and $\sigma : \mathbb{R} \to (0, \infty)$ are known, and the derivatives $\partial\sigma/\partial x$ and $\partial^2 b/\partial\theta_j \partial x$ are assumed to exist for all $j = 1, \ldots, p$. Note that σ does not depend on θ.

We assume that for any θ, (I.1) has a unique, strong solution X and that the range of X does not depend on θ. Assume furthermore that there exists a unique invariant distribution $\mu_\theta = \mu(x, \theta)dx$ such that a solution to (I.1) with $X_0 \sim \mu_\theta$ (instead of $X_0 = x_0$) is strictly stationary. Sufficient conditions for these assumptions to hold can be found in Karatzas & Shreve (1991).

The invariant density is given by

$$\mu(x, \theta) = \big(C(\theta)s(x, \theta)\sigma^2(x)\big)^{-1} \tag{I.2}$$

where $C(\theta)$ is a normalizing constant and $s(\cdot, \theta)$ is the density of the scale measure, *i.e.* $\log s(x, \theta) = -2\int^x b(y, \theta)/\sigma^2(y)\,dy$.

For all $\theta \in \Theta$, the distribution of X is denoted P_θ if $X_0 = x_0$ (as in (I.1)) and P_θ^μ if $X_0 \sim \mu_\theta$.[1] Under P_θ^μ all $X_t \sim \mu_\theta$. E_θ^μ is the expectation wrt. P_θ^μ.

The objective is to estimate θ from observations of X at discrete time-points $t_1 < \cdots < t_n$. Define $t_0 = 0$ and $\Delta_i = t_i - t_{i-1}$ and let θ_0 be the true parameter.

Finally, we need some matrix notation: Vectors in $\mathbb{R}^p$ are considered as $p \times 1$ matrices, and A^T is the transpose of A. For a function $f = (f_1, \ldots, f_q)^T : \mathbb{R} \times \Theta \to \mathbb{R}^q$ we let $f'(x, \theta)$ be the $q \times 1$ matrix of partial derivatives with respect to x and $\dot{f}(x, \theta) = D_\theta f(x, \theta)$ be the $q \times p$ matrix of partial derivatives with respect to θ, *i.e.* $\dot{f}_{jk} = \partial f_j/\partial\theta_k$.

I.3 The estimating function

In this section we derive a simple, unbiased estimating function as an approximation of the continuous-time score function.

First a comment on the model: It is important that σ does not depend on θ. Otherwise the distributions of $(X_s)_{0 \le s \le t}$ corresponding to two different parameter values are typically singular for all $t \ge 0$. If Y is the solution to $dY_t = b(Y_t, \theta)dt + \tilde{\sigma}(Y_t, \theta)dW_t$, then the process $(\int^{Y_t} 1/\tilde{\sigma}(y, \theta)dy)_{t \ge 0}$ is the solution to (I.1) with $\sigma \equiv 1$, but this is of no help for estimation purposes since the transformation depends on the (unknown) parameter.

[1] Note the difference in notation from Chapter 2 and Paper II where P_θ is the distribution of X when X_0 is started according to the stationary distribution (and where we have no notation for the distribution of X given a particular value of X_0).

When σ is completely known as we have assumed, it follows from Lipster & Shiryayev (1977) that the likelihood function for a continuous observation $(X_s)_{0\leq s\leq t}$ exists and that the corresponding score process S^c is given by

$$S_t^c(\theta) = \int_0^t \frac{\dot{b}(X_s,\theta)}{\sigma^2(X_s)}\,dX_s - \int_0^t \frac{b(X_s,\theta)\dot{b}(X_s,\theta)}{\sigma^2(X_s)}\,ds.$$

Using (I.1) we find that

$$dS_t^c(\theta) = \frac{\dot{b}(X_t,\theta)}{\sigma(X_t)}\,dW_t. \tag{I.3}$$

This shows that $S^c(\theta)$ is a local martingale and that it is a genuine martingale if $\mathrm{E}_\theta^\mu \int_0^t (\dot{b}_j(X_s,\theta)/\sigma(X_s))^2\,ds < \infty$ for all $t \geq 0$ and all $j = 1,\ldots,p$, *i.e.* if

$$\mathrm{E}_\theta^\mu \left(\frac{\dot{b}_j(X_0,\theta)}{\sigma(X_0)}\right)^2 = \int \left(\frac{\dot{b}_j(x,\theta)}{\sigma(x)}\right)^2 \mu(x,\theta)\,dx < \infty \tag{I.4}$$

for all $j = 1,\ldots,p$. In particular, $\mathrm{E}_\theta^\mu S_t^c = 0$ for all $t \geq 0$ if (I.4) holds.

If X was observed continuously on the interval $[0,t_n]$ we would estimate θ by solving the equation $S_{t_n}^c(\theta) = 0$. For discrete observations at time-points $t_1,\ldots,t_n$, the idea is to use an approximation of $S_{t_n}^c$ as estimating function.

The most obvious approximation is obtained by simply replacing the integrals in (I.3) with the corresponding Riemann and Itô sums,

$$R_n(\theta) = \sum_{i=1}^n \frac{\dot{b}(X_{t_{i-1}},\theta)}{\sigma^2(X_{t_{i-1}})}\left(X_{t_i} - X_{t_{i-1}}\right) - \sum_{i=1}^n \Delta_i \frac{b(X_{t_{i-1}},\theta)\dot{b}(X_{t_{i-1}},\theta)}{\sigma^2(X_{t_{i-1}})}. \tag{I.5}$$

Note that this would be the score function if the conditional distributions of the increments $X_{t_i} - X_{t_{i-1}}$, given the past, were Gaussian with expectation $\Delta_i b(X_{t_{i-1}},\theta)$ and variance $\Delta_i\sigma^2(X_{t_{i-1}})$. However, usually $\mathrm{E}_\theta^\mu R_n(\theta) \neq 0$, and R_n provides inconsistent estimators unless $\sup_{i=1,\ldots,n}\Delta_i \to 0$ (Florens-Zmirou 1989).

We now propose an unbiased approximation of $S_{t_n}^c$. Let $\mathscr{A}_\theta$ denote the differential operator associated with the infinitesimal generator for X, that is,

$$\mathscr{A}_\theta f(x,\theta) = b(x,\theta) f'(x,\theta) + \frac{1}{2}\sigma^2(x) f''(x,\theta)$$

for functions $f : \mathbb{R}\times\Theta \to \mathbb{R}^p$ that are twice continuously differentiable wrt. x.

Recall that μ is the invariant density and assume that the derivatives

$$h^\star = D_\theta \log\mu : \mathbb{R}\times\Theta \to \mathbb{R}^p$$

wrt. the coordinates of θ exist and are twice continuously differentiable wrt. x, such that $\mathscr{A}_\theta h^\star$ is well-defined. The connection between $h^\star$ and S^c is given in the following proposition:

Proposition I.1 *With respect to P_θ and P_θ^μ, it holds for all $t \geq 0$ that*

$$2S_t^c(\theta) = h^\star(X_t,\theta) - h^\star(X_0,\theta) - \int_0^t \mathscr{A}_\theta h^\star(X_s,\theta)\,ds. \tag{I.6}$$

Proof We show that

$$dh^\star(X_t,\theta) = \mathscr{A}_\theta h^\star(X_t,\theta)\,dt + 2\,dS_t^c(\theta). \tag{I.7}$$

Then (I.6) follows immediately since $S_0^c = 0$. Using (I.2) we easily find the first derivative of $h^\star = D_\theta \log\mu$ in terms of b and σ;

$$h^{\star\prime}(x,\theta) = D_x D_\theta \log\mu(x,\theta) = -D_\theta D_x \log s(x,\theta) = 2\frac{\dot b(x,\theta)}{\sigma^2(x)}. \tag{I.8}$$

Now simply apply Itô's formula on $h^\star$. □

The proposition suggests that we use

$$F_n(\theta) = \frac{1}{2}\sum_{i=1}^n \Delta_i \mathscr{A}_\theta h^\star(X_{t_{i-1}},\theta)$$

as an approximation to $-S_{t_n}^c$ (since the term $h^\star(X_{t_n},\theta) - h^\star(X_0,\theta)$ is negligible when n is large) and hence solve the equation $F_n(\theta) = 0$ in order to find an estimator for θ.

The right hand side of (I.6), with an arbitrary function $h \in \mathscr{C}^2(I)$ substituted for $h^\star$, is a martingale if $\mathrm{E}_\theta^\mu(h'\sigma)^2 < \infty$. Hence,

$$\mathrm{E}_\theta^\mu\, \mathscr{A}_\theta h(X_0,\theta) = 0 \tag{I.9}$$

if furthermore h and $\mathscr{A}_\theta h$ are in $L^1(\mu_\theta)$. In particular F_n is unbiased, *i.e.* $\mathrm{E}_\theta^\mu F_n(\theta) = 0$, if (I.4) holds and if $h^\star$ and $\mathscr{A}_\theta h^\star$ are in $L^1(\mu_\theta)$.

The moment condition (I.9) was used by Hansen & Scheinkman (1995) to construct general method of moments estimators (their condition C1) and by Kessler (2000) and Jacobsen (1998) to construct unbiased estimating functions. Kessler particularly suggests choosing polynomials h of low degree — regardless of the model. Instead, we suggest the *model-dependent* choice $h = h^\star$. Intuitively, this should be good for small Δ_i's since $F_n \approx -S_{t_n}^c$. Indeed, for $\Delta_i \equiv \Delta$, F_n is small Δ-optimal in the sense of Jacobsen (1998).

It should be clear, though, that moment conditions like (I.9) cannot achieve asymptotically efficient estimators for a given $\Delta > 0$ since each term in the discrete-time score function involves *pairs* of observations. Note that if the observations were independent and identically μ_θ-distributed, then the score function would equal $\sum_{i=1}^n h^\star(X_{t_i},\theta)$ which is thus optimal for $\Delta \to \infty$. Kessler (2000) discussed this estimating function.

When $\Delta_i \equiv \Delta$, F_n is a *simple* estimating function, *i.e.* a function of the form $\sum_{i=1}^n f(X_{t_{i-1}},\theta)$ where $\mathrm{E}_\theta^\mu f(X_0,\theta) = 0$ (Kessler 2000). In general, the Δ_i's can be

interpreted as weights compensating for the dependence between observations that are close in time: an observation is given much weight if it is far in time from the previous one, and little weight if it is close in time to the previous one.

Note that (I.9) holds so that F_n is unbiased even if σ depends on θ. However, the interpretation of F_n as an approximation of minus the continuous-time score function is of course no longer valid, and the method will be non-optimal even for small Δ_i's (Jacobsen 1998).

A nice property of F_n is that it is invariant to transformations of data; the estimator does not change if we observe $\varphi(X_{t_1}), \ldots, \varphi(X_{t_n})$ instead of $X_{t_1}, \ldots, X_{t_n}$. This is not the case for the polynomial martingale estimating functions discussed by Bibby & Sørensen (1995).

To prove the invariance, we need some further notation: For a diffusion process Y satisfying a stochastic differential equation similar to (I.1), we write μ_Y and $\mathscr{A}_Y$ for the corresponding invariant density and the differential operator, and define $h_Y^\star = D_\theta \log \mu_Y$.

Proposition I.2 *Let $\varphi : I \to J \subseteq \mathbb{R}$ be a bijection from $\mathscr{C}^2(I)$ with inverse φ^{-1}, and let $Y = \varphi(X)$. Then*

$$\mathscr{A}_Y h_Y^\star(y,\theta) = \mathscr{A}_X h_X^\star\big(\varphi^{-1}(y),\theta\big). \tag{I.10}$$

Proof By Itô's formula Y is the solution to

$$dY_t = b_Y(Y_t,\theta)\,dt + \sigma_Y(Y_t)\,dW_t$$

where, with obvious notation,

$$\begin{aligned} b_Y(y,\theta) &= b_X\big(\varphi^{-1}(y),\theta\big)\varphi'\big(\varphi^{-1}(y)\big) + \frac{1}{2}\big(\sigma_X^2\varphi''\big)\big(\varphi^{-1}(y)\big), \\ \sigma_Y(y) &= \big(\sigma_X\varphi'\big)\big(\varphi^{-1}(y)\big). \end{aligned}$$

One can now either check directly from (I.11) below that (I.10) holds or argue as follows. The density for the invariant distribution of $Y = \varphi(X)$ is given by $\mu_Y(y,\theta) = \mu_X\big(\varphi^{-1}(y),\theta\big)\big|\big(\varphi^{-1}\big)'(y)\big|$ and thus

$$h_Y^\star(y,\theta) = D_\theta \log \mu_Y(y,\theta) = D_\theta \log \mu_X\big(\varphi^{-1}(y),\theta\big) = h_X^\star\big(\varphi^{-1}(y),\theta\big).$$

Finally, note that $\mathscr{A}_Y(f\circ\varphi^{-1})(y) = \mathscr{A}_X f\big(\varphi^{-1}(y)\big)$ for all $f \in \mathscr{C}^2(I)$ which concludes the proof. □

In the following we write $f^\star$ for $(\mathscr{A}_\theta h^\star)/2 = (\mathscr{A}_\theta D_\theta \log\mu)/2$. It is important to note that we can express $f^\star$ — and thus $F_n = \sum_{i=1}^n \Delta_i f^\star(X_{t_{i-1}},\cdot)$ — explicitly in terms of b and σ, even if we have no explicit expression for the normalizing constant $C(\theta)$: from (I.8) we get

$$f^\star = \mathscr{A}_\theta h^\star/2 = \left(\frac{b\dot{b}}{\sigma^2} + \frac{1}{2}\dot{b}' - \frac{\dot{b}\sigma'}{\sigma}\right). \tag{I.11}$$

I.4 Asymptotic properties

In this section we state the asymptotic results for F_n. We consider equidistant observations, $t_i = i\Delta$ where Δ does not depend on n, and let $n \to \infty$.

Under suitable regularity conditions, a solution $\hat{\theta}_n$ to $F_n(\theta) = 0$ exists with a P_{θ_0}-probability tending to 1, and $\hat{\theta}_n$ is a consistent, asymptotically normal estimator for θ. The asymptotic distribution of $\hat{\theta}_n$ is given by

$$\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{D}} N\left(0, A(\theta_0)^{-1} V(\theta_0) \left(A(\theta_0)^{-1}\right)^T\right)$$

wrt. P_{θ_0} as $n \to \infty$, where $A(\theta_0) = \mathrm{E}^{\mu}_{\theta_0} \dot{f}^{\star}(X_0, \theta_0)$ and

$$V(\theta_0) = \mathrm{E}^{\mu}_{\theta_0} f^{\star}(X_0, \theta_0) f^{\star}(X_0, \theta_0)^T + 2 \sum_{k=1}^{\infty} \mathrm{E}^{\mu}_{\theta_0} f^{\star}(X_0, \theta_0) f^{\star}(X_{k\Delta}, \theta_0)^T.$$

Conditions that ensure convergence of the sum in are given by Kessler (2000). If (I.9) holds for each $\partial h_j^{\star} / \partial \theta_k$, then

$$A(\theta_0) = 2\,\mathrm{E}^{\mu}_{\theta_0} \left(\frac{\dot{b}(X_0, \theta_0)}{\sigma(X_0)}\right)^T \left(\frac{\dot{b}(X_0, \theta_0)}{\sigma(X_0)}\right),$$

and $A(\theta_0)$ is symmetric and positive semidefinite. It must be positive definite. We will not go through the additional regularity conditions here but refer to Kessler (2000) and particularly to Sørensen (1998*b*).

I.5 Examples

As already mentioned F_n can always be expressed explicitly in terms of b and σ. When b is linear wrt. the parameter we even get explicit estimators. Assume that $b(x, \theta) = b_0(x) + \sum_{j=1}^{p} b_j(x)\theta_j$ for known functions $b_0, b_1, \ldots, b_p : \mathbb{R} \to \mathbb{R}$ such that the assumptions of Sections I.2 and I.3 hold. From (I.11) we easily deduce that the k'th coordinate of $f^{\star}$ is given by

$$f_k^{\star}(x, \theta) = \sum_{j=1}^{p} \frac{b_k(x) b_j(x)}{\sigma^2(x)} \theta_j + \frac{b_0(x) b_k(x)}{\sigma^2(x)} + \frac{1}{2} b_k'(x) - \frac{b_k(x) \sigma'(x)}{\sigma(x)}.$$

It follows that F_n is linear in θ and it is easy to show that the estimating equation has a unique, explicit solution if and only if $b_1, \ldots, b_p$ are linearly independent.

The Ornstein-Uhlenbeck model and the Cox-Ingersoll-Ross model are special cases of this setup. Several authors have studied inference for these models, see Bibby & Sørensen (1995), Gourieroux *et al.* (1993), Kessler (2000), Jacobsen (1998), Overbeck & Rydén (1997), and Pedersen (1995*b*), for example. From now on, we consider equidistant observations, $\Delta_i \equiv \Delta$.

Example *(The Ornstein-Uhlenbeck process)* Let X be the solution to

$$dX_t = \theta X_t\,dt + dW_t, \quad X_0 = x_0,$$

where $\theta < 0$. The estimator is given by $\hat{\theta}_n = -n/(2\sum_{i=1}^n X_{(i-1)\Delta}^2)$. Since $h^\star$ is an eigenfunction for $\mathscr{A}_\theta$, the simple estimating functions corresponding to $f = h^\star$ and $f = f^\star$ are proportional (and hence provide the same estimator). Kessler (2000) showed that, for all Δ, $\hat{\theta}_n$ has the least asymptotic variance among estimators obtained from simple estimating equations. This also follows from results in Jacobsen (1998). □

Example *(The Cox-Ingersoll-Ross process)* Consider the solution X to

$$dX_t = (\alpha + \beta X_t)\,dt + \sqrt{X_t}\,dW_t, \quad X_0 = x_0$$

where $\beta < 0$ and $\alpha \geq 1/2$. The estimating function F_n is given by

$$F_n(\alpha,\beta) = \begin{pmatrix} \left(\alpha - \frac{1}{2}\right)\sum_{i=1}^n 1/X_{(i-1)\Delta} + n\beta \\ \beta\sum_{i=1}^n X_{(i-1)\Delta} + n\alpha \end{pmatrix}.$$

To see how the estimator performs we have compared it to three other estimators in a simulation study. We have simulated 500 processes on the interval $[0,500]$ by the Euler scheme with time-step 1/1000. The number of observations is $n = 500$ and $\Delta = 1$. For each simulation we have calculated four estimators: those obtained from F_n, R_n given by (I.5), $H_n = \sum_{i=1}^n h^\star(X_{t_{i-1}},\cdot)$, and the martingale estimating function suggested by Bibby & Sørensen (1995).

The estimating function H_n is given by

$$H_n(\theta) = 2\begin{pmatrix} \sum_{i=1}^n \log X_{(i-1)\Delta} - n\Psi(2\alpha) + n\log(-2\beta) \\ \sum_{i=1}^n X_{(i-1)\Delta} + n\alpha/\beta \end{pmatrix}$$

where Ψ is the Digamma function, $\Psi = \partial\log\Gamma/\partial x$. Note that the second coordinates of F_n and H_n are equivalent and that $H_n(\theta) = 0$ cannot be solved explicitly.

The empirical means and standard errors of the four estimators are listed in Table I.1. The true parameter values are $\alpha_0 = 10$ and $\beta_0 = -1$.

The estimator from R_n is biased (as we knew). F_n and H_n seem to be almost equally good and are both better than the martingale estimating function. □

Finally, we consider an example where the parameter of interest enters as an exponent in the drift function.

Example *(A generalized Cox-Ingersoll-Ross model)* Let X be the solution to

$$dX_t = \left(\alpha + \beta X_t^\theta\right)\,dt + \sqrt{X_t}\,dW_t$$

Estimating function	$\hat{\alpha}_n$		$\hat{\beta}_n$	
	mean	s.e.	mean	s.e.
F_n	10.1271	0.7218	-1.0126	0.0737
H_n	10.1543	0.7151	-1.0154	0.0729
R_n	6.3691	0.4279	-0.6368	0.0430
Martingale	10.2000	1.1900	-1.0200	0.1200

Table I.1: Empirical means and standard errors for 500 realizations of various estimators for (α,β) in the Cox-Ingersoll-Ross model. The number of observations is $n = 500$ and $\Delta = 1$. The true value is $(\alpha_0,\beta_0) = (10,-1)$.

where $\alpha \geq \frac{1}{2}$ and $\beta < 0$ are known and $\theta > 0$ is the unknown parameter. Note that X is a Cox-Ingersoll-Ross process if $\theta = 1$; for $\theta \neq 1$ the mean reverting force is stronger or weaker.

From (I.11) it follows that $f^\star = (\mathscr{A}_\theta h^\star)/2$ is given by

$$f^\star(x,\theta) = \beta x^{\theta-1}\log x\Big(\alpha + \beta x^\theta + \frac{\theta}{2} - \frac{1}{2}\Big) + \frac{1}{2}\beta x^{\theta-1}.$$

The estimating equation must be solved numerically. For comparison we have also considered the simple estimating function corresponding to

$$\tilde{f}(x,\theta) = x - \mathrm{E}_\theta^\mu X_0 = x - \frac{\Gamma\big((2\alpha+1)/\theta\big)}{\Gamma\big(2\alpha/\theta\big)}\left(-\frac{\theta}{2\beta}\right)^{1/\theta} \tag{I.12}$$

As above, we have simulated 500 processes by means of the Euler scheme; $n = 500$ and $\Delta = 1$. The true value of θ is $\theta_0 = 1.5$ and $\alpha = 2$, $\beta = -1$. The means (standard errors) of the estimators are 1.5028 (0.0508) when using F_n and 1.5001 (0.0590) when using $\sum \tilde{f}(X_{(i-1)\Delta},\cdot)$.

Both estimators are very precise. The estimator obtained from (I.12) is closer to the true value but has larger standard error than the estimator obtained from F_n. □

I.6 Multi-dimensional processes

So far, we have only studied one-dimensional diffusion processes. In this section we discuss to what extend the ideas carry over to the multi-dimensional case.

We consider a d-dimensional stochastic differential equation

$$dX_t = b(X_t,\theta)\,dt + \sigma(X_t)\,dW_t, \quad X_0 = x_0. \tag{I.13}$$

The parameter θ is still p-dimensional, $\theta \in \Theta \subseteq \mathbb{R}^p$, but X and W are now d-dimensional. The functions $b : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are known, $\sigma(x)$ is regular for all $x \in \mathbb{R}^d$, and $x_0 \in \mathbb{R}^d$.

Let $\dot{b}$ be the $d \times p$ matrix of derivatives; $\dot{b}_{ij} = \partial b_i / \partial \theta_j$, and let $D_i g = \partial g / \partial x_i$ and $D_{ij}^2 g = \partial g^2 / \partial x_i \partial x_j$ for functions $g : \mathbb{R}^d \times \Theta \to \mathbb{R}$ with $g(\cdot, \theta)$ in $\mathscr{C}^2(\mathbb{R}^d)$. With this notation, the analogue of $\mathscr{A}_\theta$ is given by

$$\mathscr{A}_\theta g(x,\theta) = \sum_{i=1}^d b_i(x,\theta) D_i g(x,\theta) + \frac{1}{2} \sum_{i,j=1}^d \left(\sigma(x)\sigma^T(x)\right)_{ij} D_{ij}^2 g(x,\theta)$$

and the score process is given by

$$S_t^c(\theta) = \int_0^t \dot{b}^T(X_s,\theta)\Sigma(X_s)\, dX_s - \int_0^t \dot{b}^T(X_s,\theta)\Sigma(X_s) b(X_s,\theta)\, ds,$$

where $\Sigma(x) = \left(\sigma(x)\sigma^T(x)\right)^{-1}$, see Lipster & Shiryayev (1977). Using (I.13) we find $dS_t^c(\theta) = \dot{b}^T(X_t,\theta)(\sigma^{-1})^T(X_t)\, dW_t$.

Now, similarly to (I.7) we look for functions $h_1^\star, \ldots, h_p^\star : \mathbb{R}^d \times \Theta \to \mathbb{R}$ such that for each k, $dh_k^\star(X_t,\theta) = \mathscr{A}_\theta h_k^\star(X_t,\theta)\, dt + 2\, dS_{k,t}^c(\theta)$. Arguing as above, this leads to the equations

$$D_i h_k^\star(x,\theta) = 2\left[\dot{b}^T(x,\theta)\Sigma(x)\right]_{ki} = 2\sum_{r=1}^d \dot{b}_{rk}(x,\theta)\Sigma_{ir}(x), \quad i = 1,\ldots,d \tag{I.14}$$

and thus

$$\mathscr{A}_\theta h_k^\star = 2\sum_{i,r=1}^d \dot{b}_{rk}\Sigma_{ir} b_i + \sum_{j=1}^d \frac{\partial \dot{b}_{jk}}{\partial x_j} + \sum_{i,j,r=1}^d \left(\sigma\sigma^T\right)_{ij} \dot{b}_{rk} \frac{\partial \Sigma_{ir}}{\partial x_j} \tag{I.15}$$

The equations (I.14) may, however, not any have solutions; differentiation wrt. x_j yields

$$D_{ij}^2 h_k^\star = 2\sum_{r=1}^d \left(\frac{\partial^2 b_r}{\partial \theta_k \partial x_j}\Sigma_{ir} + \frac{\partial b_r}{\partial \theta_k}\frac{\partial \Sigma_{ir}}{\partial x_j} \right),$$

but the right hand side is not necessarily symmetric wrt. i and j, see the example below.

If there *are* solutions, then (I.15) has expectation zero and the simple estimating function with $f = \left(\mathscr{A}_\theta h_1^\star, \ldots, \mathscr{A}_\theta h_p^\star\right)^T$ may be used. Otherwise, the right hand side of (I.15) is typically biased.

Example *(Homogeneous Gaussian diffusions)* Let B be a 2×2 matrix with eigenvalues with strictly negative parts and let A be an arbitrary 2×1 matrix. Consider the stochastic differential equation

$$dX_t = (A + BX_t)\, dt + \sigma dW_t, \quad X_0 = x_0$$

where $\sigma > 0$ is known, W is a two-dimensional Brownian motion, and $x_0 \in \mathbb{R}^2$.

Solutions to all the equations (I.14) exist if and only if B is symmetric. Let $\alpha_1, \alpha_2, \beta_{11}, \beta_{22}, \beta_{12}$ denote the entries of A and B and let $S_1 = \sum X_{1,(i-1)\Delta}$, $S_2 = \sum X_{2,(i-1)\Delta}$, $S_{11} = \sum X^2_{1,(i-1)\Delta}$, $S_{22} = \sum X^2_{2,(i-1)\Delta}$, and $S_{12} = \sum X_{1,(i-1)\Delta} X_{2,(i-1)\Delta}$; all sums are from 1 to n. Then the estimating equation is given by

$$\frac{1}{\sigma^2}\begin{pmatrix} n & 0 & S_1 & 0 & S_2 \\ 0 & n & 0 & S_2 & S_1 \\ S_1 & 0 & S_{11} & 0 & S_{12} \\ 0 & S_2 & 0 & S_{22} & S_{12} \\ S_2 & S_1 & S_{12} & S_{12} & S_{11}+S_{22} \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_{11} \\ \beta_{22} \\ \beta_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -n/2 \\ -n/2 \\ 0 \end{pmatrix}.$$

We have simulated 500 processes (by exact simulation), each of a length of 500 with $\Delta = 1$ and $\sigma = \sqrt{2}$. The true matrices are

$$A_0 = \begin{pmatrix} 4 \\ 1 \end{pmatrix} \qquad \text{and} \qquad B_0 = \begin{pmatrix} -2 & 1 \\ 1 & -3 \end{pmatrix}.$$

The means and the standard errors (to the right) are

$$\hat{A}_n = \begin{pmatrix} 4.0349 \\ 1.0035 \end{pmatrix} \qquad \begin{matrix} 0.2904 \\ 0.2891 \end{matrix}$$

and

$$\hat{B}_n = \begin{pmatrix} -2.0155 & 1.0078 \\ 1.0078 & -3.0247 \end{pmatrix} \qquad \begin{matrix} 0.1248 & 0.1177 \\ 0.1177 & 0.1978 \end{matrix}$$

The estimators are satisfactory. □

Acknowledgements I am grateful to my supervisor Martin Jacobsen for valuable discussions and suggestions, and to Jens Lund, Martin Richter and the referees for comments on earlier versions of the manuscript.

II

Estimation of Diffusion Parameters for Discretely Observed Diffusion Processes

Abstract

We study estimation of diffusion parameters for one-dimensional, ergodic diffusion processes that are discretely observed. We discuss a method based on a functional relationship between the drift function, the diffusion function and the invariant density and use empirical process theory to show that the estimator is $\sqrt{n}$-consistent and in certain cases weakly convergent. We try out the method on the so-called CKLS model and compare it with other methods in a simulation study.

Key words

CKLS model; diffusion parameters; ergodic diffusion processes; empirical process theory.

Publication details

A shorter version of this paper has been submitted for publication and was also printed as Preprint no. 1, 2000 at Department of Theoretical Statistics, University of Copenhagen (Sørensen 2000). The shorter version does not include Section II.7.1, the final subsection of Section II.7.2 (*Considerations on asymptotics*) and the appendices. Furthermore, various parts have been revised slightly in the present version.

II.1 Introduction

This paper is about parametric estimation of the diffusion function for a discretely observed diffusion process. The likelihood function is only known analytically in very few cases so it is usually not possible to do maximum likelihood estimation. The method discussed in this paper is inspired by a non-parametric estimation procedure suggested by Aït-Sahalia (1996) which we shall describe shortly.

Let b be the drift function, σ the diffusion function, and μ the invariant density for a one-dimensional diffusion process with state space (l,r). Then, in many cases, there is the connection $2b\mu = (\sigma^2\mu)'$, *i.e.*

$$b(x) = \frac{1}{2}\left(\left(\sigma^2\right)'(x) + \sigma^2(x)\frac{\mu'(x)}{\mu(x)} \right), \quad x \in (l,r) \tag{II.1}$$

between b, σ and μ, a prime denoting differentiation with respect to the state variable. This relationship has been used for non-parametric estimation by several authors. For σ known Banon (1978) defines an estimator of $b(x)$ for all x in (l,r) (pointwise) by plugging in kernel estimates of $\mu(x)$ and $\mu'(x)$. For σ unknown but constant (so that $(\sigma^2)' = 0$) an estimate of σ is plugged in as well. For general unknown functions σ Jiang & Knight (1997) use local time based estimators of σ^2 and $(\sigma^2)'$, see also Florens-Zmirou (1993).

Aït-Sahalia (1996) uses a related but almost opposite estimation strategy in that he first estimates the drift and next the diffusion function. He assumes that $\sigma^2(x)\mu(x) \to 0$ as $x \to l$ and uses the integrated version

$$\sigma^2(x)\mu(x) = 2\int_l^x b(u)\mu(u)\,du, \quad x \in (l,r) \tag{II.2}$$

of (II.1). He considers linear drift only and uses conditional least squares for estimation of the drift parameters. For each x an estimator of $\sigma^2(x)$ is defined by dividing a kernel estimator of the integral in (II.2) by a kernel estimator of $\mu(x)$.

This procedure yields a non-parametric estimator of σ^2. The method seems to work well for a large sample size (Aït-Sahalia uses the method on a dataset with 5505 observations). For moderate sample sizes, however, the kernel estimators and hence the diffusion estimator will be rather variable. Also, if a non-parametric analysis indicates a certain form of σ^2 (*e.g.* that of a power function), then it is natural to specify the diffusion parametrically and estimate the parameters. For parametric specifications it is also possible to verify for which parameter values the relation (II.2) actually holds.

In this paper the relationship (II.2) is utilized for parametric estimation of the diffusion function. The idea is the following. Let $f = \sigma^2\mu$. As we shall see, it is easy for each x to define a consistent estimator $\hat{f}(x)$ of $f(x,\theta)$. We also have an analytical expression for $f(x,\theta)$, and we estimate θ such that the "true" function $f(\cdot,\theta)$ is close to the estimated version $\hat{f}$ in the sense that the uniform distance $\sup_{x\in(l,r)} \left| f(x,\theta) - \hat{f}(x) \right|$ is minimal. In order for a simple estimator $\hat{f}(x)$ to exist it is crucial that f converges to zero at at least one of the endpoints, l and r, of the

state space. We distinguish between three cases: f converges to zero (*i*) at l only; (*ii*) at r only; (*iii*) at both l and r. We use different pointwise consistent estimators of f in case (*i*) and (*ii*) and a suitable average of the two in case (*iii*).

The corresponding estimators are consistent under weak regularity conditions and $\sqrt{n}$-consistent under somewhat stronger conditions. In case (*iii*) the estimator is weakly convergent. The asymptotic results are proved by means of empirical process theory.

We use the Ornstein-Uhlenbeck process and the so-called CKLS model (Chan, Karolyi, Longstaff & Sanders 1992) for illustration. For the CKLS model given by $dX_t = (\alpha + \beta X_t)\,dt + \sigma X_t^{\gamma} dW_t$, we compare the method with other estimation methods (generalized method of moments, IID estimation, and simple, explicit estimating equations) in a simulation study. The method seems to work well.

The paper is organized as follows. The model and the basic assumptions are presented in Section II.2. We discuss the estimation approach in Section II.3 and prove asymptotic properties in Sections II.4 and II.5. While in Sections II.3– II.5 it is assumed that the drift function is completely known, in Section II.6 we discuss what to do if the drift must be estimated as well. In Section II.7 we study the Ornstein-Uhlenbeck process and the CKLS model. Conclusions are drawn in Section II.8. In Appendix II.A we give a brief review of the theory of empirical processes which we use to show asymptotic properties for our estimator. Finally, Appendix II.B gives a proof that the Ornstein-Uhlenbeck process is β-mixing at an exponential rate.

II.2 Model and notation

In this section we define the diffusion model and list basic assumptions that ensure nice properties of the model. For details on diffusion processes see Karatzas & Shreve (1991), for example.

We consider the one-dimensional, time-homogeneous stochastic differential equation

$$dX_t = b(X_t)\,dt + \sigma(X_t, \theta)\,dW_t \tag{II.3}$$

where θ is an unknown p-dimensional parameter from the parameter space $\Theta \subseteq \mathbb{R}^p$, W is a one-dimensional Brownian motion and $b : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \times \Theta \to \mathbb{R}$ are known continuous functions.

The objective is estimation of θ from observations $X_\Delta, \ldots, X_{n\Delta}$ at discrete, equidistant time-points. Let θ_0 be the true parameter. Note that the drift function, b, does not depend on the parameter. We will relax this unrealistic condition later and instead assume that the drift parameters can be estimated without information on the diffusion parameters, see Section II.6.

We assume that a unique, strong solution to (II.3) exists for all $\theta \in \Theta$ and all initial distributions of X_0, that the state space, denoted by I, is the same for all $\theta \in \Theta$ and that I is open. Since X is continuous, I is an interval, and we write $I = (l, r)$ where $-\infty \leq l < r \leq +\infty$.

We furthermore make assumptions ensuring stationarity of the process: assume that $\sigma(x,\theta) > 0$ for all $(x,\theta) \in I \times \Theta$ and define

$$s(x,\theta) = \exp\left(-2\int_{x_0}^{x} \frac{b(y)}{\sigma^2(y,\theta)}\,dy\right), \quad (x,\theta) \in I \times \Theta \tag{II.4}$$

where $x_0 \in I$ is arbitrary but fixed. For each $\theta \in \Theta$ the function $x \to \int_{x_0}^{x} s(y,\theta)\,dy$ is called a scale function.

Assumption II.1 The diffusion function is positive, *i.e.* $\sigma(x,\theta) > 0$ for all $(x,\theta) \in I \times \Theta$, and for all $\theta \in \Theta$ the function $s(\cdot,\theta)$ satisfies

1. $\int_{x_0}^{r} s(x,\theta)\,dx = \int_{l}^{x_0} s(x,\theta)\,dx = +\infty$;
2. $\int_{l}^{r} \left(s(x,\theta)\sigma^2(x,\theta)\right)^{-1} dx < \infty$. □

With these assumptions, X is recurrent (hits any level in I almost surely), does not hit l and r, and has a unique invariant distribution $\mu_\theta(dx) = \mu(x,\theta)dx$ where

$$\mu(x,\theta) = K_0(\theta)\left(s(x,\theta)\sigma^2(x)\right)^{-1}. \tag{II.5}$$

The normalizing constant $K_0(\theta)$ is the inverse of the integral in Assumption II.1.2 and depends on the choice of x_0.

We let P_θ denote the distribution of X when $X_0 \sim \mu_\theta$ and E_θ the expectation with respect to P_θ. Under P_θ all $X_t \sim \mu_\theta$ and the ergodic theorem holds, *i.e.* $\frac{1}{n}\sum_{i=1}^{n} g(X_{i\Delta}) \to \mathrm{E}_\theta\, g(X_0)$ P_θ-almost surely as $n \to \infty$ for all $g \in L^1(\mu_\theta)$. In the following we shall use the ergodic theorem on the drift function b so we assume that it is μ_θ-integrable:

Assumption II.2 The drift function b is in $L^1(\mu_\theta)$ for all $\theta \in \Theta$, *i.e.* $\int |b|\,d\mu_\theta < \infty$ for all $\theta \in \Theta$. □

The estimation method described in this paper is based on the function $f = \sigma^2\mu : I \times \Theta \to (0,\infty)$. For θ fixed we will often write f_θ for the function $f(\cdot,\theta) : I \to \mathbb{R}$. By (II.5) and (II.4)

$$f_\theta(x) = f(x,\theta) = \frac{K_0(\theta)}{s(x,\theta)} = K_0(\theta)\exp\left(2\int_{x_0}^{x} \frac{b(u)}{\sigma^2(u,\theta)}\,du\right).$$

Differentiation of f with respect to x yields

$$\frac{\partial f}{\partial x} = 2f\frac{b}{\sigma^2} = 2\sigma^2\mu\frac{b}{\sigma^2} = 2b\mu \tag{II.6}$$

and $f(x_0,\theta) = K_0(\theta)$ so $f(x,\theta) = K_0(\theta) + 2\int_{x_0}^{x} b(u)\mu(u,\theta)\,du$ for $x \in I$ and $\theta \in \Theta$. In particular, for θ fixed the function f_θ is bounded by $K_0(\theta) + 2\,\mathrm{E}_\theta\,|b(X_0)|$; the limits

$f_\theta(l) = f(l,\theta) = \lim_{x \searrow l} f(x,\theta)$ and $f_\theta(r) = f(r,\theta) = \lim_{x \nearrow n} f(x,\theta)$ are well-defined and finite; and

$$f(x,\theta) = f(l,\theta) + 2\int_l^x b(u)\mu(u,\theta)\,du, \quad x \in I \tag{II.7}$$

$$f(x,\theta) = f(r,\theta) - 2\int_x^r b(u)\mu(u,\theta)\,du, \quad x \in I. \tag{II.8}$$

The limits $f(l,\theta)$ and $f(r,\theta)$ are non-negative for all $\theta \in \Theta$. For the estimation method below to work at least one of the limits must be zero for all $\theta \in \Theta$. Informally, since $f = \sigma^2\mu$, the assumption is that σ^2 does not grow too fast at (at least one of) the limits. More precisely we must check that $\int_{x_0}^r b(x)/\sigma^2(x,\theta)dx = -\infty$ for all $\theta \in \Theta$ and/or $\int_l^{x_0} b(x)/\sigma^2(x,\theta)dx = +\infty$ for all $\theta \in \Theta$. In both cases f_θ is bounded by $2\,\mathrm{E}_\theta\,|b(X_0)|$.

Some comments: (*i*) If $l > -\infty$ ($r < +\infty$) then automatically $f(l,\theta) = 0$ ($f(r,\theta) = 0$) because of Assumption II.1.2. In particular $f(l,\theta) = 0$ for all models with state space $(0,\infty)$. (*ii*) If $I = (-\infty,\infty)$ and $b \equiv 0$ so X is on natural scale, then f_θ is constant and the above integral assumption is *not* satisfied. (*iii*) It follows from (II.7) and (II.8) that $f(r,\theta) - f(l,\theta) = 2\,\mathrm{E}_\theta\, b(X_0)$. In particular $\mathrm{E}_\theta\, b(X_0) = 0$ if both $f(l,\theta) = f(r,\theta) = 0$. (*iv*) $f(l,\theta) = f(r,\theta) = 0$ holds *e.g.* for the Ornstein-Uhlenbeck process, the Cox-Ingersoll-Ross model and for the CKLS model if the exponent in the diffusion function is between 1/2 and 1, see Section II.7.2 for details. (*v*) If $f(l,\theta) = 0$ then (II.7) is identical to (II.2).

Finally a remark concerning identification: for two parameter values θ and θ' the functions f_θ and $f_{\theta'}$ are identical if and only if $\sigma(\cdot,\theta)$ and $\sigma(\cdot,\theta')$ are identical (even if neither $f(l,\theta)$ or $f(r,\theta)$ is zero for all $\theta \in \Theta$). Indeed, if $f_\theta = f_{\theta'}$, then $\mu(\cdot,\theta) = \mu(\cdot,\theta')$ according to (II.6) and hence $\sigma(\cdot,\theta) = \sigma(\cdot,\theta')$ since $f = \sigma^2\mu$. If $f(l,\theta)$ or $f(r,\theta)$ is zero for all $\theta \in \Theta$, then $f_\theta = f_{\theta'}$ if and only if $\mu_\theta = \mu_{\theta'}$ holds as well (use (II.7) or (II.8)). We will of course not allow parametrizations where $\sigma(\cdot,\theta) = \sigma(\cdot,\theta')$ is possible for $\theta \neq \theta'$.

II.3 Estimation

In this section we discuss how to define pointwise consistent estimators of $f_\theta = f(\cdot,\theta)$ and how to use them for estimation of θ.

The basic idea

If $f(l,\theta) = 0$ we see from (II.7) that

$$f(x,\theta) = 2\int_l^x b(u)\mu(u,\theta)\,du = 2\,\mathrm{E}_\theta\left(b(X_0)1_{\{X_0 \le x\}}\right), \quad x \in I,\ \theta \in \Theta.$$

From the right hand side and Assumption II.2 it follows that

$$\hat{f}_{1,n}(x) = \frac{2}{n}\sum_{i=1}^n \left(b(X_{i\Delta})1_{\{X_{i\Delta} \le x\}}\right) \tag{II.9}$$

is an unbiased and consistent estimator of $f(x,\theta)$ with respect to P_θ for all $x \in I$: $\mathrm{E}_\theta\, \hat{f}_{1,n}(x) = f(x,\theta)$ and $\hat{f}_{1,n}(x) \to f(x,\theta)$ almost surely as $n \to \infty$. Also, note that $\hat{f}_{1,n}(x) = 0 = f(l,\theta)$ for $x < \min\{X_{i\Delta} : i = 1,\ldots,n\}$ so we write $\hat{f}_{1,n}(l) = 0$.

Similarly, if $f(r,\theta) = 0$ then

$$\hat{f}_{2,n}(x) = -\frac{2}{n}\sum_{i=1}^{n}\Big(b(X_{i\Delta})1_{\{X_{i\Delta}>x\}}\Big) \tag{II.10}$$

is unbiased and consistent for $f(x,\theta)$ under P_θ for all $x \in I$. We write $\hat{f}_{2,n}(r) = 0$ since $\hat{f}_{2,n}(x) = 0$ for $x \geq \max\{X_{i\Delta} : i = 1,\ldots,n\}$

The functions $\hat{f}_{1,n}$ and $\hat{f}_{2,n}$ are piecewise constant with jumps at each data point $X_{k\Delta}$; the jump size is $2b(X_{k\Delta})/n$. In particular $\hat{f}_{1,n}$ and $\hat{f}_{2,n}$ are increasing (decreasing) if f_θ is increasing (decreasing) at $X_{k\Delta}$, *cf.* (II.6). Note that $\hat{f}_{1,n}(x) - \hat{f}_{2,n}(x) = \frac{2}{n}\sum_{i=1}^{n} b(X_{i\Delta})$ so the deviation between $\hat{f}_{1,n}(x)$ and $\hat{f}_{2,n}(x)$ is the same for all $x \in I$.

As indicated, the idea is to estimate θ by the value that makes the function f_θ close to its estimator, $\hat{f}_{1,n}$ or $\hat{f}_{2,n}$. More precisely we define the uniform distances

$$U_{i,n}(\theta) = \sup_{x\in I}\Big|\hat{f}_{i,n}(x) - f_\theta(x)\Big|, \quad i = 1,2$$

and suggest minimizing $U_{1,n}$ if $f(l,\theta) = 0$ and $U_{2,n}$ if $f(r,\theta) = 0$. Note that $U_{i,n}(\theta)$ is finite since $U_{i,n}(\theta) \leq \frac{2}{n}\sum_{j=1}^{n}|b(X_{j\Delta})| + 2\,\mathrm{E}_\theta\,|b(X_0)|$. One could use other measures of distance between $\hat{f}_{i,n}$ and f_θ. This and some computational aspects will be discussed in the end of the section.

Meanwhile, what if both $f(l,\theta)$ and $f(r,\theta)$ are zero? Then (II.9) and (II.10) are both unbiased, consistent estimators of $f(x,\theta)$ and it makes sense to minimize $U_{1,n}$ as well as $U_{2,n}$. Recall that $\mathrm{E}_\theta\, b(X_0) = 0$ so $\frac{2}{n}\sum b(X_{i\Delta})$ becomes close to zero as n grows and $\hat{f}_{1,n}$ and $\hat{f}_{2,n}$ — and hence $U_{1,n}$ and $U_{2,n}$ — are close. For a moderate size of n, like 500, it might however make a difference whether we use $\hat{f}_{1,n}$ or $\hat{f}_{2,n}$. Note in particular that either $\hat{f}_{1,n}$ or $\hat{f}_{2,n}$ becomes negative (close to r or l) whereas f is positive on (l,r).

Instead of using either $\hat{f}_{1,n}$ or $\hat{f}_{2,n}$ we suggest using a convex combination of the two. Define for $\lambda(x) = \big(\lambda_1(x),\lambda_2(x)\big)$ with $\lambda_1(x)+\lambda_2(x) = 1$ the estimator $\hat{f}_{\lambda,n}(x)$ by

$$\begin{aligned}\hat{f}_{\lambda,n}(x) &= \lambda_1(x)\,\hat{f}_{1,n}(x) + \lambda_2(x)\,\hat{f}_{2,n}(x)\\ &= \hat{f}_{1,n}(x) - \frac{2}{n}\lambda_2(x)\sum_{i=1}^{n} b(X_{i\Delta}).\end{aligned} \tag{II.11}$$

With this notation $\hat{f}_{\lambda,n} = \hat{f}_{1,n}$ for $\lambda \equiv (1,0)$ and $\hat{f}_{\lambda,n} = \hat{f}_{2,n}$ for $\lambda \equiv (0,1)$.

If $\lambda(x)$ is deterministic, then $\hat{f}_{\lambda,n}(x)$ is unbiased for $f(x,\theta)$ and it makes sense to choose $\lambda(x)$ such that the variance of $\hat{f}_{\lambda,n}(x)$ is minimal. In general it is not

possible to calculate the variance of $\hat{f}_{\lambda,n}(x)$ since it involves covariances between functionals of $X_{i\Delta}$ and $X_{j\Delta}$ for $i \neq j$ which we typically do not know. It is easy, however, to minimize an approximation to the variance: First, note that if $X_0 \sim \mu_\theta$, then $\mathrm{Cov}_\theta\big(2b(X_0)1_{\{X_0 \leq x\}}, 2b(X_0)1_{\{X_0 > x\}}\big) = f^2(x,\theta)$. *If* the observations $X_\Delta, \ldots, X_{n\Delta}$ were independent and identically μ_θ-distributed we would thus get

$$\mathrm{Var}_\theta\, \hat{f}_{\lambda,n}(x) = \frac{1}{n}\Big(\lambda_1^2(x)V_{\theta,1}(x) + \lambda_2^2(x)V_{\theta,2}(x) - 2\lambda_1(x)\lambda_2(x)\, f^2(x,\theta)\Big)$$

where $V_{\theta,1}(x)$ and $V_{\theta,2}(x)$ are given by

$$\begin{aligned} V_{\theta,1}(x) &= \mathrm{Var}_\theta\big(2b(X_0)1_{\{X_0 \leq x\}}\big) = 4\,\mathrm{E}_\theta\, b^2(X_0)1_{\{X_0 \leq x\}} - f^2(x,\theta), \\ V_{\theta,2}(x) &= \mathrm{Var}_\theta\big(2b(X_0)1_{\{X_0 > x\}}\big) = 4\,\mathrm{E}_\theta\, b^2(X_0)1_{\{X_0 > x\}} - f^2(x,\theta). \end{aligned}$$

Easy calculations show that the minimal variance is

$$\frac{1}{n}\Big(4\lambda_{\theta,1}(x)\lambda_{\theta,2}(x)\,\mathrm{E}_\theta\, b^2(X_0) - f^2(x,\theta)\Big) \tag{II.12}$$

which is obtained for

$$\lambda_{\theta,1}(x) = \frac{V_{\theta,2}(x) + f^2(x,\theta)}{V_{\theta,1}(x) + V_{\theta,2}(x) + 2f^2(x,\theta)} = \frac{\mathrm{E}_\theta\, b^2(X_0)1_{\{X_0 > x\}}}{\mathrm{E}_\theta\, b^2(X_0)} \tag{II.13}$$

$$\lambda_{\theta,2}(x) = 1 - \lambda_{\theta,1}(x) = \frac{\mathrm{E}_\theta\, b^2(X_0)1_{\{X_0 \leq x\}}}{\mathrm{E}_\theta\, b^2(X_0)}. \tag{II.14}$$

Of course, the observations are *not* independent so these weights are only approximately optimal. Also, we do not know the expectations above, but we can use their empirical counterparts and consider

$$\hat{\lambda}_{1,n}(x) = \frac{\sum_{i=1}^n b^2(X_{i\Delta})1_{\{X_i > x\}}}{\sum_{i=1}^n b^2(X_{i\Delta})} \quad \text{and} \quad \hat{\lambda}_{2,n}(x) = \frac{\sum_{i=1}^n b^2(X_{i\Delta})1_{\{X_i \leq x\}}}{\sum_{i=1}^n b^2(X_{i\Delta})}.$$

The corresponding estimator $\hat{f}_n(x) = \hat{f}_{\hat{\lambda}_n,n}(x)$ is given by

$$\frac{2}{n\sum b^2(X_{i\Delta})}\Big\{\Big(\sum b^2(X_{i\Delta})1_{\{X_{i\Delta} > x\}}\Big)\Big(\sum b(X_{j\Delta})1_{\{X_{j\Delta} \leq x\}}\Big) - \Big(\sum b(X_{i\Delta})1_{\{X_{i\Delta} > x\}}\Big)\Big(\sum b^2(X_{j\Delta})1_{\{X_{j\Delta} \leq x\}}\Big)\Big\}$$

(all sums are from 1 to n). Note that $\hat{\lambda}_n$ and hence $\hat{f}_n(x)$ are well-defined even if b is not in $L^2(\mu_\theta)$.

For x close to l we have $\hat{\lambda}_1(x)$ close to 1 and hence $\hat{f}_n(x)$ close to $\hat{f}_{1,n}(x)$. Similarly $\hat{f}_n(x)$ is close to $\hat{f}_2(x)$ when x is close to r. In particular, $\hat{f}_n(x) = 0$ for x outside

the range of the observations. Note that $\hat{f}_n(x)$ is consistent for $f(x,\theta)$ but that it can be biased although $\hat{f}_{1,n}(x)$ and $\hat{f}_{2,n}(x)$ are unbiased.

Like $\hat{f}_{1,n}$ and $\hat{f}_{2,n}$, the estimator $\hat{f}_n$ is piecewise constant with jumps at each data point $X_{k\Delta}$. The jump size is

$$\hat{f}_n(X_{k\Delta}) - \lim_{x\uparrow X_{k\Delta}} \hat{f}_n(x) = \frac{2}{n} b(X_{k\Delta}) \left(1 - \frac{\sum b(X_{i\Delta})}{\sum b^2(X_{i\Delta})} b(X_{k\Delta})\right) \tag{II.15}$$

cf. (II.11). Since X is ergodic and $\mathrm{E}_\theta\, b(X_0) = 0$, the parenthesis in (II.15) will typically be positive in which case $\hat{f}_n$ increases (decreases) at $X_{k\Delta}$ if f_θ is increasing (decreasing) at $X_{k\Delta}$, *cf.* (II.6). In particular, if the parenthesis in (II.15) is positive for all $k = 1,\ldots,n$ and b is decreasing from some positive value (or limit) at l to some negative value (or limit) at r, then $\hat{f}_n$ is increasing as long as b is positive, decreasing thereafter and strictly positive between the smallest and the largest observation.

For estimation of θ the idea is of course to minimize the uniform distance

$$U_n(\theta) = \sup_{x\in I} \left| \hat{f}_n(x) - f_\theta(x) \right|. \tag{II.16}$$

between $\hat{f}_n$ and f_θ. We let $\hat{\theta}_n$ denote the corresponding estimator.

Important comments

Below follows important remarks on the three estimators of f_θ and the corresponding U-distances.

First an illustration of the difference between the three estimators of f_θ. Figure II.1 shows graphs of $\hat{f}_{1,n}$, $\hat{f}_{2,n}$ and $\hat{f}_n$ for 100 hypothetical data points. The data are simulated from the model $dX_t = (0.04 - 0.6X_t)\,dt + 0.2X_t^\gamma dW_t$ with true parameter value $\gamma_0 = 0.75$ and $\Delta = 1$. The model is discussed in detail in Section II.7.2. For this particular simulation $\sum_{i=1}^n b(X_{i\Delta}) > 0$ so the graph of $\hat{f}_{1,n}$ lies over the graph of $\hat{f}_{2,n}$. The graph of $\hat{f}_n$ is in between; close to $\hat{f}_{1,n}$ for small data values and close to $\hat{f}_{2,n}$ for large data values.

Second, note that neither $\hat{f}_{1,n}$, $\hat{f}_{2,n}$ or $\hat{f}_n$ would change if the order of observations was changed. In other words, the observations are treated as if they were independent. This is of course unfortunate since they come from a diffusion model with built-in dependence.

For "large" values of Δ the dependence between observations is minor and we would thus expect the method to perform better for "large" Δ than for "small" Δ. Still, it turns out that the proposed estimators are consistent as $n \to \infty$ for *any* fixed value of $\Delta > 0$ (Section II.4). Intuitively, this is because θ can be identified through the invariant distribution only (recall that $\mu_\theta = \mu_{\theta'}$ if and only if $f_\theta = f_{\theta'}$ if and only if $\theta = \theta'$). However, we *do* loose the information originating from the dependence between the observations. In the more realistic case of a parameter dependent (rather than known) drift function, we will use the joint distribution of

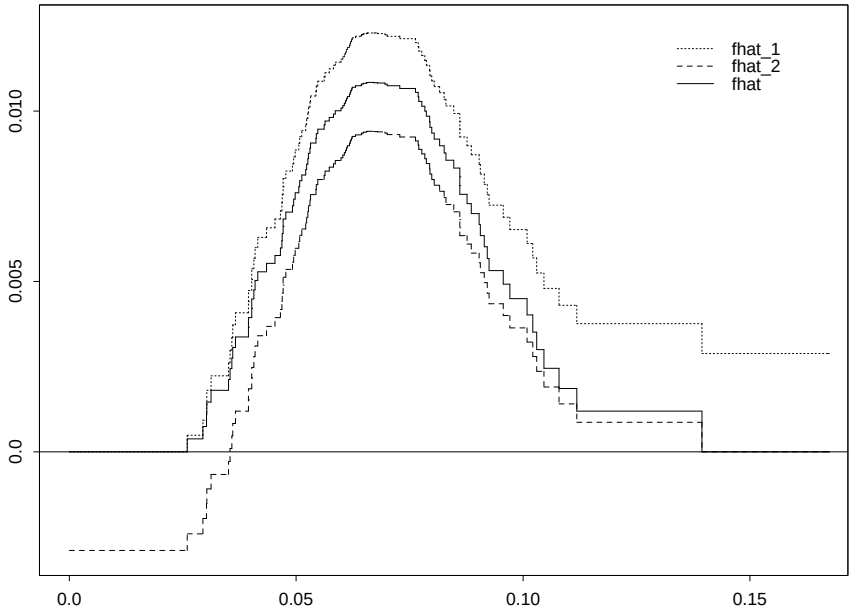

Figure II.1: Graphs for the estimators $\hat{f}_{1,n}$, $\hat{f}_{2,n}$ and $\hat{f}_n$ for 100 simulated data from the model $dX_t = (0.04 - 0.6X_t)\,dt + 0.2X_t^{\gamma}\,dW_t$ with true value $\gamma_0 = 0.75$. The value of Δ is 1.

two consecutive observations to estimate the drift parameters, see Section II.6 for details.

Third, an important practical remark. Despite the definition of $U_n(\theta)$ as a supremum over the whole state space I, we can calculate $U_n(\theta)$ from the values of f_θ and $\hat{f}_n$ at data points and points where b is zero. To be specific, let $\tilde{X}_1 \leq \cdots \leq \tilde{X}_n$ be the observations ordered according to size and $\tilde{X}_0 = l$. Then, because f_θ is continuous and has a derivative with same sign as b, and because $\hat{f}_n$ is piecewise constant, $U_n(\theta) = \max(N_0, N_1, N_2)$ where

$$
\begin{aligned}
N_1 &= \max_{k=1,\ldots,n} \left| \hat{f}_n(\tilde{X}_k) - f_\theta(\tilde{X}_k) \right| \\
N_2 &= \max_{k=1,\ldots,n} \left| \hat{f}_n(\tilde{X}_{k-1}) - f_\theta(\tilde{X}_k) \right| \\
N_0 &= \sup_{x_0 : b(x_0)=0} \left| \hat{f}_n\big(\tilde{X}(x_0)\big) - f_\theta(x_0) \right|.
\end{aligned}
$$

Here $\tilde{X}(x_0) = \max_{k=0,\ldots,n}\{\tilde{X}_k : \tilde{X}_k \leq x_0\}$ is the largest observation smaller than x_0 (or l if all observations are larger than x_0). For the most commonly used models b is only zero at very few points. In particular, if b is decreasing from some positive value (or limit) at l to some negative value (or limit) at r, then b is zero at a single point x_0 and $N_0 = \left| \hat{f}_n(\tilde{X}(x_0)) - f_\theta(x_0) \right|$.

Of course similar formulas apply to $U_{1,n}(\theta)$ ($U_{2,n}(\theta)$) as long as $f(l,\theta) = 0$ ($f(r,\theta) = 0$) for all $\theta \in \Theta$; simply substitute $\hat{f}_n$ by $\hat{f}_{1,n}$ ($\hat{f}_{2,n}$) and remember also to

compare $\hat{f}_{1,n}$ ($\hat{f}_{2,n}$) with f_θ at the endpoint r (l).

An alternative measure of distance

Finally, some comments on the nature of U_n and an alternative measure of distance between $\hat{f}_n$ and f_θ. It is easier to adjust f_θ to $\hat{f}_n$ in areas of I with many observations than in areas with few observations so the supremum in (II.16) will usually be attained for very small or very large data points. Consequently, the function $f_{\hat{\theta}_n}$ corresponding to the estimator $\hat{\theta}_n$ need not fit well with $\hat{f}_n$ for data points in the central area of the distribution but has (by definition) the least possible maximum distance.

The opposite effect is obtained if we define the distance between $\hat{f}_n$ and f_θ as a weighted sum of squares,

$$S_n^w(\theta) = \sum_{i=1}^{n} \Big(f(X_{i\Delta}, \theta) - \hat{f}_n(X_{i\Delta}) \Big)^2 w_i \tag{II.17}$$

with a contribution from each observation. It is natural to choose the i'th weight, w_i, as the inverse of the variance of $\hat{f}_n(X_{i\Delta})$, or rather an estimate of it, *e.g.*

$$\frac{1}{w_i} = \frac{4}{n} \left(\hat{\lambda}_{1,n}(X_{i\Delta}) \hat{\lambda}_{2,n}(X_{i\Delta}) \frac{1}{n} \sum_{j=1}^{n} b^2(X_{j\Delta}) - \frac{1}{4} \Big(\hat{f}_n(X_{i\Delta}) \Big)^2 \right) \tag{II.18}$$

cf. formula (II.12). Note that the variance of $\hat{f}_n(X_{i\Delta})$ is small when $X_{i\Delta}$ is close to l or r so small and large observations are given relatively large weights. In particular the variance estimate of $\hat{f}_n$ is zero for the largest observation since $\hat{\lambda}_{1,n}$ and $\hat{f}_n$ are both zero. Hence, (II.18) does not make sense for this observation. Instead we could give it same weight as the smallest observation, for example.

There are however only few observations near the endpoints and despite their large weights their contributions are in general negligible compared to the contributions from the many observations in the middle of the distribution (if the largest observation is given the weight suggested above). In effect, if $\tilde{\theta}_n$ is minimizing S_n^w, then $f_{\tilde{\theta}_n}$ and $\hat{f}_n$ fit almost perfect in areas with many observations but can differ considerably for extreme values.

The difference between the two criteria U_n and S_n^w is evident from Figure II.2 which shows $f_{\hat{\theta}_n} = f_{0.786}$ and $f_{\tilde{\theta}_n} = f_{0.770}$ for the 100 simulated data points used in Figure II.1: $\hat{f}_n$ is closer to $f_{\tilde{\theta}_n}$ than to $f_{\hat{\theta}_n}$ for average (and small) observations whereas $\hat{f}_n$ is closer to $f_{\hat{\theta}_n}$ than to $f_{\tilde{\theta}_n}$ for large observations.

In conclusion, U_n takes the tails of the distribution more into account than S_n^w. This is advantageous since there is often much information about the (diffusion) parameter contained in the tail behaviour. On the other hand, possible outliers are too influential (but could be discarded by taking supremum over a subset of I only). Note that U_n by definition compares f_θ and $\hat{f}_n$ at *all* points in I whereas S_n makes the comparison at the (random) data points only.

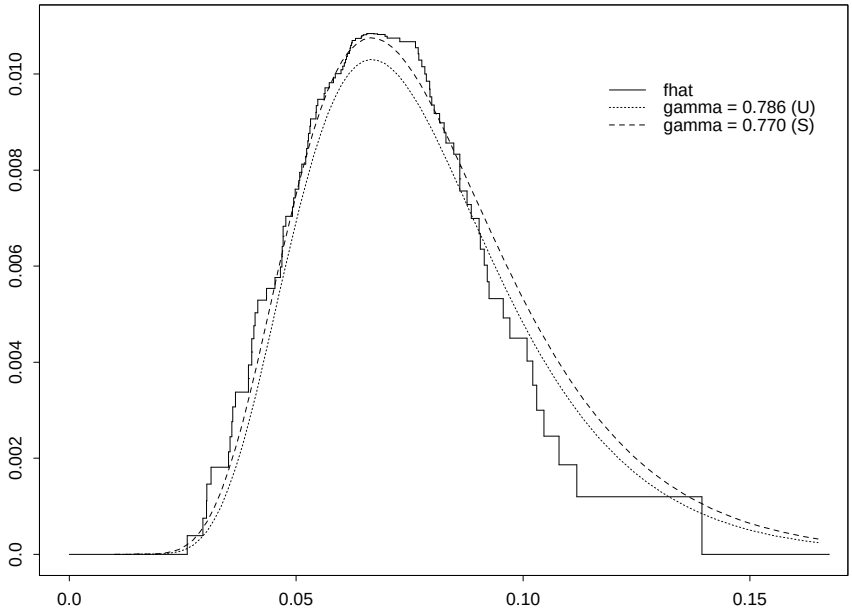

Figure II.2: The graph for $\hat{f}_n$ based on the data from Figure II.1 and graphs for f_θ for $\theta = \hat{\theta}_n = 0.786$ (minimizing U_n) and $\theta = \tilde{\theta}_n = 0.770$ (minimizing S_n^w).

In Section II.4 and II.5 we prove that $\hat{\theta}_n$ minimizing U_n is consistent and converges in distribution (when normed by $\sqrt{n}$). We have no such results for $\tilde{\theta}_n$, but in the simulation study in Section II.7.2 we calculate both estimators and there seems to be only little difference.

II.4 Consistency

In this section we prove that the estimators $\hat{\theta}_{1,n}$, $\hat{\theta}_{2,n}$ and $\hat{\theta}_n$ obtained by minimizing the supremum distances $U_{1,n}$, $U_{2,n}$ and U_n are consistent as $n \to \infty$ for any fixed $\Delta > 0$. It is implicitly assumed that the estimators exist (for n large enough).

Let $U(\theta) = \sup_{x \in I} \left| f_\theta(x) - f_{\theta_0}(x) \right|$ denote the uniform distance between f_θ and f_{θ_0}. Then $U(\theta) = 0$ if and only if $\theta = \theta_0$. We shall assume that θ_0 is well-separated as a minimum of U in following sense.

Assumption II.3 For all $\delta > 0$ it holds that

$$C(\delta) = \inf\{U(\theta) : ||\theta - \theta_0|| > \delta\} > 0. \qquad \square$$

The assumption is for example satisfied (*i*) if $\theta \to f_\theta(x)$ is increasing or decreasing for all $x \in I$ which will often be the case (this makes sense for one-dimensional parameters only); or (*ii*) if U is continuous and Θ is either open with U bounded

away from zero at the boundary or compact. A sufficient condition for continuity of U is that $\theta \to f(x,\theta)$ is continuous, uniformly in $x \in I$.

Theorem II.4 *Assume that Assumptions II.1, II.2 and II.3 hold and that b changes sign at most countably many times on I. If $f(l,\theta) = 0$ ($f(r,\theta) = 0$) for all $\theta \in \Theta$ then $\hat{\theta}_{1,n}$ ($\hat{\theta}_{2,n}$) is consistent for θ, and if $f(l,\theta) = f(r,\theta) = 0$ for all $\theta \in \Theta$ then $\hat{\theta}_n$ is consistent for θ as well.*

Proof It follows from van der Vaart & Wellner (1996, Corollary 3.2.2) that it is sufficient to show that the uniform distances converge in P_{θ_0}-probability (or almost surely with respect to P_{θ_0}) to $U(\theta)$, uniformly in θ.

First assume that $f(l,\theta) = 0$ for all $\theta \in \Theta$. By the triangle inequality for the uniform metric, it holds that $|U_{1,n}(\theta) - U(\theta)| \leq U_{1,n}(\theta_0)$ for all $\theta \in \Theta$ so it suffices to show

$$U_{1,n}(\theta_0) = \sup_{x \in I} \left| \hat{f}_{1,n}(x) - f_{\theta_0}(x) \right| \to 0 \tag{II.19}$$

P_{θ_0}-almost surely. Note that pointwise convergence follows from the ergodic theorem and Assumption II.2.

We can write $I = \cup_{j \in J} I_j$ where J is at most countable, each I_j has the form $[z_1, z_2]$ for some $z_1, z_2 \in I$ or $(l, z]$ or $[z, r)$ for some $z \in I$, and b is either non-positive or non-negative on I_j. To prove (II.19) it is enough to show that $\sup_{x \in I_j} \left| \hat{f}_{1,n}(x) - f_{\theta_0}(x) \right| \to 0$ for all $j \in J$.

Consider a $j \in J$ and assume for example that $I_j = [z_1, z_2]$ and that $b \geq 0$ on I_j. Then f_{θ_0} and $\hat{f}_{1,n}$ are non-decreasing on I_j since (II.6) holds and $\hat{f}_{1,n}$ is piecewise constant with jump size $b(X_{k\Delta})$ at $X_{k\Delta}$.

Since f_{θ_0} is continuous it takes all values in $[m, M]$ where $m = f_{\theta_0}(z_1)$ and $M = f_{\theta_0}(z_2)$. For $K \in \mathbb{N}$ given we choose $z_1 = x_0 < \ldots < x_{K-1} < x_K = z_2$ such that $f_{\theta_0}(x_k) = m + k(M-m)/K$ for all $k = 0, \ldots, K$. Then, for $k = 0, \ldots, K-1$ and $x_k \leq x \leq x_{k+1}$,

$$\begin{aligned}
\hat{f}_{1,n}(x) - f_{\theta_0}(x) &\leq \hat{f}_{1,n}(x_{k+1}) - f_{\theta_0}(x_k) \\
&= \hat{f}_{1,n}(x_{k+1}) - f_{\theta_0}(x_{k+1}) + f_{\theta_0}(x_{k+1}) - f_{\theta_0}(x_k) \\
&= \hat{f}_{1,n}(x_{k+1}) - f_{\theta_0}(x_{k+1}) + (M-m)/K
\end{aligned}$$

since $\hat{f}_{1,n}$ and f_{θ_0} are non-decreasing. Also

$$\begin{aligned}
f_{\theta_0}(x) - \hat{f}_{1,n}(x) &\leq f_{\theta_0}(x_{k+1}) - \hat{f}_{1,n}(x_k) \\
&\leq f_{\theta_0}(x_{k+1}) - f_{\theta_0}(x_k) + f_{\theta_0}(x_k) - \hat{f}_{1,n}(x_k) \\
&= f_{\theta_0}(x_k) - \hat{f}_{1,n}(x_k) + (M-m)/K.
\end{aligned}$$

Hence,

$$\sup_{x \in I_j} \left| \hat{f}_{1,n}(x) - f_{\theta_0}(x) \right| \leq \max_{k=1,\ldots,K} \left| \hat{f}_{1,n}(x_k) - f_{\theta_0}(x_k) \right| + (M-m)/K.$$

Now, choose $\tilde{A}_1, \ldots, \tilde{A}_K$ such that $P_{\theta_0}(\tilde{A}_k) = 1$ and $\bigl|\hat{f}_{1,n}(x_k) - f_{\theta_0}(x_k)\bigr| \to 0$ on A_k for all $k = 1, \ldots, K$. Then $\max_{k=1,\ldots,K}\bigl|\hat{f}_{1,n}(x_k) - f_{\theta_0}(x_k)\bigr| \to 0$ on $A_K = \tilde{A}_1 \cap \cdots \tilde{A}_K$ and $\sup_{x\in I_j}\bigl|\hat{f}_{1,n}(x) - f_{\theta_0}(x)\bigr| \to 0$ on $\cap_{K=1}^{\infty} A_K$, hence P_{θ_0}-almost surely.

Similar arguments apply if $I_j = (l, z]$ or $I_j = [x, r)$ and if $b \le 0$ on I_j. We have now proved (II.19) and thus uniform convergence of $U_{1,n}(\theta)$ to $U(\theta)$ and consistency of $\hat{\theta}_{1,n}$. Consistency of $\hat{\theta}_{2,n}$ follows similarly if $f(r, \theta) = 0$ for all $\theta \in \Theta$.

Finally assume that $f(l, \theta) = f(r, \theta) = 0$ for all $\theta \in \Theta$. Recall that $\hat{f}_n(x) = \hat{f}_{1,n}(x) - \frac{2}{n}\hat{\lambda}_{2,n}(x)\sum_{i=1}^n b(X_{i\Delta})$ and $0 \le \hat{\lambda}_{2,n}(x) \le 1$. By the triangle inequality for the supremum metric,

$$\begin{aligned}\bigl|U_n(\theta) - U(\theta)\bigr| &\le \sup_{x\in I}\bigl|\hat{f}_n(x) - f_{\theta_0}(x)\bigr| \\ &\le \sup_{x\in I}\bigl|\hat{f}_{1,n}(x) - f_{\theta_0}(x)\bigr| + 2\Bigl|\frac{1}{n}\sum_{i=1}^n b(X_{i\Delta})\Bigr|\end{aligned}$$

which converges uniformly in θ to zero P_{θ_0}-almost surely since $\mathrm{E}_{\theta_0} b(X_0) = 0$. This proves consistency of $\hat{\theta}_n$. □

II.5 Further asymptotic results

In this section we show that $\hat{\theta}_{1,n}$, $\hat{\theta}_{2,n}$, and $\hat{\theta}_n$ are $\sqrt{n}$-consistent and furthermore that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges weakly as $n \to \infty$. For simplicity we only list the assumptions for a one-dimensional parameter but the convergence result holds for multi-dimensional parameters under similar conditions.

Consider first $\hat{\theta}_{i,n}$, $i = 1, 2$. Proposition II.6 below claims that

$$M_{i,n}(h) = \sup_{x\in I}\Bigl|n^{1/2}\bigl(\hat{f}_{i,n}(x) - f_{\theta_0 + h/\sqrt{n}}(x)\bigr)\Bigr|$$

converges weakly, uniformly in $h \in H$ for any compact set $H \subseteq \mathbb{R}$. Write $M_{i,n}(h) = \sup_{x\in I}\bigl|M'_{i,n}(x) - M''_n(h,x)\bigr|$ where

$$\begin{aligned}M'_{i,n}(x) &= n^{1/2}\bigl(\hat{f}_{i,n}(x) - f_{\theta_0}(x)\bigr) \\ M''_n(h,x) &= n^{1/2}\bigl(f_{\theta_0 + h/\sqrt{n}}(x) - f_{\theta_0}(x)\bigr).\end{aligned}$$

Note that the processes $M''_{i,n}$ and $M_{i,n}$ are well-defined for n large enough provided that θ_0 is an inner point of Θ.

Recall that $|\hat{f}_{i,n}(x)| \le \frac{2}{n}\sum_{j=1}^n |b(X_{j\Delta})|$ for all $x \in I$ and that $|f_\theta(x)| \le 2\,\mathrm{E}_\theta\,|b(X_0)|$ for all $(x, \theta) \in I \times \Theta$. It follows that M''_n takes values in $l^\infty(H \times I)$ (since H is compact), $M'_{i,n}$ in $l^\infty(I)$, and thus $M_{i,n}$ in $l^\infty(H \times I)$. Here we have used the notation $l^\infty(T)$ for the set of uniformly bounded, real functions on T; $l^\infty(T) = \{g : \sup_{t\in T}|g(t)| < \infty\}$.

The process M''_n is non-stochastic and $M''_n(h,x) \to \dot{f}_{\theta_0}(x)h$ pointwise if $\theta \to f_\theta(x)$ is differentiable in θ_0 with derivative $\dot{f}_{\theta_0}(x)$. Assumption II.5.2 below ensures that

the convergence is suitably uniform. Note that it also ensures continuity of U in θ_0, *cf.* the remark below Assumption II.3. A sufficient condition for Assumption II.5.2 is that $\theta \to f_\theta(x)$ is twice differentiable in a neighbourhood Θ_0 of θ_0 for all $x \in I$ with the second derivative $\ddot{f}_\theta(x)$ bounded in $\Theta_0 \times I$, *i.e.* $\sup_{(x,\theta)\in I\times\Theta_0} |\ddot{f}_\theta(x)| < \infty$.

For convergence of M'_n we will use empirical process theory (Arcones & Yu 1994). See Appendix II.A for a brief introduction to the theory of empirical processes and the results used in the following. We assume that the drift has finite absolute p'th moment for some $p > 2$ (Assumption II.5.3) and that the temporal dependence in X decays fast enough. More precisely we assume that the β-mixing coefficients decrease at an exponential rate (Assumption II.5.4). As usual for stationary Markov processes, we define the β-mixing coefficients

$$\beta_k = \int \sup_A \left| p_{k\Delta,\theta_0}(x,A) - \mu_{\theta_0}(A) \right| d\mu_{\theta_0}(x)$$

where $p_{k\Delta,\theta_0}$ is the transition probability from time 0 to time $k\Delta$.

Assumption II.5 The true parameter value θ_0 is an inner point of Θ and for any $x \in I$ the function $\theta \to f_\theta(x) = f(x,\theta)$ is continuously differentiable in a neighbourhood of θ_0 with first partial derivative $\dot{f}_\theta = \partial f_\theta / \partial\theta$ satisfying

1. $\dot{f}_{\theta_0}$ is bounded, *i.e.* $\sup_{x\in I} |\dot{f}_{\theta_0}(x)| < \infty$;

2. $\sup_{x\in I} \left| \dot{f}_\theta(x) - \dot{f}_{\theta_0}(x) \right| \to 0$ as $\theta \to \theta_0$.

Furthermore,

3. $\mathrm{E}_{\theta_0} |b(X_0)|^p < \infty$ for some $p > 2$;

4. there exist constants $c_1, c_2 > 0$ such that $\beta_k \le c_1 e^{-c_2 k\Delta}$ for all $k \ge 1$. □

Proposition II.6 *Let H be an arbitrary compact subset of $\mathbb{R}$, and assume that Assumptions II.1, II.2 and II.5 hold. Then $\{M_{1,n}(h)\}_{h\in H}$ converges weakly if $f(l,\theta) = 0$ for all $\theta \in \Theta$ and $\{M_{2,n}(h)\}_{h\in H}$ converges weakly if $f(r,\theta) = 0$ for all $\theta \in \Theta$.*

Proof Assume first that $f(l,\theta) = 0$ for all $\theta \in \Theta$. We will use Theorem 2.1 from Arcones & Yu (1994) to show that $\{M'_{1,n}(x)\}_{x\in I}$ converges weakly to a Gaussian process. By Assumption II.5.4 the required mixing condition is satisfied: with p from Assumption II.5.3. it holds that $k^{p/(p-2)}(\log k)^{2(p-1)/(p-2)}\beta_k \to 0$ as $k \to \infty$.

Define for $x \in I$ the function $F_x : I \to \mathbb{R}$ by $F_x(y) = 2b(y)1_{\{y\le x\}}$ and let $\mathscr{F} = \{F_x\}_{x\in I}$. Then, $\mathrm{E}_\theta F_x(X_0) = f_\theta(x)$ and by definition of $\hat{f}_{1,n}$,

$$M'_{1,n}(x) = n^{-1/2} \sum_{i=1}^{n} \left(F_x(X_{i\Delta}) - \mathrm{E}_{\theta_0} F_x(X_0) \right).$$

(II.14)

The function $F_x(y)$ is jointly measurable in (x,y) and the envelope function of $\mathscr{F}$, $\sup_{x\in I}|F_x| = 2|b|$, has finite p'th moment by Assumption II.5.3. Furthermore, it follows from Lemma II.11 in the appendix that $\mathscr{F}$ is a so-called Vapnik-Červonenkis subgraph class of functions.

We conclude (Arcones & Yu 1994) that $M'_{1,n}$ converges weakly in $l^\infty(I)$ to a tight, Gaussian process with P_{θ_0}-almost all paths uniformly bounded and uniformly continuous (with respect to the metric d on I given by $d(x,y)^2 = \int (F_x - F_y)^2 d\mu_{\theta_0}$).

Convergence of M''_n follows from Assumption II.5.2, and the limit process M'' given by $M''(h,x) = \dot{f}_{\theta_0}(x)h$ is in $l^\infty(H\times I)$ by Assumption II.5.1. It now follows from Slutsky's Theorem that $M'_{1,n} - M''_n$ converges weakly in $l^\infty(H\times I)$ and finally, convergence of $M_{1,n}$ in $l^\infty(H)$ follows by the Continuous Mapping Theorem.

Similarly for $M_{2,n}$ if $f(r,\theta) = 0$ for all $\theta \in \Theta$. □

We have just established convergence of $M_{1,n}(h)$ and $M_{2,n}(h)$, uniformly in $h \in H$ for compact sets H. Note however that the limit processes are *not* Gaussian (except perhaps for very special cases). In the much simpler case where the observations are independent and identically uniformly distributed on $(0,1)$ and $b \equiv 1$ (so that $\hat{f}_{1,n}(x)$ is simply the empirical distribution function) one has a — rather unpleasant — expression for the distribution function of the limit $M_1(0) = \sup|M'_1(x)|$, see Billingsley (1968, Chapter 13). In the more complicated case under consideration in this paper it is not possible to identify the distribution of the limit.

By the above convergence results for $M_{i,n}$ we can now show $\sqrt{n}$-consistency of $\hat{\theta}_{i,n}$, $i = 1,2$.

Theorem II.7 *Assume that Assumptions II.1, II.2, II.3 and II.5 hold and furthermore that $\dot{f}_{\theta_0}(x_0) \neq 0$ for an $x_0 \in I$. Then $\sqrt{n}(\hat{\theta}_{1,n} - \theta_0)$ is $O_p(1)$ if $f(l,\theta) = 0$ for all $\theta \in \Theta$ and $\sqrt{n}(\hat{\theta}_{2,n} - \theta_0)$ is $O_p(1)$ if $f(r,\theta) = 0$ for all $\theta \in \Theta$.*

Proof Recall that $\hat{\theta}_{i,n}$ minimizes $U_{i,n}(\theta) = \sup_{x\in I}|\hat{f}_{i,n}(x) - f_\theta(x)|$ and that $U_{i,n}(\theta) \to U(\theta) = \sup_{x\in I}|f_{\theta_0}(x) - f_\theta(x)|$ P_{θ_0}-almost surely as $n \to \infty$. It is easy to see that $\sqrt{n}U(\hat{\theta}_{i,n})$ is $O_p(1)$: By the triangle inequality

$$\sqrt{n}U(\hat{\theta}_{i,n}) \le \sqrt{n}U_{i,n}(\hat{\theta}_{i,n}) + \sqrt{n}U_{i,n}(\theta_0) \le 2\sqrt{n}U_{i,n}(\theta_0)$$

and $\sqrt{n}U_{i,n}(\theta_0) = M_{i,n}(0)$ converges weakly and is hence $O_p(1)$.

Recall the definition of $C(\delta)$ from Assumption II.3 and note that $P(\sqrt{n}|\hat{\theta}_{i,n} - \theta_0| > \delta) \le P\big(\sqrt{n}U(\hat{\theta}_{i,n}) \ge \sqrt{n}C(\delta/\sqrt{n})\big)$ for all $\delta > 0$. Hence, if

$$\sqrt{n}C(\delta/\sqrt{n}) > c\delta \tag{II.20}$$

for all $\delta > 0$, some constant $c > 0$ not depending on δ and n large enough, then $\sqrt{n}\big(\hat{\theta}_{i,n} - \theta_0\big)$ is $O_p(1)$.

To prove (II.20), choose $c, \eta > 0$ such that $U(\theta) > c|\theta - \theta_0|$ for all θ with $|\theta - \theta_0| \leq \eta$. This is possible by differentiability of $\theta \to f_\theta(x_0)$ (use *e.g.* $c = |\dot{f}_{\theta_0}(x_0)|/2$). For $n > \delta^2/\eta^2$,

$$\begin{aligned} C(\delta/\sqrt{n}) &= \inf\{U(\theta) : |\theta - \theta_0| > \delta/\sqrt{n}\} \\ &= \min\Big(\inf\{U(\theta) : \delta/\sqrt{n} < |\theta - \theta_0| \leq \eta\}, \inf\{U(\theta) : |\theta - \theta_0| > \eta\}\Big) \\ &= \min\Big(\inf\{U(\theta) : \delta/\sqrt{n} < |\theta - \theta_0| \leq \eta\}, C(\eta)\Big). \end{aligned}$$

Now, $C(\eta) > 0$ by Assumption II.3 and $\inf\{U(\theta) : \delta/\sqrt{n} < |\theta - \theta_0| < \eta\} \to 0$ as $n \to \infty$ since $U(\theta_0) = 0$ and U is continuous in θ_0. Hence, for n large enough

$$C(\delta/\sqrt{n}) = \inf\{U(\theta) : \delta/\sqrt{n} < |\theta - \theta_0| < \eta\} > c\delta/\sqrt{n}$$

which proves (II.20) and thus $\sqrt{n}$-consistency of $\hat{\theta}_{i,n}$. □

We now consider the situation where $f(l,\theta) = f(r,\theta) = 0$ for all θ and show that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is $O_p(1)$ and even converges weakly.

Define $M_n'(x) = n^{1/2}\big(\hat{f}_n(x) - f_{\theta_0}(x)\big)$ and $M_n(h) = \sup_{x\in I} |M_n'(x) - M_n''(h,x)|$. We first give a uniform convergence result for M_n. As in the proof of proposition II.6 we use empirical process theory to show convergence of M'. In this case it is however not immediate that the relevant class of functions is a Vapnik-Červonenkis subgraph class, and rather than showing that it is (which is indeed the case, see Lemma II.12 in the appendix), we choose to work with *covering numbers* directly.

Proposition II.8 *Assume that Assumptions II.1, II.2 and II.5 hold and that $f(l,\theta) = f(r,\theta) = 0$ for all $\theta \in \Theta$. Then $\{M_n(h)\}_{h\in H}$ converges weakly for any compact set $H \subseteq \mathbb{R}$.*

Proof Recall that $\hat{f}_n = \hat{\lambda}_{1,n}\hat{f}_{1,n} + \hat{\lambda}_{2,n}\hat{f}_{2,n}$ where $\hat{\lambda}_{j,n}$ converges pointwise (and uniformly as we shall argue below) P_{θ_0}-almost surely to $\lambda_j := \lambda_{\theta_0,j}$, $j = 1,2$. We first argue that it suffices to consider $\lambda_1\hat{f}_{1,n} + \lambda_2\hat{f}_{2,n}$ instead of $\hat{f}_n$: By adding and subtracting $\lambda_1\hat{f}_{1,n}$ and $\lambda_2\hat{f}_{2,n}$ we get

$$\begin{aligned} \hat{f}_n &= (\hat{\lambda}_{1,n} - \lambda_1)\hat{f}_{1,n} + (\hat{\lambda}_{2,n} - \lambda_2)\hat{f}_{2,n} + \lambda_1\hat{f}_{1,n} + \lambda_2\hat{f}_{2,n} \\ &= (\hat{\lambda}_{1,n} - \lambda_1)(\hat{f}_{1,n} - f_{\theta_0}) + (\hat{\lambda}_{2,n} - \lambda_2)(\hat{f}_{2,n} - f_{\theta_0}) + \lambda_1\hat{f}_{1,n} + \lambda_2\hat{f}_{2,n} \end{aligned}$$

and hence,

$$M_n' = (\hat{\lambda}_{1,n} - \lambda_1)M_{1,n}' + (\hat{\lambda}_{2,n} - \lambda_2)M_{2,n}' + M_{\lambda,n}' \tag{II.21}$$

where $M_{\lambda,n}'(x) = n^{1/2}\big(\lambda_1(x)\hat{f}_{1,n}(x) + \lambda_2(x)\hat{f}_{2,n}(x) - f_{\theta_0}(x)\big)$.

Since λ_1 is continuous and decreasing from one to zero, it takes all values in the unit interval $(0,1)$. From arguments almost identical to those leading to

the uniform convergence (II.19) of $\hat{f}_{1,n}$ to f_{θ_0}, it follows that $\hat{\lambda}_{1,n}(x) \to \lambda_1(x)$ and hence also $\hat{\lambda}_{2,n}(x) \to \lambda_2(x)$ P_{θ_0}-almost surely, uniformly in $x \in I$. In the proof of Proposition II.6 we showed that $M'_{1,n}$ and $M'_{2,n}$ converge weakly and it now follows from Slutsky's Theorem that M'_n converges in $l^\infty(I)$ if $M'_{\lambda,n}$ does.

Now, let $\mathscr{F} = \{F_x\}_{x\in I}$ where $F_x : I \to \mathbb{R}$ is defined by

$$\begin{aligned} F_x(y) &= 2\lambda_1(x)b(y)1_{\{y\le x\}} - 2\lambda_2(x)b(y)1_{\{y>x\}} \\ &= 2b(y)\big(\lambda_1(x) - 1_{\{y>x\}}\big), \quad y \in I. \end{aligned}$$

Then $\mathrm{E}_\theta F_x(X_0) = f_\theta(x)$ and $M'_{\lambda,n}(x) = n^{-1/2}\sum_{i=1}^n \big(F_x(X_{i\Delta}) - f_{\theta_0}(x)\big)$. The function $F_x(y)$ is jointly measurable in (x,y) and the envelope function $\sup_{x\in I}|F_x| = 2|b|$ of $\mathscr{F}$ has finite p'th moment by Assumption II.5.3.

Let Q be a probability measure on I with $b \in L^2(Q)$, let $||\cdot||_Q$ be the $L^2(Q)$-norm and define $\bar{B}_Q = \int b^2 dQ$. We show that the $||\cdot||_Q$-covering number $N(\varepsilon, \mathscr{F}, ||\cdot||_Q)$, which is the minimal number of $||\cdot||_Q$-balls of radius ε needed to cover $\mathscr{F}$, is at most $32\bar{B}_Q/\varepsilon^2$ (at least for small ε).

First, note that for all $x, z \in I$

$$\begin{aligned} ||F_x - F_z||_Q^2 &= \int (F_x - F_z)^2 dQ \\ &= 4\int b^2\big(\lambda_1(x) - 1_{(x,r)} - \lambda_1(z) + 1_{(z,r)}\big)^2 dQ \\ &\le 8\int b^2\big(\lambda_1(x) - \lambda_1(z)\big)^2 dQ + 8\int b^2\big(1_{(x,r)} - 1_{(z,r)}\big)^2 dQ. \end{aligned}$$

Define $B_Q(x) = \int_l^x b^2 dQ = \int b^2 1_{(l,x]} dQ$ and use the notation $\wedge$ for minimum and $\vee$ for maximum. Then, $\big(1_{(x,r)} - 1_{(z,r)}\big)^2 = 1_{(l,x\vee z]} - 1_{(l,x\wedge z]}$ and

$$||F_x - F_z||_Q^2 \le 8\big(\lambda_1(x) - \lambda_1(z)\big)^2\bar{B}_Q + 8B_Q(x\vee z) - 8B_Q(x\wedge z).$$

Next, for $0 < \varepsilon < 4\bar{B}_Q^{1/2}$ given, let $K = 16\bar{B}_Q/\varepsilon^2$ (or rather the smallest integer larger than this number). The functions λ_1 and B_Q are continuous, λ_1 decreases from 1 to 0 and B_Q increases from 0 to $\bar{B}_Q$ so we can choose $u_1, \dots, u_{K-1}$ and $v_1, \dots, v_{K-1}$ such that

$$8\bar{B}_Q\lambda_1(u_k) = 8B_Q(v_k) = k\varepsilon^2/2, \quad k = 1, \dots, K-1.$$

For $k = 2, \dots, 2K-1$, define y_k as the $(k-1)$'st smallest of the $2(K-1)$ points $\{u_k, v_k\}_{k=1,\dots,K-1}$. Also, let $y_1 = l$ and $y_{2K} = r$. Then $||F_x - F_z||_Q < \varepsilon$ for $x, z \in [y_k, y_{k+1}]$ for some $k = 1, \dots, 2K-1$. Indeed, let $y_k \le x \le z \le y_{k+1}$ and let $\underline{u} = \max\{u_j \le y_k, j = 1, \dots, K-1\}$ and $\bar{u} = \min\{u_j \ge y_k, j = 1, \dots, K-1\}$ be the u_j's that are closest to y_k

(and smaller/larger respectively). Define $\underline{v}$ and $\bar{v}$ similarly. Then,

$$\begin{aligned}||F_z - F_z||_Q^2 &\leq 8\big(\lambda_1(x) - \lambda_1(z)\big)^2 \bar{B}_Q + 8B_Q(z) - 8B_Q(x) \\ &\leq 8\big(\lambda_1(\underline{u}) - \lambda_1(\bar{u})\big)^2 \bar{B}_Q + 8B_Q(\bar{v}) - 8B_Q(\underline{v}) \\ &\leq \frac{\big(\varepsilon^2/2\big)^2}{8B_Q} + \varepsilon^2/2 \\ &< \varepsilon^2,\end{aligned}$$

and $\mathscr{F}$ can be covered by $2K$ balls (with respect to $||\cdot||_Q$) of radius ε. Hence,

$$N(\varepsilon, \mathscr{F}, ||\cdot||_Q) \leq 2K = 32\bar{B}_Q/\varepsilon^2 = 32||b||^2_{L^2(Q)}/\varepsilon^2 \tag{II.22}$$

for any Q with $b \in L^2(Q)$ (and ε small enough). In particular (II.22) holds for $Q = \mu_{\theta_0}$ and hence $\int_0^\infty \big(\log N(\varepsilon, \mathscr{F}, ||\cdot||_{\mu_{\theta_0}})\big)^{1/2} d\varepsilon < \infty$.

It follows (Arcones & Yu 1994, Lemma 2.1) that $M'_{\lambda,n}$ converges in $l^\infty(I)$ and hence from (II.21) that M'_n converges in $l^\infty(I)$. Finally, weak convergence of M''_n and M_n follows as in the proof of Proposition II.6. □

Theorem II.9 *Assume that Assumptions II.1, II.2, II.3, and II.5 hold and $f(l,\theta) = f(r,\theta) = 0$ for all $\theta \in \Theta$. If, in addition, $\dot{f}_{\theta_0}(x_0) \neq 0$ for some $x_0 \in I$ then $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is $O_p(1)$ and if furthermore $\dot{f}_{\theta_0}(x) \neq 0$ for all $x \in I$, then $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges weakly.*

Proof The $\sqrt{n}$-consistency follows exactly as in the proof of Theorem II.7.

For the weak convergence it then suffices to show that P_{θ_0}-almost all paths of the limit M of M_n has a unique minimum (van der Vaart & Wellner 1996, Theorem 3.2.2).

The limit process $\{M(h)\}_{h\in\mathbb{R}}$ has the form $M(h) = \sup_{x\in I} |M'(x) - \dot{f}_{\theta_0}(x)h|$ where M' is the Gaussian limit of M'_n. We first show that $M'(x) \to 0$ P_{θ_0}-almost surely as $x \searrow l$ and $x \nearrow r$, that is $P_{\theta_0}(M' \in A) = 1$ where $A = \{\varphi = (\varphi_x)_{x\in I} \in l^\infty(I) : \lim_{x\searrow l} \varphi_x = \lim_{x\nearrow r} \varphi_x = 0\}$.

It is easy to see that A is closed with respect to the uniform metric $d(\varphi,\varphi') = \sup_{x\in I} |\varphi_x - \varphi'_x|$. Indeed, let (φ^n) be a sequence from A with $\varphi^n \to \varphi$ and let $\varepsilon > 0$. Choose N such that $d(\varphi^n,\varphi) < \varepsilon/2$ for all $n \geq N$. In addition, choose x_l and x_r such that $\big|\varphi^N_x\big| < \varepsilon/2$ for $x \leq x_l$ and for $x \geq x_r$. Then, for $x \leq x_l$ and $x \geq x_r$,

$$|\varphi_x| \leq |\varphi_x - \varphi^N_x| + \big|\varphi^N_x\big| \leq \sup_{x\in I} \big|\varphi_x - \varphi^N_x\big| + \big|\varphi^N_x\big| < \varepsilon$$

so $\varphi \in A$ and A is closed.

For every $n \geq 1$ all paths of M'_n are in A since $\hat{f}_n(x) = 0$ for all $x < \min\{X_{i\Delta} : i = 1,\ldots,n\}$ and all $x \geq \max\{X_{i\Delta} : i = 1,\ldots,n\}$ and $\lim_{x\to l} f_{\theta_0}(x) = \lim_{x\to r} f_{\theta_0}(x) = 0$. It now follows from Portmanteau's theorem that $P_{\theta_0}(M' \in A) \geq \limsup_{n\to\infty} P_{\theta_0}(M'_n \in A) = 1$.

Now, all paths $h \to M(h)$ satisfy $M(h) \to \infty$ as $h \to \pm\infty$ since $M(h) \geq \left|M'(x) - \dot{f}_{\theta_0}(x)h\right|$ and $\dot{f}_{\theta_0}(x) \neq 0$ for any fixed $x \in I$ (for this it suffices that $\dot{f}_{\theta_0}(x_0) = 0$ for some $x_0 \in I$). All paths are continuous since $|M(h_2) - M(h_1)| \leq |h_2 - h_1| \sup_{x\in I} |\dot{f}_{\theta_0}(x)|$ for all $h_1, h_2 \in \mathbb{R}$ and hence have a minimum. We must show that the minimum is unique.

It is easy to see that all paths of M are (weakly) convex: for $h_1, h_2 \in \mathbb{R}$ and $\alpha \in (0,1)$

$$\begin{aligned} &M\big(\alpha h_1 + (1-\alpha)h_2\big) \\ &\quad = \sup_{x\in I}\Big|\alpha\big(M'(x) - \dot{f}_{\theta_0}(x)h_1\big) + (1-\alpha)\big(M'(x) - \dot{f}_{\theta_0}(x)h_2\big)\Big| \\ &\quad \leq \alpha \sup_{x\in I}\big|M'(x) - \dot{f}_{\theta_0}(x)h_1\big| + (1-\alpha)\sup_{x\in I}\big|M'(x) - \dot{f}_{\theta_0}(x)h_2\big| \\ &\quad = \alpha M(h_1) + (1-\alpha)M(h_2). \end{aligned}$$

It holds P_{θ_0}-almost surely that M' is continuous and belongs to the set A from above. Consider a path $h \to M(h)$ for which this is the case and assume that $h_1 < h_2$ both minimize M. Let $m = M(h_1) = M(h_2)$ be the minimum value. By convexity $M(\bar{h}) = m$ where $\bar{h} = (h_1 + h_2)/2$ is the mid point between h_1 and h_2.

By definition, $M(\bar{h}) = \sup_{x\in I}\big|M'(x) - \dot{f}_{\theta_0}(x)\bar{h}\big|$. Choose a sequence (x_n) from I such that $\big|M'(x_n) - \dot{f}_{\theta_0}(x_n)\bar{h}\big| \geq m - 1/n$ for each $n \geq 1$. For $j = 1, 2$ and all $n \geq 1$,

$$m = M(h_j) \geq \big|M'(x_n) - \dot{f}_{\theta_0}(x_n)h_j\big|$$

implying that $\big|\dot{f}_{\theta_0}(x_n)\big|(h_2 - h_1)/2 \leq 1/n$ and hence $\big|\dot{f}_{\theta_0}(x_n)\big| \to 0$ as $n \to \infty$.

Since $\dot{f}_{\theta_0}(x) \neq 0$ for all $x \in I$ it thus holds for any $l < x_1 < x_2 < r$ that $x_n \notin [x_1, x_2]$ for n large enough and hence $M'(x_n) \to 0$ as $n \to \infty$. It follows that

$$m = M(\bar{h}) = \lim_{n\to\infty}\big|M'(x_n) - \dot{f}_{\theta_0}(x_n)\bar{h}\big| = 0 \tag{II.23}$$

so $M(h_1) = M(h_2) = m = 0$. This is not possible, though, since for any $x \in I$ at least one of the values $|M'(x) - \dot{f}_{\theta_0}(x)h_1|$ and $|M'(x) - \dot{f}_{\theta_0}(x)h_2|$ is strictly positive.

We conclude that M has a unique minimum P_{θ_0}-almost surely and hence that $\sqrt{n}\big(\hat{\theta}_n - \theta_0\big)$ converges weakly. □

We have just shown that $\hat{\theta}_n$ converges weakly, but there is no reason to believe that the limit distribution is Gaussian. Simulation studies indicate however that the limit distribution might be close to normal.

Parts of the above proof could be repeated with M_1 or M_2 substituted for M. If $\bar{h}$ and (x_n) are as above with M replaced by M_1, say, then it would still hold that x_n could be made arbitrarily close to l or r by choosing n large enough. But $M_1'(x)$ does not converge to zero as $x \to r$ so $\lim_{n\to\infty}\big|M_1'(x_n) - \dot{f}_{\theta_0}(x_n)\bar{h}\big|$, corresponding to (II.23), need not be zero and cannot be rejected as the minimum value of M_1.

Similarly, $M_2'(x)$ does not converge to zero as $x \searrow l$ and we cannot rule out the possibility that M_1 and M_2 have several minimum points.

Note that the limits of $M_1'(x)$ and $M_2'(x)$ as x tends to l and r *do* exist with $\lim_{x\nearrow r} M_1'(x)$ and $\lim_{x\searrow l} M_2'(x)$ Gaussian and $\lim_{x\searrow l} M_1'(x) = \lim_{x\nearrow r} M_2'(x) = 0$ P_{θ_0}-almost surely. This can be proved via Portmanteau's theorem and the continuous mapping theorem.

Finally, a comment on Assumption II.5: The boundedness conditions may seem rather restrictive but in practice we can take a supremum over a (very large) compact subset of I rather than over I when forming the criterion functions $U_{1,n}$, $U_{2,n}$ and U_n. Then, by continuity, the boundedness conditions are automatically satisfied.

II.6 When the drift is not known

So far, we have assumed that the drift is completely known which is of course unrealistic. When this is not the case we follow the approach of Aït-Sahalia (1996) in that we suggest estimating the drift beforehand and then simply pretend that the drift is equal to its estimator when estimating the diffusion parameters.

More precisely we assume that the drift has a parametric specification $b(x) = b(x,\xi)$ and that the parameter ξ can be estimated consistently without any knowledge of the diffusion parameter θ. This is the case if b is linear and the martingale part of X is a genuine martingale: then we can use martingale estimation functions as suggested by Bibby & Sørensen (1995). Let $\hat{\xi}$ be the estimator of ξ and redefine $\hat{f}_{1,n}$ in the obvious way

$$\hat{f}_{1,n}(x) = \frac{2}{n}\sum_{i=1}^{n}\Big(b(X_{i\Delta},\hat{\xi})1_{\{X_{i\Delta}\le x\}}\Big).$$

Similarly for $\hat{f}_{2,n}$ and $\hat{f}_n$. The true function f of course also depends on ξ. Again we just plug in the estimator and minimize $\sup_{x\in(l,r)}\big|f(x,\hat{\xi},\theta) - \hat{f}_n(x)\big|$.

II.7 Examples

We now consider two particular models, namely the Ornstein-Uhlenbeck process (or Vasicek model) and the CKLS model. Of course, for the Ornstein-Uhlenbeck process the estimation problem is already solved since the transition probabilities are known and we can do maximum likelihood estimation. We study it briefly anyway since we get some qualitative results on the improvement caused by using $\hat{f}_n$ rather than $\hat{f}_{1,n}$ or $\hat{f}_{2,n}$. For the CKLS model we discuss various estimation methods and compare them to our estimation approach in a simulation study.

II.7.1 The Ornstein-Uhlenbeck process

Consider the stochastic differential equation

$$dX_t = \beta X_t\, dt + \sigma\, dW_t$$

where β is a known constant and $\sigma > 0$ is the unknown parameter. A solution X exists for all combinations of $\beta \in \mathbb{R}$ and $\sigma > 0$. The transition probabilities are normal,

$$X_t | X_0 \sim N\left(e^{\beta t} X_0, -\frac{\sigma^2}{2\beta}\left(1 - e^{2\beta t}\right)\right), \quad \beta \neq 0$$

and the state space is $\mathbb{R}$.

We will only consider $\beta < 0$. Then X is stationary and ergodic with invariant distribution $\mu_\sigma = N(0, -\sigma^2/2\beta)$. The function f is thus given by

$$f(x,\sigma) = \sigma\sqrt{-\frac{\beta}{\pi}}\exp\left(\beta x^2/\sigma^2\right), \quad x \in \mathbb{R}, \sigma > 0,$$

and $f(x,\sigma) \to 0$ as $x \to \pm\infty$ for all $\sigma > 0$. Figure II.3 shows the graph of $f(\cdot,\sigma)$ for $\beta = -1$ and various values of σ.

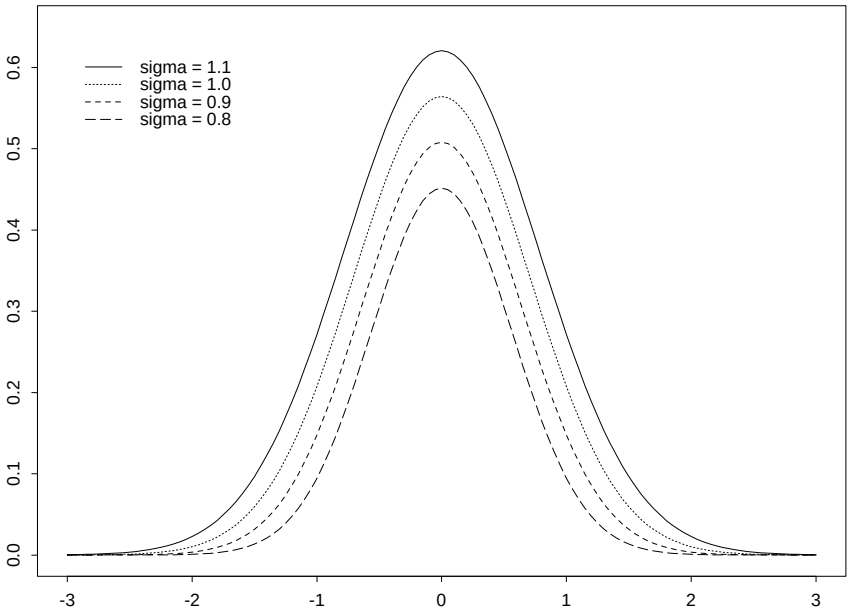

Figure II.3: The graph of the function $x \to f(x,\sigma)$ for the Ornstein-Uhlenbeck process for $\beta = -1$ and various values of σ.

The function f is twice differentiable with respect to σ with derivatives

$$\dot{f}(x,\sigma) = \frac{\partial f(x,\sigma)}{\partial\sigma} = \sqrt{-\frac{\beta}{\pi}}\left(1 - \frac{2\beta x^2}{\sigma^2}\right)\exp\left(\beta x^2/\sigma^2\right) > 0 \tag{II.24}$$

$$\ddot{f}(x,\sigma) = \frac{\partial^2 f(x,\sigma)}{\partial\sigma^2} = \sqrt{-\frac{\beta}{\pi}}\left(\frac{2\beta x^2}{\sigma^3} + \frac{4\beta^2 x^4}{\sigma^5}\right)\exp\left(\beta x^2/\sigma^2\right).$$

In particular $\dot{f}$ and $\ddot{f}$ are bounded on $\mathbb{R} \times (0,\infty)$ so Assumptions II.5.1 and II.5.2 are satisfied. Note that $\dot{f}(x,\sigma) \to 0$ as $x \to \pm\infty$ for all $\sigma > 0$. Assumption II.5.3 holds because all moments of μ_σ exist, and we prove exponential decay of the β-mixing coefficient (Assumption II.5.4) in Appendix II.B. Also, σ_0 is well-separated as a minimum of $U_\sigma = \sup_{x\in\mathbb{R}} |f_\sigma(x) - f_{\sigma_0}(x)|$ since $\sigma \to f_\sigma(x)$ is increasing for all $x \in \mathbb{R}$, *cf.* (II.24), so Assumption II.3 holds.

Hence, by the theorems in Section II.5, the estimators obtained by minimizing $U_{n,1}$, $U_{n,2}$ and U_n are all $\sqrt{n}$-consistent and the estimator obtained by minimizing U_n is even weakly convergent (when centered and scaled by $\sqrt{n}$, of course).

Now, let us consider three even simpler estimators for which we can (at least partly) determine the limit distribution. Choose $x_0 \in \mathbb{R}$ and solve the estimating equation, $\hat{f}_{1,n}(x_0) = f(x_0,\sigma)$. Denote the solution by $\bar{\sigma}_{n,1}$ and define $\bar{\sigma}_{n,2}$ and $\bar{\sigma}_n$ by substituting $\hat{f}_{2,n}$ and $\hat{f}_n$ for $\hat{f}_{1,n}$. In other words: we estimate σ by the value that makes the function $f(\cdot,\sigma)$ and its estimator ($\hat{f}_{1,n}$, $\hat{f}_{2,n}$ or $\hat{f}_n$) fit perfect in x_0 — without taking into account at all how they fit in other points.

Since the uniform criterion functions from the previous sections take the whole state space into account one would expect the corresponding estimators of σ to be more precise than the $\bar{\sigma}_n$'s just defined. The reason for considering the $\bar{\sigma}_n$'s at all, is that we for a particular x_0 are able to compare the limit distributions of $\bar{\sigma}_{n,1}$, $\bar{\sigma}_{n,2}$ and $\bar{\sigma}_n$ and hence give qualitative statements on the improvement on the variance caused by using $\hat{f}_n$ rather than $\hat{f}_{1,n}$ or $\hat{f}_{2,n}$.

First, for x_0 arbitrary, $\bar{\sigma}_{n,1}$ and $\bar{\sigma}_{n,2}$ solve

$$0 = n\big(\hat{f}_{1,n}(x_0) - f(x_0,\sigma)\big) = \sum_{i=1}^{n} \big(2b(X_{i\Delta})1_{\{X_{i\Delta}\le x_0\}} - f(x_0,\sigma)\big)$$

$$0 = n\big(\hat{f}_{2,n}(x_0) - f(x_0,\sigma)\big) = \sum_{i=1}^{n} \big(-2b(X_{i\Delta})1_{\{X_{i\Delta}> x_0\}} - f(x_0,\sigma)\big)$$

respectively. These equation are examples of so-called simple, unbiased estimating equations, *i.e.* equations on the form $\sum_{i=1}^{n} g(X_{i\Delta},\sigma) = 0$ where $\mathrm{E}_\sigma\, g(X_0,\sigma) = 0$. Under regularity conditions one can show that solutions to simple, unbiased estimating equations are consistent and asymptotically normal. See Kessler (2000), for example, for further details, proofs and the expression for the asymptotic variance (which can usually not be computed explicitly).

Let us turn to the special case $x_0 = 0$. Then the expression for $f(x_0,\sigma)$ is particularly simple, $f(x_0,\sigma) = f(0,\sigma) = \sigma(-\beta/\pi)^{1/2}$. The above estimating equations are then linear and can be solved explicitly

$$\bar{\sigma}_{n,1} = \frac{1}{n}\sum_{i=1}^{n} cX_{i\Delta}1_{\{X_{i\Delta}\le 0\}}$$

$$\bar{\sigma}_{n,2} = -\frac{1}{n}\sum_{i=1}^{n} cX_{i\Delta}1_{\{X_{i\Delta}> 0\}}$$

where $c = 2\beta(-\beta/\pi)^{-1/2} = -2(-\beta\pi)^{1/2}$. The (approximately) optimal convex combination of $\hat{f}_{1,n}(x_0) = \hat{f}_{1,n}(0)$ and $\hat{f}_{2,n}(x_0) = \hat{f}_{2,n}(0)$ is the simple average, see

(II.13) and (II.14). Hence,

$$\bar{\sigma}_n = \frac{1}{2}\big(\bar{\sigma}_{n,1} + \bar{\sigma}_{n,2}\big) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\big(cX_{i\Delta}1_{\{X_{i\Delta}\le 0\}} - cX_{i\Delta}1_{\{X_{i\Delta}>0\}}\big).$$

It follows immediately from the ergodic theorem that all three estimators are consistent for σ (for $\bar{\sigma}_{n,1}$ and $\bar{\sigma}_{n,2}$ we indeed knew this already from above.) Also, all three estimators are asymptotically normal: If g_1 is defined by $g_1(x) = cx1_{\{x\le 0\}} - \sigma_0$, then

$$\sqrt{n}\big(\bar{\sigma}_{n,1} - \sigma_0\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_1(X_{i\Delta}) \to N(0, V_{12})$$

weakly, *cf.* Florens-Zmirou (1989). The variance is given by

$$V_{12} = \mathrm{E}_{\sigma_0} g_1(X_0)^2 + 2\sum_{k=1}^{\infty} \mathrm{E}_{\sigma_0} g_1(X_0) g_1(X_{k\Delta}). \tag{II.25}$$

Simple (but tedious) calculations yield $\mathrm{E}_{\sigma_0} g_1(X_0)^2 = \sigma_0^2(\pi - 1)$ and

$$\begin{aligned}&\mathrm{E}_{\sigma_0} g_1(X_0) g_1(X_{k\Delta})\\ &\quad = \sigma_0^2\Big(\big(1 - e^{2\beta\Delta k}\big)^{3/2} - 1\Big) + c^2 e^{\beta\Delta k}\,\mathrm{E}_{\sigma_0}\big(X_0\Phi(-e^{\beta\Delta k}/\tau_k)\big)\end{aligned}$$

where Φ is the distribution function for the standard normal distribution and $\tau_k^2 = -\sigma_0^2(1 - e^{2\beta\Delta k})/(2\beta)$ is the conditional variance of X_k given X_0. There is no explicit formula for the expectation appearing in the above formula.

By symmetry, $\sqrt{n}\big(\bar{\sigma}_{n,2} - \bar{\sigma}_0\big) \to N(0, V_{12})$ as well. For $\bar{\sigma}_n$, note that $\sqrt{n}\big(\bar{\sigma}_n - \sigma_0\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_{i\Delta})$ where $g(x) = \frac{1}{2}\big(cx1_{\{x\le 0\}} - cx1_{\{x>0\}}\big) = g_1(x) - \frac{1}{2}cx$. Hence $\sqrt{n}\big(\bar{\sigma}_n - \sigma_0\big) \to N(0, V)$, where the variance V is the given by (II.25) with g_1 replaced by g. We can easily express V in terms of V_{12}: it holds that

$$\mathrm{E}_{\sigma_0} g(X_0)^2 = \mathrm{E}_{\sigma_0} g_1(X_0)^2 - \sigma_0^2\pi/2 = \sigma_0^2(\pi/2 - 1)$$

and

$$\mathrm{E}_{\sigma_0} g(X_0) g(X_{k\Delta}) = \mathrm{E}_{\sigma_0} g_1(X_0) g_1(X_{k\Delta}) - e^{\beta\Delta k}\pi\sigma_0^2/2.$$

so it follows that

$$V = V_{12} - \sigma_0^2 \frac{\pi\big(1 + e^{\beta\Delta}\big)}{2\big(1 - e^{\beta\Delta}\big)}.$$

Hence, the asymptotic variance of $\bar{\sigma}_n$ is indeed smaller than the asymptotic variance of $\bar{\sigma}_{n,1}$ and $\bar{\sigma}_{n,2}$.

Note that $V \ge \mathrm{E}_{\sigma_0} g(X_0)^2 = \sigma_0^2(\pi/2 - 1) \approx 0.57\sigma_0^2$. Of course, the above estimators cannot compete with the maximum likelihood estimator

$$\check{\sigma}_n = \left\{ -\frac{2\beta}{n\big(1 - e^{2\beta\Delta}\big)} \sum_{i=1}^{n} \big(X_{i\Delta} - e^{\beta\Delta} X_{(i-1)\Delta}\big)^2 \right\}^{1/2}$$

(II.23)

which satisfies $\sqrt{n}(\check{\sigma}_n - \sigma_0) \to N(0, \sigma_0^2/2)$. However, as argued above we would expect the estimator $\hat{\sigma}_n$ based on the supremum distance U_n be more precise than $\bar{\sigma}_n$ so the above comparison is not quite fair to the estimation approach discussed in this paper.

II.7.2 The CKLS model

In this section we study the model given by the stochastic differential equation

$$dX_t = (\alpha + \beta X_t)\,dt + \sigma X_t^\gamma\, dW_t. \tag{II.26}$$

We use the method from this paper on simulated data from the model and compare with from various other methods.

In the econometric literature the model is often called the CKLS-model after the paper by Chan et al. (1992) where the model was first discussed in this generality. It includes important and much favoured models as special cases: the geometric Brownian motion (or Black-Scholes model) for $(\alpha, \gamma) = (0, 1)$; the Ornstein-Uhlenbeck process (or Vasicek model) for $\gamma = 0$; and the square root process (or Cox-Ingersoll-Ross model) for $\gamma = 1/2$.

Let $\xi = (\alpha, \beta)$ vary in $\Xi = (0, \infty) \times (-\infty, 0)$ and let $\theta = (\gamma, \sigma)$. If $\gamma < 1/2$, then Assumption II.1.1 is not satisfied since $\int_{x_0}^r s(x, \xi, \theta)dx < +\infty$ and the process may hit zero. If $\gamma > 1$, then $f(+\infty, \xi, \theta) \neq 0$ and the locale martingale part of X is not a genuine martingale. Hence, to be able to estimate α and β by least squares and to use $\hat{f}_n$ we must assume $1/2 \leq \gamma \leq 1$. Note that $f(0, \xi, \theta) = 0$ even if $\gamma > 1$ so we could use $\hat{f}_{1,n}$ for estimation of γ and σ in that case (if estimates of ξ are available). The expression for the invariant density is different for $\gamma = 1/2$ and $\gamma = 1$, than for $1/2 \leq \gamma \leq 1$, so for simplicity we let $\theta = (\gamma, \sigma)$ vary in $\Theta = (1/2, 1) \times (0, \infty)$ only.

For $(\xi, \theta) \in \Xi \times \Theta$ the process is positive and stationary and has $f(0, \xi, \theta) = f(+\infty, \xi, \theta) = 0$; the invariant density is proportional to

$$\frac{1}{\sigma^2 x^{2\gamma}} \exp\left(\frac{2\alpha}{\sigma^2(1-2\gamma)} x^{1-2\gamma} + \frac{\beta}{\sigma^2(1-\gamma)} x^{2-2\gamma} \right), \quad x > 0; \tag{II.27}$$

and the function f is given by

$$K_0(\xi, \theta) \exp\left(\frac{2\alpha}{\sigma^2(1-2\gamma)} x^{1-2\gamma} + \frac{\beta}{\sigma^2(1-\gamma)} x^{2-2\gamma} \right), \quad x > 0.$$

There is no explicit expression for the normalizing constant, $K_0(\xi, \theta)$, but we can calculate it numerically (at least when γ is not very close to 1/2 and 1).

Estimation strategies

Recall that $\xi = (\alpha, \beta)$ is the drift parameter and $\theta = (\gamma, \sigma)$ the diffusion parameter, and let ξ_0 and θ_0 denote the true values. In the simulation study below we consider three situations: (A) α, β and σ are known so that only γ need to be estimated;

(B) σ is known and α, β and γ must be estimated; (C) all four parameters are unknown and must be estimated. The first two situations are of course unrealistic but they provide insight to the estimation problem.

As for the method discussed in this paper, the strategy is as follows. In case (A) γ is estimated as described in Section II.3, *i.e.* by minimizing

$$\sup_{x\in I}\left|\hat{f}_n(x) - f(x,\xi_0,\theta)\right|. \tag{II.28}$$

with respect to γ for $\sigma = \sigma_0$ known. In cases (B) and (C) the drift parameters are estimated by conditional least squares: For $(\xi,\theta) \in \Xi\times\Theta$ the martingale part of X is a genuine martingale; hence the conditional expectation one step ahead

$$\varphi(x,\xi) = \varphi(x,\xi,\theta) = \mathrm{E}_\theta\left(X_\Delta | X_0 = x\right) = e^{\beta\Delta}\Big(x+\frac{\alpha}{\beta}\Big) - \frac{\alpha}{\beta} \tag{II.29}$$

does not depend on θ. The drift parameters α and β are estimated by minimization of $\sum_{i=1}^n \left(X_{i\Delta} - \varphi(X_{i\Delta},\xi)\right)^2$, that is, by solving the two (martingale) estimating equations obtained by differentiation. The outcoming estimators are consistent and asymptotically normal (but note that the estimating functions could be improved if γ and σ were known (Bibby & Sørensen 1995)). Next, the diffusion parameter is estimated as described in Section II.6, that is, by substituting the estimator of ξ for ξ_0 and minimizing (II.28) with respect to γ in case (B) and (γ,σ) in case (C).

In Section II.3 we also briefly discussed the distance measure S_n^w given by (II.17). We will use it below for comparison. In practice the weights do not seem to make much difference so we have used $w_i \equiv 1$. We also compare with a few simple standard methods, namely generalized method of moments (GMM), IID estimation and simple estimating functions. The methods will be described shortly.

Honoré (1997) uses "simulated maximum likelihood estimation" on treasury bill yield data and simulated CKLS data with good results. The method is developed by Pedersen (1995*b*) and is based on approximations of the likelihood function calculated by simulation. Poulsen (1999) obtains estimators in the CKLS model via numerical solutions of the Fokker-Planck equation. Both methods are computationally rather demanding and they will not be used in this study.

GMM based on simple discretizations. This is the method used by Chan et al. (1992). It is based on simple approximations of the conditional moments of $X_{i\Delta} - X_{(i-1)\Delta}$ given $X_{(i-1)\Delta}$, namely $(\alpha+\beta X_{(i-1)\Delta})\Delta$ as approximation to the mean and $\Delta\sigma^2 X_{(i-1)\Delta}^{2\gamma}$ as approximation to the variance. These approximations are good when Δ is "small" but can be bad when Δ is "large", leading to considerable bias of the estimator.

To be specific, define $\varepsilon_i = X_{i\Delta} - X_{(i-1)\Delta} - (\alpha+\beta X_{(i-1)\Delta})\Delta$ and

$$G_n(\xi,\theta) = \sum_{i=2}^n \left(\varepsilon_i, \varepsilon_i X_{(i-1)\Delta}, \varepsilon_i^2 - \Delta\sigma^2 X_{(i-1)\Delta}^{2\gamma}, \varepsilon_i^2 X_{(i-1)\Delta} - \Delta\sigma^2 X_{(i-1)\Delta}^{1+2\gamma}\right)^T$$

(II.25)

and minimize $G_n^T(\theta,\xi)W(\theta,\xi)G_n(\theta,\xi)$ where $W(\theta,\xi)$ for all (ξ,θ) is a positive definite weight matrix. In cases (A) and (B) we of course plug in the known parameter values. Hence, G_n has larger dimension than the unknown parameter, and the estimators depend on the choice of $W(\xi,\theta)$. It is reasonable to use (an estimator of) the particular $W(\xi,\theta)$ that gives the least asymptotic variance for the estimator, see Chan et al. (1992). In case (C) the estimator simply solves $G_n(\xi,\theta)=0$ and does not depend on the weight matrix.

We know the true conditional expectation from (II.29) so alternatively we could use $\tilde{\varepsilon}_i = X_{i\Delta} - e^{\beta\Delta}\big(X_{(i-1)\Delta}+\alpha/\beta\big)+\alpha/\beta$. In case (B) and (C) it does not change the estimation of γ and σ, though, since we get the same estimators of the conditional expectations. Also, in case (A) the difference between the two corresponding γ-estimates is very small.

IID estimation. If the observations were independent, identically $\mu_{\xi,\theta}$-distributed, then the log-likelihood function would be

$$l_n(\xi,\theta)=\sum_{i=1}^{n}\log\mu(X_{i\Delta},\xi,\theta)$$

which we would maximize in order to estimate (ξ,θ). The observations are *not* independent but the estimators so obtained are nevertheless consistent and asymptotically normal (but not efficient), see Kessler (2000). Since $f=\sigma^2\mu_{\xi,\theta}$ we would expect the IID estimators and the estimators obtained by minimizing U_n (and S_n) to be highly correlated.

Note that we cannot distinguish two parameter vectors $(\xi,\theta)=(\alpha,\beta,\gamma,\sigma)$ and $(\tilde{\xi},\tilde{\theta})=(k\alpha,k\beta,\gamma,k^{1/2}\sigma)$ for $k>0$ since $\mu_{\xi,\theta}=\mu_{\tilde{\xi},\tilde{\theta}}$. Hence, we cannot use IID estimation in case (C). However, we could estimate the drift parameters by least squares as above and next use the IID approach for estimation of the diffusion parameters. We will do this in both case (B) and (C).

Simple estimating functions based on the generator. Hansen & Scheinkman (1995) and Kessler (2000) discuss estimating functions of the form

$$\begin{aligned}H_n(\xi,\theta)&=\sum_{i=1}^{n}\mathscr{A}_{\xi,\theta}h(X_{i\Delta},\xi,\theta)\\&=\sum_{i=1}^{n}(\alpha+\beta X_{i\Delta})h'(X_{i\Delta},\xi,\theta)+\frac{1}{2}\sigma^2(X_{i\Delta})X_{i\Delta}^{2\gamma}h''(X_{i\Delta},\xi,\theta).\end{aligned}$$

Here, $h'=\partial h/\partial x$ and $h''=\partial^2 h/\partial x^2$ are derivatives of $h:(0,\infty)\times\Xi\times\Theta\to\mathbb{R}$ with respect to the state variable, and $\mathscr{A}_{\xi,\theta}$ is the differential operator associated with the infinitesimal generator for the diffusion process. If h and $\mathscr{A}_{\xi,\theta}h$ are in $L^1(\mu_{\xi,\theta})$ and if $\mathrm{E}_{\xi,\theta}\big(h'(X_0,\xi,\theta)X^\gamma\big)^2<\infty$, then H_n is an unbiased estimating function, *i.e.* $\mathrm{E}_{\xi,\theta}H_n(\xi,\theta)=0$.

In case (A) we use $h(x) = x^2$. In case (B) we use $h(x) = (x, x^2, x^3)$ and define the corresponding estimating H_n coordinate-wise. The estimating functions are easy to solve. In case (C) we cannot use this approach since simple estimating functions can only be used to identify parameters from the invariant distribution.

Results of the simulation study

Now, let us turn to the details and the results of the simulation study. It is based on 100 realizations of the model (II.26) with parameters

$$\alpha_0 = 0.04; \quad \beta_0 = -0.6; \quad \gamma_0 = 0.75; \quad \sigma_0 = 0.2.$$

The simulated paths are constructed by means of the Euler scheme with time step 1/1000. Each realization consists of $n = 500$ observations and $\Delta = 1$. One of the simulated paths is shown in Figure II.4.

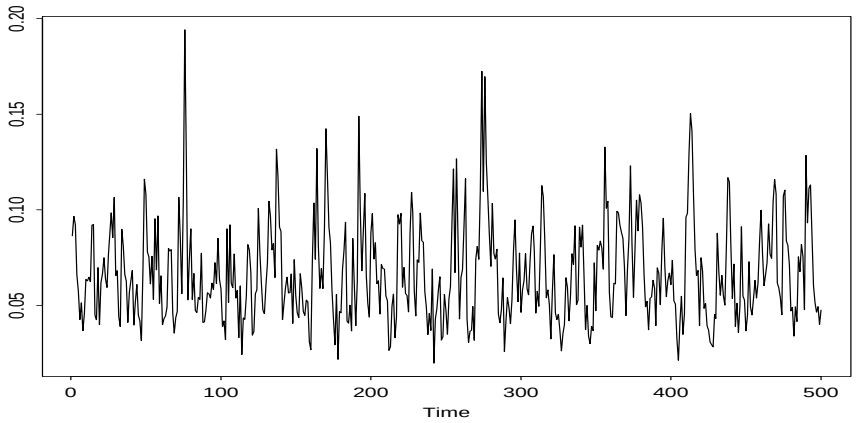

Figure II.4: A typical simulation of the CKLS model with $(\alpha, \beta, \gamma, \sigma) = (0.04, -0.60, 0.75, 0.20)$. There are $n = 500$ observations and the value of Δ is 1.

The means and standard errors of the estimators are listed in Table II.1. The first five lines are for case (A), the next six for case (B) and the last four for case (C). As explained below, not all the optimization problems are well-behaved and for some methods an optimum in the parameter space did not exist for all simulations. The number of failures is given in the notes to the table.

In *case (A)* the GMM estimator is biased but the other three estimators seem to be unbiased. The estimators based on f have slightly larger standard errors than the IID estimator, and the estimator obtained from the simple estimating function based on $h(x) = x^2$ has standard error about four times as big as the IID estimator. Figure II.5 shows $\hat{f}_n$ and $f(\cdot, \alpha_0, \beta_0, \gamma, \sigma_0)$ with γ equal to the true value (0.75) and γ equal to the estimators obtained by minimizing U_n (0.737) and S_n (0.730) for the simulated data from Figure II.4.

In case (B) and (C) the least squares estimators of α and β are by far the best — unbiased with small standard errors.

Method	$\hat{\alpha}_n$ (0.04)		$\hat{\beta}_n$ (-0.60)		$\hat{\gamma}_n$ (0.75)		$\hat{\sigma}_n$ (0.20)	
	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
$\min U_n$	—	—	—	—	0.7550	0.0167	—	—
$\min S_n$	—	—	—	—	0.7548	0.0150	—	—
IID	—	—	—	—	0.7505	0.0129	—	—
Simple	—	—	—	—	0.7597	0.0561	—	—
GMM	—	—	—	—	0.8431	0.0176	—	—
LS-min U_n	0.0411	0.0050	-0.6166	0.0785	0.7487	0.0186	—	—
LS-min S_n	0.0411	0.0050	-0.6166	0.0785	0.7482	0.0188	—	—
IID$^{(1)}$	0.0478	0.0257	-0.7192	0.3944	0.7442	0.0881	—	—
LS-IID	0.0411	0.0050	-0.6166	0.0785	0.7481	0.0187	—	—
Simple	0.0646	0.0486	-0.9734	0.7490	0.7068	0.1237	—	—
GMM	0.0278	0.0024	-0.4198	0.0377	0.8503	0.0192	—	—
LS-min $U_n$$^{(2)}$	0.0411	0.0050	-0.6166	0.0785	0.7386	0.0958	0.2009	0.0531
LS-min $S_n$$^{(3)}$	0.0411	0.0050	-0.6166	0.0785	0.7286	0.0962	0.1958	0.0514
LS-IID$^{(4)}$	0.0411	0.0050	-0.6166	0.0785	0.7467	0.0800	0.2039	0.0439
GMM$^{(5)}$	0.0306	0.0027	-0.4586	0.0422	0.5076	0.1328	0.0862	0.0352

Table II.1: Empirical means and standard errors of various estimators for 100 realizations of the CKLS model. The true parameters are given in the top line, $n = 500$, and $\Delta = 1$. A "—" means that the corresponding parameter is considered known. Notes: (1) 1 failure; (2) 6 failures (3) 3 failures; (4) 7 failures; (5) 49 estimates less than 1/2.

In *case (B)* the γ-estimates obtained from LS-IID estimation (*i.e.* maximization of l_n with α and β equal to the least squares estimates), and minimization of U_n and S_n (also with α and β equal to the least squares estimates) are equally good. The pure IID estimator ignores the dependence among observations and has standard error more than four times larger than the LS-IID estimator. The estimator obtained from simple estimating functions has an even larger standard error and the GMM estimator is biased.

In *case (C)* the LS-IID estimators for γ and σ seem to be a little better than those obtained from U_n and S_n. The GMM estimator is still biased but note that the mean of the estimator is now *smaller* than the true value. Half the γ-estimates are less than 1/2 which is in fact outside the parameter space!

The estimation results are very similar whether we minimize U_n or S_n. The empirical correlation between the two γ-estimates is 0.90 in case (A), 0.93 in case (B) and 0.83 in case (C) and the empirical correlation between the two σ-estimates in case (C) is 0.82. As one would expect the standard errors of the γ-estimates are smallest in case (A) and largest in case (C). For the estimators obtained by minimizing U_n and S_n the standard error is five times larger in case (C) compared to case (B), for example.

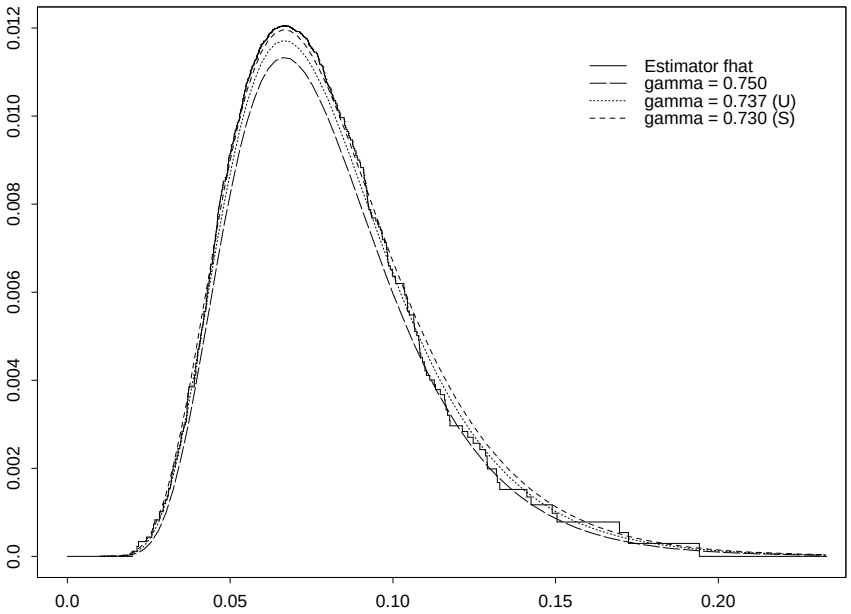

Figure II.5: Graphs of $\hat{f}_n$ and $f(\cdot, \alpha_0, \beta_0, \gamma, \sigma_0)$ for $\gamma = \gamma_0 = 0.75$ (the true value), $\gamma = 0.737$ (minimizing U_n) and $\gamma = 0.733$ (minimizing S_n). The data are those from Figure II.4.

Identification problems in case (C)

The distribution of X only depends on (γ, σ) through the values of the diffusion function σx^γ. Figure II.6 shows the graph of this function for three different values of (γ, σ). The solid line corresponds to the true value (0.75,0.20) and the two dotted lines to (0.65,0.15) and (0.85,0.26) respectively; the values of σ are chosen so all three curves intersect at $x = -\alpha_0/\beta_0 = 0.0667$. The range of x is from 0 to 0.20 which is about typical for the simulated paths. The graphs are close in the central area of the invariant distribution so it is difficult to distinguish between different values of (γ, σ) as long as the values of $\sigma(-\alpha/\beta)^\gamma$ are close.

This explains why the standard error of $\hat{\gamma}_n$ is *much* larger in case (C) compared to case (B) and implies that the estimators of γ and σ are highly correlated (for the estimators obtained by minimizing U_n the empirical correlation is 0.97, for example). It also explains why the level of the GMM-estimates of γ changes from case (B) to case (C); the average estimated value of the diffusion function evaluated at $-\alpha_0/\beta_0$ are almost the same (0.0200 versus 0.0204) in the two cases.

Of course, the identification problem also gave rise to some practical problems. Figure II.7 shows a contour plot for U_n (for the data from Figure II.4). The level curves are very oblong corresponding to a valley of local minima and the minimization routine had difficulties finding the global minimum. We solved the problem as follows: The simple estimating function corresponding to $h(x) = x^{2-2\gamma}$

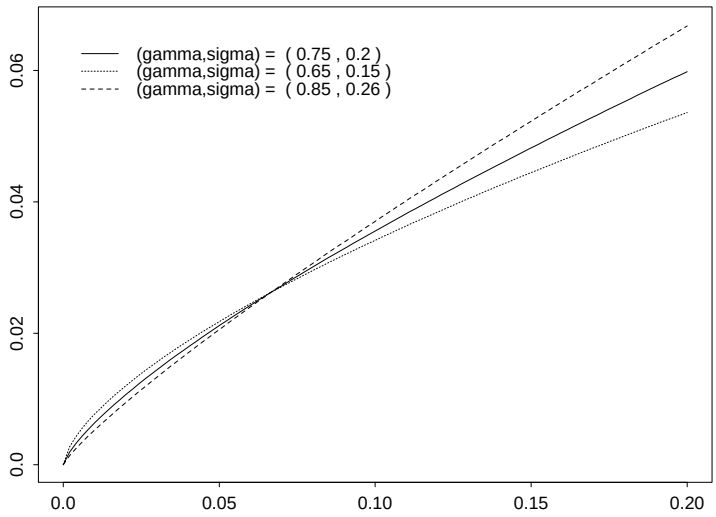

Figure II.6: The diffusion function $x \to \sigma x^\gamma$ for three different values of (γ, σ). The three graphs intersect at $x = -\alpha_0/\beta_0 = 0.0667$.

is

$$H_n(\xi, \theta) = \alpha S_1(\gamma) + \beta S_2(\gamma) + \frac{1}{2} n\sigma^2(1-2\gamma)$$

where $S_1(\gamma) = \sum_{i=1}^n X_{i\Delta}^{1-2\gamma}$ and $S_2(\gamma) = \sum_{i=1}^n X_{i\Delta}^{2-2\gamma}$. Solving the equation $H_n(\hat{\xi}_n, \theta) = 0$ where $\hat{\xi}_n$ is the estimator of the drift parameter, gives us σ as a function of γ,

$$\sigma = \sigma(\gamma) = \left(-\frac{2\hat{\alpha} S_2(\gamma) + 2\hat{\beta} S_3(\gamma)}{n(1-2\gamma)} \right)^{1/2} \tag{II.30}$$

The curve $(\gamma, \sigma(\gamma))$ is superimposed on Figure II.7. It is almost parallel to the level curves and runs relatively close to the global minimum point (denoted by a circle). Nevertheless, the minimum point on the curve (denoted by a triangle) is relatively far from the the global minimum point.

We use the curve as an indication of which area is relevant to search for the minimum. We calculate the values of U_n in a fine grid around the curve and finally use the minimum point on the grid as initial values in a numerical procedure. We use the same technique when the criterion function is S_n or l_n (and a similar technique for IID estimation in case (B)).

Considerations on asymptotics

In case (A) where α and β are known, $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$ converges weakly if the assumptions of Theorem II.9 hold. There is no a priori reason to believe that the limit distribution is Gaussian, but then what does the distribution of $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$ look like?

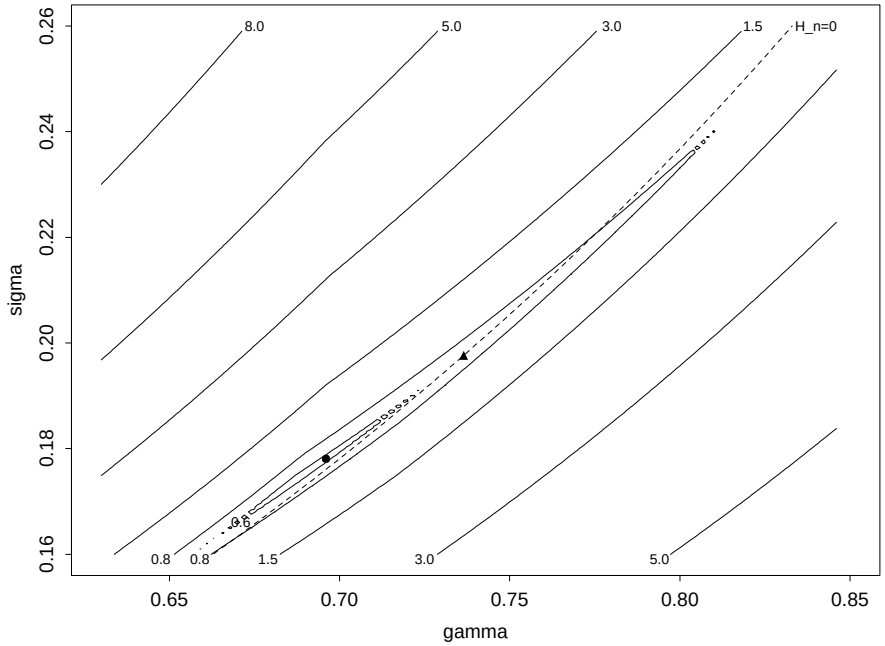

Figure II.7: Contour plot for U_n and the data from Figure II.4 together with the curve given by $H_n(\hat{\alpha},\hat{\beta},\gamma,\sigma) = 0$, that is the curve $(\gamma,\sigma(\gamma))$ given by (II.30). The triangle denotes the minimum point (0.737,0.198) on the curve whereas the circle denotes the global minimum point (0.696,0.178).

The left hand side of Figure II.8 shows a QQ-plot for $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$ with the quantiles of the standard normal distribution on the x-axis and the empirical quantiles of $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$ on the y-axis. The normal distribution fits well in the central area of the distribution but also not too badly in the tails. The right hand side is a QQ-plot of $\sqrt{n}(\tilde{\gamma}_n - \gamma_0)$ where $\tilde{\gamma}_n$ is the minimizer of S_n. We did not show any asymptotic results for $\tilde{\gamma}_n$. The QQ-plot indicates that the distribution of $\sqrt{n}(\tilde{\gamma}_n - \gamma_0)$ has slightly heavier tails than the normal distribution.

Of course, 100 simulations are far too few to judge about the distribution of the estimators. Also, one could ask how large n should be before the distribution of $\sqrt{n}(\hat{\gamma}_n - \hat{\gamma}_0)$ is close to its limit. For a further investigation we have simulated 1000 paths of the process up to time 1000 (with the same values of Δ and the parameters as before). For known values of α, β and σ we have calculated $\hat{\gamma}_n$ for the first 250 observations, the first 500 observations and all 1000 observations respectively.[1]

Table II.2 shows empirical means and standard errors of the "raw" estimates and the standardized estimates. For the standardized estimates the mean decreases as n grows but the variance is quite stable. Figure II.9 shows QQ-plots of

[1]For about 2% of the simulations U_n did not have a minimum in $(1/2,1)$ when we used the first 250 observations only. To simplify computations we did not use use these simulations at all — neither for 500 or 1000 observations. Instead, we drew new simulated paths until we had 1000 paths for which $\hat{\gamma}_{250}$, $\hat{\gamma}_{500}$ and $\hat{\gamma}_{1000}$ all existed.

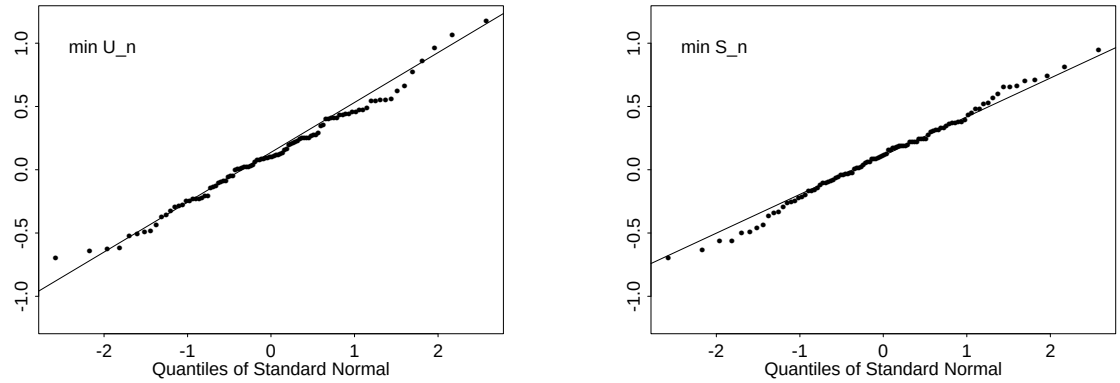

Figure II.8: QQ-plots for normalized estimators of γ from case (A). The plot to the left is for $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$ and the plot to the right is for $\sqrt{n}(\tilde{\gamma}_n - \gamma_0)$. The QQ-plots have the quantiles of the standard normal distribution on the x-axis and the quantiles of the variable under consideration on the y-axis.

n	$\hat{\gamma}_n$		$\sqrt{n}(\hat{\gamma}_n - \gamma_0)$	
	mean	s.e.	mean	s.e.
250	0.7579	0.0251	0.1256	0.3975
500	0.7549	0.0179	0.1086	0.4000
1000	0.7523	0.0129	0.0716	0.4071

Table II.2: Empirical means, variances and standard errors of $\hat{\gamma}_n$ and $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$ in case (A) for 1000 simulated paths and three different values of n, the number of observations. The true value of γ is 0.75.

the standardized estimators. The normal distribution fits rather well for $n = 1000$ (and $n = 250$ if a single very small estimate is ignored). For $n = 500$ the distribution is somewhat further from the normal distribution. We conclude that although we could not show that the limit distribution is Gaussian, a Gaussian approximation would presumably be satisfactory for practical purposes.

II.8 Concluding remarks

In this paper we have discussed a method for estimation of parameters in the diffusion function. It provides consistent and in some cases also weakly convergent estimators. The usual limit theory does not apply; instead we used empirical process theory for proving the asymptotic results. The drift parameters must be estimated before the new technique is employed. This is possible using martingale estimating functions if the drift is linear but can otherwise be difficult. We applied the method to simulated data from the difficult CKLS model and obtained satisfactory (though presumably not efficient) estimators. From a theoretical point of view the application of empirical process theory is perhaps most interesting.

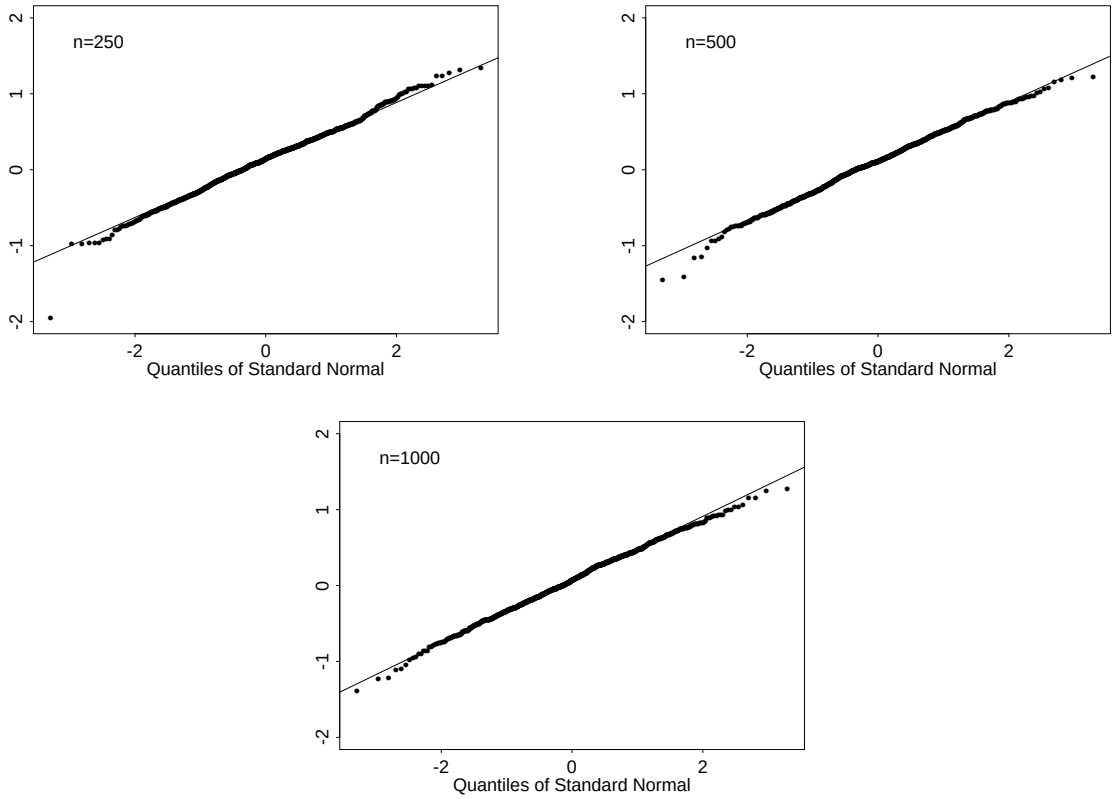

Figure II.9: QQ-plots for $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$ for 1000 simulated paths and three different values of n. The values of α, β and σ are considered as known.

II.A Appendix: On empirical process theory

The asymptotic results in this paper (Sections II.4 and II.5) are proved using empirical process theory which — in short — provides uniform versions of the classical limit theorems. The first part of this appendix is a cursory review of the theory that we use and its statistical applications for so-called M-estimation. It can be read independently from the rest of the paper. It is by no means a complete overview of the theory of empirical processes. No proofs are included either and we refer to the textbook by van der Vaart & Wellner (1996) for precise definitions and further details. The textbook by Pollard (1984) is an excellent reference as well. None of the results are new. However, we do not know of any applications of empirical process theory for statistics on diffusion processes. The second part is concerned with the application in this paper. In particular we show that certain classes of function are so-called Vapnik-Červonenkis subgraph classes.

II.A.1 General Theory

In this section we give a hasty overview of some main results from the theory of empirical processes and discuss briefly an application to M-estimators.

Glivenko-Cantelli and Donsker classes

First of all, we point out what is meant by uniform limit theorems. Let $Z, Z_1, Z_2, \ldots$ be independent, identically distributed random variables defined on some probability space $(E, \mathscr{E}, Pr)$ with values in $\mathscr{Z}$ (equipped with some σ-algebra) and common distribution $P = Z(Pr)$. Furthermore, let $f : \mathscr{Z} \to \mathbb{R}$ be measurable. Then, according to the classical law of large numbers, the average of $f(Z_1), \ldots, f(Z_n)$ converges, that is

$$\frac{1}{n}\sum_{i=1}^{n} f(Z_i) \to \mathrm{E}\, f(Z) \tag{II.31}$$

almost surely and in L^1 provided that $f \in L^1(P)$. The classical central limit theorem asserts that the centered and scaled sum converges weakly to a normal distribution, that is

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(f(Z_i) - \mathrm{E}\, f(Z)\big) \to N\Big(0, \mathrm{E}\big(f(Z) - \mathrm{E}\, f(Z)\big)^2\Big) \tag{II.32}$$

weakly provided that $f \in L^2(P)$ with $\mathrm{E}\, f^2(Z) > 0$.

The corresponding uniform theorems claim that (II.31) and (II.32) hold uniformly for f varying in suitably small classes $\mathscr{F}$. To be specific, the uniform version of the law of large numbers states that

$$\sup_{f\in\mathscr{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\big(f(Z_i) - \mathrm{E}\, f(Z)\big)\Big| \to 0 \tag{II.33}$$

almost surely, and $\mathscr{F}$ is called a *Glivenko-Cantelli class* if (II.33) holds.

To define a uniform central limit theorem, assume that $\sup_{f\in\mathscr{F}}\big|f(z) - \mathrm{E}\, f(Z)\big| < \infty$ for all $z \in \mathscr{Z}$ and let $l^\infty(\mathscr{F})$ be the set of functionals $G : \mathscr{F} \to \mathbb{R}$ for which $\sup_{f\in\mathscr{F}}|G(f)| < \infty$. Equip $l^\infty(F)$ with the uniform topology. Then, for each n, the functional $G_n : \mathscr{F} \to \mathbb{R}$ defined by

$$f \to G_n(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(f(Z_i) - \mathrm{E}\, f(Z)\big) \tag{II.34}$$

is an element of $l^\infty(\mathscr{F})$ and one can ask whether G_n converges weakly to a Gaussian limit $G \in l^\infty(\mathscr{F})$. The process G_n is called the *empirical process indexed by* $\mathscr{F}$ and $\mathscr{F}$ is called a *Donsker class* if G_n converges weakly in $l^\infty(\mathscr{F})$.

Covering numbers, entropy and VC classes

Now, when is a certain class a Glivenko-Cantelli or Donsker class? Informally, one has to measure the size of the class and decide whether it is small enough for the convergence results hold for all $f \in \mathscr{F}$ simultaneously.

Covering and entropy numbers are very important in this context. For a given norm $||\cdot||$ on $\mathscr{F}$ and $\varepsilon > 0$ the *covering number* $N(\varepsilon, \mathscr{F}, ||\cdot||)$ is defined as the

minimal number of $||\cdot||$-balls of radius ε needed to cover $\mathscr{F}$. The logarithm of the covering number is called the *entropy*. Obviously, if $\mathscr{G} \subseteq \mathscr{F}$ (and the same norm is used on both $\mathscr{F}$ and $\mathscr{G}$) then $N(\varepsilon, \mathscr{G}, ||\cdot||) \leq N(\varepsilon, \mathscr{F}, ||\cdot||)$. Hence, the covering number (or entropy number) makes sense as a measure of size of a given class.

Note that $N(\varepsilon, \mathscr{F}, ||\cdot||)$ is one for large ε if $\sup_{f,g\in\mathscr{F}} ||f-g|| < \infty$. Also, $N(\varepsilon, \mathscr{F}, ||\cdot||)$ increases as ε decreases and the crucial point is how fast it increases for small ε. In fact, $\mathscr{F}$ is a Glivenko-Cantelli class if certain measurability conditions are satisfied, if the envelope function $F = \sup_{f\in\mathscr{F}} |f|$ is in $L^1(P)$ and if

$$\sup_Q N(\varepsilon ||F||_{Q,1}, \mathscr{F}, ||\cdot||_{Q,1}) < \infty \tag{II.35}$$

for all $\varepsilon > 0$. Here $||\cdot||_{Q,r}$ denotes the $L^r(Q)$-norm, $||f||_{Q,r}^r = \int |f|^r dQ$, and the supremum is taken over all probability measures Q on $\mathscr{X}$ with $0 < ||F||_{Q,1} = \mathrm{E}_Q F < \infty$. This result is a corollary to Theorem 2.4.3 in van der Vaart & Wellner (1996).

Furthermore, $\mathscr{F}$ is a Donsker class if some further measurability conditions are met, if $F \in L^2(P)$ and if

$$\int_0^\infty \sup_Q \sqrt{\log N\big(\varepsilon ||F||_{Q,2}, \mathscr{F}, ||\cdot||_{Q,2}\big)}\, d\varepsilon < \infty \tag{II.36}$$

where the supremum is taken over all probability measures Q with $0 < ||F||_{Q,2}^2 = \mathrm{E}_Q F^2 < \infty$ (van der Vaart & Wellner 1996, Theorem 2.5.2). Note that convergence at $+\infty$ is automatic since $N\big(\varepsilon ||F||_{Q,2}, \mathscr{F}, L_2(Q)\big) = 1$ for $\varepsilon > 2$. Indeed, let $f, g \in \mathscr{F}$ be arbitrary. Then

$$||f-g||_{Q,2}^2 = \int |f-g|^2 dQ \leq \int 4|F|^2 dQ = 4||F||_{Q,2}^2$$

so for $\varepsilon > 2$, it holds that $||f-g||_{Q,2} \leq 2||F||_{Q,2} < \varepsilon ||F||_{Q,2}$ and only one ball of radius $\varepsilon ||F||_{Q,2}$ is needed to cover $\mathscr{F}$.

The above entropy conditions are automatically met for so-called Vapnik-Červonenkis subgraph classes (VC subgraph classes) of functions — a terminology that we will now introduce and later use for our application.

Definition II.10 Let $\mathscr{C}$ be a collection of subsets of a set $\mathscr{Y}$. Then $\mathscr{C}$ is said to *shatter* a finite subset $\{y_1, \ldots, y_n\}$ of $\mathscr{Y}$ if each of its 2^n subsets has the form $C \cap \{y_1, \ldots, y_n\}$ for some $C \in \mathscr{C}$, and $\mathscr{C}$ is a *VC class* (of sets) if there is a n_0 such that no subset of $\mathscr{Y}$ of size n_0 is shattered by $\mathscr{C}$. (Then the same holds for all $n \geq n_0$.) The least n_0 with this property is called the *VC index* of $\mathscr{C}$. A collection $\mathscr{F}$ of functions $f : \mathscr{X} \to \mathbb{R}$ is a *VC subgraph class* if the subgraphs $\mathscr{G} = \{(y,t) \in \mathscr{X} \times \mathbb{R} : t < f(y)\}$ form a VC class of sets in $\mathscr{X} \times \mathbb{R}$. The VC index of $\mathscr{F}$ is defined as the VC index of $\mathscr{G}$. □

Intuitively, a class of functions cannot separate many points in $\mathscr{X} \times \mathbb{R}$ if the functions are "too much alike" so that the class is small in some sense. For example, a VC subgraph class with VC index 1 consists of one single function. A

non-trivial example arises when $E = \mathbb{R}$ and $\mathscr{F} = \{f_t\}_{t\in\mathbb{R}}$ is the class of left half-lines $f_t(x) = 1_{\{x\le t\}}$. This class has VC index 2.

Formally, VC subgraph classes are useful because the corresponding covering numbers are bounded by a polynomial in $1/\varepsilon$: if $\mathscr{F}$ is a VC subgraph class then for any $r \ge 1$ there exist constants K and a such that for all probability measures Q with $||F||_{Q,r} > 0$,

$$N(\varepsilon||F||_{Q,r}, \mathscr{F}, ||\cdot||_{Q,r}) \le K\Big(\frac{1}{\varepsilon}\Big)^a.$$

This follows from Theorem 2.6.7 in van der Vaart & Wellner (1996). In particular, both (II.35) and (II.36) are satisfied so a VC subgraph class of functions that meets certain measurability conditions is a Glivenko-Cantelli class if its envelope function $F \in L^1(P)$ and a Donsker class as well if $F \in L^2(P)$.

Extension to stationary processes

All the above was for independent, identically distributed random variables. In our application we need stronger theorems since our observations originate from a stochastic differential equation and are thus not independent. Fortunately, the convergence results can be extended to cover the case of strictly stationary and sufficiently strong mixing random variables, just as for the classical theory.

Let now $\tilde{Z} = (Z_0, Z_1, \dots)$ be a strictly stationary sequence defined on $(E, \mathscr{E}, Pr)$ with invariant distribution P and β-mixing coefficients β_k, defined in the usual way

$$\beta_k = \frac{1}{2}\sup\sum_{i=1}^{I}\sum_{j=1}^{J}|Pr(A_i \cap B_j) - Pr(A_i)Pr(B_j)|.$$

The supremum is taken over $l \ge 0$ and all pairs of partitions $\{A_1, \cdots, A_I\}$ and $\{B_1, \dots, B_J\}$ of $\mathscr{E}$ such that all A_i are in the σ-algebra generated by $Z_0, \dots, Z_l$ and all B_j are in the σ-algebra generated by $Z_{l+k}, Z_{l+k+1}, \dots$. Note that if $\tilde{Z}$ is strictly stationary and Markov (which is the case in our application), then the supremum is attained for $l = 0$ (Bradley 1986, Theorem 4.1) and it also holds that

$$\beta_k = \int \sup_A \big|p_{k\Delta,\theta_0}(x, A) - \mu_{\theta_0}(A)\big|\, d\mu_{\theta_0}(x)$$

where $p_{k\Delta,\theta_0}$ is the transition probability from time 0 to time $k\Delta$ (Doukhan 1994, Chapter 2.4). Arcones & Yu (1994) prove that G_n — still defined by (II.34) — converges in $l^\infty(\mathscr{F})$ if $\mathscr{F}$ is a VC subgraph class and there is a $p > 2$ such that $F \in L^p(P)$ and

$$k^{p/(p-2)}(\log k)^{2(p-1)/(p-2)}\beta_k \to 0 \quad \text{as} \quad k \to \infty.$$

Note that while for independent observations F should only be square integrable, F should in the stationary case be in $L^p(P)$ for some p strictly larger than 2. Arcones & Yu (1994) also state a result in terms of covering numbers which we will not repeat here.

Application to M-estimators

Following van der Vaart & Wellner (1996, Section 3.2), let us turn to a statistical application of the above theory. Consider the situation where a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ is estimated by minimizing some functional $U_n(\theta)$, that is $\hat{\theta}_n = \operatorname{argmin}_{\theta\in\Theta} U_n(\theta)$. In other words, $\hat{\theta}_n$ is a *M-estimator*. Let P_0 denote the probability corresponding to the true parameter value θ_0.

First, assume that $U_n(\theta) \to U(\theta)$ in P_0-probability, uniformly in θ and that the limit process U is deterministic and has θ_0 as unique minimum point. Then *if* the argmin-functional is continuous at U, the convergence in probability of $\hat{\theta}_n$ to θ_0, that is consistency of $\hat{\theta}_n$, follows immediately. In fact, the argmin-functional *is* continuous at functions U for which the unique minimum is well-separated in the sense that $\inf_{\theta\notin\Theta_0} U(\theta) > U(\theta_0)$ for all neighbourhoods Θ_0 of θ_0. Hence, this is what we shall assume about U.

Next, assume that we have somehow established the rate of convergence and consider the "local parameters" $\theta_0 + h/r_n$ and the "localized criterion function"

$$h \to M_n(h) = U_n(\theta_0 + h/r_n) - U_n(\theta_0)$$

instead of θ and U_n themselves. Again, if $M_n \to M$ weakly with respect to P_0 in $l^\infty(\mathbb{R}^d)$ and M P_0-almost surely has a well-separated (now stochastic) minimum, $\hat{h}$, then $\hat{h}_n = r_n\big(\hat{\theta}_n - \theta_0\big)$ converges weakly to $\hat{h}$. (Of course, for a set T, $l^\infty(T)$ is the set of bounded, real functions defined on T.)

Convergence of M_n on all of $l^\infty(\mathbb{R}^d)$ is often not satisfied but fortunately less can do: if $\hat{h}_n$ is tight, then it suffices that M_n converges in $l^\infty(H)$ for all compact subsets $H \subseteq \mathbb{R}^d$ (van der Vaart & Wellner 1996, Theorem 3.2.2).

II.A.2 The application in this paper

For the applications in this paper it is easy to see that the criterion functions ($U_{n,1}$, $U_{n,2}$ and U_n) converge uniformly in θ to a deterministic limit $U(\theta)$. Hence, the corresponding estimators are consistent (under the assumption that θ_0 is well-separated as a minimum of U), see Theorem II.4.

For the convergence results in Section II.5 the hardest part is to obtain weak convergence of the centered and scaled criterion functions ($M_{n,1}$, $M_{n,2}$ and M_n). With this result, $\sqrt{n}$-consistency and weak convergence of the estimator (properly centered and scaled) follows relatively easily. The lemmas below claim that the relevant classes of functions are in fact VC classes so weak convergence can be obtained via theorems from Arcones & Yu (1994). For the precise application of the VC-property, we refer to the proof of Proposition II.6.

Lemma II.11 *The sets $\mathscr{F} = \{F_x\}_{x\in I}$ and $\tilde{\mathscr{F}} = \{\tilde{F}_x\}_{x\in I}$ where $F_x(y) = b(y)1_{\{y\le x\}}$ and $\tilde{F}_x(y) = -b(y)1_{\{y>x\}}$ are VC subgraph classes of functions with VC index 2.*

Proof Consider $\mathscr{F}$ first. By definition, we must show that the class $\mathscr{G} = \{G_x\}_{x\in I}$ of subgraphs G_x defined by

$$G_x = \{(s,t) \in I \times \mathbb{R} : t < F_x(s)\}.$$

is a VC class with index 2 on $I \times \mathbb{R}$.

We show that no subset $\{(s_1,t_1),(s_2,t_2)\}$ of $I \times \mathbb{R}$ with two elements is shattered by $\mathscr{G}$. If $\{(s_1,t_1)\}$ is picked out by $\mathscr{G}$ then $x_1 \in I$ exists such that

$$G_{x_1} \cap \{(s_1,t_1),(s_2,t_2),(s_3,t_3)\} = \{(s_1,t_1)\}.$$

This implies that $(s_1,t_1) \in G_{x_1}$ and $(s_2,t_2) \notin G_{x_1}$ (since $(s_2,t_2) \in G_{x_1}$ would imply $\{(s_2,t_2)\} = G_{x_1} \cap \{(s_2,t_2)\} \subseteq G_{x_1} \cap \{(s_1,t_1),(s_2,t_2)\} = \{(s_1,t_1)\}$). By definition of G_{x_1},

$$\begin{aligned} t_1 &< F_{x_1}(s_1) = b(s_1)1_{\{s_1 \le x_1\}} \\ t_2 &\ge F_{x_1}(s_2) = b(s_2)1_{\{s_2 \le x_1\}}. \end{aligned}$$

Similarly, if $\{(s_2,t_2)\}$ is picked out then

$$\begin{aligned} t_1 &\ge F_{x_2}(s_1) = b(s_1)1_{\{s_1 \le x_2\}} \\ t_2 &< F_{x_2}(s_2) = b(s_2)1_{\{s_2 \le x_2\}}. \end{aligned}$$

for an $x_2 \in I$. Hence, if both $\{(s_1,t_1)\}$ and $\{(s_2,t_2)\}$ are picked out then

$$b(s_1)1_{\{s_1 \le x_2\}} \le t_1 < b(s_1)1_{\{s_1 \le x_1\}}$$

implying that either $s_1 \le x_2$, $s_1 > x_1$ and $b(s_1) < 0$ or $s_1 > x_2$, $s_1 \le x_1$ and $b(s_1) > 0$. Similarly, either $s_2 \le x_1$, $s_2 > x_2$ and $b(s_2) < 0$ or $s_2 > x_1$, $s_2 \le x_2$ and $b(s_2) > 0$.

If $b(s_1)$ and $b(s_2)$ are both positive then $x_2 < s_1 \le x_1$ and $x_1 < s_2 \le x_2$ which cannot both hold. Similarly if $b(s_1)$ and $b(s_2)$ are both negative. We conclude that one of the values $b(s_1)$ and $b(s_2)$ is positive and the other negative. If $b(s_2) < 0 < b(s_1)$ then the empty set cannot be picked out: if $x_0 \in I$ exists with $t_i \ge b(s_i)1_{\{s_i \le x_0\}}$ for $i = 1,2$ then $s_2 \le x_0 < s_1$ in contradiction to the assumption that $s_1 \le s_2$. See Figure II.10 for illustration. Similarly, the two-point set $\{(s_1,t_1),(s_2,t_2)\}$ cannot be picked out if $b(s_1) < 0 < b(s_2)$.

It follows that $\mathscr{G}$ does not shatter $\{(s_1,t_1),(s_2,t_2)\}$ implying that $\mathscr{G}$ is VC with index 2 (since obviously the index is larger than 1).

For $\tilde{F}$, one can either use similar arguments or note that $\tilde{F}_x = -b + F_x$ so the subgraph for $\tilde{F}_x$ is given by

$$\tilde{G}_x = \{(s,t) : t < \tilde{F}_x(s)\} = \{(s,t) : t + b(s) < F_x(s)\}.$$

Consequently, a subset $\{(s_1,t_1),(s_2,t_2)\}$ is shattered by $\tilde{\mathscr{F}}$ if and only if $\{(s_1,t_1 + b(s_1)),(s_2,t_2 + b(s_2))\}$ is shattered by $\mathscr{F}$. The latter is not possible, *cf.* the proof above. □

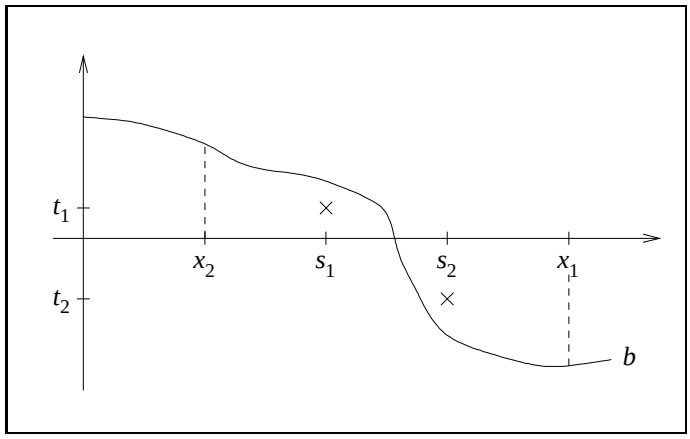

Figure II.10: The singletons $\{(s_1,t_1)\}$ and $\{(s_2,t_2)\}$ are picked out (by x_1 and x_2 respectively) but the empty set is not picked out.

In the proof above we chose to show directly that $\mathscr{F}$ and $\tilde{\mathscr{F}}$ are VC subgraph classes. One could also use lemmas 2.6.16 and 2.6.18 from van der Vaart & Wellner (1996): the indicator functions

$$H_x(y) = 1_{\{y \le x\}} = 1_{(-\infty,0]}(y-x)$$

form a VC subgraph class of functions (Lemma 2.6.16), $F_x = bH_x$ and $\tilde{F}_x = -b + bH_x$; now use Lemma 2.6.18.

In the proof of Proposition II.8 we consider functions F_x defined by $F_x(y) = 2b(y)\big(\lambda_1(x) - 1_{\{y>x\}}\big)$. By Lemma 2.6.18 from van der Vaart & Wellner (1996), $\mathscr{F}$ is a VC subgraph class if $\mathscr{H} = \{H_x\}_{x\in I}$ where

$$H_x(y) = \lambda_1(x) - 1_{\{y>x\}}, \quad y \in I$$

is a VC subgraph class. See Figure II.11 for graphs of H_x for various x's.

To our best knowledge it is not trivial that $\mathscr{H}$ is a VC subgraph class. In fact, we found it easier to give a direct proof that the covering numbers are bounded by a polynomial in $1/\varepsilon$ than proving the VC property. We refer to the proof of Proposition II.8 for the argument. For completeness, however, we now prove that $\mathscr{H}$ is a VC subgraph class with index 3. Following the approach from above, we show that no subset with three elements is shattered by the class of subgraphs $\mathscr{G} = \{G_x\}_{x\in I}$ where $G_x = \{(s,t) \in I \times \mathbb{R} : t < H_x(s)\}$. The proof is somewhat more tiresome than the one above, though not difficult, since we must take the empty set as well as all subsets with one and two elements into account. In fact, it is possible to find three-point sets for which all subsets with one and two elements are picked out.

Recall that $0 \le \lambda_1(x) \le 1$ and that λ_1 is non-increasing. This will be used frequently in the following.

Lemma II.12 *The set $\mathscr{H} = \{H_x\}_{x\in I}$ where $H_x(y) = \lambda_1(x) - 1_{\{y>x\}}$ is a VC subgraph class of functions with index 3. Consequently $\mathscr{F} = \{F_x\}_{x\in I}$ with $F_x(y) = 2b(y)\big(\lambda_1(x) - 1_{\{y>x\}}\big)$ is a VC subgraph class.*

Proof First, choose $s_1, s_2, x_0, x_1, x_2, x_{12} \in I$ and $t_1, t_2 \in \mathbb{R}$ such that

$$x_0 < s_1 < x_1 < s_2 < x_{12} < x_2$$
$$\lambda_1(x_0) - 1 < t_2 < \lambda_1(x_2) < t_1 < \lambda_1(x_{12}) < \lambda_1(x_1) < \lambda_1(x_0).$$

By inspection (or by Figure II.11), it follows that $(s_1, t_1) \in G_{x_0}^c \cap G_{x_1} \cap G_{x_2}^c \cap G_{x_{12}}$ and $(s_2, t_2) \in G_{x_0}^c \cap G_{x_1}^c \cap G_{x_2} \cap G_{12}$ so that the set $\{(s_1, t_1), (s_2, t_2)\}$ is shattered by $\mathscr{G}$. Hence the VC index of $\mathscr{H}$ is at least 3.

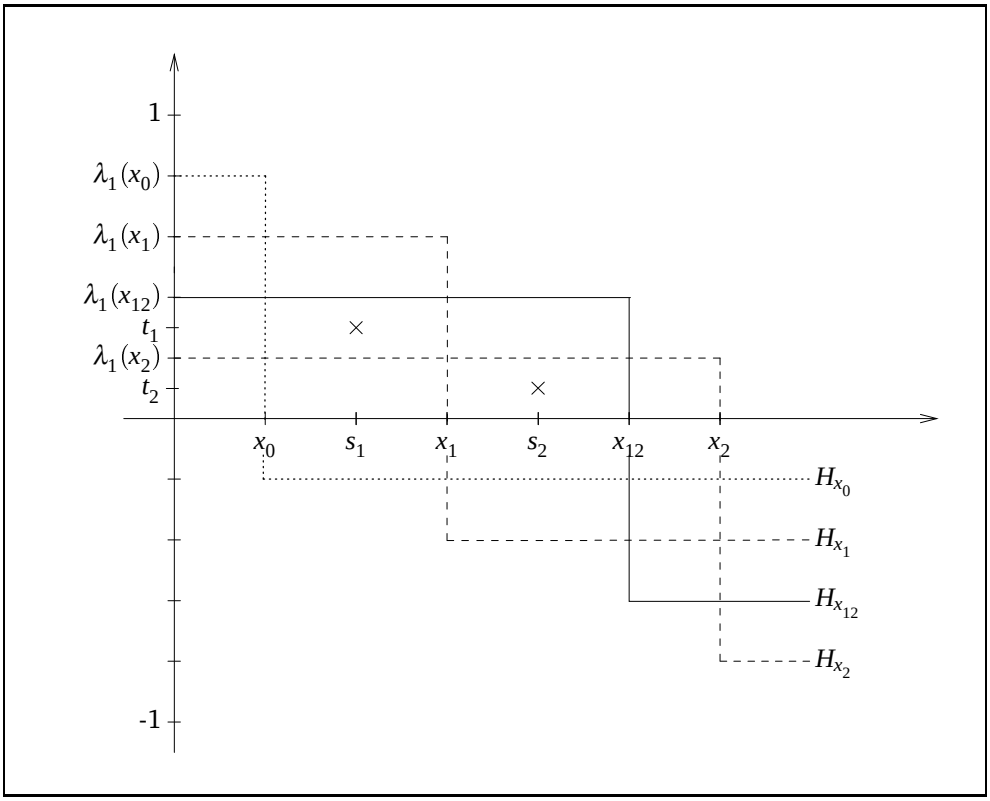

Figure II.11: Graphs of H_x for various values of x. The two-point set $\{(s_1, t_1), (s_2, t_2)\}$ is shattered by $\mathscr{G}$ so the VC index is at least 3.

Next, let $\{(s_1, t_1), (s_2, t_2), (s_3, t_3)\}$ with $s_1 \leq s_2 \leq s_3$ be a subset of $I \times \mathbb{R}$ that is shattered by $\mathscr{G}$. In particular the singletons $\{(s_1, t_1)\}$ and $\{(s_2, t_2)\}$ are picked out, so $x_1, x_2 \in I$ exist such that

$$t_1 < \lambda_1(x_1) - 1_{\{s_1 > x_1\}}; \qquad t_2 \geq \lambda_1(x_1) - 1_{\{s_2 > x_1\}}$$
$$t_1 \geq \lambda_1(x_2) - 1_{\{s_1 > x_2\}}; \qquad t_2 < \lambda_1(x_2) - 1_{\{s_2 > x_2\}}$$

It follows that

$$\lambda_1(x_2) - 1_{\{s_1 > x_2\}} \leq t_1 < \lambda_1(x_1) - 1_{\{s_1 > x_1\}} \tag{II.37}$$
$$\lambda_1(x_1) - 1_{\{s_2 > x_1\}} \leq t_2 < \lambda_1(x_2) - 1_{\{s_2 > x_2\}} \tag{II.38}$$

Hence, if $s_2 \leq x_1$ then $\lambda_1(x_1) < \lambda_1(x_2) - 1_{\{s_2 > x_2\}}$ implying $s_2 \leq x_2$ since λ_1 is non-negative. Then (II.37) and (II.38) yields $\lambda_1(x_2) < \lambda_1(x_1)$ and $\lambda_1(x_1) < \lambda_1(x_2)$ respectively. Similarly if $s_1 > x_1$ so we conclude that $s_1 \leq x_1 < s_2$.

By symmetry we obtain

$$s_1 \leq x_1 < s_2 \leq x_2 < s_3$$

if all three singletons $\{(s_1,t_1)\}$, $\{(s_2,t_2)\}$ and $\{(s_3,t_3)\}$ are picked out. Hence,

$$\begin{array}{lll} t_1 < \lambda_1(x_1); & t_2 \geq \lambda_1(x_1) - 1; & t_3 \geq \lambda_1(x_1) - 1 \\ t_1 \geq \lambda_1(x_2); & t_2 < \lambda_1(x_2); & t_3 \geq \lambda_1(x_2) - 1 \end{array}$$

and

$$t_1 \geq \lambda_1(x_3) - 1_{\{s_1 > x_3\}}; \quad t_2 \geq \lambda_1(x_3) - 1_{\{s_2 > x_3\}}; \quad t_3 < \lambda_1(x_3) - 1_{\{s_3 > x_3\}}.$$

If $s_1 \leq x_3$ then $\lambda_1(x_3) \leq t_1 < \lambda_1(x_1)$ and if $x_3 < s_3$ then $\lambda_1(x_1) - 1 \leq t_3 < \lambda_1(x_3) - 1$. Hence, we cannot have $s_1 \leq x_3 < s_3$. There are thus two possibilities; either

$$x_3 \geq s_3; \qquad t_1 \geq \lambda_1(x_3); \qquad t_2 \geq \lambda_1(x_3); \qquad t_3 < \lambda_1(x_3) \tag{II.39}$$

or

$$x_3 < s_1; \qquad t_1 \geq \lambda_1(x_3) - 1; \qquad t_2 \geq \lambda_1(x_3) - 1; \qquad t_3 < \lambda_1(x_3) - 1. \tag{II.40}$$

First, assume that (II.39) holds. Then

$$\begin{gathered} s_1 \leq x_1 < s_2 \leq x_2 < s_3 \leq x_3 \\ \lambda_1(x_1) - 1 \leq t_3 < \lambda_1(x_3) \leq t_2 < \lambda_1(x_2) \leq t_1 < \lambda_1(x_1). \end{gathered}$$

Also assume that any subsets of $\{(s_1,t_1),(s_2,t_2),(s_3,t_3)\}$ with two elements is picked out. Hence $x_{12},x_{13},x_{23} \in I$ exist such that

$$t_1 < \lambda_1(x_{12}) - 1_{\{s_1 > x_{12}\}}; \; t_2 < \lambda_1(x_{12}) - 1_{\{s_2 > x_{12}\}}; \; t_3 \geq \lambda_1(x_{12}) - 1_{\{s_3 > x_{12}\}}$$

and similarly for x_{13} and x_{23}.

Then necessarily $s_2 \leq x_{12} < x_2$, $s_1 \leq x_{13} < x_1$ and $s_3 \leq x_{23} < x_3$. This follows because, with short notation:

$$\begin{gathered} x_{12} < s_2 \Rightarrow t_2 < \lambda_1(x_{12}) - 1 \leq 0; \\ x_{12} \geq x_2 \Rightarrow \lambda_1(x_{12}) > t_1 > \lambda_1(x_2) \Rightarrow x_{12} < x_2; \\ x_{13} < s_1 \Rightarrow t_1 < \lambda_1(x_{13}) - 1 < 0; \\ x_{13} \geq x_1 \Rightarrow \lambda_1(x_{13}) > t_1 > \lambda_1(x_2) \Rightarrow x_{13} < x_2 < s_3 \\ \Rightarrow \lambda_1(x_{13}) - 1 > t_3 > \lambda_1(x_1) - 1 \Rightarrow x_{13} < x_1; \\ x_{23} \geq x_3 \Rightarrow \lambda_1(x_{23}) > t_2 > \lambda_1(x_3) \Rightarrow x_{23} < x_3; \\ x_{23} < s_3 \Rightarrow \lambda_1(x_{23}) - 1 > t_3 > \lambda_1(x_1) - 1 \\ \Rightarrow x_{23} < x_1 < s_2 \Rightarrow t_2 < \lambda_1(x_{23}) - 1 < 0 \end{gathered} \tag{II.41}$$

where the right hand sides of the implications are all contradictory to the assumptions. Hence, we have established that

$$\begin{aligned} &s_1 \le x_{13} < x_1 < s_2 \le x_{12} < x_2 < s_3 \le x_{23} < x_3; \\ &\lambda_1(x_1) - 1 \le t_3 < \lambda_1(x_{13}) - 1 \le \lambda_1(x_3) \le t_2 \\ &\qquad < \lambda_1(x_{23}) \le \lambda_1(x_2) \le t_1 < \lambda_1(x_{12}) \le \lambda_1(x_1), \end{aligned}$$

see Figure II.12 for illustration.

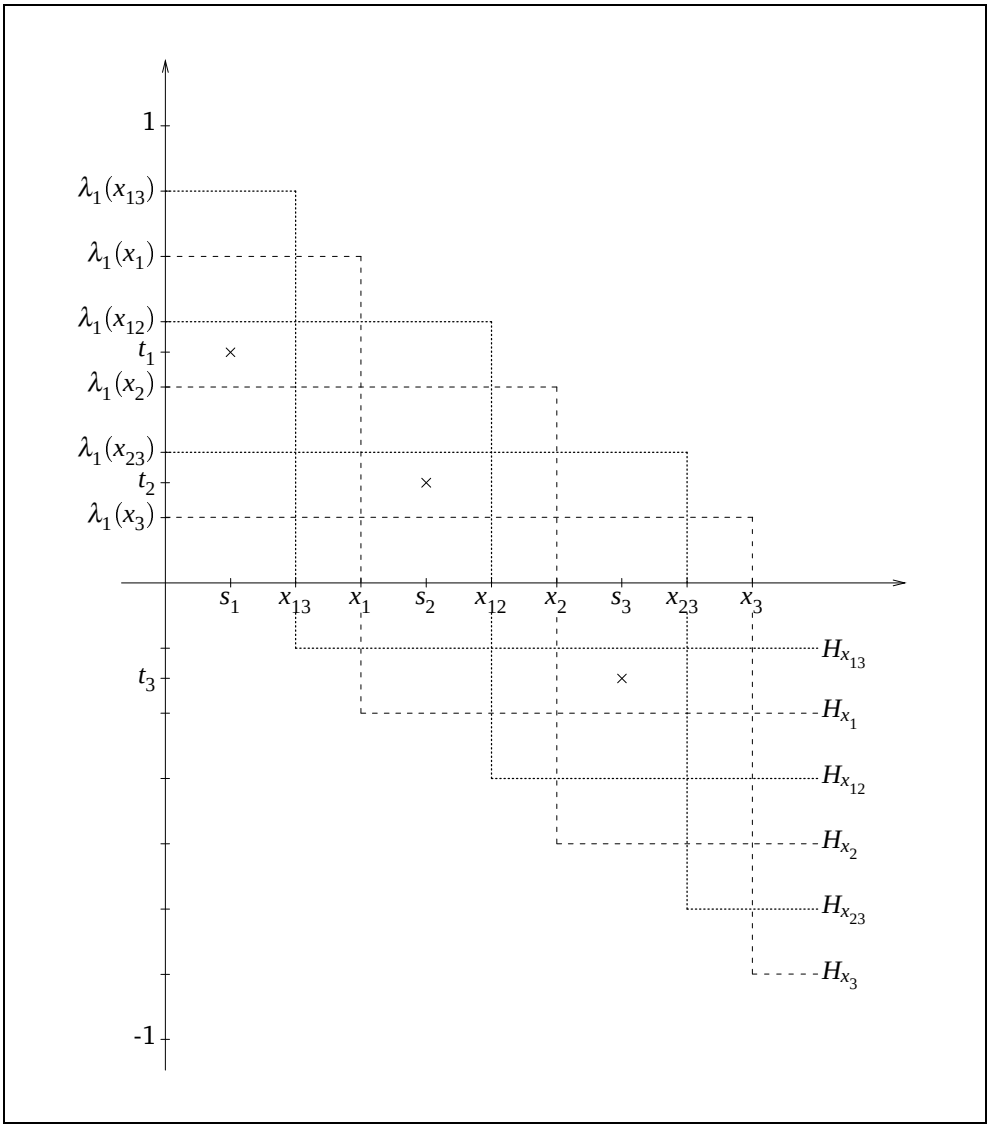

Figure II.12: Graphs of H_x for various x's. All one-point subsets and two-point subsets of $\{(s_1,t_1),(s_2,t_2),(s_3,t_3)\}$ are picked out, but the empty set cannot be picked out.

Assume furthermore that the empty set is picked out, that is, $x_0 \in I$ exists such that $t_i \ge \lambda_1(x_0) - 1_{\{s_i > x_0\}}$ for $i = 1,2,3$. Since $x_0 \ge s_3$ implies $\lambda_1(x_0) \le t_3 < 0$, we must

have $x_0 < s_3$ and thus $\lambda_1(x_0) - 1 \leq t_3 < \lambda_1(x_{13}) - 1$ implying $x_0 > x_{13} \geq s_1$. Hence, $s_1 < x_0 < s_3$. But this cannot hold since $x_0 < s_3 \leq x_{23}$ implies $\lambda_1(x_0) > \lambda_1(x_{23}) > t_2$ and $s_1 < x_0$ implies $\lambda_1(x_0) \leq t_1 < \lambda_1(x_{12})$ and hence $x_0 > x_{12} \geq s_2$ and $t_2 \geq \lambda_1(x_0)$. We conclude that the empty set cannot be picked out if (II.39) holds.

Next, assume that (II.40) holds and that all three two-point subsets are picked out. Then, by arguments as above,

$$x_3 < s_1 \leq x_{13} < x_1 < s_2 \leq x_{12} < x_2 < s_3$$

but both $x_{23} < x_3$ and $x_{23} \leq s_3$ are possible. In both cases the assumption that the empty set can be picked out, leads to a contradiction as above.

We conclude that no three-point set $\{(s_1,t_1),(s_2,t_2),(s_3,t_3)\}$ is shattered by the subgraphs and hence that $\mathscr{H}$ is a VC class of index 3. □

II.B Appendix: A mixing result for the OU-process

The proposition below claims that the Ornstein-Uhlenbeck process has β-mixing coefficients that decrease at an exponential rate.

Proposition II.13 *There exist constants $c_1, c_2 > 0$ such that the β-mixing coefficients β_k for the Ornstein-Uhlenbeck process satisfy $\beta_k \leq c_1 e^{-c_2 k\Delta}$.*

Proof Recall that

$$\beta_k = \int \sup_A \left| p_{k\Delta,\sigma_0}(x_0, A) - \mu_{\sigma_0}(A) \right| d\mu_{\sigma_0}(x_0),$$

where $p_{k\Delta,\sigma_0}$ is the transition probability from time 0 to time $k\Delta$. Consequently, if $p_{k\Delta,\sigma_0}(x_0,\cdot)$ has density $\pi_k(x_0,\cdot)$ then

$$\beta_k \leq \int_{\mathbb{R}} \int_{\mathbb{R}} \left| \pi_k(x_0,x) - \pi_0(x)) \right| \pi_0(x)\, dx\, dx_0$$

where $\pi_0 = \mu(\cdot,\sigma_0)$ is short for the true invariant density.

Let $\tau^2 = -\sigma_0^2/2\beta$ and $\tau_k^2 = \tau^2(1 - e^{2\beta k\Delta}) < \tau^2$ and let $\xi_k = e^{\beta\Delta k} x_0$ for $x_0 \in \mathbb{R}$ given. Then π_0 is the density for $N(0,\tau^2)$ and $\pi_k(x_0,\cdot)$ is the density for $N(\xi_k,\tau_k^2)$. If furthermore $\tilde{\pi}_k$ is the density for $N(0,\tau_k^2)$, then

$$\begin{aligned} &\int_{\mathbb{R}} \left| \pi_k(x_0,x) - \pi_0(x)) \right| dx \\ &\qquad \leq \int_{\mathbb{R}} \left| \pi_k(x_0,x) - \tilde{\pi}_k(x)) \right| dx + \int_{\mathbb{R}} \left| \tilde{\pi}_k(x) - \pi_0(x)) \right| dx \end{aligned} \tag{II.41}$$

for all $x_0 \in \mathbb{R}$. The integrals are L^1-distances between densities for normal distributions with same variance or same mean. The integrals are represented by the shaded areas in Figures II.13 and II.14.

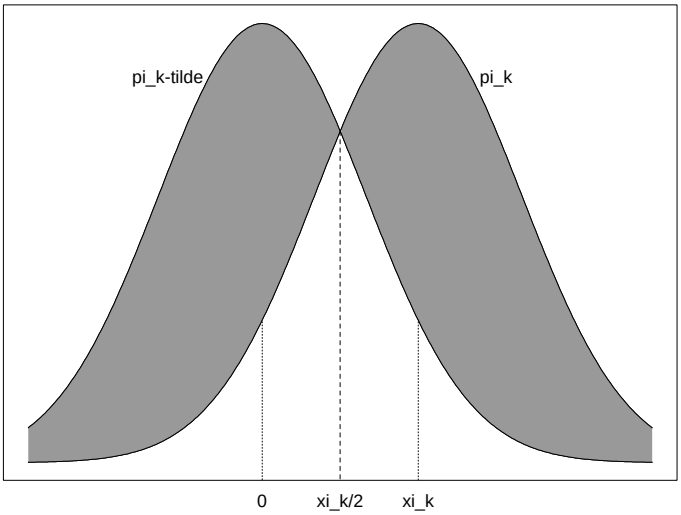

Figure II.13: Densities of two normal distributions with same variance but different expectations. The size of the shaded areas corresponds to the first integral in (II.41).

For the first integral, let $x_0 > 0$ be arbitrary and let U be a standard Gaussian random variable. Then

$$\begin{aligned}\int_{\mathbb{R}} \left|\pi_k(x_0,x) - \tilde{\pi}_k(x)\right| dx &= 2\int_{\xi_k/2}^{\infty} \left(\pi_k(x_0,x) - \tilde{\pi}_k(x)\right) dx \\ &= 2P\Big(U > -\frac{\xi_k}{2\tau_k}\Big) - 2P\Big(U > \frac{\xi_k}{2\tau_k}\Big) \\ &\leq K_1 \frac{\xi_k}{\tau_k}\end{aligned}$$

where K_1 is a constant that does not depend on k and x_0. Similarly, it holds that $\int_{\mathbb{R}} \left|\pi_k(x_0,x) - \tilde{\pi}_k(x)\right| dx \leq -K_1\xi_k/\tau_k$ if $x_0 < 0$, so for all $x_0 \in \mathbb{R}$ (recall that $\pi_k(0,\cdot) = \tilde{\pi}_k$),

$$\int_{\mathbb{R}} \left|\pi_k(x_0,x) - \tilde{\pi}_k(x)\right| dx \leq K_1 \frac{\xi_k}{\tau_k} = K_1 \frac{e^{\beta\Delta k}|x_0|}{\tau\left(1-e^{2\beta\Delta k}\right)^{1/2}}.$$

For a new constant $K_2 = K_1\tau^{-1}\,\mathrm{E}^{\mu}_{\sigma_0}|X_0| = K_1\sqrt{2/\pi}$, it follows that

$$\int_{\mathbb{R}}\int_{\mathbb{R}} \left|\pi_k(x_0,x) - \tilde{\pi}_k(x)\right| \pi_0(x_0)\, dx\, dx_0 \leq \frac{K_2 e^{\beta\Delta k}}{\left(1-e^{2\beta\Delta k}\right)^{1/2}}$$

which decreases at an exponential rate as k increases.

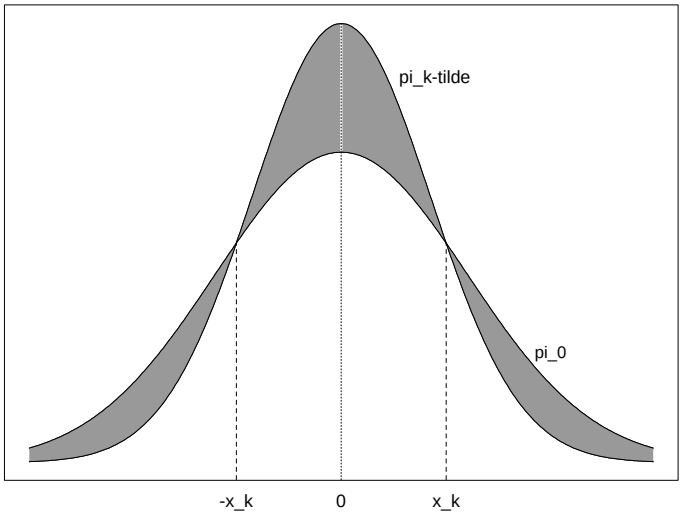

Figure II.14: Densities of two normal distributions with same expectations but different variances. The size of the shaded areas corresponds to the second integral in (II.41).

For the second integral, note that it does not depend on x_0 and let x_k be the positive point of intersection between $\tilde{\pi}_k$ and π_0, see Figure II.14. Then

$$\begin{aligned}\int_{\mathbb{R}} \left|\tilde{\pi}_k(x) - \pi_0(x))\right| dx \\ &= 2\int_0^{x_k} \left(\tilde{\pi}_k(x) - \pi_0(x)\right) dx + 2\int_{x_k}^{\infty} \left(\pi_0(x) - \tilde{\pi}_k(x)\right) dx \\ &= 4\int_{x_k}^{\infty} \pi_0(x)\, dx - 4\int_{x_k}^{\infty} \tilde{\pi}_k(x)\, dx.\end{aligned}$$

With U as above we thus get

$$\begin{aligned}\int_{\mathbb{R}} \left|\tilde{\pi}_k(x) - \pi_0(x))\right| dx &= 4P\Big(U > \frac{x_k}{\tau}\Big) - 4P\Big(U > \frac{x_k}{\tau_k}\Big) \\ &\le K_1\Big(\frac{x_k}{\tau_k} - \frac{x_k}{\tau}\Big) \\ &\le K_1 \frac{1 - \left(1 - e^{2\beta\Delta k}\right)^{1/2}}{\tau\left(1 - e^{2\beta\Delta k}\right)^{1/2}} x_k\end{aligned}$$

which tends to zero at exponential rate if x_k is bounded. By solving the equation $\tilde{\pi}_k(z) = \pi_0(z)$ one finds that

$$x_k^2 = -\tau^2 \frac{1 - e^{2\beta\Delta k}}{e^{2\beta\Delta k}} \log\left(1 - e^{2\beta\Delta k}\right) \le -\tau^2 \frac{\log\left(1 - e^{2\beta\Delta k}\right)}{e^{2\beta\Delta k}}$$

which tends to 1 as $k \to \infty$. If follows that $\int_{\mathbb{R}} \int_{\mathbb{R}} \left|\tilde{\pi}_k(x) - \pi_0(x)\right| \pi_0(x_0)\, dx\, dx_0$ and hence β_k tends to zero at an exponential rate as $k \to \infty$. □

(II.45)

Acknowledgements Thanks to Søren Feodor Nielsen who provided ideas for the asymptotic results, and to my advisor Martin Jacobsen.

III

Simulated Likelihood Approximations for Stochastic Volatility Models

Abstract

The objective of this paper is approximate maximum likelihood estimation for stochastic volatility models. We consider a two-dimensional diffusion process (X,V) where V is ergodic while X has drift and diffusion coefficient completely determined by V. The distribution of V — and thereby also the distribution of X — depends on an unknown parameter θ, and our concern is estimation of θ from discrete-time observations of X. The volatility process V remains unobserved. We consider approximate maximum likelihood estimation. For the k'th order approximation we pretend that the observations form a k'th order Markov chain, find the corresponding approximate log-likelihood function, and maximize it with respect to θ. There is no explicit expression for the approximate log-likelihood function, but it can be calculated by simulation. For each k the method yields consistent and asymptotically normal estimators. Simulations of the model where V is a Cox-Ingersoll-Ross model are used for illustration.

Key words

Approximate maximum likelihood estimation; asymptotic normality; consistency; Cox-Ingersoll-Ross process; discretely observed diffusion processes; stochastic volatility models.

Publication details

A shorter version of this paper (without Appendix III.A.2, with fewer details on the Cox-Ingersoll-Ross model and the simulation study and with fewer proofs included) will be submitted for publication shortly.

III.1 Introduction

We are concerned with inference for continuous-time stochastic volatility models. By stochastic volatility models we will mean models for a pair of processes (X,V) where V is a latent, positive diffusion process and the observable process X solves a stochastic differential equation with diffusion term $\sqrt{V}$ and drift determined by V as well. The process V is called the *volatility process*. We consider parametric specifications of the drift and the diffusion function for V, and the objective is approximate maximum likelihood estimation based on *discrete-time* observations of X.

For a start, consider the classical Black-Scholes model (or geometric Brownian motion)

$$dP_t = \alpha P_t\,dt + \tau P_t\,dW_t \tag{III.1}$$

which is (or rather was) often used to model stock prices. The classical option pricing formula was derived for this model (Black & Scholes 1973). If P solves (III.1) then $\log P$ has constant volatility (squared diffusion) and independent, normally distributed increments. It is well-known that these properties are inconsistent with empirical findings: studies have revealed that stock returns (and other financial data) often are dependent, have strongly leptokurtic marginal distributions and exhibit signs of randomly varying variance over time.

In the discrete-time setting ARCH-type models and discrete-time stochastic volatility models have been used for modeling such phenomena. See Shephard (1996) for an overview of both model types. However, for derivative pricing (and related problems) it may be advantageous to use continuous-time models, retaining the Black-Scholes machinery at our disposal. Also, irregularly sampled data are easier to handle for continuous-time models than for discrete-time models.

Of course one could generate the above features by simply allowing for non-linear drift and diffusion functions for the price process. In the stochastic volatility framework, however, the linear structure of the equation for P is retained, but an additional source of variability is introduced: the constant τ in (III.1) is replaced by the value of a latent diffusion process $\sqrt{V}$. The modified equation for P is thus

$$dP_t = \alpha P_t\,dt + \sqrt{V_t}P_t\,dW_t. \tag{III.2}$$

In this paper we shall consider models given by

$$\begin{aligned} dX_t &= \xi(V_t)\,dt + \sqrt{V_t}\,dW_t \\ dV_t &= b(V_t,\theta)\,dt + \sigma(V_t,\theta)\,d\tilde{W}_t \end{aligned}$$

where V is stationary and ergodic. With $P = e^X$ it follows by Itô's formula that this model is equivalent to (III.2) if $\xi(v) = \alpha - v/2$ (and α is known). Hence, a possible application of the model is for the logarithm of a stock price.

The drift and diffusion for V are parameter dependent, and we shall be interested in estimation of θ from equidistant observations $X_0, X_\Delta, \ldots, X_{n\Delta}$ of X. The

volatility process V remains unobserved. Conditional on V the above model is very simple as increments of X are independent and Gaussian: $Z_i = X_{i\Delta} - X_{(i-1)\Delta} \sim N(M_i, S_i)$ where $M_i = \int_{(i-1)\Delta}^{i\Delta} \xi(V_s)\,ds$ and $S_i = \int_{(i-1)\Delta}^{i\Delta} V_s\,ds$.

For the above model to make sense, we must model V as a positive process. Various models were suggested in the late eighties and early nineties: V was modeled as a geometric Brownian motion (Hull & White 1987), as a Cox-Ingersoll-Ross process (Hull & White 1988, Heston 1993), as the exponential of a Ornstein-Uhlenbeck process (Wiggins 1987, Chesney & Scott 1989) and as a squared Ornstein-Uhlenbeck process (Scott 1987, Stein & Stein 1991).

The above papers all focus on pricing of a European call option written on a stock with price process P. Pricing is investigated for fixed value of the parameter θ in the equation for V, and the majority of the papers pay no or little attention to estimation of θ. Only Scott (1987) and Chesney & Scott (1989) address the problem seriously and derive moment-like estimators for the parameters. More recently, several estimation approaches have been suggested, some of which have earlier been applied for the discrete-time versions of the models; see Shephard (1996) and Ghysels *et al.* (1996) for surveys.

Genon-Catalot *et al.* (1999) consider the approximation that the increments $Z_1, \ldots, Z_n$ are independent and identically distributed with conditional distribution of Z_1 given V equal to $N(\Delta\xi(V_0), \Delta V_0)$. The estimators are consistent as $n \to \infty$ only if the time-step Δ decreases to zero as n increases. For (large) fixed values of Δ the bias may be considerable. Also, only estimation of parameters from the stationary distribution of V is possible. In another paper, Genon-Catalot *et al.* (1998*b*) consider mean-reverting models for V. Then calculation of various moments of the joint distribution of the increments is possible, and estimation is carried out by matching theoretical and empirical moments. For any fixed Δ the estimators so obtained are consistent and asymptotically normal as n increases. However, the simulation study in Section III.7 indicates that there may be serious existence problems in practice. The two above methods require no hard numerical computations or simulations and are thus quick in practice. As opposed to this most other methods (and the one suggested in this paper) are quite computationally intensive.

Nielsen *et al.* (2000) use a filtering approach where values of V are estimated together with the parameter. This requires that n (that is, the number of observations) differential equations are solved by numerical methods. Eraker (1998) use a Bayesian approach which requires Markov Chain Monte Carlo simulation of values of V and X at time-points in between those where X is observed as well as of values of θ; see also Elerian *et al.* (2000). The so-called efficient method of moments (Gallant & Tauchen 1996) is applied to a stochastic volatility model by Andersen & Lund (1997). Finally, Sørensen (1999) studies prediction-based estimating functions. Particular attention is paid to the case where for a function f and an integer k, each term in the estimating function is given in terms of the value $f(Z_i)$ and its projection on some space determined by the previous k increments $Z_{i-k+1}, \ldots, Z_i$. Typically, the projections must be calculated by simulation.

The method suggested in this paper is somewhat related since we also choose a number $k \geq 0$ and base inference on k lags of the increments. For a given value of k the idea is to *pretend that* $(Z_1, Z_2, \dots)$ *is k'th order Markov*, find the corresponding approximate likelihood function, and maximize it with respect to θ. In particular $k = 0$ corresponds to pretending that observations are independent, drawn from the stationary distribution (and may thus be interpreted as an improvement of the method by Genon-Catalot *et al.* (1999) who use an approximation to the stationary density), and $k = 1$ corresponds to pretending that observations are Markov.

There is no explicit expression for the k-lag conditional density, but we can express it in terms of expectations with respect to the distribution of $(V_t)_{0 \leq t \leq (k+1)\Delta}$ and therefore calculate it by simulation of V on the interval from zero to $(k+1)\Delta$. For any fixed Δ and any $k \geq 0$ the approximate score function in unbiased and (under regularity conditions, of course) the estimator is consistent and asymptotically normal as the number of observations increases.

We use the model where $\xi \equiv 0$ and V is a Cox-Ingersoll-Ross process as example and use the method on simulated data. If the parameter in the diffusion function is considered known we obtain satisfactory estimates even for $k = 0$, whereas we for all three parameters unknown must use a larger k, say 4, to get reasonable estimates.

The paper is organized as follows. In Section III.2 we discuss the model and its probabilistic properties. We introduce the likelihood approximations and the estimation method in Section III.3 and discuss the computational aspects in Section III.4. The efficiency of the estimators is briefly discussed in Section III.6. In Section III.7 we discuss the Cox-Ingersoll-Ross model for V in detail, try out the estimation method on simulated data and compare with simple moment estimators. We conclude in Section III.8.

III.2 Model and basic assumptions

In this section we discuss the model and the basic assumptions in detail. Let $(W, \tilde{W}) = \{(W_t, \tilde{W}_t)\}_{t \geq 0}$ be a standard two-dimensional Brownian motion defined on a filtered probability space $(\Omega, \mathscr{F}, \mathscr{F}_t, Pr)$ satisfying the usual conditions and let $U_X, U_V : \Omega \to \mathbb{R}$ be $\mathscr{F}_0$-measurable random variables, mutually independent and independent of $(W, \tilde{W})$. Furthermore, let $(X, V) = \{(X_t, V_t)\}_{t \geq 0}$ be a two-dimensional diffusion process governed by the stochastic differential equations

$$dX_t = \xi(V_t)\,dt + \sqrt{V_t}\,dW_t, \quad X_0 = U_X \tag{III.3}$$

$$dV_t = b(V_t, \theta)\,dt + \sigma(V_t, \theta)\,d\tilde{W}_t, \quad V_0 = U_V. \tag{III.4}$$

Here θ is an unknown p-dimensional parameter from the parameter space $\Theta \subseteq \mathbb{R}^p$ and V is positive Pr-almost surely. The functions $\xi : (0,\infty) \to \mathbb{R}$, $b : (0,\infty) \times \Theta \to \mathbb{R}$ and $\sigma : (0,\infty) \times \Theta \to (0,\infty)$ are known and continuous (for b and σ: with respect to the state variable).

The parameter θ determines the distribution of V and thereby also the distribution of X, and our concern is estimation of θ from equidistant observations $X_0, X_\Delta, \ldots, X_{n\Delta}$ of X. The volatility process V remains unobserved.

We shall assume that ξ, b and σ are such that a solution (X,V) to (III.3)–(III.4) exists for all $\theta \in \Theta$ with V positive, stationary and ergodic. For the latter we need some further notation. Introduce

$$\begin{aligned}\alpha(\mathscr{A},\mathscr{B}) &= \sup\big\{\big|Pr(A\cap B) - Pr(A)Pr(B)\big| : A\in\mathscr{A}, B\in\mathscr{B}\big\} \\ &= \sup\big\{\big|\mathrm{Cov}(U_A,U_B)\big| : \sigma(U_A)\subseteq\mathscr{A}, \sigma(U_B)\subseteq\mathscr{B}, 0\le U_A,U_B\le 1\big\}\end{aligned} \tag{III.5}$$

as a measure of "dependence" between σ-algebras $\mathscr{A}\subseteq\mathscr{F}$ and $\mathscr{B}\subseteq\mathscr{F}$. The inequality $\le$ above is trivial, the other follows because $\big|\mathrm{Cov}(2U_A-1,2U_B-1)\big|$ is at most four times the expression in (III.5), see Doukhan (1994, Lemma 3, page 10). For a stochastic process $Y=\{Y_t\}_{t\in T}$ in discrete time ($T=\mathbb{N}\cup\{0\}$) or continuous time ($T=[0,\infty)$), define the *α-mixing coefficients* by

$$\alpha_Y(t) = \sup_{s\in T}\alpha\Big(\sigma\big(\{Y_u\}_{u\le s}\big), \sigma\big(\{Y_u\}_{u\ge s+t}\big)\Big)$$

and say that Y is α-mixing if $\alpha_Y(t)\to 0$ as $t\to\infty$. It is well-known that α-mixing implies ergodicity (Doukhan 1994, page 21). One can think of the α-mixing coefficients as measures of the temporal dependence in Y. See Doukhan (1994), for example, for the general theory of mixing and Genon-Catalot *et al.* (1998*b*) for an overview of mixing for diffusion processes.

We are now ready to specify the basic assumption.

Assumption III.1 For any value of $\theta\in\Theta$ there exist

(A1) a unique strong solution (X,V) to (III.3)–(III.4) with state space $\mathbb{R}\times(0,\infty)$;

(A2) a probability μ_θ on $(0,\infty)$ such that V is strictly stationary and α-mixing if $U_V\sim\mu_\theta$. □

Simple integral conditions on b and σ are known to imply stationarity and α-mixing of the diffusion V: define the *scale density* s_θ and the *speed density* $\tilde\mu_\theta$ for V by $s_\theta(v)=\exp\big(-2\int_1^v b(u,\theta)/\sigma^2(u,\theta)\,du\big)$ and $\tilde\mu_\theta(v)=(\sigma^2(v,\theta)s_\theta(v))^{-1}$. With this notation, if $\int_0^1 s_\theta(v)\,dv=\int_1^\infty s_\theta(v)\,dv=+\infty$ and $K_\theta=\int_0^\infty\tilde\mu_\theta(v)\,dv<+\infty$, then V has invariant distribution $\mu_\theta(dv)=\tilde\mu_\theta(v)/K_\theta\,dv$ and condition (A2) is satisfied. See Karlin & Taylor (1981, Section 15.6) or Karatzas & Shreve (1991, Section 5.5), for example, for the above integral conditions, and Genon-Catalot *et al.* (1998*b*) for the mixing result.

Because of the structure of the model it is natural to consider increments of X. Define for $i\in\mathbb{N}$

$$Z_i = X_{i\Delta} - X_{(i-1)\Delta};\quad M_i = \int_{(i-1)\Delta}^{i\Delta}\xi(V_s)\,ds;\quad S_i = \int_{(i-1)\Delta}^{i\Delta} V_s\,ds$$

and $H_i = (M_i, S_i)$. Let $\mathbb{R}^\infty$ be the space of real sequences and let $\mathscr{B}^\infty$ be the σ-algebra on $\mathbb{R}^\infty$ generated by the finite-dimensional rectangles. Then $Z = (Z_1, Z_2, \dots)$ is a random variable with values in $(\mathbb{R}^\infty, \mathscr{B}^\infty)$.

The following proposition states some probabilistic properties of the distribution of Z. The results are well-known but are proved below for completeness.

Proposition III.2 *Assume that condition (A1) holds. Then, conditional on $\{V_t\}_{t\geq 0}$, the increments $Z_1, Z_2, \dots$ of X are independent and the conditional distribution of Z_i is Gaussian with expectation M_i and variance S_i. If furthermore condition (A2) holds and $V_0 = U_V \sim \mu_\theta$, then $H = (H_1, H_2, \dots)$ and $Z = (Z_1, Z_2, \dots)$ are strictly stationary and Z is α-mixing.*

Proof Let $(\mathscr{C}_+, \mathscr{B}(\mathscr{C}_+))$ be the space of positive, continuous functions defined on $[0,\infty)$, equipped with the σ-algebra generated by the coordinate projections x_t°, $t \in [0,\infty)$ given by $x_t^\circ(c) = c(t)$ for $c \in \mathscr{C}_+$. With this notation V takes values in $(\mathscr{C}_+, \mathscr{B}(\mathscr{C}_+))$.

For each $v \in \mathscr{C}_+$, define the process (that is, the random variable with values in the space of real, continuous functions)

$$F(v,W) = \big(F_t(v,W)\big)_{t\geq 0} = \Big(\int_0^t \xi(v_s)\,ds + \int_0^t \sqrt{v_s}\,dw_s\Big)_{t\geq 0}$$

which is well-defined by condition (A1).

It follows (by approximation and localizations arguments) that $X - X_0$ is indistinguishable from the process $F(V,W)$ which is defined path-wise by $F(V,W)(\omega) = F(V(\omega),W)$. In particular, the conditional distribution of $X - X_0$ given $V = v$ is the same as the distribution of $F(v,W)$. The first assertion follows immediately since $F(v,W)$ has independent, Gaussian increments: $F_{t_2}(v,W) - F_{t_1}(v,W) \sim N(m(v), s(v))$ for $t_1 < t_2$ where $m(v) = \int_{t_1}^{t_2} \xi(v_s)\,ds$ and $s(v) = \int_{t_1}^{t_2} v_s\,ds$.

For the second assertion, let $j, l \in \mathbb{N}$ be arbitrary. If $V_0 \sim \mu_\theta$ then $\{V_t\}_{0\leq t\leq l\Delta}$ and $\{V_t\}_{j\Delta\leq t\leq (l+j)\Delta}$ have same distribution. Hence, $(H_1, \dots, H_l)$ and $(H_{j+1}, \dots, H_{j+l})$ have same distribution, and by the distributional result above it the follows that $(Z_1, \dots, Z_l)$ and $(Z_{j+1}, \dots, Z_{j+l})$ have same distribution. Since j and l are arbitrary, it follows that H and Z are stationary.

We finally show that the α-mixing coefficients for Z and V (corresponding to an arbitrary θ which is omitted from the notation) satisfy $\alpha_Z(j) \leq \alpha_V\big((j-1)\Delta\big)$ for all integers $j \geq 2$ so that α-mixing of V implies α-mixing of Z. Let $j \geq 2$ and $l \geq 1$ be arbitrary but fixed. Also, let $0 \leq U_1 \leq 1$ be measurable wrt. the σ-algebra generated by $(Z_1, \dots, Z_l)$ and $0 \leq U_2 \leq 1$ be measurable wrt. the σ-algebra generated by $(Z_{l+1}, Z_{l+j+1}, \dots)$. Then, by the distribution result above,

$$\mathrm{E}\,U_1U_2 = \int_\Omega \mathrm{E}\big(U_1U_2|\mathscr{G}\big)\,dPr = \int_\Omega \mathrm{E}\big(U_1|\mathscr{G}_0^{l\Delta}\big)\,\mathrm{E}\big(U_2|\mathscr{G}_{(l+j-1)\Delta}^{\infty}\big)\,dPr$$

where $\mathscr{G} = \sigma(V_t : t \geq 0)$, $\mathscr{G}_0^{l\Delta} = \sigma(V_t : 0 \leq t \leq l\Delta)$ and $\mathscr{G}_{(l+j-1)\Delta}^{\infty} = \sigma(V_t : t \geq (l+j-$

$1)\Delta)$. Hence,

$$\left|\mathrm{Cov}(U_1,U_2)\right| = \left|\mathrm{Cov}\Big(\mathrm{E}\big(U_1|\mathscr{G}_0^{l\Delta}\big), \mathrm{E}\big(U_2|\mathscr{G}_{(l+j-1)\Delta}^{\infty}\big)\Big)\right| \leq \alpha_V\big((j-1)\Delta\big) \tag{III.6}$$

since $\mathrm{E}\big(U_1|\mathscr{G}_0^{l\Delta}\big)$ is $\mathscr{G}_0^{l\Delta}$-measurable and $\mathrm{E}\big(U_2|\mathscr{G}_{(l+j-1)\Delta}^{\infty}\big)$ is $\mathscr{G}_{(l+j-1)\Delta}^{\infty}$-measurable. The inequality (III.6) holds for arbitrary $j \geq 2$ and $l \geq 1$ so it follows that $\alpha_Z(j) \leq \alpha_V\big((j-1)\Delta\big)$ and that Z is α-mixing as claimed. □

By Proposition III.2 we easily derive moments of Z in terms of moments of M and S. For example, if the relevant moments exist,

$$\begin{aligned}
\mathrm{E}_\theta Z_i &= \mathrm{E}_\theta M_i \\
\mathrm{Var}_\theta Z_i &= \mathrm{E}_\theta S_i + \mathrm{Var}_\theta M_i \\
\mathrm{E}_\theta Z_i^4 &= 3\,\mathrm{E}_\theta S_i^2 + \mathrm{E}_\theta M_i^4 + 6\,\mathrm{E}\, M_i^2 S_i
\end{aligned}$$

for $i \in \mathbb{N}$ and

$$\begin{aligned}
\mathrm{Cov}_\theta(Z_i,Z_j) &= \mathrm{Cov}_\theta(M_i,M_j) \\
\mathrm{Cov}_\theta(Z_i^2,Z_j^2) &= \mathrm{Cov}_\theta(S_i+M_i^2, S_j+M_j^2)
\end{aligned} \tag{III.7}$$

for all $i \neq j$. In particular the Z's are uncorrelated — but not independent — if $\xi \equiv 0$. For simple models of V the above moments may be calculated explicitly; for more complicated models they must be computed by simulation.

In the following we shall always assume that Assumption III.1 is satisfied and that V is started stationarily, $V_0 \sim \mu_\theta$. We let $P_\theta = Z(Pr)$ denote the distribution of $Z = (Z_1, Z_2, \dots)$ when the parameter is θ. For $d \geq 1$, let furthermore $P_\theta^d = (Z_1, \dots, Z_d)(Pr)$ be the distribution of d consecutive increments.

Note that Z is a *hidden Markov model* with continuous state space of the hidden chain: Let $\tilde{H}_i = (V_{i\Delta}, M_i, S_i)$. Then $\tilde{H} = (\tilde{H}_1, \tilde{H}_2, \dots)$ is stationary Markov (because V is stationary Markov and $\tilde{H}_i$ is a function of $(V_t)_{(i-1)\Delta \leq t \leq i\Delta}$), and conditionally on $\tilde{H}$ the increments $Z_1, Z_2, \dots$ are independent with conditional distribution of Z_i depending on $(i, \tilde{H})$ via $\tilde{H}_i$ only. Hence, the second part of the above proposition is a special case of Proposition 2.1 in Genon-Catalot *et al.* (1998*b*) which claims that a hidden Markov model inherits stationarity and ergodicity from the hidden chain. See Genon-Catalot *et al.* (1998*b*) for formal definitions and proofs of the hidden Markov properties.

The following proposition shows that Z is *reversible* in the sense that $(Z_1, \dots, Z_n)$ and $(Z_n, \dots, Z_1)$ are identically distributed.

Proposition III.3 *Under Assumption III.1,* $(Z_1, \dots, Z_n)$ *and* $(Z_n, \dots, Z_1)$ *have same distribution for all* $n \geq 1$*, i.e.* $(Z_n, \dots, Z_1) \sim P_\theta^n$ *for all* $n \geq 1$

Proof We first show that $(H_1, \dots, H_n)$ and $(H_n, \dots, H_1)$ have same distribution, next that $(Z_1, \dots, Z_n)$ and $(Z_n, \dots, Z_1)$ have the same densities.

First, for each $i = 1, \dots, n$

$$M_i = \int_{(i-1)\Delta}^{i\Delta} \xi(V_s)\, ds = \int_{(n-i)\Delta}^{(n-i+1)\Delta} \xi(V_{n\Delta - s})\, ds$$

and similarly for S_i. Define a function $f = (f_1, \dots, f_{2n})$ from the space of positive continuous functions defined on $[0, n\Delta]$ to $\mathbb{R}^{2n}$ coordinate-wise by

$$f_{2i-1}\big(\{v_s\}_{0\le s\le n\Delta}\big) = \int_{(i-1)\Delta}^{i\Delta} \xi(v_s)\, ds, \quad i = 1, \dots, n$$
$$f_{2i}\big(\{v_s\}_{0\le s\le n\Delta}\big) = \int_{(i-1)\Delta}^{i\Delta} v_s\, ds, \quad i = 1, \dots, n.$$

Then $(H_1, \dots, H_n) = \big((M_1, S_1), \dots, (M_n, S_n)\big) = f\big(\{V_s\}_{0\le s\le n\Delta}\big)$. From the theory of one-dimensional diffusion processes it is well-known that V is time reversible in the sense that the processes $\{V_{t-s}\}_{0\le s\le t}$ and $\{V_s\}_{0\le s\le t}$ are identically distributed for all $t \ge 0$. Hence,

$$(H_1, \dots, H_n) = f\big(\{V_s\}_{0\le s\le n\Delta}\big) \stackrel{\mathscr{D}}{=} f\big(\{V_{n\Delta - s}\}_{\le s\le n\Delta}\big) = (H_n, \dots, H_1),$$

that is, $(H_1, \dots, H_n)$ and $(H_n, \dots, H_1)$ are identically distributed.

Second, recall from Proposition III.2 that conditional on $\{V_t\}_{t\ge 0}$ the variables $Z_1, \dots, Z_n$ are independent and $Z_i \sim N(M_i, S_i)$. Hence, the density $p_{(Z_1,\dots,Z_n)}$ of $(Z_1, \dots, Z_n)$ at a point $(z', \dots, z'') \in \mathbb{R}^n$ is given by

$$p_{(Z_1,\dots,Z_n)}(z', \dots, z'') = \int \varphi(z', h') \cdots \varphi(z'', h'')\, d\pi_{(H_1,\dots,H_n)}(h', \dots, h'') \tag{III.8}$$

where we for $h = (m, s)$ have used the notation $\varphi(\cdot, h)$ for the Gaussian density with mean m and variance s and the notation $\pi_{(H_1,\dots,H_n)}$ for the distribution of $(H_1, \dots, H_n)$. Note that we have omitted the parameter dependence from the notation. The density of the reversed sequence $(Z_n, \dots, Z_1)$ at the same point $(z', \dots, z'')$ is (with obvious notation)

$$\begin{aligned} p_{(Z_n,\dots,Z_1)}(z', \dots, z'') &= \int \varphi(z', h') \cdots \varphi(z'', h'')\, d\pi_{(H_n,\dots,H_1)}(h', \dots, h'') \\ &= \int \varphi(z', h') \cdots \varphi(z'', h'')\, d\pi_{(H_1,\dots,H_n)}(h', \dots, h'') \\ &= p_{(Z_1,\dots,Z_n)}(z', \dots, z'') \end{aligned}$$

where the second equality holds because $\pi_{(H_n,\dots,H_1)} = \pi_{(H_1,\dots,H_n)}$ and the third equality follows from (III.8). Since $z', \dots, z''$ were arbitrary the sequences $(Z_1, \dots, Z_n)$ and $(Z_n, \dots, Z_1)$ have same distribution. □

Finally some comments on possible generalizations of the model. Under (III.3) and (III.4) the distribution of X is completely determined by V. This is no longer

true if ξ and the diffusion function for X is allowed to depend on X or if the Brownian motions W and $\tilde{W}$ are correlated. Both generalizations destroy the nice conditional distribution result in Proposition III.2 and make estimation in the model very difficult.

One could also generalize the model by allowing ξ and the diffusion function of X to depend on an unknown parameter η. The increments of X would still be independent and Gaussian, but the mean and variance of the Gaussian distributions would depend on η. Estimation of η is easily built into the estimation method below, see Section III.8 for further remarks.

III.3 Approximations to the likelihood function

We aim at estimation of θ from discrete-time observations $X_0, X_\Delta, \ldots, X_{n\Delta}$. In this section we describe a class of approximations to the likelihood function. Later we discuss computational aspects (Section III.4) and show that maximization of *any* of the approximations leads to a consistent and asymptotically normal estimator of θ (Section III.5).

III.3.1 The fundamental idea

Motivated by the distributional result in Proposition III.2 we consider the vector of increments $(Z_1, \ldots, Z_n)$. For an observation $(z_1, \ldots, z_n)$ the likelihood function is given by

$$\begin{aligned} L_n(\theta) &= \int \prod_{i=1}^n \frac{1}{\sqrt{2\pi s_i}} \exp\Big(-\frac{(z_i - m_i)^2}{2s_i}\Big)\, d\pi_\theta^n(h^n) \\ &= \mathrm{E}_{\pi_\theta^n} \prod_{i=1}^n \varphi(z_i, M_i, S_i), \end{aligned} \tag{III.9}$$

where h^n is short for $(h_1, \ldots, h_n) = \big((m_1, s_1), \ldots, (m_n, s_n)\big)$, $\pi_\theta^n = H^n(Pr)$ is the distribution of H^n and $\varphi(\cdot, m, s)$ is the density of $N(m, s)$.

The likelihood is the expectation with respect to the distribution of H^n of a certain functional. In principle, this expectation could be calculated to any precision as follows: (*i*) simulate a number of paths V up to time $n\Delta$ according to (III.4); (*ii*) calculate for each simulation (approximations to) the integrals M_i and S_i and the above product; (*iii*) calculate the average of the simulated product values. Finally the (simulated) likelihood function should be maximized in order to obtain an estimator of θ. However, this approach is not feasible in practice because one needs a *huge* number of simulated paths of V just to calculate the likelihood function for a *single* parameter value. This is not strange since two paths of V over a large time interval may be very different.

Our approach will be to consider suitable approximations to L_n rather than L_n itself. The approximations under consideration are easier to simulate, but of course this is at the expense of loss of efficiency.

Introduce some further notation on the distribution of Z: let $p_\theta^k(z_1,\dots,z_k)$ denote the density at $(z_1,\dots,z_k)$ of the simultaneous distribution of $Z_1,\dots,Z_k$, $k\in\mathbb{N}$. It follows from Proposition III.2 that $p_\theta^k>0$ so the k-lag conditional density

$$p_\theta^{c,k}(z_{k+1}|z_1,\dots,z_k)=\frac{p_\theta^{k+1}(z_1,\dots,z_{k+1})}{p_\theta^k(z_1,\dots,z_k)}$$

at z_{k+1} of Z_{k+1} given $(Z_1,\dots,Z_k)=(z_1,\dots,z_k)$ is well-defined and positive for all $z_1,\dots,z_{k+1}$. For $k=0$ we let $p_\theta^{c,0}=p_\theta^1$. Furthermore, introduce the notation z_i^j for the vector $(z_i,\dots,z_j)$, $i\le j$. With this notation the likelihood has the form

$$L_n(\theta)=p_\theta^n(z)=\prod_{i=0}^{n-1}p_\theta^{c,i}(z_{i+1}|z_1,\dots,z_i)=\prod_{i=0}^{n-1}p_\theta^{c,i}(z_{i+1}|z_1^i) \tag{III.10}$$

since Z is strictly stationary (Proposition III.2).

Recall that the increments form an α-mixing sequence, that is $\alpha_Z(k)\to 0$ as $k\to\infty$. Intuitively, this means that the dependence between Z_i and $(Z_1,\dots,Z_j)$ is small when i is much larger than j. It thus makes good sense to approximate the conditional densities in (III.10) by k-lag conditional densities for some k large enough. To be specific, leave for $0\le k<n$ fixed the first $k+1$ terms in (III.10) unchanged but approximate for $i=k+1,\dots,n-1$ the conditional density $p_\theta^{c,i}(z_{i+1}|z_1^i)$ by $p_\theta^{c,k}(z_{i+1}|z_{i-k+1}^i)$. The corresponding approximation of the likelihood is

$$\begin{aligned}L_n^k(\theta)&=\prod_{i=0}^{k}p_\theta^{c,i}(z_{i+1}|z_1,\dots,z_i)\prod_{i=k+1}^{n-1}p_\theta^{c,k}(z_{i+1}|z_{i-k+1},\dots,z_i)\\&=p_\theta^{k+1}(z_1,\dots,z_{k+1})\prod_{i=k+1}^{n-1}p_\theta^{c,k}(z_{i+1}|z_{i-k+1},\dots,z_i)\end{aligned}$$

and the idea is to use the approximation L_n^k instead of the true likelihood function, that is, maximize $L_n^k(\theta)$ in order to obtain an estimator $\hat\theta_n^k$ of θ. In particular $k=1$ corresponds to a Markov approximation:

$$L_n^1(\theta)=p_\theta^1(z_1)\prod_{i=1}^{n-1}p_\theta^{c,1}(z_{i+1}|z_i)$$

and $k=0$ corresponds to independence of $Z_1,\dots,Z_n$:

$$L_n^0(\theta)=\prod_{i=1}^{n}p_\theta^1(z_i).$$

No approximation is made for $k=n-1$, but the idea is to use a small value of k. Note that L_n^k would be the true likelihood function *if* Z was k'th order Markov.

It is important to realize that, although we use approximations of the likelihood function, *no bias is introduced* and the estimators are consistent, see Section III.5. The reason is that we use the *true* k-lag conditional k-lag densities and

not some approximation. For example, assume for a moment that Z is a strictly stationary auto-regressive process of order 2 with $N(\theta_1 Z_1 + \theta_2 Z_2, \sigma^2)$ as the conditional distribution of Z_3 given (Z_1, Z_2). For $k = 1$ we would *not* just put $\theta_1 = 0$! Instead we would use that stationarity implies that the conditional distribution of Z_2 given Z_1 is Gaussian with mean αZ_1 where $\alpha = \theta_2/(1-\theta_1)$ and variance $\sigma_1^2 = \sigma^2/(1-\theta_1^2)$. Similarly, for $k = 0$ we would use the true stationary marginal distribution $N(0, \sigma_1^2/(1-\alpha^2))$ (rather than $N(0,\sigma^2)$ corresponding to $\theta_1 = \theta_2 = 0$).

Another important property is that the k'th order approximate maximum likelihood estimator is *invariant to data transformations*: if g is a bijective function from $\mathbb{R}$ to some subset of $\mathbb{R}$ then the estimator based on $g(Z_1), \ldots, g(Z_n)$ is the same as that based on $Z_1, \ldots, Z_n$.

In practice we shall of course minimize $U_n^k = -\log L_n^k/n$ rather than maximize L_n^k. Define $u_\theta^k = -\log p_\theta^k$, $u_\theta^{c,k} = -\log p_\theta^{c,k}$. With this notation

$$U_n^k(\theta) = -\frac{1}{n}\log L_n^k(\theta) = \frac{1}{n}u_\theta^{k+1}(z_1^{k+1}) + \frac{1}{n}\sum_{i=k+1}^{n-1} u_\theta^{c,k}(z_{i+1}|z_{i-k+1}^i) \tag{III.11}$$

$$= \frac{1}{n}\sum_{i=k}^{n-1} u_\theta^{k+1}(z_{i-k+1}^{i+1}) - \frac{1}{n}\sum_{i=k+1}^{n-1} u_\theta^k(z_{i-k+1}^i). \tag{III.12}$$

III.3.2 Comments on the number of lags needed

Now some comments on how to choose k. Further remarks follow in Section III.6. First note that is does no harm to use a larger k than the actual dependence structure in data calls for. For example, if the dependence on lag k is negligible, then U_n^{k-1} and U_n^k should be indistinguishable.

In Section III.5 we show that, for each k, the estimator $\hat{\theta}_n^k$ obtained by minimizing U_n^k is consistent and asymptotically normal as n increases. From this point of view, choosing k is a question of efficiency. Intuitively we should prefer large k's to small k's since further characteristics of the dependence structure are taken into account as k increases. However, we have not been able to show that asymptotic efficiency (measured as one divided by the asymptotic variance of the estimator in case of a one-dimensional parameter) is in fact increasing in k; see Section III.6 for further comments. Also, one should take into account that the computing time increases with k, see Section III.4.

It is of course crucial that the parameter is identifiable from the conditional distribution of Z_{k+1} given Z_1^k:

$$\mathscr{L}_\theta(Z_{k+1}|Z_1^k) \neq \mathscr{L}_{\theta'}(Z_{k+1}|Z_1^k), \quad \theta \neq \theta'. \tag{III.13}$$

The distribution $\{V_t\}_{0\le t\le\Delta}$ depends on all parameters (otherwise the model is over-parametrized). Typically, this implies that the distributions of H_1 and Z_1 depend on all parameters as well, such that the identifiability condition (III.13) is satisfied for $k = 0$. Note that for $\xi \equiv 0$ (implying $M_i \equiv 0$) it is easy to see that $\mathscr{L}_\theta(Z_1) = \mathscr{L}_{\theta'}(Z_1)$ *if and only* if $\mathscr{L}_\theta(S_1) = \mathscr{L}_{\theta'}(S_1)$: indeed, the characteristic function at $x \in \mathbb{R}$ of the

stationary distribution of Z_1 is given by

$$\mathrm{E}_\theta\, e^{ixZ_1} = \mathrm{E}_\theta\, \mathrm{E}_\theta\big(e^{ixZ_1}|S_1\big) = \mathrm{E}_\theta\, e^{-S_1x^2/2} \tag{III.14}$$

which is the Laplace transform of the distribution of S_1 evaluated at $x^2/2$.

In principle it could happen that (III.13) holds for some k_0 but not for all $k > k_0$. For example, the conditional distribution of Z_2 given Z_1 (corresponding to $k = 1$) need not depend on θ just because the stationary distribution does ($k_0 = 0$). We believe, however, that this problem is not likely to appear for the diffusion models considered in this paper.

In practice, it might be very difficult (or impossible) to check that (III.13) is satisfied. However, we may be able to check that the parameter is determined from the simultaneous distribution of $(Z_1, \ldots, Z_{k+1})$, that is,

$$\mathscr{L}_\theta(Z_1^{k+1}) \neq \mathscr{L}_{\theta'}(Z_1^{k+1}), \quad \theta \neq \theta', \tag{III.15}$$

for example via moment considerations. Note that (III.15) is necessary, but not sufficient for (III.13). Also note that the first sum in (III.12) can be interpreted as a sum of $k+1$ (minus) log-likelihoods, each of which is obtained by pretending that $(k+1)$-tuples with no overlap are independent, see Appendix III.A.1 for details. Hence, in case the parameter is determined from the simultaneous, but not from the conditional distributions, one could consider minimizing the first sum in (III.12) rather than (III.12) itself.

Although (III.13) — or (III.15) — holds, we might have problems identifying the parameters in practice. For example, consider the model where X has no drift ($\xi \equiv 0$) and V is a Cox-Ingersoll-Ross model. We study this model in detail in Section III.7. It turns out that although the distributions of (Z_1, Z_2) for two different parameter values are not the same, they can be very much alike, even for parameters far from each other. This makes estimation of all parameters in the model practically impossible for $k = 1$. However, the identifiability problems seem to diminish as k increases, and $k = 4$ yields acceptable estimates for seven of ten simulated datasets considered in Section III.7.

Finally some more specific guidelines on how to choose k for concrete data. Since for increasing k, U_n^k takes more of the dependence structure of the model into account, it might be useful to plot the autocorrelation functions for various transformations of the data (like the data squared or the absolute values of the data). If the empirical autocorrelation coefficients from lag k_0 and onwards are negligible then it seems reasonable not to use k much larger than k_0. As noted above, if we for some k_0 have caught the important features of the distribution then U_n^k should be close to $U_n^{k_0}$ for $k > k_0$. Hence, so should the corresponding estimates and one can try increasing values of k until the parameter estimates and the minimal values of U_n^k stabilize.

III.4 Computational aspects

In this section we discuss how to compute $U_n^k(\theta)$ in practice for a fixed but arbitrary value of θ. Let us first focus on calculation of $p_\theta^{k+1}(\tilde{z}_1^{k+1})$ for arbitrary $\tilde{z}_1,\dots,\tilde{z}_{k+1} \in \mathbb{R}$. An expression for $U_n^k(\theta)$ follows almost immediately.

Replace n in formula (III.9) by $k+1$ in order to write $p_\theta^{k+1}(\tilde{z}_1^{k+1})$ as an expectation

$$p_\theta^{k+1}(\tilde{z}_1^{k+1}) = \mathrm{E}_{\pi_\theta^{k+1}} \prod_{j=1}^{k+1} \varphi(\tilde{z}_j, M_j, S_j) \tag{III.16}$$

with respect to the distribution of $(H_1,\dots,H_{k+1})$. Again, $\varphi(\cdot,m,s)$ is the density of $N(m,s)$. We compute (III.16) as an average of R simulated values,

$$\frac{1}{R}\sum_{r=1}^{R}\prod_{j=1}^{k+1}\varphi\Big(\tilde{z}_j, M_j^{(r)}, S_j^{(r)}\Big)$$

where for each $r = 1,\dots,R$

$$\Big(H_1^{(r)},\dots,H_{k+1}^{(r)}\Big) = \Big(\big(M_1^{(r)},S_1^{(r)}\big),\dots,\big(M_{k+1}^{(r)},S_{k+1}^{(r)}\big)\Big)$$

is a simulation of $(H_1,\dots,H_{k+1})$. We can compute (III.16) to any accuracy by choosing R large enough. Of course $p_\theta^k(\tilde{z}_1^k)$ is calculated similarly; simply replace the above product from 1 to $k+1$ by the product from 1 to k. Note that we can use the *same simulations* of $(H_1,\dots,H_k)$ when we calculate p_θ^k and p_θ^{k+1}.

The r'th simulation of $(H_1,\dots,H_{k+1})$ is calculated via a simulation, $V^{(r)}$, of the volatility process V from time zero to time $(k+1)\Delta$ as follows. First, the initial value of $V^{(r)}$ is chosen according to the stationary distribution,

$$V_0^{(r)} \sim \mu_\theta.$$

Next, split the interval $[0,(k+1)\Delta]$ into $N(k+1)\Delta$ subintervals of length $\delta = 1/N$ and calculate values $V_{l\delta}^{(r)}$, $l \le N(k+1)\Delta$ recursively by the Millstein scheme

$$\begin{aligned} V_{l\delta}^{(r)} &= V_{(l-1)\delta}^{(r)} + b\big(V_{(l-1)\delta}^{(r)},\theta\big)\,\delta + \sigma\big(V_{(l-1)\delta}^{(r)},\theta\big)\,\varepsilon_l^{(r)} \\ &\quad + \frac{1}{2}\sigma\big(V_{(l-1)\delta}^{(r)},\theta\big)\,\sigma'\big(V_{(l-1)\delta}^{(r)},\theta\big)\Big(\big(\varepsilon_l^{(r)}\big)^2 - \delta\Big), \quad l \le N(k+1)\Delta \end{aligned}$$

where $\sigma' = \partial_v\sigma$ is the derivative of σ with respect to the state variable and the innovations $\varepsilon_1^{(r)},\dots,\varepsilon_{(k+1)N}^{(r)}$ are independent, identically $N(0,\delta)$-distributed random variables. We could of course use the simpler Euler scheme (that is, the above recursive scheme without the last term) instead of the Millstein scheme.

Finally, recall that $M_j = \int_{(j-1)\Delta}^{j\Delta}\xi(V_s)\,ds$ and $S_j = \int_{(j-1)\Delta}^{j\Delta} V_s\,ds$ and let for $j = 1,\dots,k+1$

$$M_j^{(r)} = \frac{1}{\delta}\sum_{l=(j-1)N}^{jN-1}\xi\Big(V_{l\delta}^{(r)}\Big), \qquad S_j^{(r)} = \frac{1}{\delta}\sum_{l=(j-1)N}^{lN-1} V_{l\delta}^{(r)}$$

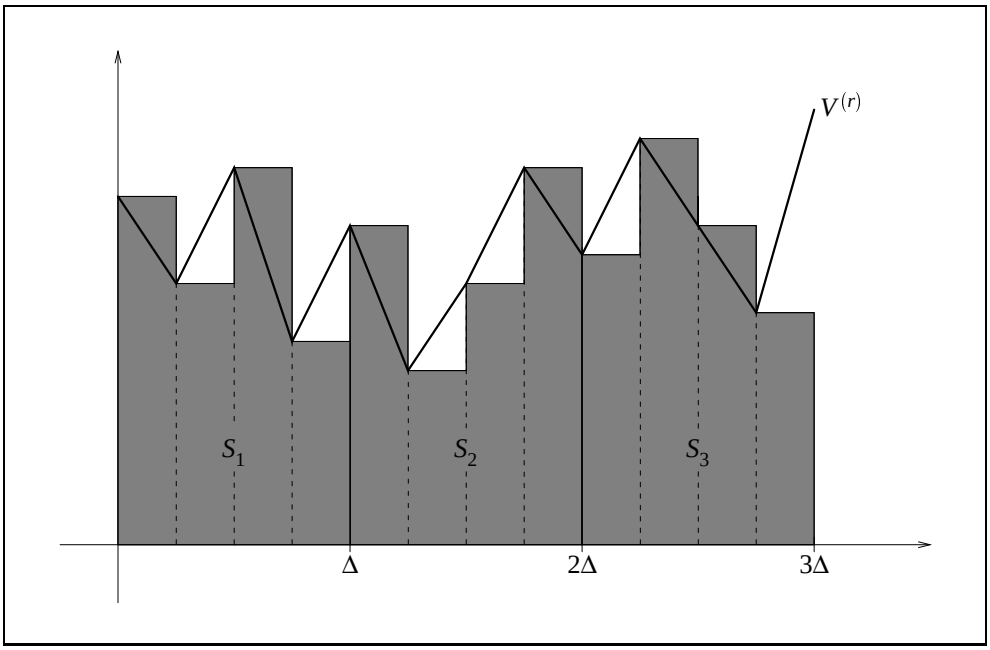

Figure III.1: Calculation of $S_1^{(r)},\dots,S_{k+1}^{(r)}$ from simulated values $V_{l\delta}^{(r)}$ as a left Riemann sum. The thick line shows (the linear interpolation of) the path $V^{(r)}$ and $S_1,\dots.S_{k+1}$ are the volumes of the shaded areas. In the Figure, $k+1=3$ and $N=4$.

be the simple left Riemann approximations. The calculation of $S_1,\dots,S_{k+1}$ from the discrete-time simulation $V^{(r)}$ is illustrated in Figure III.1 for $k+1=3$ and $N=4$. The thick line shows the simulated V-path (where we have used linear interpolation between partition points $l\delta$), and S_1,S_2 and S_3 are the sizes of the shaded areas.[1]

As noted we can use the same simulations $(H_1^{(r)},\dots,H_k^{(r)})$ of $(H_1,\dots,H_k)$ for computation of p_θ^k and p_θ^{k+1}. Even more important, we can use the same simulations of $(H_1,\dots,H_{k+1})$ for *all arguments* $\tilde{z}_1^{k+1}$. In other words we calculate $U_n^k(\theta)$ as

$$-\frac{1}{n}\sum_{i=k}^{n-1}\log\frac{1}{R}\sum_{r=1}^{R}\prod_{j=1}^{k+1}\varphi_{i-k+j,j}^{(r)}+\frac{1}{n}\sum_{i=k+1}^{n-1}\log\frac{1}{R}\sum_{r=1}^{R}\prod_{j=1}^{k}\varphi_{i-k+j,j}^{(r)} \tag{III.17}$$

where $\varphi_{i,j}^{(r)}$ is short for $\varphi(z_i,M_j^{(r)},S_j^{(r)})$.

There are several "parameters" to choose: the number R of repetitions, the number N of subintervals per Δ-interval, and of course the number of lags k. We already discussed how to choose k in the end of Section III.3. The parameters N

[1]Of course, one could use better approximations to the integrals; for example the size of the areas under the thick line. It would probably not improve the calculation much though, since (*i*) the simple approximation introduces no systematic error, and (*ii*) we do not know how the simulated path would behave had we simulated it at points in between the $l\delta$'s.

and (in particular) R determine how accurately the values of U_n^k are determined and must be large enough that the calculation of $U_n^k(\theta)$ is suitably stable, that is, the simulated vales of $U_n^k(\theta)$ are "sufficiently close" for different simulations.

The number of calculations needed to compute one single value of $U_n^k(\theta)$ increases approximately linearly in both R and $k+1$, and if computing time is limited one must compromize between stability and the number of lags involved. Note that it might be necessary to increase R as k increases since we must simulate longer paths of V and thus might need more simulations to obtain numerical stability.

So-called *antithetic variables* may increase computational stability. Here, it means that we make simulations of V in pairs where we in one simulation use the randomly generated ε's in the Millstein scheme and in the other one use *minus* the ε's. For R sets of randomly generated ε's we thus compute $2R$ simulated paths of V, compute the $\varphi^{(r)}$-values in (III.17) for each of the $2R$ simulated paths of V, and average over all $2R$ simulations. The two $\varphi^{(r)}$-values corresponding to the same set of ε's (plus and minus) tend to be negatively correlated. The computing time is approximately doubled when we use antithetic variables, but hopefully we need R less than half as big as without antithetic variables in order to obtain same precision.

It is possible to compute suitably accurate values of U_n^k in reasonable time: for $n = 500$ observations from the model where $\xi \equiv 0$ and V is a Cox-Ingersoll-Ross process, it takes somewhat less than a minute to compute a value of U_n^4 with $N = 10$ and $R = 10.000$ on a Digital alpha running at 500 MHz. This is only to give an idea of the computational burden — no attempts have been made as to optimize the routine.

Finally a very important remark: As always when criterion functions (or estimating functions) are simulated, it is crucial to use the *same random numbers* for different values of θ. Otherwise R must be chosen extremely large for the simulated criterion function to behave continuously.

III.5 Asymptotic results

In this section we prove consistency and asymptotic normality (as $n \to \infty$) of the estimator $\hat{\theta}_n^k$ satisfying $U_n^k(\hat{\theta}_n^k) = \inf_{\theta\in\Theta} U_n^k(\theta)$. The results hold for *any fixed values of k and Δ*. The true parameter is denoted by θ_0, and all results are with respect to P_{θ_0}.

It is essential for the results below that limit theorems hold for the sequence Z. As already mentioned, the ergodic theorem holds under Assumption III.1 since α-mixing implies ergodicity. This, together with some regularity conditions, is sufficient to show consistency. For the asymptotic normality we furthermore need a central limit theorem for Z. We use a version of the central limit theorem involving further assumptions on the α-mixing coefficients. Both limit theorems are formulated and proved in the appendix (Theorem III.12) although the results are well-known.

The first term in (III.11) is negligible as n increases so we focus on the sum $\frac{1}{n}\sum_{i=k+1}^{n-1} u_\theta^{c,k}(z_{i+1}|z_{i-k+1}^i)$. In the following we let $||\cdot||$ denote the usual Eucledian norm on $\mathbb{R}^p$, and for a function $g:\mathbb{R}^p \to \mathbb{R}$ and a probability Q on $\mathbb{R}^p$ we write $Q(g)$ for the integral $\int g\,dQ$.

III.5.1 Consistency

Apart from Assumption III.1 we need the following regularity conditions for consistency of $\hat{\theta}_n^k$.

Assumption III.4 Fix $k \geq 0$ and assume that the following conditions hold:

(B1) the parameter space Θ is a compact subset of $\mathbb{R}^p$;

(B2) for all $\theta \in \Theta$ there are a constant $\delta_\theta > 0$ and a function $\bar{u}_\theta : \mathbb{R}^{k+1} \to \mathbb{R}$ in $L^1(P_{\theta_0}^{k+1})$ such that $\sup_{\theta' \in T_{\theta,\delta_\theta}} \left|u_{\theta'}^{c,k}(z_{k+1}|z_1^k)\right| \leq \bar{u}_\theta(z_1^{k+1})$ for all $z_1,\ldots,z_{k+1} \in \mathbb{R}$ where $T_{\theta,\delta} = \{\theta' \in \Theta : ||\theta - \theta'|| \leq \delta\}$;

(B3) the functions $\theta \to u_\theta^{c,k}(z_{k+1}|z_1,\ldots,z_k)$ from Θ to $\mathbb{R}$ are continuous for all $z_1,\ldots,z_{k+1} \in \mathbb{R}$;

(B4) the conditional distributions of Z_{k+1} given $Z_1^k = z_1^k$ with respect to P_θ^{k+1} and $P_{\theta'}^{k+1}$ are different for $\theta \neq \theta'$ and all $z_1,\ldots,z_k \in \mathbb{R}$. □

Note that conditions (B1) and (B3) ensure that a minimum of U_n^k exists, but the minimum could be attained at the boundary of Θ and need not be unique. Condition (B2) expresses that $u_\theta^{c,k}$ is locally dominated integrable wrt. $P_{\theta_0}^{k+1}$ and implies that $u_\theta^{c,k}$ is in $L^1(P_{\theta_0}^{k+1})$ for all $\theta \in \Theta$. The ergodic theorem thus yields

$$U_n^k(\theta) \to P_{\theta_0}^{k+1}(u_\theta^{c,k}) = \mathrm{E}_{\theta_0} u_\theta^{c,k}\big(Z_{k+1}|Z_1,\ldots,Z_k\big) \tag{III.18}$$

as $n \to \infty$ in P_{θ_0}-probability (even P_{θ_0}-a.s and in $L^1(P_{\theta_0})$). Denote the limit by $J^k(\theta)$. Conditions (B2) and (B3) make U_n^k and J^k continuous and ensure that the convergence (III.18) holds uniformly in θ (Lemma III.6). Condition (B4) is an identifiability condition ensuring that J^k has unique minimum for $\theta = \theta_0$ as asserted in the following lemma.

Lemma III.5 *Assume that Assumption III.1 holds. If furthermore (B2) and (B4) hold then $J^k(\theta) \geq J^k(\theta_0)$ for all $\theta \in \Theta$ with equality if and only if $\theta = \theta_0$.*

Proof By definition of J^k and Jensen's inequality we get for $\theta \in \Theta$

$$\begin{aligned} J^k(\theta_0) - J^k(\theta) &= \mathrm{E}_{\theta_0} u^{c,k}_{\theta_0}(Z_{k+1}|Z_1^k) - \mathrm{E}_{\theta_0} u^{c,k}_{\theta}(Z_{k+1}|Z_1^k) \\ &= \mathrm{E}_{\theta_0} \log\left(\frac{p^{c,k}_{\theta}(Z_{k+1}|Z_1^k)}{p^{c,k}_{\theta_0}(Z_{k+1}|Z_1^k)}\right) \\ &\leq \log \mathrm{E}_{\theta_0}\left(\frac{p^{c,k}_{\theta}(Z_{k+1}|Z_1^k)}{p^{c,k}_{\theta_0}(Z_{k+1}|Z_1^k)}\right) \end{aligned}$$

with equality if and only if $p^{c,k}_{\theta}(z_{k+1}|z_1^k) = p^{c,k}_{\theta_0}(z_{k+1}|z_1^k)$ for P_{θ_0}-almost all $z_1, \ldots, z_{k+1}$, that is, if and only if $\theta = \theta_0$ by condition (B4). The density of $(Z_1, \ldots, Z_{k+1})$ wrt. P_{θ_0} at $(z_1, \ldots, z_{k+1})$ is $p^k_{\theta_0}(z_1^k) p^{c,k}_{\theta_0}(z_{k+1}|z_1^k)$. Hence,

$$\begin{aligned} \mathrm{E}_{\theta_0}\left(\frac{p^{c,k}_{\theta}(Z_{k+1}|Z_1^k)}{p^{c,k}_{\theta_0}(Z_{k+1}|Z_1^k)}\right) &= \int_{\mathbb{R}^{k+1}} p^{c,k}_{\theta}(z_{k+1}|z_1^k)\, p^k_{\theta_0}(z_1^k)\, d(z_1^{k+1}) \\ &= \int_{\mathbb{R}^k} p^k_{\theta_0}(z_1^k) \int_{\mathbb{R}} p^{c,k}_{\theta}(z_{k+1}|z_1^k)\, dz_{k+1}\, dz_1^k \\ &= \int_{\mathbb{R}^k} p^k_{\theta_0}(z_1^k)\, dz_1^k \\ &= 1 \end{aligned}$$

where we have used that $p^{c,k}_{\theta}(\cdot|z_1^k)$ and $p^k_{\theta_0}(\cdot)$ are densities on $\mathbb{R}$ and $\mathbb{R}^k$ respectively. It follows that $J^k(\theta_0) - J^k(\theta) \leq \log 1 = 0$ with equality if and only if $\theta = \theta_0$. □

The next lemma claims that the convergence (III.18) is uniform in $\theta \in \Theta$. It is of course important that Θ is compact. The proof is almost identical to the proof of Lemma 3.3 in Bibby & Sørensen (1995) but is given here for completeness.

Lemma III.6 *Under Assumption III.1 and conditions (B1), (B2), and (B3), J^k is continuous and $\sup_{\theta\in\Theta}\left|U^k_n(\theta) - J^k(\theta)\right| \to 0$ as $n \to \infty$ in probability wrt. P_{θ_0}.*

Proof We first show continuity of J^k: Let $\theta_n \to \theta$ and choose δ_θ and $\bar{u}_\theta$ according to (B2). Then $||\theta_n - \theta|| < \delta_\theta$ and hence $\left|u^{c,k}_{\theta_n}\right| \leq \bar{u}_\theta$ for n large enough. Dominated convergence now yields $J^k(\theta_n) \to J^k(\theta)$.

Next, recall that $T_{\theta,\delta} = \{\theta' \in \Theta : ||\theta - \theta'|| \leq \delta\}$ and define the function $w : \Theta \times (0,\infty) \times \mathbb{R}^{k+1}$ by

$$w(\theta, \delta, z_1^{k+1}) = \sup_{\theta' \in T_{\theta,\delta}} \left|u^{c,k}_{\theta}(z_{k+1}|z_1^k) - u^{c,k}_{\theta'}(z_{k+1}|z_1^k)\right|.$$

Then $w(\theta, \delta, z_1^{k+1}) \to 0$ as $\delta \to 0$ for all $\theta \in \Theta$ and all $z_1, \ldots, z_{k+1} \in \mathbb{R}$. This follows from condition (B3) on continuity of $\theta \to u^{c,k}_{\theta}(z_{k+1}|z_1^k)$. Also, $w(\theta, \delta, \cdot)$ is dominated

by $2\bar{u}_\theta$ for all $\theta \in \Theta$, all $\delta < \delta_\theta$ and all $z_1, \ldots, z_{k+1} \in \mathbb{R}$. Of course δ_θ and $\bar{u}_\theta$ are chosen according to condition (B2). Hence, by dominated convergence $w(\theta, \delta, \cdot)$ is in $L^1(P_{\theta_0}^{k+1})$ for $\delta < \delta_\theta$ and $\mathrm{E}_{\theta_0} w(\theta, \delta, Z_1^{k+1}) \to 0$ as $\delta \to 0$ for all $\theta \in \Theta$.

Now let $\varepsilon > 0$. For each $\theta \in \Theta$ choose $\lambda_\theta \in (0, \delta_\theta]$ such that $w(\theta, \delta, \cdot)$ is in $L^1(P_{\theta_0}^{k+1})$ with $\mathrm{E}_{\theta_0} w(\theta, \delta, Z_1^{k+1}) < \varepsilon/4$ for all $\delta < \delta_\theta$ and $|J^k(\theta) - J^k(\theta')| < \varepsilon/4$ if $||\theta - \theta'|| < \lambda_\theta$ (recall that J^k is continuous). Let $B(\theta, \lambda) = \{\theta' \in \mathbb{R}^p : ||\theta - \theta'|| < \lambda\}$ be the ball with centre θ and radius λ. Then $\Theta \subseteq \cup_{\theta \in \Theta} B(\theta, \lambda_\theta)$ and since Θ is compact the open covering of Θ has a finite sub-covering. That is, $\theta_1, \ldots, \theta_m$ exist such that $\Theta \subseteq \cup_{j=1}^m B(\theta_j, \lambda_{\theta_j})$.

Consider a fixed $\theta \in \Theta$ and choose $j \in \{1, \ldots, m\}$ such that $\theta \in B(\theta_j, \lambda_{\theta_j})$. Then $||\theta - \theta_j|| < \lambda_{\theta_j}$ and

$$\left|U_n^k(\theta) - J^k(\theta)\right| \leq \left|U_n^k(\theta_j) - J^k(\theta_j)\right| + \left|U_n^k(\theta) - U_n^k(\theta_j)\right| + \left|J^k(\theta_j) - J^k(\theta)\right|.$$

Here, the first term only depends on θ_j and the third term is smaller than $\varepsilon/4$. For the second term, note that

$$\begin{aligned}
\left|U_n^k(\theta) - U_n^k(\theta_j)\right| &= \left|\frac{1}{n} \sum_{i=k+1}^{n-1} \left(u_\theta^{c,k}(Z_{i+1}|Z_{i-k+1}^i) - u_{\theta_j}^{c,k}(Z_{i+1}|Z_{i-k+1}^i)\right)\right| \\
&\leq \frac{1}{n} \sum_{i=k+1}^{n-1} \left|u_\theta^{c,k}(Z_{i+1}|Z_{i-k+1}^i) - u_{\theta_j}^{c,k}(Z_{i+1}|Z_{i-k+1}^i)\right| \\
&\leq \frac{1}{n} \sum_{i=k+1}^{n-1} w\left(\theta_j, \lambda_{\theta_j}, Z_{i-k+1}^{i+1}\right) \\
&\leq \left|\frac{1}{n} \sum_{i=k+1}^{n-1} w\left(\theta_j, \lambda_{\theta_j}, Z_{i-k+1}^{i+1}\right) - \mathrm{E}_{\theta_0} w\left(\theta_j, \lambda_{\theta_j}, Z_1^{k+1}\right)\right| + \varepsilon/4
\end{aligned}$$

which only depends on θ_j. It follows that the supremum of $|U_n^k(\theta) - J^k(\theta)|$ over Θ is bounded by the maximum over $\{\theta_1, \ldots, \theta_m\}$:

$$\begin{aligned}
\sup_{\theta \in \Theta} \left|U_n^k(\theta) - J^k(\theta)\right| &\leq \max_{j=1,\ldots,m} \left|U_n^k(\theta_j) - J^k(\theta_j)\right| \\
&\quad + \max_{j=1,\ldots,m} \left|\frac{1}{n} \sum_{i=k+1}^{n-1} w\left(\theta_j, \lambda_{\theta_j}, Z_{i-k+1}^{i+1}\right) - \mathrm{E}_{\theta_0} w\left(\theta_j, \lambda_{\theta_j}, Z_1^{k+1}\right)\right| + \varepsilon/2.
\end{aligned}$$

Recall that λ_{θ_j} is chosen such that $w(\theta_j, \lambda_{\theta_j})$ is in $L^1(P_{\theta_0}^{k+1})$. Also, $u_{\theta_j}^{c,k}$ is in $L^1(P_{\theta_0}^{k+1})$ by condition (B2). Hence, by the ergodic theorem, the two first terms converge to zero in P_{θ_0}-probability and the lemma follows immediately. □

With these lemmas in hand it is easy to prove consistency of $\hat{\theta}_n^k$:

Theorem III.7 *Under Assumptions III.1 and III.4, $\hat{\theta}_n^k$ is consistent for θ_0, that is, $\hat{\theta}_n^k \to \theta_0$ in probability wrt. P_{θ_0} as $n \to \infty$.*

Proof By assumption, Θ is compact and $\theta \to U_n^k(\theta)$ is continuous. The function J^k is well-defined under (B2) and continuous under (B3), see Lemma III.6. Since J^k is defined on a compact set J^k is even uniformly continuous.

For $\eta > 0$ define $W_n(\eta) = \sup_{||\theta-\theta'||\leq\eta} \left|U_n^k(\theta) - U_n^k(\theta')\right|$. Then $\hat{\theta}_n^k$ is consistent if

$$P_{\theta_0}\big(W_n(\eta) \geq 2\psi(\eta)\big) \to 0, \quad n \to \infty \tag{III.19}$$

where $\psi : [0,\infty) \to \mathbb{R}$ satisfies $\lim_{\eta\to 0} \psi(\eta) = 0$ (Dacunha-Castelle & Duflo 1986, Theorem 3.2.8).

By the triangle inequality

$$\begin{aligned} W_n(\eta) &\leq \sup_{||\theta-\theta'||\leq\eta} \Big(\left|U_n^k(\theta) - J^k(\theta)\right| + \left|J^k(\theta) - J^k(\theta')\right| + \left|U_n^k(\theta') - J^k(\theta')\right| \Big) \\ &\leq 2 \sup_{\theta\in\Theta} \left|U_n^k(\theta) - J^k(\theta)\right| + \sup_{||\theta-\theta'||\leq\eta} \left|J^k(\theta) - J^k(\theta')\right|. \end{aligned}$$

Here, the first term converges to zero in P_{θ_0}-probability, cf. Lemma III.6 above. The second term is deterministic and converges to zero as $\eta \to 0$ since J^k is uniformly continuous. Hence, with $\psi(\eta) = \sup_{||\theta-\theta'||\leq\eta} \left|J^k(\theta) - J^k(\theta')\right|$, condition (III.19) and thus consistency of $\hat{\theta}_n^k$ follows. □

III.5.2 Asymptotic normality

We now turn to asymptotic normality of $\hat{\theta}_n^k$. Assume that the criterion function U_n^k is continuously differentiable (Assumption (C2) below) and let $\dot{U}_n^k$ denote the p-vector of first derivatives. Then any minimizer of U_n^k is either on the boundary of Θ or solves the equation $\dot{U}_n^k(\theta) = 0$. In the latter case the theory of estimating functions applies, see Sørensen (1998*b*), for example. Theorem III.9 below claims that (with a probability tending to one) there exists a solution to $\dot{U}_n^k(\theta) = 0$ and that the solution is asymptotically normal. Let us be more specific about the regularity conditions:

Assumption III.8 Let Θ° denote the set of inner points of Θ and assume that

(C1) the true parameter θ_0 is an inner point of Θ, *i.e.* $\theta_0 \in \Theta^\circ$;

(C2) the functions $\theta \to p_\theta^{c,k}(z_{k+1}|z_1^k)$ are twice continuously differentiable for all $z_1, \ldots, z_{k+1} \in \mathbb{R}$.

Then $\theta \to u_\theta^{c,k}(z_{k+1}|z_1^k)$ is twice continuously differentiable as well. Let $\dot{u}_\theta^{c,k} = (\dot{u}_{\theta,j}^{c,k})_{j=1,\ldots,p} = (\partial_{\theta_j} u_\theta^{c,k})_{j=1,\ldots,p}$ denote the p-vector of first derivatives and let $\ddot{u}_\theta^{c,k} = (\ddot{u}_{\theta,jl}^{c,k})_{j,l=1,\ldots,p} = (\partial_{\theta_j}\partial_{\theta_l} u_\theta^{c,k})_{j,l=1,\ldots,p}$ be the symmetric $p \times p$-matrix of second derivatives of $u_\theta^{c,k}$. Assume furthermore that

(C3) there exists an $\eta > 0$ such that $\dot{u}^{c,k}_{\theta_0,j}$ is in $L^{2+\eta}(P^{k+1}_{\theta_0})$ for all $j = 1,\ldots,p$ and such that the α-mixing coefficients for Z corresponding to θ_0 satisfy $\sum_{m=1}^{\infty} \alpha_Z(m)^{2/(2+\eta)} < \infty$;

(C4) there is a neighbourhood T_0 of θ_0 such that for all $\theta \in T_0$ and all $j,l = 1,\ldots,p$ there is a constant $\delta_{\theta,jl} > 0$ and a function $\bar{u}_{\theta,jl} : \mathbb{R}^{k+1} \to \mathbb{R}$ in $L^1(P^{k+1}_{\theta_0})$ such that for all $z_1,\ldots,z_{k+1} \in \mathbb{R}$, $\sup_{\theta' \in T_{\theta,\delta_{\theta,jl}}} \left|\ddot{u}^{c,k}_{\theta',jl}(z_{k+1}|z_1^k)\right| \le \bar{u}_{\theta,jl}(z_1^{k+1})$ where, as before, $T_{\theta,\delta} = \{\theta' \in \Theta : ||\theta - \theta'|| \le \delta\}$;

(C5) the symmetric $p \times p$ matrix

$$A^k(\theta_0) = P^{k+1}_{\theta_0}\left(\ddot{u}^{c,k}_{\theta_0}\right) = \mathrm{E}_{\theta_0}\, \ddot{u}^{c,k}_{\theta_0}(Z_{k+1}|Z_1^k)$$

is positive definite. □

Under (C2), U_n^k is twice continuously differentiable with first derivative given by the p-vector $\dot{U}_n^k(\theta) = \frac{1}{n}\sum_{i=k+1}^{n-1} \dot{u}^{c,k}_{\theta}(z_{i+1}|z_{i-k+1}^i)$ and second derivative given by the $p \times p$ matrix $\ddot{U}_n^k = \frac{1}{n}\sum_{i=k+1}^{n-1} \ddot{u}^{c,k}_{\theta}(z_{i+1}|x_{i-k+1}^i)$. Any minimizer of U_n^k is either on the boundary of Θ or solves $\dot{U}_n^k(\theta) = 0$. In particular, under (C1), any minimizer of U_n^k that is consistent for θ_0 solves the estimating equation (with a probability tending to one).

Note that the estimating function $\dot{U}_n^k$ is unbiased, that is, $\mathrm{E}_\theta \dot{U}_n^k(\theta) = 0$ for all $\theta \in \Theta^\circ$. Indeed,

$$\mathrm{E}_\theta\, \dot{u}^{c,k}_{\theta,j}(Z_{k+1}|Z_1^k) = \mathrm{E}_\theta\, \mathrm{E}_\theta\left(\dot{u}^{c,k}_{\theta,j}(Z_{k+1}|Z_1^k)\middle|Z_1^k\right)$$

and, with obvious notation for the derivatives of $p^{c,k}_\theta$ (and if differentiation wrt. θ_j and integration wrt. z_{k+1} are interchangeable),

$$\mathrm{E}_\theta\left(\dot{u}^{c,k}_{\theta,j}(Z_{k+1})\middle|Z_1^k = z_1^k\right) = -\int \dot{p}^{c,k}_{\theta,j}(z|z_1^k)\,dz = -\frac{\partial}{\partial\theta_j}\int p^{c,k}_\theta(z|z_1^k)\,dz = 0$$

for all $z_1,\ldots,z_k \in \mathbb{R}$ and all $j = 1,\ldots,p$.

It is essential for the proof below that the estimating function itself evaluated at the true parameter value and scaled by $n^{1/2}$ converges in distribution. Under condition (C3) this follows from the central limit theorem for α-mixing processes in the appendix (Theorem III.12.2). To be specific let $\zeta_j(z_{i-k+1}^{i+1})$ be short for $\dot{u}^{c,k}_{\theta_0,j}(z_{i+1}|z_{i-k+1}^i)$; then Theorem III.12.2 claims that the $p \times p$ matrix Γ^k defined coordinate-wise by

$$\begin{aligned}\Gamma^k_{jl}(\theta_0) &= \mathrm{E}_{\theta_0}\left(\zeta_j(Z_1^{k+1})\zeta_l(Z_1^{k+1})\right) \\ &\quad + \sum_{m=1}^{\infty}\left\{\mathrm{E}_{\theta_0}\left(\zeta_j(Z_1^{k+1})\zeta_l(Z_{m+1}^{k+m+1})\right) + \mathrm{E}_{\theta_0}\left(\zeta_l(Z_1^{k+1})\zeta_j(Z_{m+1}^{k+m+1})\right)\right\}\end{aligned}$$

is well-defined and that $n^{1/2}\dot{U}_n^k(\theta_0) \to N(0,\Gamma^k(\theta_0))$.

Condition (C4) ensures integrability of $\ddot{u}_{\theta_0}^{c,k}$ and suitably uniform convergence in probability of $\ddot{U}_n^k(\theta_0)$ to $A^k(\theta_0)$. Note that if integration and twice differentiation can be interchanged, then $P_{\theta_0}^{k+1}(\ddot{p}_{\theta_0}^{c,k}/p_{\theta_0}^{c,k}) = 0$ and

$$A^k(\theta_0) = P_{\theta_0}^{k+1}\big(\ddot{u}_{\theta_0}^{c,k}\big) = P_{\theta_0}^{k+1}\Big(\big(\dot{u}_{\theta_0}^{c,k}\big)\big(\dot{u}_{\theta_0}^{c,k}\big)^T\Big),$$

that is, $A^k(\theta_0)$ equals the first term in Γ^k. The condition that $A^k(\theta_0)$ is positive definite is an identifiability condition.

Theorem III.9 *Suppose that Assumptions III.1 and III.8 hold. Then a solution $\hat{\theta}_n^k$ to $\dot{U}_n^k(\theta) = 0$ exists with a probability tending to 1 as $n \to \infty$. Moreover*

$$\sqrt{n}\big(\hat{\theta}_n^k - \theta_0\big) \to N\big(A^k(\theta_0)^{-1}\Gamma^k(\theta_0)A^k(\theta_0)^{-1}\big). \tag{III.20}$$

Proof It follows from Corollary 2.5 and Theorem 2.8 in Sørensen (1998*b*) that it is sufficient to show

$$n^{1/2}\dot{U}_n^k(\theta_0) \to N\big(0,\Gamma^k(\theta_0)\big) \tag{III.21}$$

in distribution wrt. P_{θ_0} as $n \to \infty$ and

$$\sup_{\theta \in T_{\theta_0,\eta/\sqrt{n}}} \big|\ddot{U}_{n,jl}^k(\theta) - A^k(\theta_0)\big| \to 0 \tag{III.22}$$

in probability wrt. P_{θ_0} as $n \to \infty$ for all $\eta > 0$ and all $j,l \in \{1,\dots,p\}$.

As already noted (III.21) follows immediately from condition (C3) and Theorem III.12.2. In order to show (III.22) define $A^k(\theta) = P_{\theta_0}^{k+1}\big(\ddot{u}_\theta^{c,k}\big)$ for $\theta \in T_0$ and let $j,l \in \{1,\dots,p\}$ and $\eta > 0$ be fixed. By the triangle inequality

$$\big|\ddot{U}_{n,jl}^k(\theta) - A^k(\theta_0)\big| \le \big|\ddot{U}_{n,jl}^k(\theta) - A^k(\theta)\big| + \big|A^k(\theta) - A^k(\theta_0)\big|.$$

Choose N large enough that $T_{\theta_0,\eta/\sqrt{N}} \subseteq T_0$. Then, for $n \ge N$, $A^k(\theta)$ is well-defined for all $\theta \in T_{\theta_0,\eta/\sqrt{n}}$. By arguments almost identical to those in the proof of Theorem III.6 it now follows that

$$\sup_{\theta \in T_{\theta_0,\eta/\sqrt{N}}} \big|\ddot{U}_{n,jl}^k(\theta) - A^k(\theta)\big| \to 0$$

in P_{θ_0}-probability as $n \to \infty$ (recall that $T_{\theta_0,\eta/\sqrt{N}}$ is compact).

Also, A^k is continuous in θ_0. The convergence (III.22) follows immediately. This proves both the existence assertion and the convergence result (III.20). □

Note that although asymptotic normality is indeed a nice property of the estimator, it is difficult to use in practice as we are not able to compute the asymptotic variance. Also, the above conditions are all expressed in terms of the distribution of Z and are thus in general difficult (if possible at all) to check. The condition on the α-mixing coefficients in (C3) is an exception: we showed in the proof of Proposition III.2 that $\alpha_Z(j) \leq \alpha_V\big((j-1)\Delta\big)$ for all $j \geq 2$ so it is sufficient that the condition holds for the α-mixing coefficients for V. See Genon-Catalot *et al.* (1998*b*), for example, for conditions ensuring exponential decay of the α-mixing coefficients for V.

Recall from (III.14) that for $\xi \equiv 0$ and $k = 0$, the identifiability condition (B4) holds if and only if the distributions of S_1 corresponding to two values θ and θ' differ when $\theta \neq \theta'$. We have no similar results for larger values of k. For the remaining conditions recall that

$$u_\theta^{c,k}(z_{k+1}|z_1^k) = -\log \mathrm{E}_{\pi_\theta^{k+1}} \prod_{i=1}^{k+1} \varphi(z_i, H_i) + \log \mathrm{E}_{\pi_\theta^k} \prod_{i=1}^{k} \varphi(z_i, H_i)$$

where $H_i = (M_i, S_i)$, π_θ^l is the distribution of $H^l = (H_1, \ldots, H_l)$ and $\varphi(\cdot, h) = \varphi(\cdot, m, s)$ is the density of $N(m, s)$ for $h = (m, s)$. Hence, the continuity, differentiability and local integrability conditions imposed on $u_\theta^{c,k}$ would follow from roughly similar conditions on the densities of H_1^k and H_1^{k+1}. This is not very helpful though, since we have no explicit expression for the latter densities either.

Finally, it is important to stress that the above results hold for fixed value of k (and Δ) as $n \to \infty$. In particular, the above results do *not* imply nice asymptotic behaviour of the maximum likelihood estimator (which corresponds to $k = k(n) = n-1$). The problem is of course that the terms in the log-likelihood function U_n^{n-1} originate from *different* functions ($p_\theta^{c,i}$ for observation z_{i+1}) such that the usual limit theorems do not apply.

As noted in Section III.2 we can think of the model as a hidden Markov model with continuous, unbounded state space of the hidden chain $\tilde{H}$ given by $\tilde{H}_i = (V_{i\Delta}, M_i, S_i)$. Asymptotic results for the maximum likelihood estimator have been proved for hidden Markov models where the state space for the hidden chain is finite (Bickel & Ritov 1996, Bickel, Ritov & Rydén 1998) or compact (Jensen & Petersen 1999). Neither approach can be applied in our setting and there are in fact no results in the literature concerning asymptotic properties of the maximum likelihood estimator for the models considered in this paper.

III.6 Efficiency considerations

In this section we briefly discuss how the number of lags k influence the quality of the estimators. The subject is essential but unfortunately we have not been able to prove very powerful results.

Intuitively we would expect estimators to improve as the number of lags increases. With the asymptotic normality from Theorem III.9 in hand we could

compare estimators for different k's by their asymptotic variances and hope that the variance is decreasing as a function of k (for symmetric positive semi-definite matrices A and B we write $A \leq B$ if and only if the difference $B-A$ is positive semi-definite). We have not been able to prove results like this! The problem is of course that the expression for the asymptotic variance is so complicated that comparison between different k's is impossible, even for a one-dimensional parameter.

The simulation study in Section III.7 indicates that minimization of U_n^k in practice may give rise to identification problems even if the k-lag conditional distribution uniquely determines the parameter (theoretically). In the simulation study this is reflected in very oblong level curves corresponding to small values of certain linear combinations of the coordinates in $\mathrm{E}_{\theta_0} \ddot{U}_n^k(\theta_0)$ and thereby (ignoring the matrix $\Gamma(\theta_0)$ in (III.20)) to large asymptotic variance of the estimator. In the simulation study the problem seems to diminish as we use larger values of k suggesting that estimation in fact improves as k increases. On the other hand, in a simpler situation with no identification problems for any value of k we did not find any substantial differences among the estimators for different values of k.

Note that we in principle could improve estimation by introducing weight functions. To be specific, consider estimating functions on the form

$$D_n^k(\theta) = \frac{1}{n} \sum_{i=k}^{n-1} d_i(Z_{i-k+1}^i, \theta)\, \dot{u}_\theta^{c,k}(Z_{i+1}|Z_{i-k+1}^i)$$

where $d_k, \ldots, d_{n-1}$ are function from $\mathbb{R}^k \times \Theta$ to $\mathbb{R}$. Note that we for simplicity have left out the contribution from the first k observations and that $\dot{U}_n^k$ (except for the first term in U_n^k) corresponds to $d_i \equiv 1$, $i = k, \ldots, n-1$.

The estimating function D_n^k is unbiased since for each $i = k \ldots, n-1$

$$\begin{aligned}
&\mathrm{E}_{\theta_0} d_i(Z_{i-k+1}^i, \theta_0) u_{\theta_0}^{c,k}(Z_{i+1}|Z_{i-k+1}^i) \\
&\quad = \mathrm{E}_{\theta_0} d_i(Z_{i-k+1}^i) \Big(\mathrm{E}_{\theta_0} u_{\theta_0}^{c,k}(Z_{i+1}|Z_{i-k+1}^i)|Z_{i-k+1}^i \Big) = 0.
\end{aligned}$$

Under regularity conditions similar to those of Assumption III.8 the solution to $D_n^k(\theta) = 0$ is a consistent and asymptotically normal estimator of θ. By choosing the functions d_i cleverly we can obtain smaller asymptotic variance than is the case for $\hat{\theta}_n^k$, see Sørensen (1999) for similar considerations. This is only of theoretical interest, though! In order to find the optimal weight functions one must invert an $(n-k) \times (n-k)$ matrix (which depends on θ and whose entries we do not even know explicitly).

Finally, we prove a result concerning the approximate log-likelihood functions U_n^k rather than the corresponding estimators: the limit, in probability, of $U_n^k(\theta_0)$ is decreasing in k. It holds only for U_n^k evaluated at the true parameter and is thus not very useful in practice. Nevertheless it tells us that the approximations of the likelihood improve in this sense.

Proposition III.10 *Let $0 \leq k' \leq k''$ and assume that Condition (B2) is satisfied for $\theta = \theta_0$ and $k = k'$ and $k = k''$. Then*

$$\mathrm{E}_{\theta_0} u_{\theta_0}^{c,k''}\big(Z_{k''+1}|Z_1^{k''}\big) \leq \mathrm{E}_{\theta_0} u_{\theta_0}^{c,k'}\big(Z_{k'+1}|Z_1^{k'}\big).$$

Consequently, $\mathrm{E}_{\theta_0} U_n^{k''}(\theta_0) \leq \mathrm{E}_{\theta_0} U_n^{k'}(\theta_0)$ *and* $\lim_{n\to\infty} U_n^{k''}(\theta_0) \leq \lim_{n\to\infty} U_n^{k'}(\theta_0)$ *where convergence means convergence in P_{θ_0}-probability.*

Proof It will suffice to consider $k' = k$ and $k'' = k+1$ for $k \geq 0$ arbitrary. By stationarity it follows that it is sufficient to show that

$$\mathrm{E}_{\theta_0} u_{\theta_0}^{c,k+1}\big(Z_{k+2}|Z_1^{k+1}\big) \leq \mathrm{E}_{\theta_0} u_{\theta_0}^{c,k}\big(Z_{k+2}|Z_2^{k+1}\big). \tag{III.23}$$

By definition,

$$u_{\theta_0}^{c,k+1}\big(Z_{k+2}|Z_1^{k+1}\big) - u_{\theta_0}^{c,k}\big(Z_{k+1}|Z_1^{k}\big) = \log \frac{p_{\theta_0}^{c,k}\big(Z_{k+2}|Z_2^{k+1}\big)}{p_{\theta_0}^{c,k+1}\big(Z_{k+2}|Z_1^{k+1}\big)}$$

so Jensen's equality yields

$$\mathrm{E}_{\theta_0}\Big(u_{\theta_0}^{c,k+1}\big(Z_{k+2}|Z_1^{k+1}\big) - u_{\theta_0}^{c,k}\big(Z_{k+2}|Z_2^{k+1}\big)\Big) \leq \log \mathrm{E}_{\theta_0} \frac{p_{\theta_0}^{c,k}\big(Z_{k+2}|Z_2^{k+1}\big)}{p_{\theta_0}^{c,k+1}\big(Z_{k+2}|Z_1^{k+1}\big)}.$$

Calculations similar to those in the proof of Lemma III.5 show that the latter expectation is one, which yields (III.23). The expectation assertion follows immediately by

$$U_n^{k+1}(\theta_0) - U_n^{k}(\theta_0) = \frac{1}{n}\sum_{k=1}^{n-1}\Big(u_{\theta_0}^{c,k+1}\big(Z_{i+1}|Z_{i-k}^{i}\big) - u_{\theta_0}^{c,k}\big(Z_{i+1}|Z_{i-k+1}^{i}\big)\Big)$$

and the convergence result follows by the ergodic theorem. □

III.7 Example: The Cox-Ingersoll-Ross process

In this section we discuss a particular model, namely the model where the observed X-process has no drift, and the volatility process V is a *Cox-Ingersoll-Ross process*. This specification of the volatility process was first considered by Hull & White (1987) and later by Heston (1993).

III.7.1 Basic properties

The model is given by the stochastic differential equations

$$\begin{aligned} dX_t &= \sqrt{V_t}\, dW_t, \quad X_0 = U_X \\ dV_t &= \alpha(\beta - V_t)\, dt + \sigma\sqrt{V_t}\, d\tilde{W}_t, \quad V_0 = U_V. \end{aligned} \tag{III.24}$$

with parameter $\theta = (\alpha, \beta, \sigma)$. Let $\Theta = \{(\alpha, \beta, \sigma) : \alpha, \beta, \sigma > 0, \sigma^2 \leq 2\alpha\beta\}$. It is well-known that for $(\alpha, \beta, \sigma) \in \Theta$, V is positive, stationary and α-mixing, that is, Assumption III.1 is satisfied. Actually, the α-mixing coefficients decrease at exponential rate (Genon-Catalot *et al.* 1998*b*, Corollary 1.1) so the condition on the α-mixing coefficients in condition (C3) is satisfied for $\theta \in \Theta$.

The invariant distribution is the Gamma distribution with shape parameter $2\alpha\beta/\sigma^2$ and scale parameter $\sigma^2/(2\alpha)$. The transition probabilities are known to be non-central χ^2-distributions. The parameter β is simply the mean value of V whereas the "mean reverting parameter" α can be interpreted as the size of the force pulling the process back to its mean.

Figure III.2 shows simulated data from the model with $\Delta = 1$ and parameter $(\alpha, \beta, \sigma) = (0.1, 1, 0.35)$. The bottom figure shows a simulated path of the V-process from time 0 to time 500 and the top figure shows increments $Z_i = X_{i\Delta} - X_{(i-1)\Delta}$ of X for $i = 1, \ldots, 500$. Clearly the increments are more volatile in periods with relatively large values of the volatility process V than in periods with low values of V.

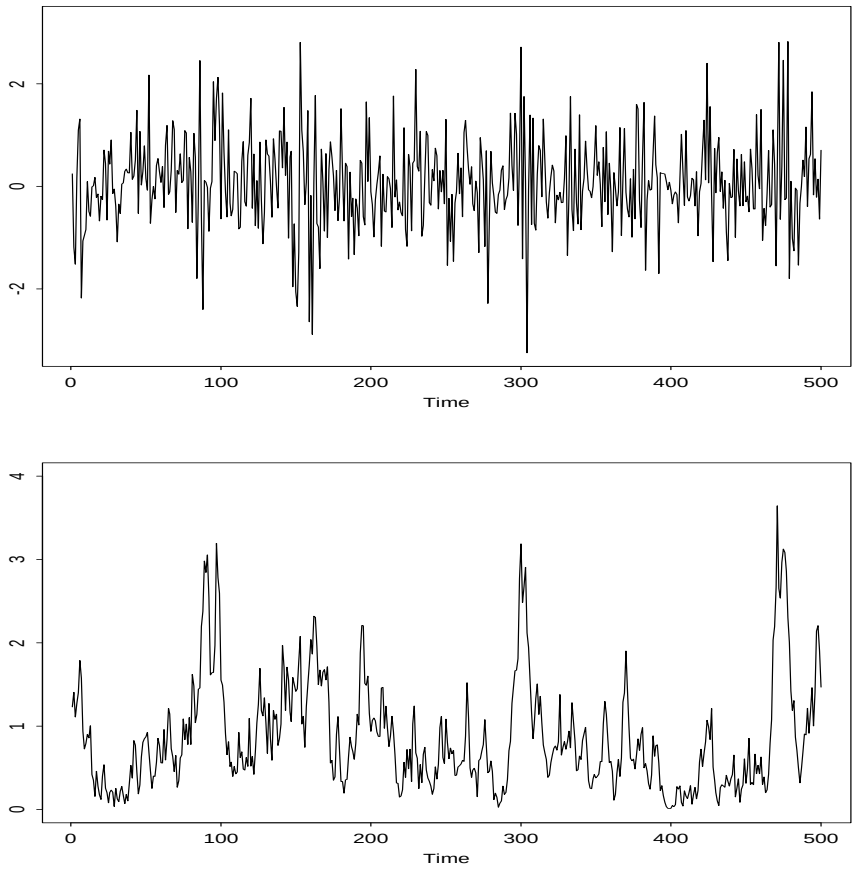

Figure III.2: Simulated values of $Z_i = X_{i\Delta} - X_{(i-1)\Delta}$ (top) and $V_{i\Delta}$ (bottom) from the Cox-Ingersoll-Ross model for $\Delta = 1$ and $i = 1, \ldots, n$ where $n = 500$. The model parameter is $(\alpha, \beta, \sigma) = (0.1, 1, 0.35)$.

Figure III.3 is a QQ-plot of the increments and we see that they are far too

heavy-tailed to be Gaussian. Figure III.4 shows the correlogram for the incre-

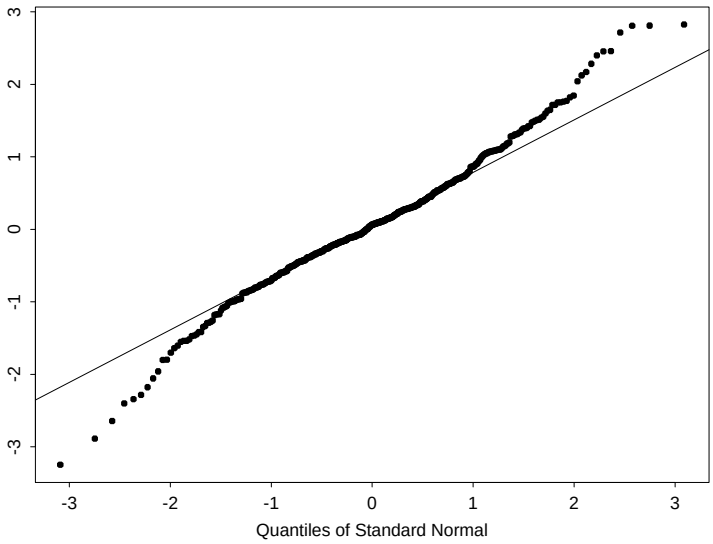

Figure III.3: QQ-plot for the data in the top of Figure III.2; quantiles of the standard normal distribution at the *x*-axis, quantiles of data at the *y*-axis.

ments to the left and for the squared increments to the right. The dashed lines provide approximate 95%-confidence intervals. Recall from (III.7) that Z_i and Z_j are uncorrelated for $i \neq j$ since $\xi \equiv 0$. From the right figure we see that correlation between squared observations is small from lag 9, say, and onwards.

If V is started stationarily, $V_0 = U_V \sim \Gamma(2\alpha\beta/\sigma^2, \sigma^2/(2\alpha))$, then it is easy to calculate various moments in the model. Most of the results in the following proposition are known from Genon-Catalot *et al.* (1998*b*).

Proposition III.11 *Let $\theta = (\alpha, \beta, \sigma) \in \Theta$ and assume that V is started according to the invariant distribution: $V_0 = U_V \sim \Gamma(2\alpha\beta/\sigma^2, \sigma^2/(2\alpha))$. For the unobserved V-process it holds for $s, t \geq 0$ that*

$$\mathrm{E}_\theta V_t = \beta; \quad \mathrm{Var}_\theta V_t = \frac{\beta\sigma^2}{2\alpha}; \quad \mathrm{Cov}_\theta(V_s, V_t) = \frac{\beta\sigma^2}{2\alpha} e^{-\alpha|t-s|}.$$

For the unobserved, integrated variables S_i, $i \in \mathbb{N}$:

$$\mathrm{E}_\theta S_i = \beta\Delta \tag{III.24}$$

$$\mathrm{Var}_\theta S_i = \frac{\beta\sigma^2}{\alpha^3}\left(\alpha\Delta - 1 + e^{-\alpha\Delta}\right) \tag{III.25}$$

$$\mathrm{Cov}_\theta(S_i, S_j) = \frac{\beta\sigma^2}{2\alpha^3} e^{-\alpha\Delta(j-i-1)}\left(1 - e^{-\alpha\Delta}\right)^2, \quad j > i. \tag{III.26}$$

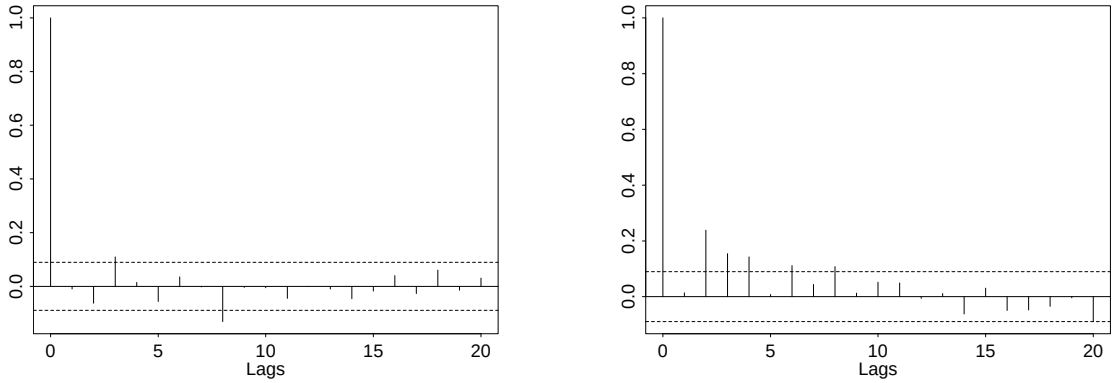

Figure III.4: Correlogram for the data from the top of Figure III.2 (to the left) and the same data squared (to the right). The dashed lines give approximate 95%-confidence intervals.

For the observed increments Z_i, $i \in \mathbb{N}$:

$$\begin{aligned}
\mathrm{E}_\theta Z_i &= 0 \\
\mathrm{Var}_\theta Z_i = \mathrm{E}_\theta Z_1^2 &= \mathrm{E}_\theta S_1 = \beta\Delta \\
\mathrm{Var}_\theta Z_i^2 = 3\,\mathrm{Var}_\theta S_1 + 2(\mathrm{E}_\theta S_1)^2 &= 2\beta^2\Delta^2 + \frac{3\beta\sigma^2}{\alpha^3}\left(\alpha\Delta - 1 + e^{-\alpha\Delta}\right) \\
\mathrm{Cov}_\theta(Z_i^2, Z_j^2) = \mathrm{Cov}_\theta(S_i, S_j) &= \frac{\beta\sigma^2}{2\alpha^3} e^{-\alpha\Delta(j-i-1)}\left(1 - e^{-\alpha\Delta}\right)^2, \quad j > i.
\end{aligned}$$

Proof The expressions for V follow immediately by the Gamma distribution, stationarity and the well-known formula $\mathrm{E}_\theta(V_t | V_0 = v) = e^{-\alpha t}v + \beta\left(1 - e^{-\alpha t}\right)$ for the conditional expectation.

Recall that $S_i = \int_{(i-1)\Delta}^{i\Delta} V_s\,ds$. Stationarity of $(S_1, S_2, \dots)$ follows by stationary of V, and $\mathrm{E}_\theta S_1 = \int_{(i-1)\Delta}^{i\Delta} \mathrm{E}_\theta V_s\,ds = \beta\Delta$. For $l \geq 1$,

$$\begin{aligned}
\mathrm{E}_\theta S_1 S_l &= \mathrm{E}_\theta\left(\int_0^\Delta V_u\,du\right)\left(\int_{(l-1)\Delta}^{l\Delta} V_s\,ds\right) \\
&= \int_0^\Delta \int_{(l-1)\Delta}^{l\Delta} \mathrm{E}_\theta V_s V_u\,du\,ds \\
&= \int_0^\Delta \int_{(l-1)\Delta}^{l\Delta} \left(\frac{\beta\sigma^2}{2\alpha} e^{-\alpha|u-s|} + \beta^2\right) du\,ds,
\end{aligned}$$

and straightforward calculations and subtraction of $\beta^2\Delta^2$ yield the variance of S_1 for $l = 1$ and the covariance between S_1 and S_l for $l \geq 2$.

The expressions for the moments of Z_i follow immediately by Theorem III.2 and the moments of S_i. For example,

$$\mathrm{Var}_\theta Z_1^2 = \mathrm{E}_\theta Z_1^4 - \left(\mathrm{E}_\theta Z_1^2\right)^2 = 3\,\mathrm{E}_\theta S_1^2 - (\mathrm{E}_\theta S_1)^2 = 3\,\mathrm{Var}_\theta S_1 + 2(\mathrm{E}_\theta S_1)^2. \qquad \square$$

(III.27)

Note that $\mathrm{Var}_\theta Z_1^2 > 3\,\mathrm{Var}_\theta S_1$ and that $\mathrm{Cov}_\theta(Z_1^2, Z_j^2) = \mathrm{Cov}_\theta(S_1, S_j) > 0$ is decreasing at exponential rate. The correlation between Z_1^2 and Z_j^2 is thus positive, exponentially decreasing and at most 1/3 for all $j \geq 2$. In fact it is at most 1/5 which can be seen as follows:

$$\mathrm{Corr}_\theta(Z_1^2, Z_j^2) = \frac{\sigma^2\big(1-e^{-\alpha\Delta}\big)^2}{4\alpha^3\beta\Delta^2 + 6\sigma^2\big(\alpha\Delta - 1 + e^{-\alpha\Delta}\big)} e^{-\alpha\Delta(j-2)}$$

which is increasing in σ^2. For fixed α and β, the correlation is hence maximal for $\sigma^2 = 2\alpha\beta$, with

$$\mathrm{Corr}_{(\alpha,\beta,(2\alpha\beta)^{1/2})}(Z_1^2, Z_j^2) = \frac{\big(1-e^{-\alpha\Delta}\big)^2}{2\alpha^2\Delta^2 + 6\big(\alpha\Delta - 1 + e^{-\alpha\Delta}\big)} e^{-\alpha\Delta(j-2)}.$$

The right hand side does not depend on β and is decreasing as a function of $\alpha\Delta$ with limit 1/5 as $\alpha\Delta \to 0$. Also note that the excess kurtosis $\mathrm{E}_\theta Z_1^4 / (\mathrm{E}_\theta Z_1^2)^2 - 3$ is at most 3. Hence, the model is not appropriate for data with very heavy tails or with large correlations between squared observations Z_1^2 and Z_j^2 for some $j \geq 2$.

III.7.2 A small simulation study

In the following we present a small simulation study. We have simulated 10 datasets of increments, $Z^1, \ldots, Z^{10}$, each consisting of $n = 500$ observations. The model parameters are

$$(\alpha, \beta, \sigma) = (\alpha_0, \beta_0, \sigma_0) = (0.1, 1, 0.35)$$

and the value of Δ is 1. Each dataset was simulated as follows: a V-process was simulated by the Millstein scheme with each interval $[(i-1)\Delta, i\Delta]$ split into 1000 subintervals; the integrals $S_1, \ldots, S_n$ were approximated as described in Section III.4; and $Z_1, \ldots, Z_n$ were finally drawn independently, Z_i from $N(0, S_i)$. One of the simulated datasets, Z^4, was shown in Figure III.2 together with the corresponding simulated V-values, and we shall use this dataset as example throughout the section.

In a real-world application we would not have observed the V-process at all, but in this simulation study we have saved the simulated values $V_0, V_\Delta, \ldots, V_{n\Delta}$ for each simulation. Hence, we can estimate the parameters from the V-process as well as from the Z's and thus get an idea of how much information is lost when Z rather than V is observed (see comments below).

By Proposition III.11 it is easy to calculate various moments of V, S and Z for the chosen values of α, β, σ and Δ. For example,

$$\begin{gathered}
\mathrm{E}_{\theta_0} V_0 = 1; \quad \mathrm{Var}_{\theta_0} V_0 = 0.613; \quad \mathrm{Corr}_{\theta_0}(V_0, V_\Delta) = 0.905 \\
\mathrm{E}_{\theta_0} S_1 = 1; \quad \mathrm{Var}_{\theta_0} S_1 = 0.593; \quad \mathrm{Corr}_{\theta_0}(S_1, S_2) = 0.936 \\
\mathrm{Var}_{\theta_0} Z_1 = \mathrm{E}_{\theta_0} Z_1^2 = 1; \quad \mathrm{E}_{\theta_0} Z_1^4 = 4.778; \quad \mathrm{Corr}_{\theta_0}(Z_1^2, Z_2^2) = 0.147.
\end{gathered}$$

We see that values of V at two consecutive time points $(i-1)\Delta$ and $i\Delta$ as well as two consecutive S's are strongly correlated, and that the excess kurtosis of Z is 1.778.

For later use, define $m_2(\theta) = \mathrm{E}_\theta Z_1^2$, $m_4(\theta) = \mathrm{E}_\theta Z_1^4$ and $m_{1,2}(\theta) = \mathrm{E}_\theta Z_1^2 Z_2^2$ and let for a given dataset

$$\tilde{m}_2 = \frac{1}{n}\sum_{i=1}^{n} Z_i^2; \quad \tilde{m}_4 = \frac{1}{n}\sum_{i=1}^{n} Z_i^4; \quad \tilde{m}_{1,2} = \frac{1}{n-1}\sum_{i=2}^{n} Z_{i-1}^2 Z_i^2$$

be the corresponding empirical moments. Also, let $\tilde{c}_{1,2} = (\tilde{m}_{1,2} - \tilde{m}_2^2)/(\tilde{m}_4 - \tilde{m}_2^2)$ be the first empirical autocorrelation coefficient. Table III.1 in Appendix III.B lists the average and $\tilde{m}_2$, $\tilde{m}_4$, $\tilde{m}_{1,2}$ and $\tilde{c}_{1,2}$ for the simulated datasets.

In the rest of this section we shall estimate the parameters α, β and σ from each of the ten simulated datasets. We consider three different set-ups; (A) only one parameter, say α, is considered unknown whereas the two others are known; (B) two parameters are considered unknown; (C) all three parameters are considered unknown. Cases (A) and (B) are of course not realistic but provide insight to the estimation problem.

In case (A) we compute the estimators $\hat{\alpha}_n^k$ for $k = 0, \ldots, 4$ although it turns out that even $k = 0$ yields satisfactory estimates. A comparison of the five values of k with respect to mean and variance (over the ten simulations) shows that $k = 1$ is the best choice and $k = 0$ the worst, but the difference between the five estimators is not substantial. In case (B) we use only $k = 0$ and $k = 1$; both values yield acceptable estimates as long as β is not the unknown parameter. The estimation problem in case (C) is more difficult, and we must use larger values of k, say $k = 4$. Still, the estimators are not completely satisfactory.

In each of the three cases we compare (i) with "method of moments estimators" (Genon-Catalot *et al.* 1998*b*), that is, estimators obtained by matching various empirical and theoretical moments; and (ii) with simple martingale estimators based on the V-data (Bibby & Sørensen 1995, Sørensen 1997). The latter would of course not be possible in practice. The moment estimators are quite bad and there are often existence problems. The estimators based on V are not surprisingly quite good. In case (A) the difference between the estimators based on V and the approximate maximum likelihood estimators based on Z is moderate, whereas it is very substantial in case (C)

Of course, the above results are only indications of the relations between the estimators. We cannot draw final conclusions from the simulation study, since it is based on only ten simulations. However, the study confirms that the method is indeed applicable in practice!

Now, let us go through the three cases in detail. For all the below computations of U_n^k we have used $N = 10$ and $R = 10.000$, *cf.* Section III.4. We start out gently and consider estimation of one parameter only.

Case (A): Estimation of one parameter

We choose α as the unknown parameter and consider $\beta = \beta_0 = 1$ and $\sigma = \sigma_0 = 0.35$ known. Recall that the true value of α is $\alpha_0 = 0.1$.

Figure III.5 shows the graphs of U_n^k for $k = 0, \ldots, 4$ and data Z^4 in the interval from 0.06 to 0.16. To see the curvature of the curves more clearly, we have plotted

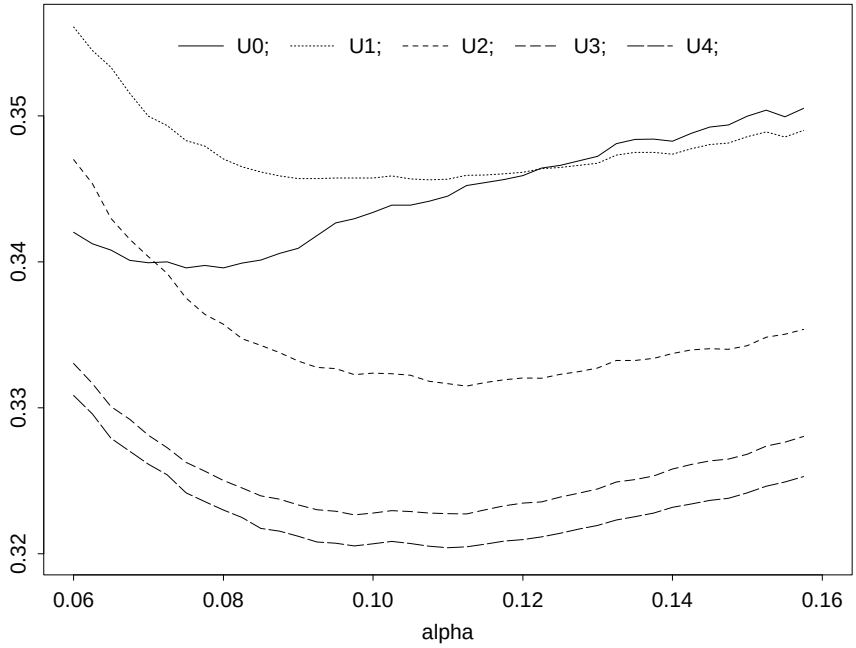

Figure III.5: Graphs of $\alpha \to U_n^k(\alpha, \beta_0, \sigma_0)$ for data Z^4, $k = 0, \ldots, 4$, $\beta_0 = 1$ and $\sigma_0 = 0.35$. The true value of α is $\alpha_0 = 0.1$.

the difference between the functions and their respective minima in Figure III.6. For this particular simulation, the curvatures of U_n^3 and U_n^4 are almost identical, and very similar to the curvature of U_n^2 and U_n^1. Hence, the corresponding estimates are close, around 0.105–0.110. The function U_n^0 has different curvature and minimum below 0.08.

The estimation results are graphically illustrated in the first five columns in Figure III.7. All five values of k yield reasonable estimators, with averages from 0.1027 ($k = 1$) to 0.1101 ($k = 0$) and standard errors from 0.0169 ($k = 1$) to 0.0281 ($k = 0$). In particular, the estimator $\hat{\alpha}_n^1$ is the best — and $\hat{\alpha}_n^0$ the worst — in this study both with respect to bias and variance. The difference between the five estimators is not substantial, though, and it is difficult to find any patterns in the differences. The values of the estimators are listed in columns two through six in Table III.2 in Appendix III.B.

For comparison we have also calculated the moment estimators $\tilde{\alpha}_n^4$ and $\tilde{\alpha}_n^{1,2}$, that is the estimators obtained by solving the equations $\tilde{m}_4 = m_4(\alpha, \beta_0, \sigma_0)$ and $\tilde{m}_{1,2} = m_{1,2}(\alpha, \beta_0, \sigma_0)$ respectively. The estimators are listed in the seventh and

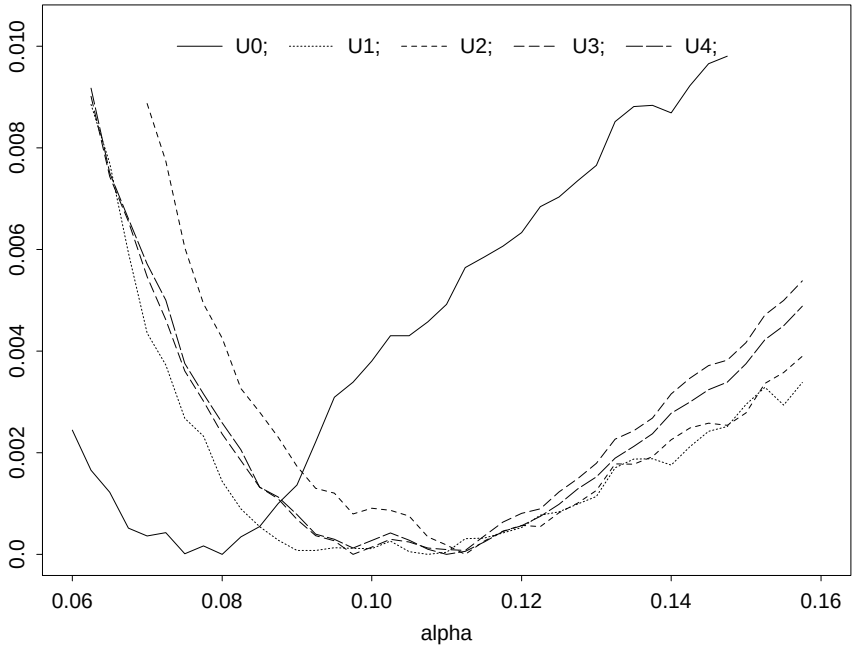

Figure III.6: Graphs of $\alpha \to U_n^k(\alpha, \beta_0, \sigma_0) - \min_\alpha U_n^k(\alpha, \beta_0, \sigma_0)$ for data Z^4, $k = 0, \ldots, 4$, $\beta_0 = 1$ and $\sigma_0 = 0.35$. The true value of α is $\alpha_0 = 0.1$.

eighth column of Table III.2 in Appendix III.B. For the datasets Z^3 and Z^4 the equations have no solution. The averages for the remaining eight datasets are 0.4111 and 0.1472 respectively so there is a considerable bias. The standard errors are large; 0.6117 and 0.2648 respectively. Of course, one could have chosen to match other moments, but note that neither the first three moments of Z nor $\mathrm{E}_\theta Z_1 Z_j$ depend on α. Hence, they cannot be used for estimation in case (A), and we are forced to use higher order moments like m_4 and $m_{1,2}$ as above.

Finally, we have estimated α from the volatility data $V_0, V_\Delta, \ldots, V_{n\Delta}$. Maximum likelihood estimation is in principle possible since the transition probabilities are known (non-central χ^2-distributions), but for simplicity we have used the martingale estimating equation

$$\sum_{i=1}^{n} \frac{\partial_\alpha F(V_{(i-1)\Delta}, \alpha, \beta_0)}{\Phi(V_{(i-1)\Delta}, \alpha, \beta_0)} \Big(V_{i\Delta} - F(V_{(i-1)\Delta}, \alpha, \beta_0) \Big) = 0$$

instead (Bibby & Sørensen 1995). Here, we have let $F(v, \alpha, \beta) = e^{-\alpha\Delta}(v - \beta) + \beta$ and $\sigma^2\Phi(v, \alpha, \beta) = \sigma^2\big((\beta - 2v)e^{-2\alpha\Delta} - 2(\beta - v)e^{-\alpha\Delta} + \beta\big)/(2\alpha)$ denote the expectation and the variance of the conditional distribution of V_Δ given $V_0 = v$. The weight function $\partial_\alpha F/\Phi$ is optimal in the sense that the corresponding estimator has the least asymptotic variance among martingale estimators based on the first order conditional moments (Bibby & Sørensen 1995). Note however that the maximum likelihood estimator would have even smaller asymptotic variance.

(III.31)

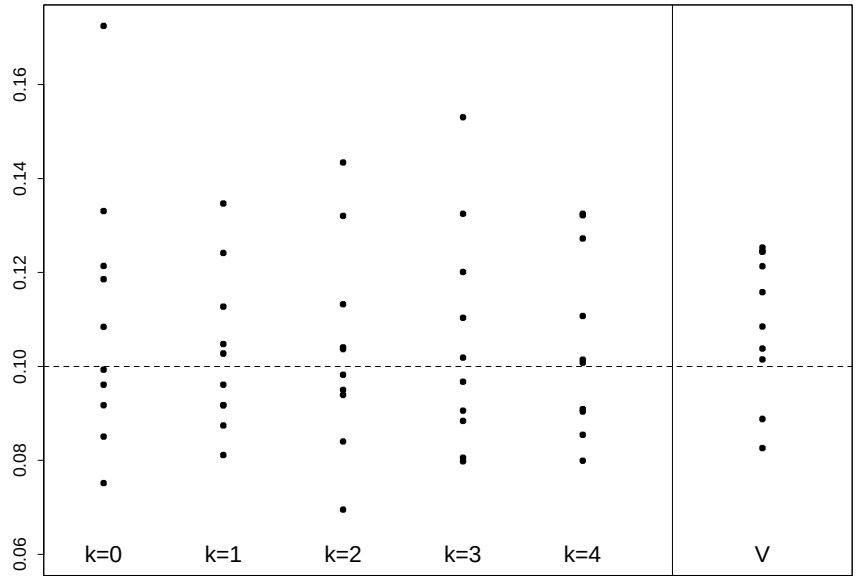

Figure III.7: The estimators $\hat{\alpha}_n^k$ for $k = 0, \ldots, 4$ (the first five columns) and the martingale estimator $\hat{\alpha}_n^V$ based on V (the last column). The true value of α is $\alpha_0 = 0.1$ (shown by the dashed line).

The martingale estimators are plotted in the last column of Figure III.7, and listed in the last column of Table III.2 in Appendix III.B. The average of $\hat{\alpha}_n^V$ is 0.1097. As one would expect, $\hat{\alpha}_n^V$ has smaller standard error (0.0154) than the estimators based on Z. It is slightly surprising that the standard error is only roughly 10% lower than that of $\hat{\alpha}_n^1$.

Case (B): Estimation of two parameters

We now very briefly consider estimation of (α, β) for $\sigma = \sigma_0 = 0.35$ known and estimation of (β, σ) for $\alpha = \alpha_0 = 0.1$ known. The combination with β known and (α, σ) unknown is much more diffucult as will be clear from the below discussion of case (C).

We use approximate maximum likelihood with $k = 0$ and $k = 1$, moment estimation (based on m_2 and $m_{1,2}$) and martingale estimation based on V. The estimators for β are very much alike. This is expectable as β is simply the variance of Z which is easy to estimate. For α and σ, respectively, the conclusions are essentially as in case (A) and we omit the details: the approximate maximum likelihood estimates are fine, the moments estimators are quite poor and the estimators based on V are superior.

Case (C): Estimation of all three parameters

Estimation of one or two parameters was succesful even for $k = 0$ and $k = 1$ (as long as β was one of the unknown parameters). The estimation problem is far more delicate when all three parameters are unknown, and a larger k is necessary in order to obtain reasonable estimates.

At first glance it seems promising to use $k = 1$ for estimation of all three parameters as well: by the moment considerations in Proposition III.11 it follows that the three-dimensional parameter is uniquely determined by the distribution of the pair (Z_1, Z_2) — and thereby presumably also by the conditional distribution of Z_2 given Z_1. Hence, U_n^1 should be able to distinguish between different parameter values.

In practice it turns out that U_n^1 is almost constant — and very close to its minimum — on a curve in $\mathbb{R}^3$. In other words: U_n^1 has difficulties distinguishing between parameters on this curve. This is perhaps not too surprising, though. One could suspect that only the marginal (invariant) distribution of V is easily determined. The invariant distribution of V is determined completely by two parameter functions, namely the shape parameter $2\alpha\beta/\sigma^2$ and the scale parameter $\sigma^2/(2\alpha)$. One could thus imagine these parameter functions — but not the parameters α, β and σ themselves — to be easy to estimate.

It is easy to get an estimate of the product β of the shape and the scale parameter; simply use the empirical second moment (or the empirical variance) divided by Δ,

$$\tilde{\beta}_n = \tilde{m}_2/\Delta = \frac{1}{\Delta}\sum_{i=1}^{n} Z_i^2. \tag{III.26}$$

But for given β the distribution of (Z_1, Z_2) wrt. $P_{(\alpha',\beta,\sigma')}$ and $P_{(\alpha'',\beta,\sigma'')}$ can be very much alike, though not the same, for (α', σ') and (α'', σ'') far from each other — as long as $(\sigma')^2/(2\alpha')$ is close to $(\sigma'')^2/(2\alpha'')$.

This is illustrated by Figure III.8 where we consider level curves for the moment functions

$$(\alpha, \sigma^2) \to m_4(\alpha, \beta_0, \sigma) = \mathrm{E}_{(\alpha,\beta_0,\sigma)} Z_1^4$$
$$(\alpha, \sigma^2) \to m_{1,j}(\alpha, \beta_0, \sigma) = \mathrm{E}_{(\alpha,\beta_0,\sigma)} Z_1^2 Z_j^2, \quad j = 2, 3, 4, 5;$$

$\beta_0 = 1$ being the true value of the parameter β. The two solid curves are level curves, one for m_4 and one for $m_{1,2}$. Both go through the true value $(\alpha_0, \sigma_0) = (0.1, 0.35)$, that is, all parameters on the curve for m_1, say, have same value of m_1 as the true parameter values. The two level curves are very close, suggesting that $m_4(\alpha, \beta_0, \sigma)$ is "close" to $m_4(\alpha_0, \beta_0, \sigma_0)$ if and only if $m_{1,2}(\alpha, \beta_0, \sigma)$ is "close" to $m_{1,2}(\alpha_0, \beta_0, \sigma_0)$. Although the distribution of (Z_1, Z_2) is not determined by these two moments alone, the figure indicates that is hard to distinguish between different parameter values around the two curves. The three dashed curves in Figure III.8 are level curves for the moments $(\alpha, \sigma^2) \to m_{1,j}(\alpha, \beta_0, \sigma)$ for $j = 3, 4, 5$; they indicate that identification might be easier for larger value of k.

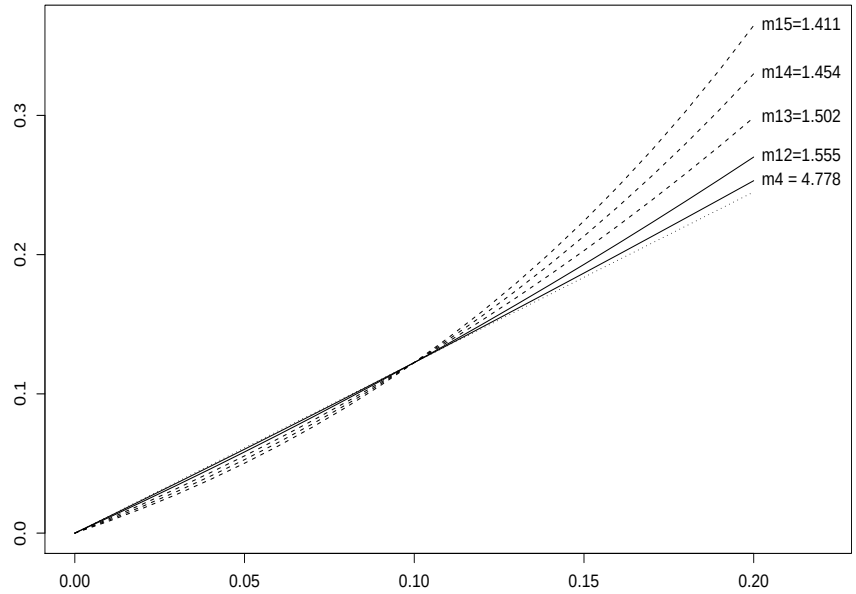

Figure III.8: The solid and dashed curves are level curves for $m_4 = \mathrm{E}_{(\alpha,\beta_0,\sigma)} Z_1^4$ and $m_{1,j} = \mathrm{E}_{(\alpha,\beta_0,\sigma)} Z_1^2 Z_j^2$, $j = 2,3,4,5$; α on the x-axis, σ^2 on the y-axis. The value of β is fixed and equal to the true value 1, and the levels are those corresponding to the true values (0.1,0.1225) of (α, σ^2). The dotted line is the line through (0,0) with slope $\sigma_0^2/\alpha_0 = 1.225$.

We choose $k = 4$. It is important to find good starting points for the numerical minimization routine. At first glance an obvious choice would be moment estimators since they are easily computable. However, we know from cases (A) and (B) that they are quite bad and that there may be problems with existence of solutions. The existence problem is even worse in case (C): a solution to the equation

$$(\tilde{m}_2, \tilde{m}_4, \tilde{m}_{1,2}) = \big(m_2(\theta), m_4(\theta), m_{1,2}(\theta)\big)$$

only exists for two of the ten simulated datasets (Z^1 and Z^2). Since we shall use the result as starting point for minimization of U_n^4, it would be natural to use $m_{1,5}$ rather than $m_{1,2}$. Then we get solutions for five of the ten datasets (but the estimates are still quite bad).

We are thus forced to come up with better alternatives. The following account of our approach may be somewhat tedious, but is included since it is an important part of our numerical procedure and since we believe that it provides a better understand of the problems involved.

The distribution of $(Z_1, \ldots, Z_5)$ is determined by the distribution of $(S_1, \ldots, S_5)$. Probably, the marginal (invariant) distribution of S is fairly well-determined. We do not know the invariant distribution of S, but for the moment we approximate it by a Γ-distribution with shape parameter λ and scale parameter τ. With this approximation $\mathrm{E}_{\lambda,\tau} S_1 = \lambda\tau$ and $\mathrm{Var}_{\lambda,\tau} S_1 = \lambda\tau^2$, and we establish a link between

(λ,τ) and the original parameters (α,β,σ) by fitting the expectation and variance, that is,

$$\beta\Delta = \lambda\tau; \quad \frac{\beta\sigma^2}{\alpha^3}\big(\alpha\Delta - 1 + e^{-\alpha\Delta}\big) = \lambda\tau^2, \tag{III.27}$$

see (III.24)–(III.25). In particular, this determines σ as a function of (α,β,λ),

$$\sigma^2 = \sigma^2(\alpha,\beta,\lambda) = \frac{\alpha^3\beta\Delta}{\lambda\big(\alpha\Delta - 1 + e^{-\alpha\Delta}\big)}. \tag{III.28}$$

For $(\alpha,\beta,\sigma) = (0.1,1,0.35)$ we have $(\lambda,\tau) = (1.6875, 0.5926)$.

The estimation strategy now is the following: (*i*) estimate β by $\tilde\beta_n$ given by (III.26); (*ii*) find an estimate $\tilde\lambda_n$ of λ as described below; (*iii*) minimize U_n^4 *along the curve* given by (III.28) with $\beta = \tilde\beta_n$ and $\lambda = \tilde\lambda_n$, that is, find

$$\tilde\alpha_n = \operatorname{argmin}_\alpha U_n^4\Big(\alpha, \tilde\beta_n, \sigma(\alpha,\tilde\beta_n,\tilde\lambda_n)\Big) \tag{III.29}$$

and the corresponding $\tilde\sigma_n^2 = \sigma^2(\tilde\alpha_n,\tilde\beta_n,\tilde\lambda_n)$; and finally (*iv*) minimize U_n^4 on $\mathbb{R}^3$ with starting point $(\tilde\alpha_n,\tilde\beta_n,\tilde\sigma_n)$.

For step (*ii*), recall that $Z_i \sim N(0,S_i)$ conditionally on V, and let $\tilde\lambda_n$ be the minimum point of the function

$$\tilde U_n^0(\lambda) = -\frac{1}{n}\sum_{i=1}^n \log\int_0^\infty \tilde p_{\lambda,\tilde\beta_n/\lambda}(\tilde s)\varphi(z_i,0,\tilde s)\,ds = -\frac{1}{n}\sum_{i=1}^n \tilde{\mathrm{E}}_{\lambda,\tilde\beta_n\Delta/\lambda}\varphi(z_i,0,\tilde S)$$

where $\tilde p_{\lambda,\tau}$ is the density of $\Gamma(\lambda,\tau)$ and $\varphi(\cdot,m,s)$ as usual is the density of $N(m,s)$. In practice we calculate $\tilde U_n^0(\lambda)$ as

$$-\frac{1}{n}\sum_{i=1}^n \log\frac{1}{R}\sum_{r=1}^R \varphi\big(z_i,0,\tilde S^{(r)}\big) \tag{III.30}$$

where $\tilde S^{(1)},\dots,\tilde S^{(R)}$ are independent randomly generated $\Gamma(\lambda,\tilde\beta_n\Delta/\lambda)$-variables. For α, σ and λ related by (III.28) with $\beta = \tilde\beta_n$, the only difference between (III.17) with $k=0$ and (III.30) is the distribution from which the S-variables are drawn. For $\tilde U_n^0$ each $\tilde S$ is drawn according to a Γ-distribution, whereas for U_n^0, S is generated as an integral of V-values.

This has two important consequences. First, it is faster to draw directly from the Γ-distribution than to draw V-paths and calculate integrals. Second, there is no a priori reason to believe that the marginal distribution is a Γ-distribution so we may have introduced bias. This means that U_n^0 does not necessarily have its minimum on the curve given by $(\tilde\beta_n,\tilde\lambda_n)$. The minimum point of U_n^4 may be even further away from the curve. This is not really problematic, though, since we only use the curve for finding good starting points. In practice the value of U_n^4 is indeed small at the starting point and the minimization routine has no problem moving away from the curve.

(III.35)

Step (*ii*) is very much in the spirit of Genon-Catalot *et al.* (1999). They suggest approximating the marginal density of S_1 by a Γ-distribution as well. Actually, they find an explicit expression for the marginal density of Z_1 if S_1 is Gamma-distributed. It would indeed have been faster (and smarter) to use this explicit expression for computing the density rather than the above simulation procedure.

Note that Genon-Catalot *et al.* (1999) link the parameters in the Γ-distribution and the original parameters differently than we do: they use shape parameter $\lambda' = 2\alpha\beta/\sigma^2$ and scale parameter $\tau' = \sigma^2\Delta/(2\alpha)$ instead of λ and τ given by (III.27) so the variance in their approximate Γ-distribution is not equal to the actual variance of S. For small values of $\alpha\Delta$, there is not much difference between the parametrizations. The one with (λ', τ') is motivated by the approximation

$$S_1 = \int_0^\Delta V_s\, ds \approx \Delta V_0 \sim \Gamma\Big(2\frac{\alpha\beta}{\sigma^2}, \frac{\sigma^2\Delta}{2\alpha}\Big)$$

which is good for small Δ. Indeed, Genon-Catalot *et al.* (1999) show that the corresponding estimators are asymptotically well-behaved if $\Delta = \Delta(n) \to 0$ as $n \to \infty$, in which case the parametrizations (λ, τ) and (λ', τ') coincide in the limit.

For the dataset Z^4 we find $\tilde{\beta}_n = 0.7528$ and $\tilde{\lambda}_n = 1.6732$. Figure III.9 shows level curves of $(\alpha, \sigma^2) \to U_n^0(\alpha, \tilde{\beta}_n, \sigma)$. The dashed curve is given by (III.28) with $\beta = \tilde{\beta}_n$ and $\lambda = \tilde{\lambda}_n$. The level curves are very oblong and those corresponding

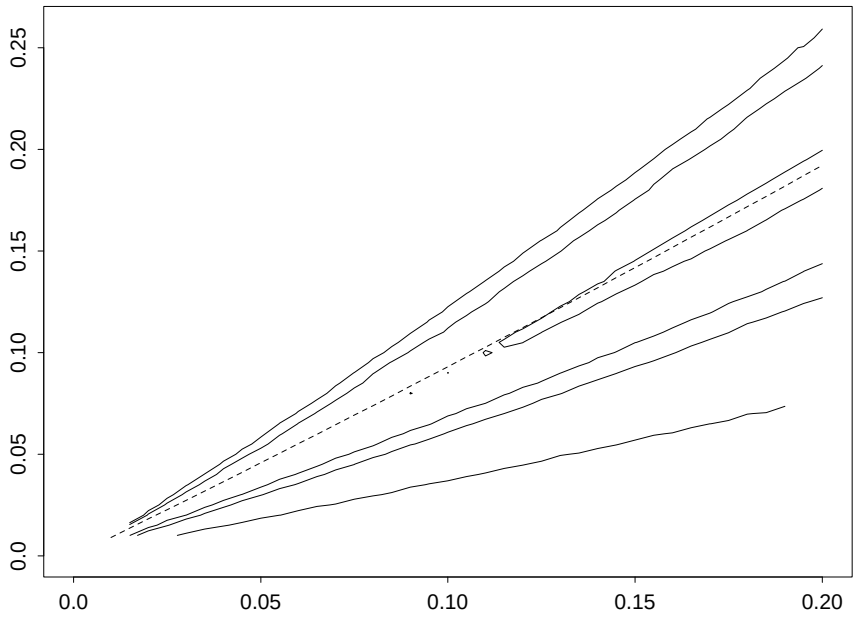

Figure III.9: Level curves of $(\alpha, \sigma^2) \to U_n^0(\alpha, \tilde{\beta}_n, \sigma)$ for Z^4; α on the x-axis and σ^2 on the y-axis. The dashed curve is given by (III.28) with $\beta = \tilde{\beta}_n = 0.7528$ and $\lambda = \tilde{\lambda}_n = 1.6732$. The true value of (α, σ^2) is $(\alpha_0, \sigma_0^2) = (0.1, 0.1225)$.

to low values are almost parallel to the dashed curve. The minimum along the

curve is far from the true values; in fact it is outside the figure, for α around 0.25. Figure III.10 shows the level curves of $(\alpha, \sigma^2) \to U_n^4(\alpha, \tilde{\beta}_n, \sigma)$. The level curves

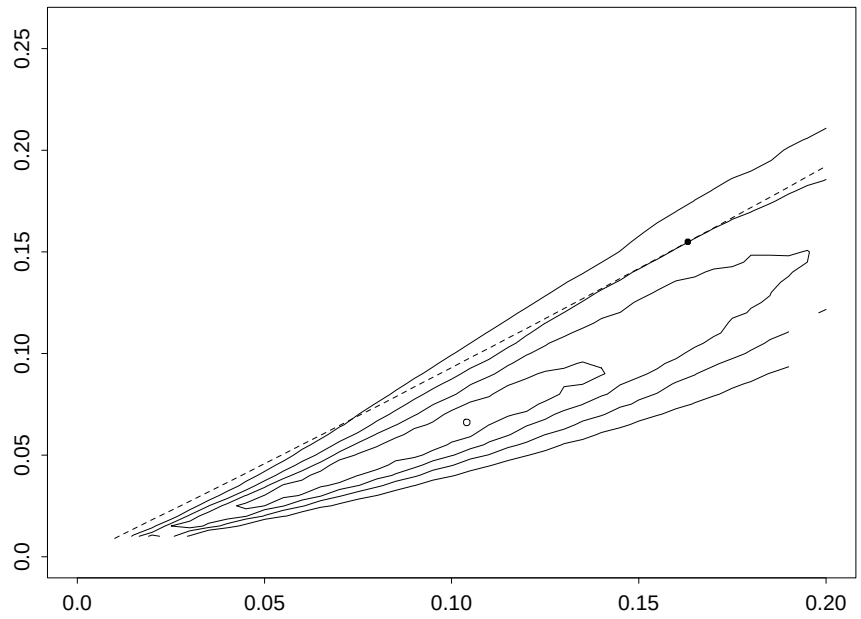

Figure III.10: Level curves of $(\alpha, \sigma^2) \to U_n^4(\alpha, \tilde{\beta}_n, \sigma)$ for Z^4; α on the x-axis and σ^2 on the y-axis. The dashed curve is given by (III.28) with $\beta = \tilde{\beta}_n = 0.7528$ and $\lambda = \tilde{\lambda}_n = 1.6732$. The solid circle denotes the minimum point $(\tilde{\alpha}_n, \tilde{\sigma}_n^2) = (0.1631, 0.1549)$ along the dashed curve, and the circle denotes the global minimum — when β varies as well — $(\hat{\alpha}_n^4, (\hat{\sigma}_n^4)^2) = (0.1040, 0.0661)$. The true value of (α, σ^2) is $(\alpha_0, \sigma_0^2) = (0.1, 0.1225)$.

are not parallel to the dashed curve (and thereby not to the level curves of U_n^0). Anyway, the value of U_n^4 is relatively low at the minimum of the dashed curve (denoted by a solid circle in the figure).

Figure III.11 shows the graph of U_n^4 along the curve, *i.e.* the criterion function in (III.29). Minimum is attained at $\tilde{\alpha}_n = 0.1631$. The corresponding value of σ is $\tilde{\sigma}_n = \sqrt{0.1549} = 0.3936$. In step (*iv*) the minimization routine moves from the starting point (0.1631,0.7528,0.3936) to the global minimum point

$$\left(\hat{\alpha}_n^4, \hat{\beta}_n^4, \hat{\sigma}_n^4\right) = (0.1040, 0.7441, 0.2571).$$

Note that the estimate of β changes (slightly) in this last step, too. The point $(\hat{\alpha}_n^4, (\hat{\sigma}_n^4)^2)$ is shown with a circle in Figure III.10.

The averages of $\hat{\alpha}_n^4$, $\hat{\beta}_n^4$ and $\hat{\sigma}_n^4$ are 0.1113, 1.0037 and 0.3036 respectively. This is not too bad. However, for three of the datasets (number 1, 5 and 6), the estimators for α and σ are very bad. This is reflected in huge standard errors: 0.1423, 0.1457 and 0.2463 respectively. If we leave out simulations 1, 5 an 6, $\hat{\alpha}_n^4$ has average 0.0866 and standard error 0.0355, and $\hat{\sigma}_n^4$ has average 0.2994 and

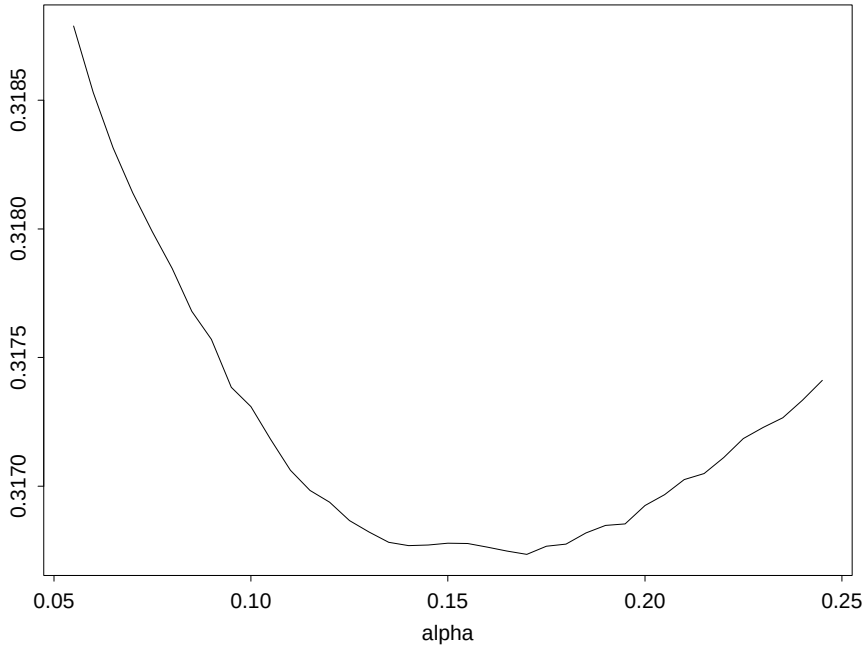

Figure III.11: Graph of $\alpha \to U_n^4(\alpha, \tilde{\beta}_n, \sigma(\alpha, \tilde{\beta}_n, \tilde{\lambda}_n))$ for Z^4 where $\sigma(\alpha, \beta, \lambda)$ is given by (III.28).

standard error 0.0887. The estimates $\tilde{\lambda}_n$, $\tilde{\alpha}_n$, $\hat{\alpha}_n^4$, $\hat{\sigma}_n^4$ and $\hat{\sigma}_n^4$ are listed in columns two through six in Table III.3 in Appendix III.B.

Again, it is easy to find estimators based on the volatility process. They are solutions to simple martingale estimating equations given in terms of the conditional mean and variance one step ahead, see Sørensen (1997) for details. The martingale estimators are listed in the last three columns in Table III.3 in Appendix III.B. The means (standard errors) are 0.1146 (0.0286) for $\hat{\alpha}_n^V$, 1.0024 (0.1485) for $\hat{\beta}_n^V$, and 0.3548 (0.0134) for $\hat{\sigma}_n^V$ so the estimators are far better than the approximate maximum likelihood estimators based on Z. This is clearly illustrated in Figure III.12 where the approximate maximum likelihood estimates are plotted in columns 1, 3, 5 and the martingale estimators are plotted in columns 2, 4 and 6. Recall however that V would not be observed in applications so martingale estimation based on V would not be an option.

Above we have used $k = 4$ which seemed to work reasonably well for seven of the ten datasets. Of course we could have used other values of k, and informal studies indicate that $k = 3$ would have worked reasonably for three of the simulations and $k = 2$ for two simulations. In other words: estimation seems to improve as k increases. This leaves us with some hope that estimation would improve for datasets 1, 5 and 6 if we used more than four lags. The hope is strengthened by inspection of the correlograms of the squared observations for the three datasets which all have relatively large correlations (compared to the other datasets) on several lags larger than four, indicating that U_n^4 does not capture all information

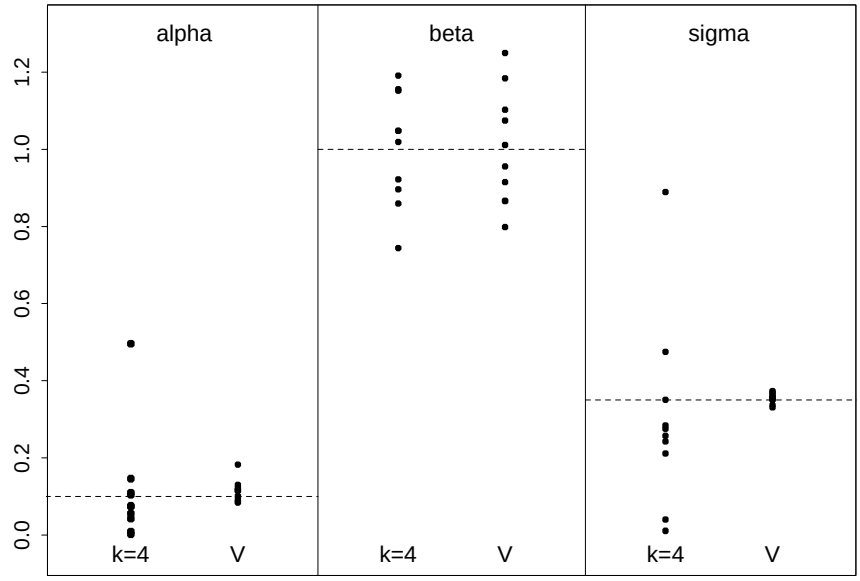

Figure III.12: The approximate maximum likelihood estimators $\hat{\alpha}_n^4$, $\hat{\beta}_n^4$, $\hat{\sigma}_n^4$ in columns 1, 3 and 5, and the martingale estimators $\hat{\alpha}_n^V$, $\hat{\beta}_n^V$, $\hat{\sigma}_n^V$ in columns 2, 4 and 6. The true values (0.1, 1 and 0.35) are shown with the dashed lines.

in data. The correlograms are omitted.

III.8 Conclusion

We have discussed approximate maximum likelihood estimation for increments $(Z_1, \ldots, Z_n)$ from a stochastic volatility model. For $k \geq 0$ the k'th order approximation to the likelihood function was obtained by pretending that $(Z_1, \ldots, Z_n)$ is k'th order Markov. Hence, the approximate likelihood is (essentially) a product of conditional densities $p_\theta^{c,k}(Z_i | Z_{i-k}, \ldots, Z_{i-1})$, $i = k+1, \ldots, n$. The corresponding estimators are consistent and asymptotically normal, essentially because we use the *true* conditional densities given the k previous observations. There are no explicit expressions for the densities but they are easy, though computationally demanding, to simulate for small values of $k \geq 0$.

Throughout the paper we have assumed that the drift and diffusion for X (of which $Z_1, \ldots, Z_n$ are increments) are determined completely by the process V and that the two Brownian motions driving V and X, respectively, are independent. The second assumption is not easily relaxed since we extensively employ that the conditional distribution of $(Z_1, \ldots, Z_n)$ given V is known. The nice properties of the conditional distribution of $(Z_1, \ldots, Z_n)$ given V are also destroyed if the drift and diffusion functions for X are functions of X as well as of V. However, it is

straightforward to generalize the method so it applies to models where the drift function for X is parameter dependent.

Also, the estimation procedure is applicable for other data types with similar properties, in particular for (other) hidden Markov models. In this respect, the important features of the models are the following: (*i*) given the values of an unobservable process, the observations $Z_1, \ldots, Z_n$ are independent with a known distribution (up to some parameter) determined by the latent process; (*ii*) the unobserved process is easy to simulate for all values of the parameter. These properties make it easy to simulate values of the approximate likelihood function.

The idea of considering approximations to the likelihood function in terms of k-lag conditional densities is of course applicable in all kinds of models with complicated dependence structures. There are other possible approximations. For example, one could split data into tuples of some length, and pretend that the tuples were independent (see Appendix III.A.1). Or one could both condition forwards and backwards in time, *i.e.* base estimation on the conditional densities $p_\theta^{c,k}(Z_i|Z_{i-k},\ldots,Z_{i-1},Z_{i+1},\ldots,Z_{i+k})$ given the k previous and the k subsequent observations. We would get asymptotically well-behaved estimators by these approximations as well. However, since time runs forward, we feel that the approximations based on conditioning backwards in time only, are the most natural.

Finally some comments on possible future work. First, in order for the method to be really useful one should be able to estimate the variance of the estimator. The expression for the asymptotic variance from Theorem III.9 is not useful in practice as it is given in terms of the unknown k-lag conditional density and its derivatives. Second, there are possibilities of model control built into the method: For each k an estimator of the same parameter is obtained. Consequently, significantly different estimators are indications of misspecification of the model. Third, when proving asymptotic properties for $\hat{\theta}_n^k$, it was implicitly assumed that the approximate likelihood function could be computed accurately. It would be interesting to see how computation of L_n^k via simulation influence the estimators. Similar work was done for martingale estimating functions (Kessler & Paredes 1999).

III.A Appendix: Miscellaneous

In this appendix we first give an interpretation of the first term of the k'th order approximation of the log-likelihood function. Next, we state and prove and ergodic theorem and a central limit theorem for the sequence Z.

III.A.1 Split data log-likelihoods

Consider the expression (III.12) for the k'th order approximation to (minus) the log-likelihood function. We show that the first term may be interpreted as a sum of "split data log-likelihoods" in the sense of Rydén (1994). Assume for simplicity that the number of observations, n, is a multiple of $k+1$, that is, $J = n/(k+1)$ is

an integer, and split $(Z_1,\dots,Z_n)$ into J tuples of length $k+1$,

$$\big(Z_1,\dots Z_{k+1}\big),\dots,\big(Z_{n-k},\dots,Z_n\big). \tag{III.31}$$

If the J tuples were independent, then minus the log-likelihood would be

$$\sum_{j=1}^{J} u_\theta^{k+1}\Big(z_{(j-1)(k+1)+1},\dots,z_{j(k+1)}\Big). \tag{III.32}$$

It would be just as natural to split the data into one of the sets of $J-1$ $(k+1)$-tuples

$$\big(Z_{a+1},\dots Z_{k+a+1}\big),\dots,\big(Z_{(J-2)(k+1)+a+1},\dots,Z_{n+a-(k+1)}\big) \tag{III.33}$$

for $a=1,\dots,k$ although it for each a would leave us with some observations ($Z_1,\dots,Z_a$ and $Z_{n+a-k},\dots,Z_n$) not included in a tuple.

Note that (III.33) with $a=0$ equals (III.31). For each $a=0,\dots,k$, we get an expression similar to (III.32) — plus extra terms originating from observations not in a tuple for $a\neq 0$ — if we pretend that the tuples (III.33) are independent. The sum over a of minus the log-likelihoods (without the extra terms) is

$$\sum_{i=k}^{n-1} u_\theta^{k+1}(z_{i-k+1}^{i+1}), \tag{III.34}$$

compare with (III.12). In other words: the first term in (III.12) can be interpreted as a sum of log-likelihoods, each of which is obtained by pretending that $(k+1)$-tuples with no overlap are independent. Note that observations z_i for $i=1,\dots,k$ and $z_{n-k+1},\dots,z_n$ appear in less than $k+1$ of the terms in (III.34); this could be corrected by including the extra terms mentioned above.

III.A.2 Limit Theorems

We now state and prove an ergodic theorem and a central limit theorem for the sequence Z. The proofs are very similar to the proofs of Theorem 2.2 and Corollary 2.1 in Genon-Catalot *et al.* (1998*b*).

In the proof of the central limit theorem we use the following result which follows immediately from Hall & Heyde (1980, Corollary A.2, page 278): Let $\mathscr{A}$ and $\mathscr{B}$ be σ-algebras included in $\mathscr{F}$ and let U_1 and U_2 be random variables which are $\mathscr{A}$- and $\mathscr{B}$-measurable, respectively. If $\mathrm{E}\,|U_1|^{r_1}<\infty$ and $\mathrm{E}\,|U_2|^{r_2}<\infty$ where $r_1,r_2>1$ and $1/r_1+1/r_2<1$, then

$$\mathrm{Cov}(U_1,U_2)\leq 8||U_1||_{r_1}||U_2||_{r_2}\alpha(\mathscr{A},\mathscr{B})^{1-1/r_1-1/r_2}. \tag{III.35}$$

Theorem III.12 *Suppose that Assumption III.1 holds and let $d\geq 1$ be arbitrary but fixed.*

1. *(Ergodic theorem) For any function* $\psi : \mathbb{R}^d \to \mathbb{R}$ *in* $L^1(P_{\theta_0}^d)$ *it holds that*

$$\frac{1}{n}\sum_{i=1}^{n-d+1} \psi(Z_i,\dots,Z_{i+d-1}) \to P_{\theta_0}^d(\psi) = \mathrm{E}_{\theta_0}\,\psi(Z_1,\dots,Z_d)$$

P_{θ_0}-almost surely and in $L^1(P_{\theta_0})$ as $n \to \infty$.

2. *(Central limit theorem) Let $q \geq 1$ and consider functions $\psi_1,\dots,\psi_q : \mathbb{R}^d \to \mathbb{R}$ from $L^1(P_{\theta_0}^d)$ with $P_{\theta_0}^d(\psi_j) = \mathrm{E}_{\theta_0}\,\psi_j(Z_1,\dots,Z_d) = 0$ for all $j = 1,\dots,q$. Suppose that there exists an $\eta > 0$ such that ψ_j is in $L^{2+\eta}(P_{\theta_0}^d)$ for all $j = 1,\dots,q$ and such that the α-mixing coefficients for Z corresponding to θ_0 satisfy the condition $\sum_{m=1}^{\infty} \alpha_Z(m)^{\eta/(2+\eta)} < \infty$. Then*

$$\begin{aligned}\Sigma_{jl} = \mathrm{E}_{\theta_0}\,\psi_j(Z_1^d)\psi_l(Z_1^d) \\ + \sum_{m=1}^{\infty}\Big(\mathrm{E}_{\theta_0}\,\psi_j(Z_1^d)\psi_l(Z_{m+1}^{m+d}) + \mathrm{E}_{\theta_0}\,\psi_l(Z_1^d)\psi_j(Z_{m+1}^{m+d})\Big)\end{aligned}$$

is well-defined for all $j,l = 1,\dots,q$ and if the $q \times q$ matrix $\Sigma = (\Sigma_{jl})_{jl}$ is positive definite then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n-d+1}\Big(\psi_1(Z_i^{i+d-1}),\dots,\psi_q(Z_i^{i+d-1})\Big)^T \to N(0,\Sigma)$$

in distribution wrt. P_{θ_0} as $n \to \infty$.

Proof Under Assumption III.1, Z is α-mixing (Proposition III.2). It is well-known that α-mixing implies ergodicity, see *e.g.* Doukhan (1994, page 21).

For the central limit theorem, first assume that $q = 1$ and define $Y_i = \psi(Z_i^{i+d-1})$, $i \geq 1$. Then the σ-algebras generated by Y satisfy

$$\begin{aligned}\sigma\big(\{Y_i\}_{i\leq l}\big) = \sigma\big(\{\psi(Z_i^{i+d-1})\}_{i\leq l}\big) \subseteq \sigma\big(\{Z_i\}_{i\leq l+d-1}\big) \\ \sigma\big(\{Y_i\}_{i\geq l+m}\big) = \sigma\big(\{\psi(Z_i^{i+d-1})\}_{i\geq l+m}\big) \subseteq \sigma\big(\{Z_i\}_{i\geq l+m}\big)\end{aligned}$$

for all $s,t \in \mathbb{N}$. Hence, the α-mixing coefficients for $Y = (Y_1, Y_2, \dots)$ satisfy $\alpha_Y(m) \leq \alpha_Z(m-d+1)$ for $m \geq d$ and thus

$$\sum_{m=1}^{\infty} \alpha^Y(m)^{\eta/(2+\eta)} \leq \sum_{m=1}^{d-1} \alpha_Y(m)^{\eta/(2+\eta)} + \sum_{m=1}^{\infty} \alpha^Z(m)^{\eta/(2+\eta)} < \infty.$$

It now follows from Hall & Heyde (1980, Corollary 5.1, page 132) that Σ (which is a real number since $q = 1$) is non-negative and finite and

$$\mathrm{Var}_{\theta_0}\Big(n^{-1/2}\sum_{i=1}^{n-d+1}\psi(Z_i)\Big) \to \Sigma \tag{III.36}$$

as $n \to \infty$. If $\Sigma > 0$ then furthermore

$$n^{-1/2} \sum_{i=1}^{n-d+1} \psi_1(Z_i^{i+d-1}) = n^{-1/2} \sum_{i=1}^{n-d+1} Y_i \to N(0,\Sigma)$$

in distribution.

It might be useful to see how $\Sigma < \infty$ and the convergence in (III.36) are obtained: the left hand side of (III.36) is equal to

$$\begin{aligned}
&\frac{1}{n} \sum_{i,j=1}^{n-d+1} \mathrm{E}_{\theta_0} \psi(Z_i^{i+d-1}) \psi(Z_j^{j+d-1}) \\
&\quad = \mathrm{E}_{\theta_0} \psi^2(Z_1^d) + \frac{1}{n} \sum_{m=1}^{n-d} (n-m) \, \mathrm{E}_{\theta_0} \left(\psi(Z_1^d) \psi(Z_{m+1}^{m+d}) + \psi(Z_{m+1}^{m+d}) \psi(Z_1^d) \right) \\
&\quad = \mathrm{E}_{\theta_0} \psi^2(Z_1^d) + \sum_{m=1}^{n-d} \left(\mathrm{E}_{\theta_0} \psi(Z_1^d) \psi(Z_{m+1}^{m+d}) + \mathrm{E}_{\theta_0} \psi(Z_{m+1}^{m+d}) \psi(Z_1^d) \right) \qquad \text{(III.37)} \\
&\qquad\quad - \frac{1}{n} \sum_{m=1}^{n-d} m \left(\mathrm{E}_{\theta_0} \psi(Z_1^d) \psi(Z_{m+1}^{m+d}) + \mathrm{E}_{\theta_0} \psi(Z_{m+1}^{m+d}) \psi(Z_1^d) \right). \qquad \text{(III.38)}
\end{aligned}$$

Let $r_1 = r_2 = 2 + \eta$. It follows by (III.35) that the expectations $|\mathrm{E}_{\theta_0} \psi(Z_1^d) \psi(Z_{m+1}^{m+d})|$ and $|\mathrm{E}_{\theta_0} \psi(Z_{m+1}^{m+d}) \psi(Z_1^d)|$ are bounded by $8 ||\psi(Z_1^d)||_{2+\eta}^2 \, \alpha_Z(m)^{\eta/(2+\eta)}$. Hence, by assumption, Σ is finite and (III.37) converges to Σ. It finally follows by Kronecker Lemma that the sum (III.38) converges to zero as $n \to \infty$ so that the convergence (III.36) holds.

Now let $q \geq 2$. Calculations similar to those above show that Σ_{jl} is well-defined and can be obtained as a limit of covariances,

$$\mathrm{Cov}_{\theta_0} \left(n^{-1/2} \sum_{i=1}^{n-d+1} \psi_j(Z_i), n^{-1/2} \sum_{i=1}^{n-d+1} \psi_l(Z_i) \right) \to \Sigma_{jl}$$

for all $j, l = 1 \ldots, q$. By Cramer-Wold's device it suffices to show that for any $y = (y_1, \ldots, y_q)^T \in \mathbb{R}^q$ the linear combination

$$n^{-1/2} \sum_{j=1}^{q} y_j \sum_{i=1}^{n-d+1} \psi_j(Z_i^{i+d-1}) = n^{-1/2} \sum_{i=1}^{n-d+1} \sum_{j=1}^{q} y_j \psi_j(Z_i^{i+d-1})$$

converges in distribution to the normal distribution with mean 0 and variance $y^T \Sigma y$. This follows immediately from the one-dimensional result. □

III.B Appendix: Results from the simulation study

In this appendix we have collected tables with estimation results from the simulation study for the Cox-Ingersoll-Ross process (Section III.7.2). Table III.1 lists empirical moments for the simulated datasets. Table III.2 lists estimators from case (A) where only α is unknown, and Table III.3 lists estimators from case (C) where all three parameters are unknown.

Data	Mean (0)	$\tilde{m}_2$ (1)	$\tilde{m}_4$ (4.778)	$\tilde{m}_{1,2}$ (1.555)	$\tilde{c}_{1,2}$ (0.147)
Z^1	0.023	1.042	4.635	1.418	0.093
Z^2	0.005	1.156	6.905	1.965	0.112
Z^3	0.009	0.850	2.774	0.960	0.115
Z^4	0.068	0.753	2.396	0.594	0.014
Z^5	0.010	1.027	4.457	1.553	0.147
Z^6	0.005	1.053	5.217	1.910	0.194
Z^7	0.075	0.880	3.088	1.030	0.110
Z^8	-0.061	1.141	5.203	2.017	0.182
Z^9	0.026	0.919	3.085	1.197	0.156
Z^{10}	-0.039	1.211	6.739	3.025	0.300

Table III.1: Various empirical quantities for the ten simulated datasets. The true values are shown in parentheses in the top line.

Data	$\hat{\alpha}_n^0$	$\hat{\alpha}_n^1$	$\hat{\alpha}_n^2$	$\hat{\alpha}_n^3$	$\hat{\alpha}_n^4$	$\tilde{\alpha}_n^4$	$\hat{\alpha}_n^{1,2}$	$\hat{\alpha}_n^V$
Z^1	0.0961	0.1241	0.1434	0.1531	0.1322	0.1084	0.1290	0.1245
Z^2	0.1725	0.1347	0.1320	0.1201	0.1325	0.0463	0.0598	0.1244
Z^3	0.1084	0.1127	0.0982	0.0967	0.1014	NA	NA	0.1253
Z^4	0.0751	0.1048	0.1132	0.1104	0.1107	NA	NA	0.0888
Z^5	0.0993	0.0917	0.0939	0.0798	0.0799	0.1212	0.1004	0.0826
Z^6	0.0917	0.0961	0.1041	0.1325	0.1272	0.0807	0.0632	0.1213
Z^7	0.0851	0.0874	0.0950	0.0906	0.0904	1.3884	0.8916	0.1085
Z^8	0.1186	0.0917	0.1037	0.1018	0.1008	0.0812	0.0569	0.1158
Z^9	0.1214	0.0811	0.0695	0.0806	0.0909	1.4138	0.2449	0.1015
Z^{10}	0.1331	0.1027	0.0840	0.0884	0.0854	0.0484	0.0294	0.1038
mean	0.1101	0.1027	0.1037	0.1054	0.1051	0.4111	0.1969	0.1097
s.e.	0.0281	0.0169	0.0217	0.0238	0.0196	0.6117	0.2886	0.0154

Table III.2: Estimators for α in case (A) where $\beta_0 = 1$ and $\sigma_0 = 0.35$ are known. The true value of α is $\alpha_0 = 0.1$: approximate maximum likelihood estimates $\hat{\alpha}_n^k$, $k = 0, \ldots, 4$; moment estimators $\tilde{\alpha}_n^4$ and $\tilde{\alpha}_n^{1,2}$ based on moments m_4 and $m_{1,2}$; and a martingale estimator $\hat{\alpha}_n^V$ based on observations of V. NA means that the moment equation has no solution.

Data	$\tilde{\lambda}_n$	$\tilde{\alpha}_n$	$\hat{\alpha}_n^4$	$\hat{\beta}_n^4$	$\hat{\sigma}_n^4$	$\hat{\alpha}_n^V$	$\hat{\beta}_n^V$	$\hat{\sigma}_n^V$
Z^1	1.5663	0.2127	0.0020	1.0479	0.0398	0.1302	0.9556	0.3507
Z^2	2.3553	0.0694	0.0733	1.1522	0.2755	0.1159	1.1844	0.3642
Z^3	2.1925	0.0569	0.0556	0.8593	0.2111	0.1826	0.7983	0.3561
Z^4	1.6732	0.1631	0.1040	0.7441	0.2571	0.0862	0.8666	0.3306
Z^5	1.5881	0.0193	0.0088	1.0194	0.0106	0.0840	1.0113	0.3368
Z^6	1.3629	0.4784	0.4959	1.0485	0.8894	0.1168	1.0747	0.3552
Z^7	1.7447	0.0790	0.0755	0.8964	0.2843	0.1225	0.8660	0.3697
Z^8	1.7060	0.1470	0.1459	1.1560	0.4751	0.0925	1.2501	0.3732
Z^9	2.1827	0.0817	0.1097	0.9220	0.3507	0.1144	0.9148	0.3508
Z^{10}	2.0514	0.0292	0.0425	1.1914	0.2424	0.1010	1.1025	0.3603
mean	1.8423	0.1337	0.1113	1.0037	0.3036	0.1146	1.0024	0.3548
s.e.	0.3288	0.1358	0.1423	0.1457	0.2463	0.0286	0.1485	0.0134

Table III.3: Estimates in case (C). The true values are $\alpha_0 = 0.1$, $\beta_0 = 1$ and $\sigma_0 = 0.35$. The second and third column list preliminary estimates of λ and α used to find starting points for the numerical routine (the preliminary estimate $\tilde{\beta}_n$ is listed in Table III.1 as $\tilde{m}_2$). Colomns four through six list the final approximate maximum likelihood estimates (for $k = 4$). The final three columns list martingale estimators based on observations of V.

Acknowledgements I wish to thank my advisor Martin Jacobsen for many valuable discussions and for many helpful comments on the manuscript.

Bibliography

Aït-Sahalia, Y. (1996), 'Nonparametric pricing of interest rate derivative securities', *Econometrica* **64**, 527–560.

Aït-Sahalia, Y. (1998), Maximum likelihood estimation of discretely sampled diffusions: a closed-form approach, Revised version of Working Paper 467, Graduate School of Business, University of Chicago.

Andersen, T. G. & Lund, J. (1997), 'Estimating continuous-time stochastic volatility models of the short-term interest rate', *Journal of Econometrics* **77**, 343–377.

Andersen, T. G. & Sørensen, B. E. (1996), 'GMM estimation of a stochastic volatility model: A Monte Carlo study', *Journal of Business and Economic Statistics* **14**, 328–352.

Arcones, M. A. & Yu, B. (1994), 'Central limit theorems for empirical and U-processes of stationary mixing sequences', *Journal of Theoretical Probability* **7**, 47–71.

Banon, G. (1978), 'Nonparametric identification for diffusion processes', *Siam J. Control and Optimization* **16**, 380–395.

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. & Shephard, N. (1999), Non-Gaussian OU based models and some of their uses in financial economics, Working Paper 37, Centre for Analytical Finance, University of Aarhus.

Bera, A. K. & Higgins, M. L. (1993), 'ARCH models: properties, estimation and testing', *Journal of Economic Surveys* **7**, 305–366.

Bibby, B. M. & Sørensen, M. (1995), 'Martingale estimation functions for discretely observed diffusion processes', *Bernoulli* **1**, 17–39.

Bibby, B. M. & Sørensen, M. (1996), 'On estimation for discretely observed diffusions: A review', *Theory of Stochastic processes* **2**, 49–56.

Bibby, B. M. & Sørensen, M. (1998), Simplified estimating functions for diffusion models with a high-dimensional parameter, Preprint 1998-10, Department of Theoretical Statistics, University of Copenhagen. To appear *in Scandinavian Journal of Statistics*.

Bickel, P. J. & Ritov, Y. (1996), 'Inference in hidden Markov models I: Local asymptotic normality in the stationary case', *Bernoulli* **2**, 199–228.

Bickel, P. J., Ritov, Y. & Rydén, T. (1998), 'Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models', *Annals of Statistics* **26**, 1614–1635.

Billingsley, P. (1961), 'The Lindeberg-Levy Theorem for martingales', *Proceedings of the American Mathematical Society* **12**, 788–792.

Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley.

Black, F. & Scholes, M. (1973), 'The pricing of options and corporate liabilities', *Journal of Political Economy* **81**, 637–654.

Bradley, R. C. (1986), Basic properties of strong mixing conditions, *in* E. Eberlein & M. S. Taqqu, eds, 'Dependence in Probability and Statistics: A Survey of Recent Results', Birkhäuser, Boston, pp. 165–192.

Le Breton, A. (1974), Parameter estimation in a linear stochastic differential equation, *in* 'Transactions of the Seventh Prague Conference and of the European Meeting of Statisticians', pp. 353–366.

Le Breton, A. (1976), 'On continuous and discrete sampling for parameter estimation in diffusion type processes', *Mathematical Programming Study* **5**, 124–144.

Chan, K. C., Karolyi, G. A., Longstaff, F. A. & Sanders, A. B. (1992), 'An empirical comparison of alternative models of the short-term interest rate', *Journal of Finance* **47**, 1209–1227.

Chesney, M. & Scott, L. (1989), 'Pricing European currency options: a comparison of the modified Black-Scholes model and a random variance model', *Journal of Financial and Quantitative Analysis* **24**, 267–284.

Dacunha-Castelle, D. & Duflo, M. (1986), *Probability and Statistics*, Vol. 2, Springer-Verlag, New York.

Dacunha-Castelle, D. & Florens-Zmirou, D. (1986), 'Estimation of the coefficients of a diffusion from discrete observations', *Stochastics* **19**, 263–284.

Danielsson, J. (1994), 'Stochastic volatility in asset prices: estimation with simulated maximum likelihood', *Journal of Econometrics* **64**, 375–400.

Danielsson, J. & Richard, J.-F. (1993), 'Accelerated Gaussian importance sampler with application to dynamic latent variable models', *Journal of Applied Econometrics* **8**, S153–S173.

Doukhan, P. (1994), *Mixing: Properties and Examples*, Lecture Notes in Statistics **85**, Springer-Verlag, New York.

Elerian, O., Chib, S. & Shephard, N. (2000), Likelihood inference for discretely observed non-linear diffusions, Economics discussion paper 146, Nuffield College, Oxford. To appear in *Econometrica*.

Eraker, B. (1998), MCMC analysis of diffusion models with application to finance, Discussion paper 1998-5, Department of Finance and Management Science, Norwegian School of Economics and Business Administration.

Florens-Zmirou, D. (1989), 'Approximate discrete-time schemes for statistics of diffusion processes', *Statistics* **20**, 547–557.

Florens-Zmirou, D. (1993), 'On estimating the diffusion coefficient from discrete observations', *Journal of Applied Probability* **30**, 790–804.

Gallant, A. R., Hsieh, D. & Tauchen, G. (1997), 'Estimation of stochastic volatility models with diagnostics', *Journal of Econometrics* **81**, 159–192.

Gallant, A. R. & Long, J. R. (1997), 'Estimating stochastic differential equations efficiently by minimum chi-squared', *Biometrika* **84**, 125–141.

Gallant, A. R. & Tauchen, G. (1996), 'Which moments to match?', *Econometric Theory* **12**, 657–681.

Genon-Catalot, V., Jeantheau, T. & Laredo, C. (1998*a*), 'Limit theorems for discretely observed stochastic volatility models', *Bernoulli* **4**, 283–303.

Genon-Catalot, V., Jeantheau, T. & Laredo, C. (1998*b*), Stochastic volatility models as hidden Markov models and statistical applications, Preprint 1998-22, Equipe d'Analyse et de Mathématiques Appliquées, Université de Marne-la-Vallée.

Genon-Catalot, V., Jeantheau, T. & Laredo, C. (1999), 'Parameter estimation for discretely observed stochastic volatility models', *Bernoulli* **5**, 855–872.

Ghysels, E., Harvey, A. C. & Renault, E. (1996), Stochastic volatility, *in* G. S. Maddala & C. R. Rao, eds, 'Statistical Methods in Finance', Vol. 14 of *Handbook of Statistics*, North-Holland, Amsterdam, pp. 119–191.

Gourieroux, C., Monfort, A. & Renault, E. (1993), 'Indirect inference', *Journal of Applied Econometrics* **8**, S85–S118.

Hall, P. & Heyde, C. C. (1980), *Martingale Limit Theory and its Application*, Academic Press, New York.

Hansen, L. P. (1982), 'Large sample properties of generalized method of moments estimators', *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.

Harvey, A., Ruiz, E. & Shephard, N. (1994), 'Multivariate stochastic variance models', *Review of Economic Studies* **61**, 247–264.

Heston, S. L. (1993), 'A closed-form solution for options with stochastic volatility with applications to bond and currency options', *The Review of Financial Studies* **6**, 327–343.

Heyde, C. C. (1997), *Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation*, Springer-Verlag, New York.

Honoré, P. (1997), Maximum likelihood estimation of non-linear continuous-time term-structure models, Working paper 1997-7, Department of Finance, Aarhus School of Business.

Hull, J. & White, A. (1987), 'The pricing of options on assets with stochastic volatilities', *The Journal of Finance* **42**, 281–300.

Hull, J. & White, A. (1988), 'An analysis of the bias in option pricing caused by a stochastic volatility', *Advances in Futures and Options Research* **3**, 29–61.

Jacobsen, M. (1998), Discretely observed diffusions: classes of estimating functions and small Δ-optimality, Preprint 1998-11, Department of Theoretical Statistics, University of Copenhagen. To appear in *Scandinavian Journal of Statistics*.

Jacquier, E., Polson, N. G. & Rossi, P. E. (1994), 'Bayesian analysis of stochastic volatility models (with discussion)', *Journal of Business and Economic Statistics* **12**, 371–417.

Jensen, J. L. & Petersen, N. V. (1999), 'Asymptotic normality of the maximum likelihood estimator in state space models', *Annals of Statistics* **27**, 514–535.

Jiang, G. J. & Knight, J. L. (1997), 'A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model', *Econometric Theory* **13**, 615–645.

Karatzas, I. & Shreve, S. E. (1991), *Brownian Motion and Stochastic Calculus*, 2nd edn, Springer-Verlag, New York.

Karlin, S. & Taylor, H. M. (1981), *A Second Course in Stochastic Processes*, Academic Press, New York.

Kessler, M. (2000), 'Simple and explicit estimating functions for a discretely observed diffusion process', *Scandinavian Journal of Statistics* **27**, 65–82.

Kessler, M. & Paredes, S. (1999), Computational aspects related to martingale estimating functions for a discretely observed diffusion, Manuscript, Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena.

Kessler, M. & Sørensen, M. (1999), 'Estimating equations based on eigenfunctions for a discretely observed diffusion process', *Bernoulli* **5**, 299–314.

Kim, S., Shephard, N. & Chib, S. (1998), 'Stochastic volatility: likelihood inference and comparison with ARCH models', *Review of Economic Studies* **65**, 361–393.

Lipster, R. S. & Shiryayev, A. N. (1977), *Statistics of Random Processes*, Vol. 1, Springer-Verlag, New York.

Lo, A. W. (1988), 'Maximum likelihood estimation of generalized Itô processes with discretely sampled data', *Econometric Theory* **4**, 231–247.

Nelson, D. B. (1990), 'ARCH models as diffusion approximations', *Journal of Econometrics* **45**, 7–38.

Nelson, D. B. (1992), 'Filtering and forecasting with misspecified ARCH models I', *Journal of Econometrics* **52**, 61–90.

Nielsen, J. N., Vestergaard, M. & Madsen, H. (2000), Estimation in continuous-time stochastic volatility models using nonlinear filters. To appear in *International Journal of Theoretical and Applied Finance*.

Overbeck, L. & Rydén, T. (1997), 'Estimation in the Cox-Ingersoll-Ross model', *Econometric Theory* **13**, 430–461.

Pedersen, A. R. (1995*a*), 'Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes', *Bernoulli* **1**(3), 257–279.

Pedersen, A. R. (1995*b*), 'A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations', *Scandinavian Journal of Statistics* **22**, 55–71.

Pollard, D. (1984), *Convergence of Stochastic Processes*, Springer-Verlag, New York.

Poulsen, R. (1999), Approximate maximum likelihood estimation of discretely observed diffusion processes, Working paper 29, Centre for Analytical Finance, Aarhus.

Rogers, L. C. G. & Williams, D. (1987), *Diffusions, Markov Processes, and Martingales*, Vol. 2: Itô Calculus, Wiley.

Rydén, T. (1994), 'Consistent and asymptotically normal parameter estimates for hidden Markov models', *Annals of Statistics* **22**, 1884–1895.

Scott, L. O. (1987), 'Option pricing when the variance changes randomly: theory, estimation and an application', *Journal of Financial and Quantitative analysis* **22**, 419–438.

Shephard, N. (1996), Statistical aspects of ARCH and stochastic volatility, *in* D. R. Cox, D. V. Hinkley & O. E. Barndorff-Nielsen, eds, 'Time Series Models in Econometrics, Finance and Other Fields', Chapman & Hall, London, pp. 1–67.

Sørensen, H. (1998*a*), Approximation of the score functions for diffusion processes, Preprint 1998-8, Department of Theoretical Statistics, University of Copenhagen.

Sørensen, H. (2000), Estimation of diffusion parameters for discretely observed diffusion processes, Preprint 2000-1, Department of Theoretical Statistics, University of Copenhagen.

Sørensen, M. (1997), Estimating functions for discretely observed diffusions: A review, *in* I. V. Basawa, V. P. Godambe & R. L. Taylor, eds, 'Selected Proceedings of the Symposium on Estimating Functions', Vol. 32, IMS Lecture notes, pp. 305–325.

Sørensen, M. (1998*b*), On asymptotics of estimating functions, Preprint 1998-6, Department of Theoretical Statistics, University of Copenhagen. To appear in *Brazilian Journal of Probability and Statistics*.

Sørensen, M. (1999), Prediction-based estimating functions, Preprint 1999-5, Department of Theoretical Statistics, University of Copenhagen.

Stein, E. M. & Stein, J. C. (1991), 'Stock price distributions with stochastic volatility: an analytic approach', *The Review of Financial Studies* **4**, 727–752.

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer-Verlag, New York.

Wiggins, J. B. (1987), 'Options values under stochastic volatility', *Journal of Financial Economics* **19**, 351–372.