Estimation in Two Classes of Semiparametric Diffusion Models[*]

Dennis Kristensen
Department of Economics/FMG, LSE

June 7, 2004

Abstract

In this paper we propose an estimation method for two classes of semiparametric scalar diffusion models driven by a Brownian motion: In the first class, only the diffusion term is parameterised while the drift is unspecified; in the second, the drift term is specified while the diffusion term is of unknown form. The estimation method is based on the assumption of stationarity of the observed process. This allows us to express the unspecified term as a functional of the parametric part and the stationary density. A MLE-like estimator for the parametric part and a kernel estimator of the nonparametric part are defined for a discrete sample with a fixed time distance between the observations. We show that the parametric part of the estimator is $\sqrt{n}$-consistent, while the nonparametric part has a slower convergence rate. Also, the asymptotic distribution of the estimator is derived. We give a brief discussion of the issue of semiparametric efficiency, and present a small simulation study of the finite-sample performance of our estimator.

[*] This paper is part of my PhD dissertation at the Department of Econonomics, LSE. I am very grateful to my supervisor Oliver Linton for his support and helpful suggestions. I thank Ayan Sen for helping me out with the data. I also wish to thank Matthias Hagmann, Antonio Mele, Peter Robinson, and participants at the European Winter Meetings 2002 of the Econometric Society, the 2003 workshop of the Danish Doctoral Education Network in Finance, and seminars at the University of Copenhagen, LSE, UCSD, Iowa State, Georgetown, Chicaco GSB, UW-Madison, Pittsburgh, Oxford, UCL and Yale for comments. I gratefully acknowledge the financial support from the Danish Research Agency and the FMG, LSE.

[†] Address: R4z19C, FMG, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. E-mail: d.kristensen@lse.ac.uk

1 Introduction

Continuous time stochastic processes are widely used in dynamic models in economics and finance. In the past three decades since the groundbreaking work by Black and Scholes (1973), Merton (1973) stochastic processes have gained a major role in finance theory where they are used in the modelling of the dynamics of economic variables over time, for example interest rates, stock prices, and exchange rates; an overview of such models can be found in Björk (1998) and Duffie (1996). To a lesser extent these have also been used to model the dynamics of macroeconomic variables, see e.g. Bergstrom (1990). Unfortunately, economic theory has very little to say about the precise specification of the processes. As a consequence, a wide range of parametric models have been suggested in the literature, for example Black and Scholes (1973), Chan et al. (1992), Cox et al. (1985), Vasicek (1977), but it is not obvious that these models are able to deliver an adequate description of the observed process. This may lead to the use of a misspecified model that are not able to capture the true dynamics of the process in consideration. This again can have serious implications on the conclusions drawn from the model. Non- and semiparametric methods may help to detect and to some extent solve such problems, since these methods allow for a high degree of flexibility and should thereby better safeguard one against possible misspecification.

We will here take a semiparametric approach to the modelling and estimation of scalar stochastic differential equations (SDE's) driven by a Brownian motion. Such processes are fully characterised by their drift and diffusion function, which we wish to model in a flexible manner. Two very general classes of models will be considered: In the first class, the drift is specified (up to an unknown parameter) while the diffusion term is left unspecified; in the second class it is the diffusion term that is parameterised while the drift term is not specified. We define an estimator for the drift and diffusion function for models in each of the two classes, and derive its asymptotic properties under regularity conditions. We also construct a simple test for parametric submodels against the semiparametric alternative. The main restriction we need to impose is that the diffusion processes in the two classes are strongly stationary since this property is used for identification of the unspecified term. This excludes for example time-inhomogenous processes, where the drift and diffusion functions are allowed to depend on time, since these are non-stationary by construction. The two classes are still very rich, and include a majority of the parametric homogeneous models proposed in the literature since these in most cases allow for stationary solutions. In particular, for any parameterisation of a stationary diffusion process, each of the two classes contains a semiparametric model which has this fully parametric model as a submodel.

Only a few studies in the existing literature have considered semiparametric diffusion models. Aït-Sahalia (1996a) proposes a semiparametric model with a linear parameterisation of the drift, while leaving the diffusion term unspecified. Conley et al. (1997) on the other hand suggest to use a simple parametric form for the diffusion term, while either applying a global series expansion or a locally linear approximation of the drift term. The model of Aït-Sahalia (1996a) belongs to the first class of models considered here, while the Conley et al. (1997) model is situated in the second one. These two models are quite general, but one may still want to allow for other, more flexible, specifications of either the drift or the diffusion term than the two proposed by the aforementioned authors. This is made possible with the two classes of semiparametric models proposed here, which allows for virtually any reasonable parameterisation of either the drift or diffusion term. In Bandi and Phillips (2000), least squares estimators for any parameterisation of either the drift or the diffusion term is proposed. Their results however depend on the time distance between observations shrinking to zero, the so-called infill assumption, while ours hold for a fixed time distance; see also Florens-Zmirou (1989) and Genon-Catalot (1990).

The semiparametric models under consideration here can be very useful as an intermediate step in model building, moving from an initial nonparametric model towards a parsimonious fully parametric one. There is a large literature on fully nonparametric estimation of the drift and diffusion function. Most of the proposed estimators are based on kernel methods, making use of the characterisation of the drift and diffusion function as the instantaneous conditional mean and variance respectively. Assuming that the time distance between observations shrinks to zero as the number of observations goes to infinity, standard kernel regression methods can be used to consistently estimate the drift and diffusion term. This approach is pursued by, for example, Bandi and Phillips (2003), Jiang and Knight (1997), and Stanton (1997). These estimators are however prone to a discretization bias if the process is in fact observed at fixed time instants, c.f. Nicolau (2003). Chen et al. (2000a), Darolles and Gouriéroux (2001) and Gobet et al. (2003) derive nonparametric estimators for univariate diffusion models by the method of sieves, allowing for a fixed time distance between observations. Their approach is based on the so-called infinitesimal operator of the diffusion model, which uniquely identifies the model. They decompose the operator into its eigenfunctions and demonstrate that from these one may recover the drift and diffusion term. Estimators of the eigenfunctions are then constructed, and thereby also estimators of the drift and diffusion function.

On particular area where the semiparametric models can be used is in the continuous time-modelling of the short term interest rate. As theory provides very little guidance about the correct specification, a very large number of parametric diffusion models have been used to model short-term interest rates; Rogers (1995) provides an overview. Most of these models appear not to be able to give a very good fit of the observed interest rates however, and so it is still an open question what the appropriate model for the short-term interest rate is. In Kristensen (2004), a specific semiparametric diffusion model is proposed as an alternative to the parametric models. The model is fitted to an interest rate data set using the estimation method proposed here. It proves to be able to pick up non-linearities in the drift term that standard parametric models are unable to capture. Using the test statistic provided here, the most flexible parametric submodel is in fact rejected when tested against the semiparametric alternative.

Our estimation method is based on the assumption that the sampled diffusion process is stationary and ergodic, thereby ensuring that an invariant density of the process exists. By using the Kolmogorov forward equation, the density can be expressed in terms of the drift and diffusion term. Inverting this expression, one can write the drift (diffusion) term as a functional of the density and the diffusion (drift) term. This allows us to uniquely identify the drift (diffusion) term given a parameterisation of the diffusion (drift) together with a nonparametric estimator of the invariant density. This idea originates from Wong (1964), and was further developed in Hansen and Scheinkman (1995), and Hansen et al. (1998). Aït-Sahalia (1996a) made use of the same link to estimate his semiparametric diffusion model. Due to the higher level of generality, our estimator becomes more involved than the one in Aït-Sahalia

(1996a) though. There, a closed form estimator for the parametric part is derived, not depending on the nonparametric part. Unfortunately, in the general case it does not appear as if one can separate the estimation of the parametric part from the nonparametric one when given discrete observations. Instead, our estimator is obtained in the following three steps: First, we obtain a nonparametric estimator of the marginal density. Then the parametric part is estimated using the log-transition density of the diffusion process with the marginal density estimator plugged in as a nuisance parameter. Finally, the nonparametric part is estimated as a functional of the nonparametric density estimator and the parametric estimator.

The benefits from using the log-transition density to estimate the parameter are twofold: First, it is more likely that the parameter is identified since the transition density gives a full description of the probability structure of the sampled process.[1] Second, assuming that the nonparametric part is known, estimation of the parametric part by the log-transition density yields the efficient MLE. One would expect the semiparametric estimator to be close to the (infeasible) fully parametric MLE, and thereby enjoy a high level of efficiency.

Since it is not possible to directly evaluate the transition density, we propose either to use approximate (e.g. Aït-Sahalia, 2002) or simulation-based methods (see e.g. Durham and Gallant, 2002) in order to implement the estimator. The estimator obtained from these methods will enjoy the same properties as the actual, but infeasible one, under suitable conditions. The finite sample properties of the estimator using approximate likelihood is investigated in a small simulation study. Here, we will see that even for moderate sample sizes, our estimator performs well, and that the approximate method does a good job.

Under regularity conditions, we derive the asymptotic properties of the estimator, showing that the parametric part is $\sqrt{n}$-consistent, while the nonparametric part has a slower convergence rate. Also, the estimator is shown to follow a normal distribution asymptotically. The asymptotics of the estimator are based on discrete observations with a fixed time distance in between. This is in contrast to the papers on nonparametric kernel estimation of the drift and diffusion cited above, and is a desirable property since a continuous time record of observations may not be available in practice. High frequency (so-called tick-by-tick) data of, e.g., stock prices and exchange rates are now widely available. One could argue that these present a (nearly) continuous record, but the data often suffers from various market microstructure effects, see for example Dunis and Zhou (1998). One may therefore be willing to sacrifice some of the available observations to avoid having to deal with such effects, and only use observations of lower frequency (e.g. daily) when estimating the diffusion model.

The paper is organised as follows: In Section 2, we set up the framework and give an informal introduction to the proposed estimation procedure. In Section 3, theoretical results concerning the nonparametric part of the estimator are given. The asymptotics of the parametric part of the estimator is derived in Section 4. We discuss the efficiency of the parametric part in Section 5, and propose a 1-step adjustment which should reach the semiparametric efficiency bound. The implementation of the estimator is discussed in Section 6, and the results of the simulation study is presented in Section 7.

[1] A related problem is the so-called aliasing-problem where discretely sampled stochastic processes are indistinguishable, c.f. Phillips (1973). Hansen and Scheinkman (1995, p. 786) show however that the aliasing problem does not exist for reversible Markov processes.

We conclude in Section 8. All proofs and lemmas are collected into the appendices.

Throughout the text, $g^{(\lambda)}(x;\theta)$ denotes the λth derivative w.r.t. x of a function $g : \mathbb{R} \times \Theta \mapsto \mathbb{R}$ with $g^{(0)} \equiv g$, while $\dot{g}(x;\theta)$ and $\ddot{g}(x;\theta)$ denote the first and second derivative w.r.t. θ. At times we shall however also denote derivatives by $\partial^{ij}_{x,\theta} g(x;\theta) \equiv \partial^i \partial^j g(x;\theta) / \partial^i x \partial^j \theta$. We shall write $\|g\|_\infty = \sup_{x \in I} |g(x)|$ and $\|g\|_2 = (\int_I |g(x)|^2 dx)^{1/2}$ for any function with domain $I \subseteq \mathbb{R}$.

2 Framework

Let $\{X_t\} = \{X_t : t \geq 0\}$ be the stochastic process solving the following homogenous SDE,

$$dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dW_t, \tag{1}$$

where $\{W_t\}$ is a standard Brownian motion. The domain of $\{X_t\}$ is denoted $I = (l, r)$ where $-\infty \leq l < r \leq \infty$. We define the scale density $s(x) = \exp\left[-2\int_{x^*}^{x} \mu(y)/\sigma^2(y)\,dy\right]$, for some x^* in the interior of I. Sufficient conditions for strong stationarity are (S1) $\int_l^{x^*} s(x)\,dx = -\infty$, $\int_{x^*}^r s(x)\,dx = +\infty$, and (S2) $1/M \equiv \int_l^r \left[s(x)\sigma^2(x)\right]^{-1} dx < \infty$, c.f. Karlin and Taylor (1981, Section 15.6) and Karatzas and Shreve (1991, Section 5.5). Under these conditions, $\{X_t\}$ is stationary and ergodic with an invariant measure π, $\pi(A) = \int_I P(X_t \in A | X_0 = x)\,d\pi(x)$ for any Borel-set A, which has a density given by[2]

$$\pi(x) = \frac{M}{s(x)\sigma^2(x)} = \frac{M}{\sigma^2(x)} \exp\left[2\int_{x^*}^{x} \frac{\mu(y)}{\sigma^2(y)} dy\right]. \tag{2}$$

In a parametric framework, models for the above diffusion process is normally constructed by specifying the drift term, μ, and the diffusion term, σ^2, up to an unknown parameter vector $\theta \in \Theta$ where $\Theta \subseteq \mathbb{R}^d$ is a finite-dimensional parameter space. We see from (2) that one then implicitly also specifies the stationary density. It is possible to revert (2) in either of the two following ways,

$$\begin{aligned} \mu(x) &= \frac{1}{2\pi(x)} \frac{\partial}{\partial x}\left[\sigma^2(x)\pi(x)\right], & (3) \\ \sigma^2(x) &= \frac{2}{\pi(x)} \int_l^x \mu(y)\pi(y)\,dy. & (4) \end{aligned}$$

So an alternative specification scheme would be to specify the marginal density together with either the drift or the diffusion term, an idea originating from Wong (1964); see also Cobb et al. (1983), Hansen and Scheinkman (1995), Hansen et al. (1998). This could be done in a fully parametric framework, but here we only specify either the drift or the diffusion term and then rely on a nonparametric estimator of π. For example, we may parameterise the diffusion term, and then plug this into (3) together with a nonparametric estimator of π. We thereby obtain a semiparametric estimator of μ, by which we mean that it depends both on a parameter, θ, and a function, π. These considerations lead us to suggest the following two semiparametric classes of diffusion models:

Class 1:

$$dX_t = \mu(X_t)\,dt + \sigma(X_t;\theta)\,dW_t, \tag{5}$$

with $\mu(\cdot)$ unknown and $\sigma^2(\cdot;\theta)$ known up to the parameter θ.

[2] We here use π to denote both the measure and the density.

Class 2:

$$dX_t = \mu(X_t;\theta)\,dt + \sigma(X_t)\,dW_t, \tag{6}$$

with $\mu(\cdot;\theta)$ known up to the parameter θ and $\sigma^2(\cdot)$ unknown.

Here and in the following, μ_0, σ_0^2 and π_0 will denote the true drift, diffusion and invariant density respectively associated with the data-generating process. To discuss the estimation of the two classes of models, let us as an example consider a model from Class 1. In this case, we are given a parameterisation of the diffusion term, $\sigma^2(\cdot;\theta)$, which we plug into the RHS of (3) together with a density π,

$$\mu(x;\theta,\pi) = \frac{1}{2\pi(x)}\frac{\partial}{\partial x}\left[\sigma^2(x;\theta)\,\pi(x)\right]. \tag{7}$$

To obtain an estimator of θ we then make use of the transition density p of $\{X_t\}$, which is characterised by $P(X_{t+\Delta} \in A|X_t = x) = \int_A p(y|x)\,dy$ for any Borel-set A. Since $\{X_t\}$ is completely characterised by μ and σ, p is a functional of these two, $p(y|x) = p(y|x;\mu(\cdot),\sigma(\cdot))$. In the following section, a precise expression of p as a functional of μ and σ is derived by utilising results of Dacunha-Castelle and Florens-Zmirou (1986). By plugging in $\sigma(x;\theta)$ and $\mu(x;\theta,\pi)$, a semiparametric version of the transition density, $p(y|x;\theta,\pi) = p(y|x;\mu(\cdot;\theta,\pi),\sigma(\cdot;\theta))$, now appears. This version of the transition density will be employed to perform MLE-like estimation of θ given a nonparametric estimator of π. Let $X_0, X_\Delta, X_{2\Delta}..., X_{n\Delta}$ be $n+1$ observations obtained from (5), where $\Delta > 0$ is the fixed time distance between observations; without loss of generality, we set $\Delta \equiv 1$ in the following.[3] The following nonparametric kernel estimator of the rth derivative, $\pi_0^{(r)}$ (assuming that it exists), is then available,

$$\hat{\pi}^{(r)}(x) = \frac{1}{nh_r^{r+1}}\sum_{i=0}^{n-1} K^{(r)}\left(\frac{x-X_i}{h_r}\right), \quad r \geq 1, \tag{8}$$

for a kernel K and a bandwidth h_r; see Silverman (1986) for an introduction to these concepts. Note that we use potentially different bandwidths to estimate each derivative. Under regularity conditions, including $h_r = h_{r,n} \to 0$ and $nh_r^{2r+1} \to \infty$, $\hat{\pi}^{(r)}(x) \to^P \pi_0^{(r)}(x)$ as $n \to \infty$. We plug $\hat{\pi}$ and $\hat{\pi}^{(1)}$ into (7), yielding $\hat{\mu}(x;\theta) = \mu(x;\theta,\hat{\pi})$, which in turn is plugged into the transition density. We then propose to estimate θ by

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L_n(\theta,\hat{\mu}(\cdot;\theta)) \tag{9}$$

where

$$L_n(\theta,\mu) = \frac{1}{n}\sum_{i=1}^{n} \log p(X_{i+1}|X_i;\mu,\sigma(\cdot;\theta)). \tag{10}$$

Once $\hat{\theta}$ has been found, the obvious pointwise estimator of $\sigma^2(x)$ is $\sigma^2(x;\hat{\theta})$ while $\mu(x)$ is estimated by plugging $\hat{\theta}$ and $\hat{\pi}$ into (7) yielding $\hat{\mu}(x) = \mu(x;\hat{\theta},\hat{\pi})$. The above procedure is also applicable for models from Class 2, only this time we are given a full parameterisation of $\mu(\cdot) = \mu(\cdot;\theta)$, which can be substituted into (4) together with a nonparametric estimator of π, thereby obtaining a semiparametric estimator of $\sigma^2(\cdot) = \sigma^2(\cdot;\theta,\pi)$.

[3] To simplify the exposition, equidistant observations over time are assumed; our results can be extended to allow for varying time distances between observations.

The dependence of the nonparametric estimators, $\hat{\mu}(x)$ in Class 1 and $\hat{\sigma}^2(x)$ in Class 2, on the smoothing parameter h (and a trimming parameter introduced later) chosen by the user is an undesirable feature, which they share with many other non- and semiparametric estimators. The sensitivity of the estimators towards h can be high, and one therefore has to be careful when choosing the bandwidth. Too small values of h can give imprecise estimates, while a too large choice can induce bias. Rules of thumb are often applied for the bandwidth choice, but data driven methods such as cross-validation may lead to better performance. In our framework, such methods are not readily available however. A further discussion of these and related issues can be found in Section 6.

The estimation procedure described above belongs to a general class of semiparametric estimation problems, where an estimator of a finite-dimensional parameter θ is obtained with the help of a preliminary estimator of an infinite-dimensional nuisance parameter (here, π). General treatments of the asymptotic properties of such profiled/concentrated semiparametric estimators can be found in e.g. Andrews (1994), Chen et al. (2003), Newey and McFadden (1994, Section 8). The estimation of the finite-dimensional parameter is performed by what we may call semiparametric MLE. There is a large literature on non- and semiparametric MLE,[4] but there the infinite-dimensional parameter is estimated together with the finite-dimensional one, while here we make use of a preliminary estimator of the former. This makes our asymptotic theory somewhat different from that strand of the literature. Instead our estimator fits nicely into a general class of semiparametric two-step estimators: In the first step a function is (nonparametrically) estimated, while in the second step this is used to obtain an estimator of a finite-dimensional parameter. So in this setting the function estimated in the first step can be seen as a nuisance parameter. In our case, the function in question is the invariant density. Chen et al. (2003) and Newey and McFadden (1994, Section 8) give general conditions for consistency and asymptotic normality for such profiled semiparametric estimators. Unfortunately, the problem at hand here cannot directly be dealt with in the framework of those two studies since we have to introduce trimming of our nonparametric estimators. We therefore have to modify their conditions in order to establish our theoretical results; Ai (1997) and Robinson (1988) contain related applications of trimming in a semiparametric framework. Furthermore, the transition density takes a very complicated form, and a careful analysis of it as a function of the drift and diffusion function is required in order to derive the asymptotic properties. In particular, the derivation of the asymptotic distribution of the parametric part is very cumbersome, and we are unable to give an explicit expression for the resulting asymptotic variance. We are however able to set up a consistent estimator of it. Finally, Chen et al. (2003) and Newey and McFadden (1994) only give conditions for i.i.d. data, while our observations are dependent. In order to handle this additional complication, we have to assume that our process is not only stationary, but weakly dependent (also known as β-mixing), and restrict the decay rate of the mixing-coefficients in a suitable manner. This should be seen as a technical assumption however used to facilitate our analysis rather than a necessary property needed for the results to carry through.

There are certain obstacles with the implementation of the proposed estimator since the transition density p for general specifications of μ and σ^2 cannot be written in an explicit form, thereby not allowing for direct evaluation. We resolve this problem by relying on either approximate methods (see e.g. Lo 1988, Aït-Sahalia 2002) or simulation-based methods (see e.g. Durham and Gallant 2002,

[4] See for example Murhpy and Van der Vaart (2000) and the references therein.

Elerian et al 2001, Hurn et al 2003, Pedersen 1995). Applying such methods in the implementation of our estimator will have an asymptotically negligible effect on $\hat{\theta}$ if the order of approximation is allowed to increase with the sample size at a fast enough rate.

We remark that other criterion functions than $\log p$ could be used to estimate θ. In Aït-Sahalia (1996a) for example OLS is used; it is not clear however if this idea can be adapted to more general cases or only works for his specific choice of parameterisation. There is a variety of other estimating procedures in the literature for diffusion models, see e.g. Duffie and Singleton (1993), Hansen and Scheinkman (1995), Gallant and Long (1997), Gallant and Tauchen (1996), Gouriéroux et al. (1993), Sørensen (1997), but the log-likelihood approach is the most natural choice, and one would expect that this would yield a near-optimal estimator.

There is also room for different estimators of π_0. Our theoretical results are based on the use of the above kernel estimator of π_0, but can be substituted with alternative estimators such as series or spline estimators, c.f. Stone (1990), as long as one is able to show uniform consistency with a sufficiently high convergence rate for this.

Observe that if π_0 was known, we would be in a fully parametric framework and $\hat{\theta}$ would be the maximum-likelihood estimator (MLE), which under regularity conditions would enjoy full efficiency. But since we have not fully specified our model, the asymptotic variance of $\hat{\theta}$ may not reach the Cramer-Rao bound. One would however expect that the asymptotic properties of $\hat{\theta}$ are closely related to the fully parametric MLE. As we shall see, the asymptotic distribution of $\hat{\theta}$ in fact equals that of the fully parametric MLE plus an additional term entering the variance; this is due to the fact that we use an estimator of π_0 instead of the unknown density itself. This is related to the issue of semiparametric efficiency, see Newey (1990) and Severini and Tripathi (2001) for overviews. It could be of interest to derive the efficiency bound for the semiparametric models of this paper, and see whether our estimator reaches it. This is non-trivial though. Most of the existing literature on semiparametric efficiency is concerned with i.i.d. data, while we work with a Markov process. Moreover, the analysis of the transition density as a functional of π is not easy, and will require a lot of additional work. In Section 5, we give a brief discussion of these issues, and propose a 1-step adjustment to our semiparametric estimator which we conjecture will reach the semiparametric efficiency bound. A rigorous treatment of the efficiency bound and the 1-step adjustment is left for future research.

As stressed earlier, we here restrict our attention to stationary diffusion processes. The above identification scheme can however be extended to a wider class of processes satisfying (S1), but not necessarily (S2). In this case, the invariant density π exists, but is not necessarily integrable, allowing for $\int_I \pi(x)\,dx = +\infty$. The density will still satisfy (2) (leaving out M), such that the relation given in (3) remains valid, while for (4) to hold one has to require $\lim_{x\to l} \pi(x)\,\sigma^2(x) = 0$.[5] In the groundbreaking work by Bandi and Phillips (2003), it is demonstrated that for this extended class of "weakly" non-stationary (so-called recurrent) processes, μ and σ^2 can be consistently estimated by kernel methods as $\Delta \to 0$; for related results, we refer to Karlsen and Tjøstheim (2001) and Park and Phillips (1998). However, it is not clear what the asymptotic behaviour of the estimators proposed above will be when (S2) does not hold; this will be investigated in future research.

Our estimation procedure cannot readily be extended to general multivariate diffusion models, since

[5] This will automatically be satisfied under (S2).

the link between the invariant density, the drift and the diffusion term utilised here does not necessarily hold in higher dimensions. If one is ready to restrict the attention to the class of multivariate models satisfying this relation,[6] the proposed estimation procedure should still work. But it would suffer from the well-known curse of dimensionality of nonparametric estimators. Moreover, the transition density in the general multivariate case is even more difficult to analyse than in the univariate one, so the task of establishing theoretical results for the parametric part of the estimator in a multivariate setting will be a rather difficult one.

3 The Nonparametric Estimator

In this section, we show that the nonparametric estimators of μ and σ^2 proposed in the previous section will be pointwise consistent and asymptotically normally distributed for any given $\sqrt{n}$-consistent estimator of θ. So we here assume the existence of such an estimator. In the next section we show that the estimator of θ proposed in the previous section is indeed $\sqrt{n}$-consistent. We also give uniform convergence rates and define a simple test statistic allowing one to test any parametric submodel against the semiparametric alternative.

In Class 1, we simply plug in the initial estimators of the marginal density π and the parameter θ, yielding

$$\hat{\mu}(x) = \frac{1}{2\hat{\pi}(x)} \frac{\partial}{\partial x} \left[\sigma^2(x;\hat{\theta})\hat{\pi}(x) \right], \tag{11}$$

where $\hat{\pi}$ is the kernel estimator in (8) and $\hat{\theta}$ is the estimator of θ. For Class 2, we observe that by the Law of Large Numbers (LLN) for stationary and ergodic sequences,

$$\frac{1}{n}\sum_{i=1}^{n} 1_{(l,x)}(X_i)\,\mu(X_i;\theta) \to^P \int_l^x \pi_0(y)\,\mu(y;\theta)\,dy,$$

for any $(x,\theta) \in \mathbb{R} \times \Theta$ given the moment exists. We then define[7]

$$\hat{\sigma}^2(x) = \frac{2}{\hat{\pi}(x)} \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x)}(X_i)\,\mu(X_i;\hat{\theta}).$$

As noted earlier, we have to assume that $\{X_t\}$ is stationary and ergodic in order to be able to identify the unspecified term. In fact, we require it to be geometrically β-mixing. The results stated in this section will actually hold under weaker mixing conditions. But since in the next section we need β-mixing in order to employ U-statistics results for dependent sequences (see Serfling, 1980; Arcones, 1995), we impose this restriction throughout for clarity. Similar conditions have been imposed elsewhere in the nonparametric literature to control the dependence structure, for example in Aït-Sahalia (1996a) and Robinson (1989). The following assumption (A0) is sufficient for $\{X_t\}$ to be well-defined, stationary and geometrically β-mixing. In particular, (A0) implies (S1)-(S2) given in the previous section.

[6] This restriction is for example imposed by Chen et al (2000b) in their nonparametric study of multivariate diffusion models.

[7] An alternative estimator would be $\hat{\sigma}^2(x) = 2\int_l^x \hat{\pi}(y)\,\mu(X_{i\Delta};\hat{\theta})dy/\hat{\pi}(x)$. The advantage of this is that it is continuous and differentiable. On the other hand, $\int_l^x \hat{\pi}(y)\,\mu(X_{i\Delta};\theta)dy$ is a biased estimator of $\int_l^x \pi_0(y)\,\mu(X_{i\Delta};\theta)dy$.

A0 (i) The drift $\mu_0(\cdot)$ and diffusion $\sigma_0^2(\cdot) > 0$ are continuously differentiable, and (ii) there exists a function $V : \mathbb{R} \mapsto \mathbb{R}_+$ satisfying $V(x) \geq |x|^{\bar{q}}$ as $x \to l^+$ and r^- with $\bar{q} > 1$, and constants $b, c > 0$ such that

$$\mu_0(x)\, V\ \ (x) + \frac{1}{2}\sigma_0^2(x)\, V\ \ (x) \leq -cV(x) + b. \tag{12}$$

Under (A0), (i), there exists a unique solution to (1), c.f. Karatzas and Shreve (1991, Theorem 5.5.15 and Corollary 5.3.23). The condition given in (12) is a so-called drift criterion, known from the ergodic theory for Markov chains. The function V is a norm-like function, and under (12), there exists $\rho \in (0, 1)$ such that $E[V(X_\Delta)\,|X_0 = x] \leq \rho V(x) + b$, ensuring that the process is mean-reverting. This condition not only implies that the process is β-mixing with exponentially decaying mixing-coefficients, but also that $E_\pi\left[|X_0|^{\bar{q}}\right] < \infty$, where $E_\pi[\cdot]$ denotes the expectations operator w.r.t. the stationary measure of X. (A0) is based on results by Meyn and Tweedie (1992); alternative conditions for mixing of diffusion processes can be found in Chen et al. (1999), Hansen and Scheinkman (1995) and Veretennikov (1997); see also Karatzas and Shreve (1991, Section 5.5). Most parametric model found in the literature can be shown to satisfy (A0): Continuity and differentiability of μ and σ^2 are normally satisfied, and with $V(x) = x^q$, $q > 1$, the second condition becomes

$$q\mu(x)\, x + \frac{q(q-1)}{2}\sigma^2(x) \leq -cx^2, \tag{13}$$

as $|x| \to \infty$ (assuming $I = \mathbb{R}$). If for example $\mu(x) = \beta(\alpha - x)$, the condition becomes

$$\begin{aligned} \sigma^2(x) &\leq c_1 x^2 + c_2, \quad |x| \to \infty, \\ c_1 &\equiv \frac{2(q\beta - c)}{q(q-1)},\ c_1 \equiv \frac{-2q\beta\alpha}{q(q-1)} \end{aligned}$$

with $0 < c < q\beta$ where we require $\beta > 0$.

In some cases, one might want to have precise expressions of the convergence rate. This is for example the case in the next section where the convergence must take place at a sufficiently high rate. To speed up the convergence, we employ so-called higher order kernels in the estimation of $\pi\ (\omega)0$

$\sqrt{n}$-consistency of θ, see e.g. Robinson (1988) for an early application of higher order kernels to semiparametric estimation. Andrews (1995) gives uniform convergence rates of the density estimator and its derivatives using this type of kernels under fairly general conditions. We apply his results here even though the convergence rates stated there are not optimal. Masry (1996) obtains optimal convergence rates but only considers convergence on compact sets, while we wish to allow for a non-compact domain I. Similarly, Bosq (1998) establishes uniform consistency with a near optimal convergence rate on the whole of $\mathbb{R}$ for Markov processes, but estimators of the derivatives of the density are not considered. One could extend their results to hold on the whole of $\mathbb{R}$ and for density derivatives, but this is not the focus of this paper and we shall simply apply the results of Andrews (1995) here. The pointwise asymptotic distribution of $\hat{\pi}^{(r)}$ has been established in a number of papers, see e.g. Robinson (1983). Given the consistency and the asymptotic distribution of $\hat{\pi}^{(r)}$, the asymptotic properties of the two nonparametric estimators can now be derived using standard delta-methods.

One might also wish to have uniform convergence of the nonparametric estimators. This is for example needed in the next section when dealing with the asymptotics of $\hat{\theta}$. We wish to show uniform consistency of the nonparametric estimators in the supremum-norm. However, since the estimators and the limits themselves potentially are unbounded functions, this is not readily possible. To circumvent this problem, we control the tail behaviour of the estimator by trimming, ensuring that the nonparametric estimator equals zero outside a compact, but growing set. We define

$$\hat{A} = \{x | \hat{\pi}(x) \geq a\} \tag{14}$$

for some sequence $a = a_n \to 0$. We then show uniform convergence on the increasing set $\hat{A}$.

In addition to (A0), we impose the following assumptions:

A1 The true density, π_0, is ω times continuously differentiable with bounded derivatives in a neighbourhood of each point $x_i \in I$ where $x_i \neq x_j$, $i, j = 1, ..., d$.

A2 $\sqrt{n}(\hat{\theta} - \theta_0) = O_P(1)$.

The condition that π_0 is ω times continuously differentiable is satisfied if μ_0 and σ_0^2 are ω times continuously differentiable, c.f. (2). Since the rate of convergence of $\hat{\pi}$ and its derivatives is slower than $\sqrt{n}$, the asymptotic distribution of $\hat{\theta}$ will not have any effect on the ones of $\hat{\mu}$ and $\hat{\sigma}^2$. In particular, the efficiency of $\hat{\theta}$ is not important in this context. Condition (A2) can be weakened to allow for slower convergence rate of $\hat{\theta}$, as long as it is faster than $\sqrt{nh^3}$ ($\sqrt{nh}$) when estimating μ_0 (σ_0^2). If this is not the case, the asymptotic distribution of $\hat{\theta}$ will influence the one of the nonparametric estimator.

In the following let $\{x_i | 1 \leq i \leq N\}$ be N distinct points in I, $x_i \neq x_j$ for $i \neq j$.

Theorem 1 (Class 1) *Assume that $K \in \mathcal{K}(\omega, 1)$, and (A0)-(A2) hold with $\omega \geq 3$; $\theta \mapsto \sigma^2(x;\theta)$ is continuously differentiable satisfying $||\partial_{x,\theta}^{ij}\sigma^2(x;\theta)|| \leq C(1 + |x|^{\bar{q}})$, $i, j = 0, 1$; and $h_i \to 0$, and $nh_i^{2i+1} \to \infty$, $i = 0, 1$. Then the nonparametric estimator of the drift is pointwise consistent and asymptotically normally distributed,*

$$\sqrt{nh_1^3}\,[\hat{\mu}(x_1) - \mu_0(x_1), ..., \hat{\mu}(x_d) - \mu$$

where $V_\mu = \mathrm{diag}(\{V_\mu(x_i)\}_{i=1}^N)$ *is a diagonal matrix and* $V_\mu(x) = \frac{1}{4}||K^{(1)}||_2^2\sigma_0^4(x)/\pi_0(x)$*. Moreover,*

$$\sup_{x\in\hat{A}} |\hat{\mu}(x) - \mu_0(x)| = \sum_{i=0}^{1}\left\{O_P(n^{-1/2}a^{i-3}h_i^{-1-i}) + O_P(a^{i-3}h_i^{\omega-i})\right\}.$$

Theorem 2 (Class 2) *Assume that* $K \in \mathcal{K}(\omega, 0)$*, and (A0)-(A2) hold with* $\omega \geq 2$*;* $\theta \mapsto \mu(x;\theta)$ *is continuously differentiable, satisfying* $||\partial_\theta^i \mu(x;\theta)|| \leq C\left(1 + |x|^{\bar{q}/2}\right)$*,* $i = 0, 1$*;* $h_0 \to 0$ *and* $nh_0 \to \infty$*. Then the nonparametric estimator of the diffusion term is pointwise consistent and asymptotically normally distributed,*

$$\sqrt{nh_0}\left[\hat{\sigma}^2(x_1) - \sigma_0^2(x_1), ..., \hat{\sigma}^2(x_d) - \sigma_0^2(x_d)\right] \xrightarrow{d} N(0, V_\sigma),$$

where $V_\sigma = \mathrm{diag}(\{V_\sigma(x_i)\}_{i=1}^N)$ *is a diagonal matrix with* $V_\sigma(x) = ||K||_2^2\sigma_0^4(x)/\pi_0(x)$*. Moreover,*

$$\sup_{x\in\hat{A}}\left|\hat{\sigma}^2(x) - \sigma_0^2(x)\right| = O_P(n^{-1/2}a^{-2}h_0^{-1}) + O_P\left(a^{-2}h_0^{\omega}\right).$$

Pointwise estimators of the asymptotic variance for $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$ respectively can be constructed as

$$\hat{V}_\mu(x) = \frac{[\int K\ (y)^2\,dy]\sigma^4(x;\hat{\theta})}{4\hat{\pi}(x)},\quad \hat{V}_\sigma(x) = \frac{[\int K\,(y)^2\,dy]\hat{\sigma}^4(x)}{\hat{\pi}(x)}. \tag{15}$$

We only state results for the estimation of μ and σ^2 but one is able to derive similar results for the estimators of the derivatives of μ and σ^2. Observe that both nonparametric estimators are asymptotically independent across the points $\{x_i\}_{i=1}^N$. This is a well-known property of kernel-estimators, c.f. Robinson (1983), which facilitates global inference, for example when constructing pointwise confidence bands, and testing hypotheses (see below).

The pointwise rate of convergence of $\hat{\mu}$ might at first appear surprisingly slow given the interpretation of μ as the (instantaneous) conditional mean. In a standard nonparametric regression model, one is able to estimate the conditional mean with rate $\sqrt{nh}$, but observe that in our case μ is not only a functional of π but also of its derivative $\pi^{(1)}$ with the nonparametric estimator $\hat{\pi}^{(1)}$ having slower convergence rate than $\hat{\pi}$, $\sqrt{nh^3}$ relative to $\sqrt{nh}$. In contrast, we obtain the standard rate of convergence as found in kernel regressions for $\hat{\sigma}^2$. This owes to the fact that $\hat{\sigma}^2$ is only a function of $\hat{\pi}$ and not any of its derivatives. Thus, the drift is more difficult to estimate than the diffusion term in a nonparametric setting. This observation has been made elsewhere in the literature. Gobet et al. (2003) report similar results for their sieve-estimator, and coin the nonparametric estimation of μ given discrete observations as an "ill-posed problem". Similarly, Bandi and Phillips (2003) demonstrate that for a stationary diffusion, it is only possible to estimate $\mu(x)$ nonparametrically with $\sqrt{n\Delta h}$-rate, while $\sigma^2(x)$ can be estimated at the faster rate$\sqrt{nh}$ as $\Delta \to 0$ and $n\Delta \to \infty$.

The first part of the result stated in Theorem 2 has already been obtained by Aït-Sahalia (1996a) for the special case $\mu(x;\theta) = \beta(\alpha - x)$. So we here extend his result to hold for a more general class of semiparametric diffusion models.

Next, we set up a simple test for a parametric diffusion submodel against our semiparametric alternative. We start out with Class 1, for which we consider a parametric specification of the drift, $\mu(\cdot;\beta)$ for $\beta \in \mathcal{B} \subseteq \mathbb{R}^d$. We then wish to test the following nested hypothesis

$$H_{10} : \mu_0(\cdot) = \mu(\cdot;\beta_0) \text{ for some } \beta_0 \in \mathcal{B}$$

against the nonparametric alternative,

$$H_{11} : \mu_0(\cdot) \neq \mu(\cdot;\beta) \text{ for all } \beta \in \mathcal{B},$$

Under the null the model is fully specified and the parameters (θ, β) can be estimated using standard methods, with the obvious one being MLE. Under regularity conditions, this will yield $\sqrt{n}$-consistent estimators of (θ, β). We base the our test statistic on the pointwise difference between the nonparametric and parametric estimate. For similar test procedures for conditional means, see Gozalo (1995, 1997) and Härdle and Mammen (1993). Under H_{10}, $\mu(x;\hat{\beta}) - \mu_0(x) = O_P\left(n^{-1/2}\right)$, when a $\sqrt{n}$-consistent estimator $\hat{\beta}$ is available[8], and smoothness conditions are imposed on $\beta \mapsto \mu(x;\beta)$, such that

$$\sqrt{nh^3}\frac{\hat{\mu}(x) - \mu(x;\hat{\beta})}{\hat{V}_\mu^{1/2}(x)} = \frac{\hat{V}_\mu^{1/2}(x)}{V_\mu^{1/2}(x)}\sqrt{nh^3}\frac{\hat{\mu}(x) - \mu_0(x)}{V_\mu^{1/2}(x)} + o_P(1) \quad {}^{d} \qquad\qquad P$$

it can be shown that the number of points N used in the test statistic for e.g. μ can grow with n as long as it does so at a rate slower than $\sqrt{nh^3}$.

Instead of relying on the asymptotic distribution as an approximation of the finite-sample properties of T_n, it may be worthwhile to use bootstrapping since nonparametric goodness-of-fit tests appear to exhibit significant differences between nominal and true size in finite samples, see e.g. Fan (1994, 1995). It should be possible to show consistency of the bootstrap in our case by following her arguments.[9]

4 The Semiparametric Estimator

In this section we construct an estimator for θ and derive its asymptotic properties in each of the two classes of models. This is done along the lines proposed in Section 2, using the log-transition density to define our criterion function. For each class, we show that $\hat{\theta}$ is consistent, and converges weakly towards a normal distribution with $\sqrt{n}$-rate.

The results stated in this section are established under the assumption that the domain $I = \mathbb{R}$. We conjecture that our results also hold for other domains by using arguments similar to those in Aït-Sahalia (2002). Allowing for such will however further complicate the proofs, since we need to give specific treatment to the boundary behaviour of $\{X_t\}$; in particular, the transition density will depend on the specified domain.

Drawing upon results of Dacunha-Castelle and Florens-Zmirou (1986), we are able to obtain an expression for $\log p$ (c.f. Lemma 38) as a functional of μ and σ^2. This characterisation was also utilised by Aït-Sahalia (2002) in his derivation of an approximation of the likelihood-function. The log-density takes the following form,

$$\log p\left(x \left|x_0; \mu, \sigma^2\right.\right) \propto -\frac{1}{4} \log\left[\sigma^2(x)\,\sigma^2(x_0)\right] - \left(\int_{x_0}^{x} \sigma^{-1}(w)\,dw\right)^2 /2 + \log\left(E_B\left[\psi\left(x|x_0\right)\right]\right), \tag{16}$$

where

$$\begin{aligned} \psi\left(x\left|x_0\right.\right) &= \exp\left[\Delta \int_0^1 \lambda_Y\left(Z_t\left(x|x_0\right)\right)dt\right], & (17)\\ Z_t\left(x|x_0\right) &= \gamma^{-1}\left(t\gamma(x) + (1-t)\,\gamma(x_0) + B_t\right), & (18)\\ \lambda_Y(z) &= -\frac{1}{2}\left[\mu_Y^2(z) + \partial_x \mu_Y(z)\,\sigma(z)\right], & (19)\\ \mu_Y(z) &= \frac{\mu(z)}{\sigma(z)} - \frac{1}{4}\frac{\partial_x \sigma^2(z)}{\sigma(z)}, & (20)\\ \gamma(z) &= \int \sigma(z)^{-1}\,dz. & (21)\end{aligned}$$

and $\{B_t | 0 \leq t \leq 1\}$ is a standard Brownian Bridge with associated expectations operator $E_B[\cdot]$.[10] As can be seen, the function ψ depends on $\left(\mu, \sigma^2\right)$ in a fairly complicated way, so the analysis of $\log p$ as a functional of these is not straight forward. The analysis is further complicated by the presence of the Brownian Bridge in the expression of ψ.

[9] Fan (1994, 1995) only consider the i.i.d. bootstrap; in our setting a different bootstrap method have to be used, for example Horowitz (2003).

[10] See Karatzas and Shreve (1991, p. 358-360) for a definition of the Brownian Bridge.

4.1 Class 1

In this subsection, we derive the asymptotic properties of $\hat{\theta}$ in Class 1. We propose the following estimator

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L_n\left(\theta, \hat{\mu}\left(\cdot;\theta\right)\right), \tag{22}$$

where

$$L_n\left(\theta,\mu\right) = \frac{1}{n}\sum_{i=1}^{n}\log p\left(X_{i+1}|X_i;\theta,\mu\right), \tag{23}$$

and $p\left(x|x_0;\theta,\mu\right) = p\left(x|x_0;\mu,\sigma^2\left(\cdot;\theta\right)\right)$ with $p\left(x|x_0;\mu,\sigma^2\right)$ given in (16). We observe that ψ only depends on μ and $\mu^{(1)}$, so when showing consistency of $\hat{\theta}$, we only need to show $||\hat{\mu}^{(i)} - \hat{\mu}_0^{(i)}||_\infty \to^P 0$, $i = 0,1$. However, due to θ appearing in $Z_t\left(x|x_0;\theta\right)$, $\partial_\theta \log p$ depends on $\mu^{(i)}$, $i = 0,1,2$, and $\partial_\theta^2 \log p$ on $\mu^{(i)}$, $i = 0,1,2,3$. So in order to derive the asymptotic distribution of $\hat{\theta}$, we have to ensure that $||\hat{\mu}^{(i)} - \hat{\mu}_0^{(i)}||_\infty \to^P 0$, $i = 0,1,2,3$, and that convergence takes place with rate $n^{1/4}$.

We are now ready to set up the conditions, which we will work under.

C1.1 (A0)-(A1) holds with $\omega \geq 6$, the kernel $K \in \mathcal{K}\left(\omega,4\right)$, and the trimming function $\hat{T} \in \mathcal{T}\left(4\right)$.

C1.2 The diffusion function $x \mapsto \sigma\left(x;\theta\right)$ is six times continuously differentiable for any $\theta \in \Theta$; $\theta \mapsto \sigma^2\left(x;\theta\right)$ is three times continuously differentiable for any $x \in I$; $\underline{\sigma}^2 \leq \sigma^2\left(x;\theta\right)$, and $||\partial_{x,\theta}^{ij}\sigma^2\left(x;\theta\right)|| \leq \bar{\sigma}^2$, $0 \leq i \leq 4$ and $0 \leq j \leq 2$.

C1.3 (i) The drift function satisfies $||\partial_x^i\partial_\theta^j\mu_0\left(x;\theta\right)|| \leq C\left(1 + |x|^q\right)$, $0 \leq i \leq 6$ and $0 \leq j \leq 2$, with $4q + 2 + \delta \leq \bar{q}$ for some $\delta > 0$ where $\bar{q}$ given in (A0); (ii) $-C\left(1 + |z|^{\bar{q}}\right) \leq \lambda_Y\left(z;\theta,\mu_0\left(\cdot;\theta\right)\right) \leq \bar{\lambda}_Y$ uniformly in $\left(x,\theta\right)$.

C1.4 The parameter space $\Theta \subseteq \mathbb{R}^d$ is compact.

C1.5 The moment $L\left(\theta,\mu_0\left(\cdot;\theta\right)\right) = E_\pi\left[\log p\left(X_1|X_0;\theta,\mu_0\left(\cdot;\theta\right)\right)\right]$ has a unique maximum at $\theta_0 \in \Theta$, such that

$$H\left(\theta_0,\mu_0\right) \equiv -E_\pi\left[\frac{\partial^2}{\partial\theta^2}\log p\left(X_1|X_0;\theta_0,\mu_0\left(\cdot;\theta_0\right)\right)\right] \tag{24}$$

is positive definite.

C1.6a (i) $na^{2(i-4)}h_i^{2(1+i)} \to \infty$, (ii) $a^{i-4}h_i^{\omega-i} \to 0$,$\lambda = 0,1,2$, as $a, h_i \to 0$, for $i = 0,1,2$.

C1.6b (i) $na^{4(i-4)}h_i^{4(1+i)} \to \infty$ (ii) $na^{4(i-4)}h_i^{4(\omega-i)} \to 0$, (iii) $na^{4(4-i)}h_i^{1+i} \to 0$, and (iv) $a^{-2}h_i^{2(\omega-i)} \to 0$, for $0 \leq i \leq 3$; (v) $na^{-4}h_4^{10} \to \infty$, and (vi) $a^{-2}h_4^{\omega-4} \to 0$; (vii) $nP_{\pi\quad B}\left(a/2 \leq \pi_0\left(Z_t\right) \leq a\right) \to 0$; (viii) $nP_\pi\left(a/2 \leq \pi_0\left(X_0\right) \leq a\right) \to 0$.

C1.7 The density $\bar{p}\left(z\right)$ given by

$$\bar{p}\left(z\right) = \frac{1}{\sqrt{\Delta}\sigma_0\left(z\right)}\int_{\mathbb{R}^2} p\left(x|x_0\right)\pi_0\left(x_0\right)\left\{\int_0^1 p_B\left(t, Z^{-1}\left(t,x,x_0,z\right)\right)dt\right\}dxdx_0. \tag{25}$$

where

$$p_B\left(t,b\right) = \phi\left(\frac{b}{\sqrt{t\left(1-t\right)}}\right) = \frac{1}{\sqrt{2\pi t\left(1-t\right)}}\exp\left[\frac{-b^2}{2t\left(1-t\right)}\right].$$

and

$$Z^{-1}(t,x,x_0,z) = \frac{\gamma_0(z) - t\gamma_0(x) - (1-t)\gamma_0(x_0)}{\sqrt{\Delta}}.$$

satisfies $\overline{p}^{(i)} = O(\pi_0^{(i)})$, $0 \leq i \leq 4$.

The smoothness criteria on π_0 in (C1.1) are used to ensure that the transition density is well-defined, and to decrease the bias from the kernel estimation. As discussed earlier, a high degree of smoothness together with the use of higher-order kernels will reduce the bias of the kernel estimator, c.f. Lemma 39.

The smoothness assumptions on $\sigma^2(x;\theta)$ in (C1.2) is needed for the first and second derivative of $\log p$ w.r.t. θ to be well-defined. The boundedness conditions on $\sigma^2(x;\theta)$ and its derivatives are very restrictive. These bounds are primarily used to establish suitable bounds for the various terms entering $\log p$, in particular $E_B[\psi(x,x_0;\theta,\mu)]$. We conjecture that it should be possible to obtain these under weaker assumptions on σ^2, but this will complicate the proofs further. In practice the boundedness assumption should not be a problem, since one can always choose a parameterisation such that $\sigma^2(x;\theta)$ is constant outside a compact set, which can be chosen arbitrarily large.

The conditions (C1.1)-(C1.3) guarantee that the transition density exists (c.f. Aït-Sahalia (2002), Proposition 2). There is some tension between (A0) and (C1.3), which both impose growth conditions on the drift function. In a fully parametric framework, (C1.4)-(C1.5) are standard assumptions when deriving the asymptotic properties of the MLE. In fact, if π_0 was known, the MLE of θ is consistent and asymptotically normally distributed under (C1.1)-(C1.5), c.f. Aït-Sahalia (2002, Proposition 3 and Theorem 2).

Condition (C1.6a) and (C1.6b) restrict the choice of bandwidths and the trimming sequences to ensure that (a) $||\hat{\mu}^{(i)}(\cdot;\theta) - \hat{\mu}_0^{(i)}(\cdot;\theta)||_\infty \to^P 0$ and (b) $E_\pi[|\hat{\mu}_0^{(i)}(X_0;\theta) - \mu_0^{(i)}(X_0;\theta)|] \to 0$, together with derivatives w.r.t. θ, at a sufficiently fast rate. To prove consistency, we merely have to show that the convergence takes place, while $\sqrt{n}$-asymptotic normality requires that the convergence takes place with rate $n^{1/4}$. The first convergence creates a tension between a and h as they go to zero. Two bias and variance terms have to be controlled for: The one incurred from using $\hat{\pi}$ instead of π_0 in the estimation, which goes to zero as $h \to 0$, the other is caused by the trimming since the trimmed version of the score function may not equal zero for $a > 0$. One then has to balance the two effects to obtain consistency and asymptotic normality. For (b) to hold one needs that $a \to 0$, while for it to happen with rate $n^{1/4}$ we need to impose (C1.6b), (vii). The condition (C1.6b) is more restrictive than (C1.6a). When showing $\sqrt{n}$-asymptotic normality, further restrictions on the set of permissible bandwidth and trimming sequences are required since we now have to ensure that the two biases in (a) and (b) go to zero with a faster rate. The trimming bias can normally be avoided by introducing the trimming in such a way that the conditional expectations of $\partial_\theta \log p$ equals zero for any a as done in Ai (1997) and Robinson (1988). This does not appear to feasible here though since μ enters p in a complicated manner.

The measure $P_{\pi \quad B}$ appearing in (C1.6b) is the product of the probability measures associated with the Brownian Bridge and $\{X_i\}$ respectively. Primitive conditions for (C1.6b), (viii), can be derived under the assumption that $\pi_0(x)$ is monotonously decreasing (increasing) for $x > R$ ($x < -R$) for some

$R > 0$. Then, for a sufficiently small,

$$\begin{aligned} P_{\pi}\left(\pi_0\left(X_0\right) \leq a\right) &= \int_{z\ \pi_0(z)\leq a} \pi_0\left(z\right) dz \\ &= \int_0^a \frac{\pi_0\left(\underline{\pi}_0^{-1}\left(z\right)\right)}{\pi_0^{(1)}\left(\underline{\pi}_0^{-1}\left(z\right)\right)} dz - \int_0^a \frac{\pi_0\left(\overline{\pi}_0^{-1}\left(z\right)\right)}{\pi_0^{(1)}\left(\overline{\pi}_0^{-1}\left(z\right)\right)} dz \\ &= \int_0^a \frac{z}{\pi_0^{(1)}\left(\underline{\pi}_0^{-1}\left(z\right)\right)} dz - \int_0^a \frac{z}{\pi_0^{(1)}\left(\overline{\pi}_0^{-1}\left(z\right)\right)} dz, \end{aligned}$$

where $\underline{\pi}_0^{-1}$ $(\overline{\pi}_0^{-1})$ is the inverse of π_0 for $x < -R$ $(x > R)$. Assume that $\mu_0\left(z\right) \propto \pm z^p$ as $z \rightarrow \mp\infty$, for some $p > 0$. Since $\sigma\left(x;\theta_0\right)$ is assumed to be bounded, $\pi_0\left(z\right) \propto \exp\left[-\left|z\right|^{p+1}\right]$. Thus, $\pi_0^{(1)}\left(z\right) \propto -\left|z\right|^p \exp\left[-\left|z\right|^{p+1}\right]$, and $\underline{\pi}_0^{-1}\left(y\right) = -\left(-\log\left(y\right)\right)^{1/(p+1)}$ and $\overline{\pi}_0^{-1}\left(y\right) = \left(-\log\left(y\right)\right)^{1/(p+1)}$, yielding

$$P_{\pi}\left(\pi_0\left(X_0\right) \leq a\right) \simeq 2\int_0^a z^{p+1} \exp\left[-\log\left(z\right)\right] dz = 2\int_0^a z^p dz = \frac{2}{p+1} a^{p+1}.$$

So the tail-thickness of π_0 determines the rate with which a is allowed to go to zero.

The density $\bar{p}$ in (25) introduced in (C1.7) is implicitly given by

$$\int f\left(z\right) \bar{p}\left(z\right) dz = \int_0^1 E_{\pi\ \ B}\left[f\left(Z_t\left(X_1|X_0\right)\right)\right] dt.$$

The restriction imposed on $\bar{p}$ is used to ensure that the asymptotic variance of $\hat{\theta}$ is finite. It appears difficult to come up with primitive conditions for this to hold since $\bar{p}$ takes a very complex form.

In order to show consistency of the parametric part of our estimator, we basically have to demonstrate that the log-likelihood function is continuous w.r.t. μ in probability and that $\hat{\mu} \rightarrow^P \mu_0$ in a normed function space. Once this is established, standard consistency results for parametric estimators can be applied to $L_n\left(\theta, \mu_0\left(\cdot;\theta\right)\right)$, see e.g. Newey and McFadden (1994, Theorem 2.1) and Chen et al.

corresponding sample version,

$$\nabla S_n\left(\theta_0, \hat{\mu}_0\right)[d\mu] = \frac{1}{n}\sum_{i=1}^{n} \nabla s\left(X_i|X_{i-1}; \theta_0, \hat{\mu}_0\right)[d\mu] .$$

and its moment, $\nabla S\left(\theta_0, \hat{\mu}_0\right)[d\mu] = E_\pi\left[\nabla s\left(X_i|X_{i-1}; \theta_0, \hat{\mu}_0\right)[d\mu]\right]$. By Lemma 17, the pathwise derivative is well-defined, and satisfies

$$S_n\left(\theta_0, \hat{\mu}\right) - S_n\left(\theta_0, \hat{\mu}_0\right) - \nabla S_n\left(\theta_0, \hat{\mu}_0\right)[\hat{\mu} - \hat{\mu}_0] = o_P(n^{-1/2}).$$

It can be shown that $\nabla S_n\left(\theta_0, \hat{\mu}_0\right)[\hat{\mu} - \hat{\mu}_0] = \nabla S\left(\theta_0, \mu_0\right)[\hat{\mu} - \hat{\mu}_0] + o_P(n^{-1/2})$ using standard U-statistics results for weakly dependent sequences (c.f. Lemma 19 and 20). Finally, $\nabla S\left(\theta_0, \mu_0\right)[\hat{\mu} - \hat{\mu}_0]$ can be written as a normed sum plus a remainder term with the latter being asymptotically negligible,

$$\nabla S\left(\theta_0, \mu_0\right)[\hat{\mu} - \hat{\mu}_0] = \frac{1}{n}\sum_{i=1}^{n} \delta\left(X_{i-1}\right) + o_P(n^{-1/2}), \tag{26}$$

where $E_\pi\left[\delta\left(X_0\right)\right] = 0$ and $E_\pi\left[\|\delta\left(X_0\right)\|^2\right] < \infty$ (c.f. Lemma 21). These results combined with the fact that $H_n(\hat{\theta}, \hat{\mu})$ converges towards $H\left(\theta_0, \mu_0\right)$ (c.f. Lemma 22) proves the following result:

Theorem 6 (Class 1) *Assume that (C1.1)-(C1.7) hold, and that* $\theta_0 \in \text{int}\left(\Theta\right)$. *Then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, H_0^{-1}\left(H_0 + V_0\right) H_0^{-1}\right),$$

where $H_0 = H\left(\theta_0, \mu_0\right)$, *and* $V_0 = \Omega_0 + 2\sum_{i=1}^{\infty} \Omega_i$ *with* $\Omega_i = E_\pi\left[\delta\left(X_0\right)\delta\left(X_i\right)\quad\right]$.

The extra term, V_0, in the variance expression is an adjustment term due to the use of $\hat{\pi}$ instead of π_0 in the estimation. If π_0 was known, $V_0 \equiv 0$, and the asymptotic variance expression would collapse to the standard inverse information matrix, H_0^{-1}. Instead, we here experience an increase in the asymptotic variance. The derivation of (26) is based on the Riesz Representation Theorem, and we therefore are not able to supply a closed form expression for δ. We are however able to show that it has mean zero and finite variance. Furthermore, it is possible to derive a consistent estimator of it, following the same strategy as in Newey (1994a). This estimator can in turn be used to obtain an estimator of the asymptotic variance by using the so-called HAC variance estimators, see e.g. Robinson and Velasco (1997). Here, we present an estimator based on the idea of Newey and West (1987).

Theorem 7 (Class 1) *Assume that (C1.1)-(C1.7) hold and* $E_\pi[\|\delta\left(X_0\right)\|^{4+\delta}] < \infty$. *Then consistent estimators of* H_0 *and* V_0 *respectively are given by* $\hat{H}_n = H_n(\hat{\theta}, \hat{\mu})$ *and*

$$\hat{V}_n = \hat{\Omega}_0 + \sum_{i=1}^{M} w_{M,i}(\hat{\Omega}_i + \hat{\Omega}_i\ ),$$

where $w_{M,i} = 1 - [i/\left(M + 1\right)]$, $\hat{\Omega}_i = n^{-1}\sum_{j=i}^{n} \hat{\omega}_j \hat{\omega}_{j-i}$, $\hat{\omega}_j = \hat{\delta}_j - n^{-1}\sum_{k=0}^{n-1} \hat{\delta}_k$,

$$\hat{\delta}_j = \frac{1}{n}\sum_{k=1}^{n} \left.\frac{\partial s(X_k|X_{k-1}; \hat{\theta}, \mu(\cdot; \hat{\theta}, \hat{\pi} + \alpha K_h\left(\cdot - X_j\right)))}{\partial\alpha}\right|_{\alpha=0},$$

and $M \to \infty$, $M/n^{1/8} \to 0$.

Observe that in the parametric framework of Newey and West (1987), it is required that $M_n/n^{1/4} \to 0$. We have to require that $M_n \to \infty$ at a slower rate due to the presence of the nonparametric part here, only exhibiting $n^{1/4}$-convergence rate. One advantage of the above variance estimator is its simple implementation; one can evaluate the variance estimator by numerical differentiation of $\log p\left(x|x_0;\theta,\mu\left(\cdot;\theta,\pi+\alpha K_h\left(\cdot-y\right)\right)\right)$ w.r.t. θ and α, instead of deriving the analytical derivatives (on the other hand, these may lead to superior numerical estimates). Since $E_\pi\left[\delta\left(X_0\right)\right]=0$, one could leave out the average appearing in the expression for $\hat{\omega}_j$, but in finite sample this adjustment may improve on the performance of the estimator. An alternative to the variance estimator suggested here would be to construct one by either bootstrapping or subsampling, which should improve on the finite sample approximation; Hall (1992) and Politis, Romano and Wolf (1999) respectively provide in-depth treatment of these two methods. The recent work by Horowitz (2003), where a bootstrap method for Markov chains is suggested based on a kernel estimator of the transition density, is very well-suited for our framework. Since the sampled observations of the process $\{X_t\}$ indeed is a Markov chain, and we have here obtained a semiparametric estimator of the transition density, one should be able to adapt the results of Horowitz (2003) to our setting. Chen et al. (2003, Theorem B) give conditions for consistency of the bootstrap for a general class of semiparametric estimators. This is done under the assumption of i.i.d. observations, but combining their approach with Horowitz's results should yield the desired result for our estimator. The verification of this claim is out of the scope of this paper however.

Having obtained the estimator $\hat{\theta}$ in either of the two classes, one could now be interested in testing hypotheses concerning the parametric part, e.g. $H_0:\theta=\theta_0$ for some given $\theta_0\in\text{int}\Theta$. An obvious choice of test statistic for this hypothesis would appear to be the likelihood ratio,

$$T_n \equiv n\left[L_n(\hat{\theta},\hat{\mu}(\cdot;\hat{\theta}))-L_n\left(\theta_0,\hat{\mu}\left(\cdot;\theta_0\right)\right)\right],$$

A general treatment of the semiparametric likelihood ratio test can be found in Murphy and Van der Vaart (1997) who show that under regularity conditions the likelihood-ratio converges towards a $\chi^2\left(p\right)$-distribution, where p is the dimension of θ. This is however not valid in our case. This owes to the fact here a preliminary estimator of the nonparametric part is used, while in Murphy and Van der Vaart (1997) the nonparametric part is estimated together with θ. This has strong implications for the asymptotic distribution of T_n. Instead one may use that $n\left(\hat{\theta}-\theta_0\right)H_0^{-1}\left(H_0+V_0\right)H_0^{-1}\left(\hat{\theta}-\theta_0\right)\to^d \chi^2\left(p\right)$, or apply a GMM-type test statistic based on the score function, $S_n\left(\theta,\mu\right)$.

Given the $\sqrt{n}$-consistency of $\hat{\theta}$ as established above, Theorem 1 now establishes consistency and asymptotic normality of the nonparametric estimator of the drift, $\hat{\mu}\left(x\right)=\hat{\mu}(x;\hat{\theta})$. Note that the bandwidths used to estimate $\hat{\mu}$ should not be chosen to satisfy (C1.6). The bandwidth restrictions there were tailored to ensure a sufficiently fast convergence rate of both $\hat{\mu}$ and its first two derivatives while taking into account the trimming; this is not needed to prove pointwise asymptotic normality of $\hat{\mu}$.

4.2 Class 2

Here, we derive theoretical results for the estimator of θ for models in Class 2. Since our conditions, strategy of proof and results are very much the same as for Class 1, we will not give any thorough discussions of these, and instead refer to the previous section.

As in the previous section, we need to trim our estimator of $\sigma_0^2(x;\theta)$ to control for the tailbehaviour. Define

$$\hat{\sigma}^2(x) = \hat{T}(x;a)\frac{2}{\hat{\pi}(x)}\frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x)}(X_i)\,\mu(X_i;\theta) + (1-\hat{T}(x;a))\underline{\sigma}^2,$$

with $\hat{T}$ defined earlier. Observe that we here make sure that $\hat{\sigma}^2(x) \geq \underline{\sigma}^2$ for some lower bound $\underline{\sigma}^2 > 0$; this is needed since $\sigma^2(x)$ enters as a denominator in p. We furthermore define

$$\hat{\sigma}_0^2(x;\theta) = \hat{T}(y;a)\,\sigma_0^2(x;\theta) + (1-\hat{T}(x;a))\underline{\sigma}^2, \quad \sigma_0^2(x;\theta) = \frac{2}{\pi_0(x)}\int_l^x \pi_0(y)\,\mu(y;\theta)dy.$$

We are now able to establish $\left\|\hat{\sigma}^2(\cdot;\theta) - \hat{\sigma}_0^2(\cdot;\theta)\right\|_\infty \to^P 0$. To estimate $\partial_x\sigma^2$, we define

$$\partial_x\hat{\sigma}^2(x;\theta) \equiv \hat{T}(x;a)\left\{2\mu(x;\theta) - \frac{\hat{\pi}^{(1)}(x)}{\hat{\pi}^2(x)}\frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x)}(X_i)\,\mu(X_i;\theta)\right\},$$

and similarly for other derivatives w.r.t. x and θ.

We write $p\left(x|x_0;\theta,\sigma^2\right) = p\left(x|x_0;\mu(\cdot;\theta),\sigma^2\right)$ with $p\left(x|x_0;\mu,\sigma^2\right)$ as given in (16), and define our estimator as

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L_n\left(\theta,\hat{\sigma}^2(\cdot;\theta)\right),$$

$$L_n\left(\theta,\sigma^2\right) = \frac{1}{n}\sum_{i=1}^{n}\log p\left(X_{i+1}|X_i;\theta,\sigma^2\right),$$

For the kernel estimator, we again use a higher order kernel of order $\omega \geq 5$. The following assumptions are imposed:

C2.1 (A0)-(A1) holds with $\omega \geq 6$, the kernel $K \in \mathcal{K}(\omega,4)$, and the trimming function $\hat{T} \in \mathcal{T}(4)$.

C2.2 (i) $\underline{\sigma}^2 \leq \sigma_0^2(x;\theta)$, and (ii) $||\partial_x^i\partial_\theta^j\sigma_0^2(x;\theta)|| \leq \overline{\sigma}^2$, $0 \leq i \leq 6$ and $0 \leq j \leq 2$.

C2.3 (i) The drift function satisfies $||\partial_x^i\partial_\theta^j\mu_0(x;\theta)|| \leq C(1+|x|^q)$, $0 \leq i \leq 6$ and $0 \leq j \leq 2$, with $4q+2+\delta \leq \overline{q}$ for some $\delta > 0$ where $\overline{q}$ given in (A0); (ii) $-C\left(1+|z|^{\overline{q}}\right) \leq \lambda_Y(z;\theta,\mu_0(\cdot;\theta)) \leq \overline{\lambda}_Y$ uniformly in (x,θ).

C2.4 The parameter space $\Theta \subseteq \mathbb{R}^d$ is compact.

C2.5 $L(\theta,\sigma_0(\cdot;\theta)) = E_\pi[\log p(X_1,X_0;\theta,\sigma_0(\cdot;\theta))]$ has a unique maximum at θ_0, and

$$H\left(\theta_0,\sigma_0^2\right) \equiv E_\pi\left[\frac{\partial^2}{\partial\theta\partial\theta}\log p\left(X_1|X_0;\theta_0,\sigma_0^2(\cdot;\theta_0)\right)\right] \tag{27}$$

is non-singular.

C2.6a (i) $na^{2(i-5)}h_i^{2(1+i)} \to \infty$, (ii) $a^{i-5}h_i^{\omega-i} \to 0$ as a and $h_i \to 0$, $i = 0,1,2$.

C2.6b (i) $na^{4(i-5)}h_i^{4(1+i)} \to \infty$ (ii) $na^{4(i-5)}h_i^{4(\omega-i)} \to 0$, (iii) $na^{4(5-i)}h_i^{1+i} \to \infty$, (iv) $a^{-1}h_i^{\omega-i} \to 0$, $0 \leq i \leq 3$; (v) $na^{-6}h_4^{10} \to \infty$, (vi) $a^{-3}h_4^{\omega-4} \to 0$;

(vii) $nP_{\pi}\ _{B}(a/2 \leq \pi_0(Z_t) \leq a) \to 0$; (viii) $nP_\pi(a/2 \leq \pi_0(X_0) \leq a) \to 0$.

C2.7 The density $\bar{p}(z)$ given in (25) satisfies $\bar{p}^{(i)} = O(\pi_0^{(i)})$, $0 \leq i \leq 4$.

The conditions are essentially the same as the ones imposed on the models in Class 1. Note that we here assume that the $\sigma_0^2(x;\theta)$ is bounded from below by $\underline{\sigma}^2$ which is known. The assumption that $\underline{\sigma}^2$ is known is used to simplify our proofs. One could allow for an unknown bound by introducing another trimming parameter $\underline{\sigma}_n^2 \to 0$. The proofs of consistency and asymptotic normality now proceed as for Class 1. First, the estimator is shown to be consistent:

Theorem 8 (Class 2) *Under (C2.1)-(C2.6a), $\hat{\theta} \to^P \theta_0$.*

We introduce the score $s(x|x_0;\theta,\sigma^2) = \partial_\theta \log p(x|x_0;\theta,\sigma^2(\cdot;\theta))$, and the pathwise derivative of the score $s(x|x_0;\theta,\sigma^2)$ w.r.t. σ^2 in the direction $d\sigma^2$, $\nabla s(x|x_0;\theta,\sigma^2)[d\sigma^2]$ which will be used in the derivation of the asymptotic distribution. Using the same notation as in the previous section, we have that $\nabla S_n(\theta_0,\hat{\sigma}^2)[\hat{\sigma}^2 - \hat{\sigma}_0^2] = \nabla S(\theta_0,\sigma_0^2)[\hat{\sigma}^2 - \hat{\sigma}_0^2] + o_P(n^{-1/2})$, c.f. Lemma 33 and 34. Furthermore, $\nabla S(\theta_0,\sigma_0^2)[\hat{\sigma}^2 - \hat{\sigma}_0^2]$ can be written as a sum and a remainder term with the latter being asymptotically negligible (Lemma 35),

$$\nabla S(\theta_0,\sigma_0^2)[\hat{\sigma}^2 - \hat{\sigma}_0^2] = \frac{1}{n}\sum_{i=1}^{n}\delta(X_i) + o_P(1/\sqrt{n}).$$

It should be noted, that the function δ here is not identical to the δ-function appearing in Class 1. Finally, the Hessian $h(x|x_0;\theta,\sigma^2) = \partial_{\theta\theta}^2 \log p(x|x_0;\theta,\sigma^2(\cdot;\theta))$ satisfies $H_n(\hat{\theta},\hat{\sigma}^2) \to^P H(\theta_0,\sigma_0^2)$, c.f. Lemma 36. We are able to conclude:

Theorem 9 (Class 2) *Assume that (C2.1)-(C2.7) hold and that $\theta_0 \in \operatorname{int}(\Theta)$. Then the conclusions of Theorem 6 hold for Class 2 with $H_0 = H(\theta_0,\sigma_0^2)$.*

We also obtain a consistent estimator of the asymptotic variance:

Theorem 10 (Class 2) *Assume that (C2.1)-(C2.7) hold and $E_\pi[\|\delta(X_0)\|^{4+\delta}] < \infty$. Then consistent estimators of H_0 and V_0 respectively are given by $\hat{H}_n = H_n(\hat{\theta},\hat{\sigma}^2)$ and $\hat{V}_n$ given as in Theorem 7 with*

$$\hat{\delta}_j = \frac{1}{n}\sum_{k=1}^{n}\left.\frac{\partial s(X_k|X_{k-1};\hat{\theta},\sigma^2(\cdot;\hat{\theta},\hat{\pi} + \alpha K_h(\cdot - X_j)))}{\partial\alpha}\right|_{\alpha=0},$$

and $M \to \infty$, $M/n^{1/8} \to 0$.

Having obtained $\sqrt{n}$-consistency of $\hat{\theta}$, Theorem 2 establishes pointwise consistency and asymptotic normality of the nonparametric estimator of $\sigma_0^2(x)$.

5 Semiparametric Efficiency

As observed in the previous section, our semiparametric estimator is not adaptive in the general case since $V_0 > 0$. A natural question to ask is whether it at least reaches the semiparametric efficiency bound. We are unfortunately not able to give a rigorous answer to this, but due to the nature of our estimator we conjecture this is not the case. We furthermore propose a one-step adjustment to our estimator $\hat{\theta}$ which we conjecture will reach the bound in any circumstance.

For a semiparametric model, Stein (1956) defined the semiparametric efficiency bound as the "least favourable" parametric subproblem of the original semiparametric problem. The Fisher information of the semiparametric problem is obviously no greater than the information of any parametric subproblem. The semiparametric efficiency bound is then defined as the lower bound of the information of all parametric subproblems.

In our setting, the nonparametric part is π, so we therefore consider any smooth parameterisation of π for which the associated Fisher information may be derived. In the following assume for simplicity that θ is one-dimensional, and consider a smooth parameterisation of π, $\theta \mapsto \pi_\theta$, with $\pi_{\theta_0} = \pi_0$. The Fisher information of this subproblem is then given by

$$I(\dot{\pi}) = E_\pi \left[s_0 (X_1 | X_0) [\dot{\pi}]^2 \right]$$

where $s_0 [\dot{\pi}] = s(\theta_0, \pi_0) [\dot{\pi}]$, $s(\theta, \pi) [\dot{\pi}] = \partial_\theta \log p(y|x; \theta, \pi_0) + \nabla_\pi \log p(y|x; \theta, \pi_0) [\dot{\pi}]$, with $\nabla_\pi \log p$ denoting the pathwise derivative of the log-density w.r.t. π at (θ, π), and $\dot{\pi} = \partial_\theta \pi_\theta |_{\theta = \theta_0}$ is the tangent vector of the curve $\theta \mapsto \pi_\theta$ at θ_0. The space of tangent vectors/nuisance scores is given by $\mathcal{S} = \left\{ \dot{\pi} \in L_2 (I) \mid \int_I \dot{\pi}(z)\, dz = 0 \right\}$, and we denote the closure of $\mathcal{S}$ by $\bar{\mathcal{S}}$. A tangent vector $\dot{\pi}^* \in \bar{\mathcal{S}}$ is then called the least favourable direction if

$$I(\dot{\pi}^*) = \inf_{\dot{\pi} \in \bar{\ }} I(\dot{\pi}),$$

and $I^{-1}(\dot{\pi}^*)$ is the semiparametric efficiency bound. The associated score function $s_0^* = s_0 [\dot{\pi}^*]$ is called the efficient score function, and any parameterisation $\theta \mapsto \pi_\theta$ which satisfies $\partial_\theta \pi_\theta |_{\theta = \theta_0} = \dot{\pi}^*$ is called a least favourable model. Observe that $\nabla_\pi \log p_0 [\dot{\pi}^*]$ is the projection of $-\partial_\theta \log p_0$ onto the space $\left\{ \nabla_\pi \log p_0 [\dot{\pi}] \,|\, \dot{\pi} \in \bar{\mathcal{S}} \right\}$. This characterisation was utilised in Severini and Tripathi (1999) to calculate the efficiency bound in a number of semiparametric problems.

But it appears problematic to find $\dot{\pi}^*$ in our case, since $s[\dot{\pi}]$ takes a form which is very difficult to analyse. And even if we were able to find $\dot{\pi}^*$, $I^{-1}(\dot{\pi}^*)$ would be difficult to compare to the variance of $\hat{\theta}$ derived in the previous section since we have no closed form expression for V_0. Instead, we construct a one-step estimator which is designed to reach the efficiency bound. Given the estimator proposed in the previous section, we perform a one-step Newton-Raphson iteration using an estimate of the efficient score. The resulting estimator will be semiparametric efficient. This procedure is very much a generalisation of the one-step Newton-Raphson estimator found in the fully parametric literature: An initial $\sqrt{n}$-consistent estimator is adjusted by the estimated score function making the resulting estimator efficient. This procedure has also been used in the semiparametric literature, see for example Drost and Klaassen and Werker (1997).

The main problem is to obtain an estimator of the efficient score. Here, we rely on the literature on semiparametric profile estimation. As mentioned earlier, a number of studies have developed a general theory for semiparametric profile likelihood estimators; see for example Wong and Severini (1991), Severini and Wong (1992) and Murphy and Van der Vaart (1997, 2000). A very nice property of these estimators is that, under regularity conditions, they reach the semiparametric efficiency bound. In the following, we first introduce the semiparametric profile estimator for our specific problem, and then define a one-step estimator of θ based on the profile likelihood which is computationally less demanding than the actual profile estimator. : There exists $\theta \mapsto \pi_\theta$ satisfying $\dot{\pi}_\theta = \dot{\pi}^*$;

The profile likelihood estimator is defined as

$$\tilde{\theta} = \arg\max_{\theta\in\Theta} L_n\left(\theta, \hat{\pi}_\theta\right),$$

where $\hat{\pi}_\theta = \arg\max_{\pi\in\Pi} L_n(\theta,\pi)$, and $\Pi \subseteq \left\{\pi \geq 0 | \int_I \pi(x)\, dx = 1\right\}$ is a subspace of all densities. Intuitively, this estimator should perform better than our estimator, $\hat{\theta}$. The latter is based on a fixed initial estimator $\tilde{\pi}$, while the profile estimator relies on an estimator $\hat{\pi}_\theta$ which adjusts to θ. The profile estimator should then reach the semiparametric efficiency bound. General conditions for this can be found in Murphy and Van der Vaart (2000). These are: (i) there exists a least favourable model, (ii) $\hat{\pi}_{\theta_n} \to^P \pi_0$, and (iii)

$$E_\pi\left[s\left(\theta_0, \hat{\pi}_{\theta_n}\right)\left[\hat{\dot{\pi}}_{\theta_n}\right]\right] = o_P\left(\|\theta_n - \theta_0\|\right) + o_P\left(n^{-1/2}\right),$$

for any any random sequence $\theta_n \to^P \theta_0$. If we start with (ii), since Π is an infinite-dimensional space, this condition is not easily verified.[11] One solution to this problem is to apply the method of sieves :[12] For each $n \geq 1$, let Π_n be a finite-dimensional space of densities with support I such that the sequence $\{\Pi_n\}$ grows dense in Π as $n \to \infty$. We then redefine $\hat{\pi}_\theta$ as $\hat{\pi}_\theta = \arg\max_{\pi\in\Pi_n} L_n(\theta,\pi)$. Under regularity conditions, $\hat{\pi}_\theta \to^P \pi_\theta$ for any given $\theta \in \Theta$, where

$$\pi_\theta = \arg\max_{\pi\in\Pi} E_\pi\left[\log p(\theta,\pi)\right]; \tag{28}$$

see for example Chen and Shen (1998). Sufficient conditions for (ii) to hold is then (a) $\theta \mapsto \pi_\theta$ is a continuous mapping and (b) $\sup_{\theta\in\Theta}\|\hat{\pi}_\theta - \pi_\theta\|_\infty \to^P 0$. These will hold if $\{\hat{\pi}_\theta | \theta \in \Theta\}$ is stochastically equicontinuous, c.f. Newey (1991). The curve defined in (28) is moreover a natural candidate for the least favourable model in (i). Condition (iii) is a smoothness condition on the score function.

While we expect $\tilde{\theta}$ to reach the efficiency bound, it is much more computationally burdensome than $\hat{\theta}$ since at each given value of θ we have to perform a high-dimensional optimisation routine over Π_n in order to obtain $\hat{\pi}_\theta$. A computationally attractive alternative to $\tilde{\theta}$ is the following one-step adjustment estimator,

$$\check{\theta} = \hat{\theta} + H_n^{-1}(\hat{\theta}, \hat{\pi}_{\hat{\theta}}) S_n(\hat{\theta}, \hat{\pi}_{\hat{\theta}}),$$

where $\hat{\pi}_\theta$ is defined as before and $\hat{\theta}$ is the estimator considered in the previous section. The adjustment term is basically a Newton-Raphson iteration. Under the regularity conditions in Murphy and Van der Vaart (2000), $S_n(\hat{\theta}, \hat{\pi}_{\hat{\theta}}) = n^{-1}\sum_{i=1}^n s_0^*\left(X_i | X_{i-1}\right) + o_P\left(n^{-1/2}\right)$. Given this, it should be possible to show that $\check{\theta}$ has the desired asymptotic properties, see Bickel et al (1993, Section 7.8). We shall not pursue this any further here, and leave the proof of this conjecture to future research.

6 Implementation

In this section we discuss the implementation of the estimator. As mentioned earlier, the transition density p does not in general have a closed form expression, and so one can not directly evaluate it.

[11] The estimator $\hat{\pi}_\theta$ might not even be well-defined and, even if it is, very difficult to compute.

[12] See Chen (2004) for an overview of this method.

Instead, a number of different suggestions for how to either approximate or simulate it have been proposed in the literature. Lo (1988) observes that p solves a linear 2nd order partial differential equation, and suggests the application of numerical methods to solve it and thereby obtain p. A closed-form approximation of p can be found in Aït-Sahalia (2002), derived by using Edgeworth-expansion-type arguments. Durham and Gallant (2002), Elerian et al. (2001), Hurn et al (2003), Nicolau (2002) and Pedersen (1995) all consider simulation-based maximum-likelihood. Either of the above methods can be applied to our estimator. As mentioned in the previous section, in the implementation of the above mentioned methods evaluation of γ is not required, except for the method of Nicolau (2002).

An important part of the estimator is the choice of bandwidths and trimming parameter. An obvious way of choosing the bandwidths would be cross-validation methods, see Härdle et al. (1990); other options are rule-of-thumb and plug-in methods, see Silverman (1986) for a discussion of these and related methods. Most existing methods however are designed to minimise the mean square error, while the conditions imposed on the set of bandwidths when deriving asymptotic normality of $\hat{\theta}$ require them to be of a different order. So the above methods do not appear to be directly applicable in our case. This is demonstrated in Härdle et al. (1992) where results for the optimal bandwidth choice for the average derivative estimator is derived; it is shown that the optimal bandwidths used in the semiparametric estimation are not equivalent to the ones minimising the mean squared error. Powell and Stoker (1996) extend their results to other semiparametric problems. Data-driven methods to obtain bandwidths in semiparametric estimation are yet to be derived however. It is outside the scope of this paper to construct a bandwidth choice method tailored to our application of the kernel estimator. Newey (1994a) suggests that in practice a good method would be to start with the standard cross-validated choice of bandwidth, and then decrease it until $\hat{\theta}$ does not change too much. Another rule-of-thumb method is the following: For $||\hat{\pi}^{(i)} - \pi_0^{(i)}||_\infty = o_P\left(n^{-1/4}\right)$ to hold, we require that $n^{\frac{1}{4(1+i)}} h_i \to \infty$ and $n^{\frac{1}{4(\omega-i)}} h_i \to 0$. Restricting h_i to $h_i = cn^{-q_i}$, these restrictions can be written as $n^{\frac{1}{4(1+i)}-q_i} \to \infty$, and $n^{\frac{1}{4(\omega-i)}-q_i} \to 0$. Thus, we require that $\frac{1}{4(\omega-i)} < q_i < \frac{1}{4(1+i)}$, which holds if $\omega > 2i+1$. If this is satisfied, the optimal choice is $h_i = cn^{-\frac{1}{4(1+i)}} \log(n)$. Using some data-driven method minimising the MSE, we obtain $h_i^* = O\left(n^{-1/(2i+5)}\right)$. One way of choosing h_i in an application is then as

$$h_i = h_0^* n^{\frac{1}{5}-\frac{1}{4(1+i)}} \log(n) = h_0^* n^{\frac{4i-1}{20(1+i)}} \log(n),$$

or alternatively $h_i = h_i^* n^{\frac{1}{2i+5}-\frac{1}{4(1+i)}} \log(n)$.

Various studies suggest that the dependence structure of the available data will affect the performance of the kernel estimators in finite samples. In particular, strong dependence will deteriorate the finite sample performance. Hall et al. (1995) give theoretical results concerning robustness of the cross-validation procedure towards dependence, while Pritsker (1998), who reconsidered the work of Aït-Sahalia (1996b), demonstrates that for the data set used there, the asymptotic distribution of the marginal density estimator provided an unsatisfactory approximation of the finite-sample distribution. It appeared that the major problem was the strong dependence between the observations, which may slow down the convergence of the kernel density estimator; in such cases the use of the asymptotic distribution is not appropriate for tests and confidence bands.[13] The same problem is reported by

[13] In a different context, Fan (1994) argues that rather the problems are caused by big differences between true and nominal sizes of the test in finite sample.

Chapman and Pearson (2000) who also give evidence of potential boundary problems of the kernel estimates.

Another potential problem with the performance of our estimator is that kernel estimators of density derivatives appear to be systematically biased in finite sample. Stoker (1993) give theoretical evidence of a systematic bias towards zero of these, and suggests a method for correcting for this bias in weighted average derivative estimators; his results are generalised by Newey et al. (1992), see also Newey et al (2004). These results indicate that great care should be taken when choosing the bandwidths, and that the use of the asymptotic distribution to approximate the finite sample distribution of non- and semiparametric estimators may not be a terribly good idea. In the worst case, the estimator of the drift and diffusion term may be heavily biased. This lends support to the application of bootstrap methods when conducting inference.

7 A Simulation Study

In this section we present results from a small simulation study. The simulation study demonstrates that the estimator performs well for moderate sample sizes, suggesting that the concerns put forward in the previous section may not be so relevant.

The estimator is implemented using the approximation of p suggested in Aït-Sahalia (2002). Let $p^{(M,N)}$ denote the approximation where $M \geq 1$ and $N \geq 1$ are integers; as $M, N \to \infty$, $p^{(M,N)}(x\,|x_0) \to p(x\,|x_0)$ uniformly over (x, x_0) on any compact set under regularity conditions on μ and σ^2, c.f. Aït-Sahalia (2002, Theorem 1). The (M, N)th approximation requires the evaluation of the N first derivatives of μ and σ^2.

We choose a model in Class 1, so in order to implement the procedure for a given data set, we perform the following three-step procedure:

1. Obtain $\hat{\pi}^{(i)}$ for some kernel K and bandwidth h_i, $0 \leq i \leq N + 1$.
2. Obtain $\hat{\theta}$ using the approximate MLE method for given (M, N)-parameter with $\hat{\pi}^{(i)}$, $0 \leq i \leq N+1$, plugged in.
3. Calculate $\hat{\mu}(x) = \mu(x; \hat{\theta}, \hat{\pi})$.

We have 4 parameters ,which have to be chosen to run the above procedure: The kernel K, the bandwidth h, the trimming parameter a, and the approximation order (M, N). One would expect that there would be a trade off between the size of (M, N) and the estimation of π: As (M, N) goes to infinity, the approximate likelihood approaches the true one; on the other hand, in the actual implementation a large value of (M, N) requires a large number of derivatives of π to be estimated. However, in the simulation study it was found that the estimator was very stable towards the choice of (M, N). It appears that the higher order terms of the approximation (and thereby the estimation of higher order derivatives) are not terribly important.[14] On the other hand, one has to be careful with the choice of the bandwidths for $\hat{\pi}^{(i)}$, $0 \leq i \leq 3$; we found that the estimator of θ was relatively sensitive towards

[14] But this may be specific to the model considered here.

choice of h_i, $0 \leq i \leq 3$. The trimming parameter was chosen such that only observations between the 2.5th and 97.5th the empirical percentile were included in the estimation of θ; the full data set was used in the preliminary estimation of π and its derivatives however. We tried out other percentiles in the range 0-5 and 95-100 respectively without any significant changes in the results. We also tried out various kernels, finding that the performance of the estimator appears to be very robust towards the choice of kernel. In particular, in practice higher order kernels did provide any significant improvement on the performance of the estimator.

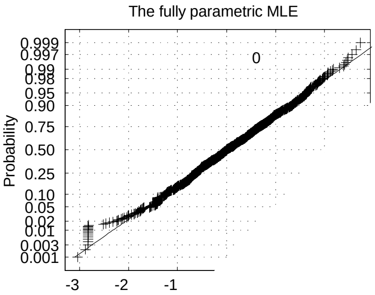
The model we simulate from is the so-called CIR- model suggested by Cox et al. (1985),

$$\begin{aligned} dX_t &= \mu(X_t)\,dt + \theta\sqrt{X_t}dW_t, \\ \mu(x) &= 0.5\,(0.08 - x), \end{aligned}$$

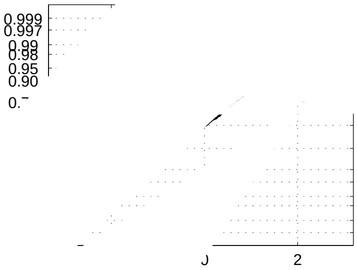
with $\theta = \sqrt{0.02} = 0.1414$. The specification of μ and σ^2 ensures that the data generating process is stationary. In the estimation, $\mu(\cdot)$ and θ are the unknown parameters of interest. We set the time distance between observations to $\Delta = 1/12$. The advantage of this model is that the transition density is known so we can perform actual MLE when we allow ourselves to use the information that $\mu(x) = 0.5\,(0.08 - x)$. This allows us to compare the semiparametric and actual MLE. We simulate n observations of the process using the standard Euler scheme. For each data set we then go through the steps 1.-3. given above. We employ the second rule-of-thumb method suggested in the previous section to choose the bandwidths for each data set. The results reported below could probably be improved upon by using data-driven bandwidth selection procedures, but our rule-of-thumb method seems to do a good job.

We simulate 1000 data sets, where each data set consists of $n = 500$ observations, and we choose $M = 3, 4, 5$ and $N = 2M$. We consider the semiparametric estimator $\hat{\theta}_{500}^{(M,N)}$ for each of the 3 choices of (M, N) and also the actual fully parametric MLE, $\hat{\theta}_{0,500}$. In Figure 1, we have for each of the estimators made a QQ-plot of its empirical distribution against a $N\left(\theta, \hat{s}^2\right)$ distribution where $\hat{s}^2$ is the estimator's empirical variance. As can be seen, there are some slight problems with $\hat{\theta}_{0,500}$ in the left tail, which owes to numerical problems in the optimisation procedure. But the semiparametric estimator performs remarkably well with the choice of (M, N) having a negligible effect on the performance. The estimators are close to being unbiased: $\hat{\theta}_{0,500}$ has empirical mean 0.1416 and std. 0.050 while $\hat{\theta}_{500}^{(M,N)}$ has empirical mean 0.1410 (0.0047), 0.1412 (0.0049) and 0.1413 (0.0049) for $M = 3, 4, 5$. This indicates that for this specific model, the adjustment term δ is small.[15] Next, we report on the nonparametric part, μ. In Figure 2, the estimated drift for $M = 5$ is plotted together with the actual drift. As can bee seen, the estimator is biased but the true drift lies within its 95% confidence bands in the major part of the domain. However, in the left tail there appears to be problems. As expected, the estimator of the nonparametric part does not perform as well as the parametric part due to the slower convergence rate of the former.

[15] This may be due to the aforementioned numerical problems experienced with the MLE though, since these increase its empirical variance.



0.06



0.02

0 0.1

0

0.1

8 Conclusion

We have considered two broad classes of diffusion models where either the drift or the diffusion term of the model is left unspecified while the other is specified up to a finite-dimensional parameter. Under the assumption of stationarity, estimators of both the parametric and nonparametric part were proposed, and their asymptotic properties were derived. We suggested that in the practical implementation of the estimator, approximate or simulation-based methods should be applied. Under suitable conditions these will have asymptotically negligible effects on the performance of the estimator. A small simulation study was carried out which supported the theoretical results. An approximation of the transition density was implemented, and we found that the choice of the approximation order had a small impact on the performance of the estimator, while the choice of bandwidths appeared to be important.

Various issues and extensions related to this study could be of interest to investigate in future research. As observed earlier, one may wish to allow for weak non-stationarity of the processes. Another important extension would be to consider multivariate diffusion models. It could also be of interest to consider other types of semiparametric continuous-time models. Here, we have restricted the noise process driving the SDE to be a Brownian motion, while allowing for an unspecified drift or diffusion term. One could take the alternative approach of specifying the two while leaving the noise process unspecified. This approach is pursued by Werker et al. (2000) where an extended version of the Vasicek (1977) model is considered. Finally, the issue of semiparametric efficiency was only discussed heuristically here; rigorous results in this area for the two classes of diffusion models are yet to be derived.

References

Ahn, D.-H. & B. Gao (1999) A Parametric Nonlinear Model of Term Structure Dynamics. *Review of Financial Studies* 12, 721-762.

Ai, C. (1997) A Semiparametric Maximum Likelihood Estimator. *Econometrica* 65, 933-963.

Aït-Sahalia, Y. (1996a) Nonparametric Pricing of Interest Rate Derivative Securities. *Econometrica* 64, 527-560.

Aït-Sahalia, Y. (1996b) Testing Continuous-Time Models of the Spot Interest Rate. *Review of Financial Studies* 9, 385-426.

Aït-Sahalia, Y. (2002) Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approximation Approach. *Econometrica* 70, 223-262.

Andrews, D.W.K. (1994) Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity. *Econometrica* 62, 43-72.

Andrews, D.W.K. (1995) Nonparametric Kernel Estimation for Semiparametric Models. *Econometric Theory* 11, 560-596.

Arcones, M.A. (1995) On the Central Limit Theorem for U-Statistics under Absolute Regularity. *Statistics and Probability Letters* 24, 245-249.

Bandi, F.M. & P.C.B. Phillips (2000) Accelerated Asymptotics for Diffusion Model Estimation. Working paper, GSB, University of Chicago.

Bandi, F.M. & P.C.B. Phillips (2003) Fully Nonparametric Estimation of Scalar Diffusion Models. *Econometrica* 71, 241-283.

Bergstrom, A.R. (1990) *Continuous Time Econometric Modelling.* Oxford: Oxford University Press.

Bibby, B. and M. Sørensen (1995) Martingale Estimating Functions for Discretely Observed Diffusion Processes. *Bernoulli* 1, 17-39.

Bickel, P.J., C.A.J. Klaassen, Y. Ritov & J.A. Wellner (1993) *Efficient and Adaptive Estimation for Semiparametric Models.* The John Hopkins University Press.

Bierens, H.J. (1987) Kernel Estimators of Regression Functions. In T.F. Bewley (ed.) *Advances in Econometrics: Fifth World Congress* Vol. 1, 99-144. Cambridge: Cambridge University Press.

Björk, T. (1998) *Arbitrage Theory in Continuous Time.* Oxford: Oxford University Press.

Black, F. & M. Scholes (1973) The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81, 637-654.

Bosq, D. (1998) *Nonparametric Statistics for Stochastic Processes.* New York: Springer-Verlag.

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.

Chen, X. (2004) Semiparametric and Nonparametric Estimation via the Method of Sieves. Forthcoming in *Handbook of Econometrics* vol. 6 (eds. J.J. Heckman & E.E. Leamer). Amsterdam: North-Holland

Chen, X., L.P. Hansen & M. Carrasco (1999) Nonlinearity and Temporal Dependence. Working paper, University of Rochester.

Chen, X., L.P. Hansen & J. Scheinkman (2000a) Shape-preserving Estimation of Diffusions. Working paper, University of Chicago.

Chen, X., L.P. Hansen & J. Scheinkman (2000b) Principal Components and the Long Run. Working paper, University of Chicago.

Chen, X., O. Linton & I. Van Keilegom (2003) Estimation of Semiparametric Models when the Criterion Function is not Smooth. *Econometrica* 71, 1591-1608.

Chen, X. & X. Shen (1998) Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica* 66, 289-314

Cobb, L., P. Koppstein & N.H. Chen (1983) Estimation and Moment Recursion Relations for Multimodal Distributions of the Exponential Family. *Journal of the American Statistical Association* 78, 124-130.

Conley, T., L.P. Hansen, E. Luttmer & J. Scheinkman (1997) Short-term Interest Rates as Subordinated Diffusions. *Review of Financial Studies* 10, 525-577.

Constantinides, G.M. (1992) A Theory of the Nominal Term Structure of Interest Rates. *Review of Financial Studies* 5, 531-552.

Cox, J.C., J.E. Ingersoll & S. Ross (1985) A Theory of the Term Structure of Interest Rates. *Econometrica* 53, 373-384.

Dacunha-Castelle, D. & D. Florens-Zmirou (1986) Estimation of the Coefficients of a Diffusion from Discrete Observations. *Stochastics* 19, 263-284.

Darolles, S. & C. Gouriéroux (2001) Truncated Dynamics and Estimation of Diffusion Equations. *Journal of Econometrics* 102, 1-22.

Doukhan, P., P. Massart & E. Rio (1994) The Central Limit Theorem for Strongly Mixing Processes. *Annales de l'Institut Henri Poincare*, Series B, Probability and Statistics 30, 63-82.

Doukhan, P., P. Massart & E. Rio (1995) Invariance Principles for Absolutely Regular Processes. *Annales de l'Institut Henri Poincare*, Series B, Probability and Statistics 31, 393-427.

Drost, F.C., C.A.J. Klaassen & B.J.M. Werker (1997). Adaptive Estimation in Time-Series Models. *Annals of Statistics* 25, 786–817.

Duffie, D. (1996) *Dynamic Asset Pricing Theory*, 3rd edition. New Jersey: Princeton University Press.

Duffie, D. & K.J. Singleton (1993) Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica* 61, 929-952.

Dunis, C. & B. Zhou (eds.) (1998) *Nonlinear Modelling of High Frequency Financial Time Series*. New York: John Wiley & Sons.

Durham, G.B. & R.A. Gallant (2002) Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes. *The Journal of Business and Economic Statistics* 20, 297-316.

Elerian, O., S. Chib, S. & N. Shephard (2001) Likelihood Inference for Discretely Observed Non-linear Diffusions. *Econometrica* 69, 959-993.

Fan, Y. (1994) Testing the Goodness-of-Fit of a Parametric Density Function by Kernel Method. *Econometric Theory* 10, 316-356.

Fan, Y. (1995) Bootstrapping a Consistent Nonparametric Goodness-of-Fit Test. *Econometric Reviews* 14, 367-382.

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. & J.R. Long (1997) Estimating Stochastic Differential Equations Efficiently by Minimum Chi-Square. *Biometrika* 84, 125-141.

Gallant, A.R. & G. Tauchen (1996) Which Moments to Match? *Econometric Theory* 12, 657-681.

Genon-Catalot, V. (1990) Maximum Contrast Estimation for Diffusion Processes from Discrete Observations. *Statistics* 21, 99-116.

Gobet, E., M. Hoffmann & M. Reiß (2003) Nonparametric Estimation of Scalar Diffusions Based on Low Frequency Data. Working paper, Universités de Paris 6 & 7 - CNRS (UMR 7599). Forthcoming in *Annals of Statistics*.

Gouriéroux, C., A. Monfort & E. Renault (1993) Indirect Inference. *Journal of Applied Econometrics* 8, 85-118

Hall, P. (1992) *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Hall, P., S.N. Lahiri & J. Polzehl (1995) On Bandwidth Choice in Nonparametric Regression with Both Short- and Long-Range Dependent Errors. *Annals of Statistics* 23, 1921-1936.

Hansen, L.P. & J.A. Scheinkman (1995) Back to the Future: Generating Moment Implications for Continuous Time Markov Processes. *Econometrica* 63, 767-804.

Hansen, L.P., J.A. Scheinkman & N. Touzi (1998) Spectral Methods for Identifying Scalar Diffusions. *Journal of Econometrics* 86, 1-32.

Härdle, W., J.D. Hart, J.S. Marron & A.B. Tsybakov (1992) Bandwidth Choice for Average Derivative Estimation. *Journal of the American Statistical Association* 87, 218-233.

Härdle, W., J.S. Marron & M.P. Wand (1990) Bandwidth Choice for Density Derivatives. *Journal of the Royal Statistical Society*, Series B, 52, 223-232.

Horowitz, J. (2003) Bootstrap Methods for Markov Processes. *Econometrica* 71, 1049-1082.

Jiang, G. & J. Knight (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. & S.E. Shreve (1991) *Brownian Motion and Stochastic Calculus,* 2nd Edition. New York: Springer Verlag.

Karlin, S. & H.M. Taylor (1981) *A Second Course in Stochastic Processes.* Academic Press.

Karlsen, H. and D. Tjøstheim (2001) Nonparametric Estimation in Null Recurrent Time Series. *Annals of Statistics* 29, 372-416.

Kristensen, D. (2004) A Semiparametric Single-Factor Model for the Term Structure. Working Paper, LSE.

Lo, A. (1988) Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data. *Econometric Theory* 4, 231-247.

Longstaff, F.A. & E.S. Schwartz (1992) Interest Rate Volatility and the Term Structure: A Two Factor General Equilibrium Model. *Journal of Finance* 47, 1259-1282.

Marsh, T.A. & E.R. Rosenfeld (1983) Stochastic Processes for Interest Rates and Equilibrium Bond Prices. *Journal of Finance* 38, 635-646.

Masry, E. (1996) Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates. *Journal of Time Series Analysis* 17, 571-599.

Merton, R.C. (1973) Theory of Rational Option Pricing. *Bell Journal of Economics and Management Science* 4, 141-183.

Merton, R.C. (1976) Option Pricing when Underlying Stock Returns Are Discontinuous. *Journal of Financial Economics* 3, 125-144.

Meyn, S.P. & R.L. Tweedie (1993) Stability of Markovian processes III: Foster-Lyapunov Criteria for Continuous-time Processes. *Advances in Applied Probability* 25, 518-548.

Murphy, S.A. and A.W. van der Vaart (1997) Semiparametric Likelihood Ratio Inference. *Annals of Statistics* 25, 1471–1509.

Murphy, S.A. and A.W. van der Vaart (2000) On Profile Likelihood. *Journal of the American Statistical Association* 95, 449-485.

Newey, W.K. (1990) Semiparametric Efficiency Bounds. *Journal of Applied Econometrics* 5, 99-135.

Newey, W.K. (1991) Uniform Convergence in Probability and Stochastic Equicontinuity. *Econometrica* 59, 1161-1167.

Newey, W.K. (1994a) Kernel Estimation of Partial Means and a General Variance Estimator. *Econometric Theory* 10, 233-253.

Newey, W.K. (1994b) The Asymptotic Variance of Semiparametric Estimators. *Econometrica* 62, 1349-1362.

Newey, W.K., F. Hsieh & J. Robins (1992) Bias Corrected Semiparametric Estimation. Working paper, MIT.

Newey, W.K., F. Hsieh & J. Robins (2004) Twicing Kernels and a Small Bias Property of Semiparametric Estimators. *Econometrica* 72, 947-962.

Newey, W.K. & D. McFadden (1994) Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics* vol. IV, ch. 36. Amsterdam: North-Holland.

Newey, W.K. & K.D. West (1987) A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55, 703-708.

Nicolau, J. (2002) A New Technique for Simulating the Likelihood of Stochastic Differential Equations. *Econometrics Journal* 5, 91-103.

Nicolau, J. (2003) Bias Reduction in Nonparametric Diffusion Coefficient Estimation. *Econometric Theory* 19, 754-777.

Parzen, E. (1962) On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics* 33, 1065-1076.

Pedersen, A.R. (1995) A New Approach to Maximum-Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations. *Scandinavian Journal of Statistics* 22, 55-71.

Phillips, P.C.B. (1973) The Problem of Identification in Finite Parameter Continuous-Time Models. *Journal of Econometrics* 4, 351-362.

Phillips, P.C.B. & J.Y. Park (1998) Nonstationary Density Estimation and Kernel Autoregression. Cowles Foundation Discussion Paper no. 1181, Yale University.

Politis, D.N, J.P. Romano and M. Wolf (1999): *Subsampling.* New York: Springer-Verlag.

Powell, J.L. & T.M. Stoker (1996) Optimal Bandwidth Choice for Density-Weighted Averages. *Journal of Econometrics* 75, 291-316.

Robinson, P.M. (1983) Nonparametric Estimators for Time Series. *Journal of Time Series Analysis* 4, 185–297.

Robinson, P.M. (1988) Root-N-Consistent Semiparametric Regression. *Econometrica* 56, 931-954.

Robinson, P.M. & C. Velasco (1997) Autocorrelation-Robust Inference. In *Robust Inference* (eds. G.S. Maddala & C.R. Rao), 267-298. Amsterdam: North-Holland.

Rogers, L.C.G. (1995) Which Model For Term-Structure of Interest Rates Should One Use?. In *Mathematical Finance* (eds. M.H.A. Davis, D. Duffie, W.H. Fleming and S.E. Shreve). New York: Springer-Verlag.

Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics.* New York: John Wiley & Sons.

Severini, T.A. & G. Tripathi (2001) A Simplified Approach to Computing Efficiency Bounds in Semiparametric Models. *Journal of Econometrics* 102, 23-66.

Severini, T. and Wong, W. (1992) Generalized Profile Likelihood and Conditionally Parametric Models. *Annals of Statistics* 20, 1768-1802.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.

Stanton, R. (1997) A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk. *Journal of Finance* 52, 1973-2002.

Stein (1956) Efficient Nonparametric Testing and Estimation. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 187-195. Berkeley: University of California Press

Stoker, T.M. (1993) Smoothing Bias in Density Derivative Estimation. *Journal of the American Statistical Association* 88, 855-863.

Stone, C.J. (1990) Large Sample Inference for Log-Spline Models. *Annals of Statistics* 18, 717-741.

Sundaresan, S.M. (2000) Continuous-Time Methods in Finance: A Review and an Assessment. *Journal of Finance* 55, 1569-1622.

Sørensen, M. (1997) Estimating Functions for Discretely Observed Diffusions: A Review. *Selected Proceedings of the Symposium on Estimating Functions*, 305-325 (eds. I.V. Basawa, V.P. Godambe and R.L. Taylor), IMS Lecture Notes 32. Hayward: Institute of Mathematical Statistics.

Tauchen, G.E. (1985) Diagnostic Testing and Evaluation of Maximum Likelihood Models. *Journal of Econometrics* 30, 415-443.

Tauchen, G.E. (1995) New Minimum Chi-Square Methods in Empirical Finance. In *Advances in Economics and Econometrics: Theory and Applications* (eds. K. Wallace & D. Kreps) vol. III, 279-317. Cambridge: Cambridge University Press.

van de Geer, S. (2000) *Empirical Processes in M-Estimation.* Cambridge: Cambridge University Press.

Vasicek, O. (1977) An Equilibrium Characterization of the Term Structure. *Journal of Financial Economics* 5, 177-188.

Veretennikov, A.Y. (1997) On Polynomial Mixing Bounds for Stochastic Differential Equations. *Stochastic Processes and their Applications* 70, 115-127.

Werker, B.J.M., M. Hallin & C. Koell (2000) Optimal Inference for Discretely Observed Semiparametric Ornstein-Uhlenbeck Processes. *Journal of Statistical Planning and Inference* 91, 323-340.

Wong, E. (1964) The Construction of a Class of Stationary Markoff Processes. *Sixteenth Symposium in Applied Mathematics - Stoch. Proc. in Math. Physics and Engineering* (ed. R. Bellman), 264-276. Providence: American Mathematical Society.

Wong, W. H. and Severini, T. A. (1991) On Maximum Likelihood Estimation in Infinite Dimensional Parameter Space. *Annals of Statististics* 19, 603-632.

A Proofs

Proof of Theorem 1. We have

$$\begin{aligned}\hat{\mu}(x)-\mu_0(x) &= \frac{1}{2}\sigma^2(x;\theta_0)\left[\frac{\hat{\pi}^{(1)}(x)}{\hat{\pi}(x)}-\frac{\pi_0^{(1)}(x)}{\pi_0(x)}\right]\\ &\quad+\frac{1}{2}\left[\partial_x\sigma^2(x;\hat{\theta})-\partial_x\sigma^2(x;\theta_0)\right]+\frac{\hat{\pi}^{(1)}(x)}{2\hat{\pi}(x)}\left[\sigma^2(x;\hat{\theta})-\sigma^2(x;\theta_0)\right],\end{aligned}$$

where

$$\partial_x^i\sigma^2(x;\hat{\theta})-\partial_x^i\sigma^2(x;\theta_0)=\partial_x^i\dot{\sigma}^2\left(x;\overline{\theta}_i\right)(\hat{\theta}-\theta_0)=O_P\left(n^{-1/2}\right),$$

for some $\overline{\theta}_i\in[\theta_0,\hat{\theta}]$, $i=0,1$, while

$$\begin{aligned}\sqrt{nh^3}\frac{\hat{\pi}^{(1)}(x)}{\hat{\pi}(x)}-\frac{\pi_0^{(1)}(x)}{\pi_0(x)} &= \frac{1}{\pi_0(x)}\sqrt{nh^3}[\hat{\pi}^{(1)}(x)-\pi_0^{(1)}(x)]\\ &\quad-\frac{\pi_0^{(1)}(x)}{\pi_0^2(x)}\sqrt{nh^3}\left[\hat{\pi}(x)-\pi_0(x)\right]\\ &\quad+\sqrt{nh^3}O\left(|\hat{\pi}^{(1)}(x)-\pi_0^{(1)}(x)|^2+|\hat{\pi}(x)-\pi_0(x)|^2\right).\end{aligned}$$

Using standard methods for kernel estimators, see Robinson (1983), we obtain

$$\sqrt{nh^3}\{\hat{\pi}^{(1)}(x_i)-\pi_0^{(1)}(x_i)\}_{i=1}^N\xrightarrow{d}N(0,V_\pi),$$

where $V_\pi=\mathrm{diag}(\{V_\pi(x_i)\}_{i=1}^N)$ with $V_\pi(x)=\pi_0(x)\,||K^{(1)}||_2^2$, while the two remainder terms are $o_P(1)$, c.f. Lemma 39. The first part of the theorem now follows from Slutsky's Theorem. The uniform convergence result is obtained by combining the proof of Lemma 11 with 39. ■

Proof of Theorem 2. By Lemma 23 and arguments similar to the ones of the previous proof,

$$\hat{\sigma}^2(x)-\sigma_0^2(x)=2\int_l^x\mu(y;\theta_0)\pi_0(y)\,dy[\frac{1}{\hat{\pi}(x)}-\frac{1}{\pi_0(x)}]+O_P(n^{-1/2}),$$

where

$$\frac{1}{\hat{\pi}(x)}-\frac{1}{\pi_0(x)}=-\frac{1}{\pi_0^2(x)}\left[\hat{\pi}(x)-\pi_0(x)\right]+\frac{[\hat{\pi}(x)-\pi_0(x)]^2}{4\left(\lambda\hat{\pi}(x)+(1-\lambda)\pi_0(x)\right)^3},$$

for some $\lambda\in[0,1]$. Using standard results for kernel estimators, see e.g. Robinson (1983), we obtain

$$\sqrt{nh}\left\{\hat{\pi}(x_i)-\pi_0(x_i)\right\}_{i=1}^N\xrightarrow{d}N(0,V_\pi),$$

where $V_\pi=\mathrm{diag}(\{V_\pi(x_i)\}_{i=1}^N)$ with $V_\pi(x)=\pi_0(x)\,||K||_2^2$, while

$$\hat{\pi}(x)-\pi_0(x)=O_P(n^{-1/2}h^{-1})+O_P(h^\omega).$$

Slutsky's Theorem now gives the claimed asymptotic distribution. The uniform convergence result is established by combining Lemma 24 and 39. ■

Proof of Theorem 5. We are allowed to disregard the term $\sqrt{\pi(x)/\pi(x_0)}$ appearing in (34), since this does not depend on θ. Thus,

$$\hat{\theta}=\arg\min_{\theta\in\Theta}Q_n(\theta,\hat{\mu}(\cdot;\theta))$$

where $Q_n(\theta,\mu) = \frac{1}{n}\sum_{i=1}^{n} q(X_i|X_{i-1};\theta,\mu)$ and

$$q(x|x_0;\theta,\mu) = -\frac{1}{4}\log\left(\sigma^2(x;\theta)\,\sigma^2(x_0;\theta)\right) - \frac{1}{2\Delta}\left(\int\limits_{x_0}^{x}\sigma(w;\theta)^{-1}\,dw\right)^2 + \log E_B\left[\psi(x|x_0;\theta,\mu)\right].$$

We wish to show that 1) $\sup_{\theta\in\Theta}|Q_n(\theta,\hat\mu) - Q_n(\theta,\hat\mu_0)| \to^P 0$; 2) $\sup_{\theta\in\Theta}|Q_n(\theta,\hat\mu_0) - Q_n(\theta,\mu_0)| \to^P 0$; 3) $\sup_{\theta\in\Theta}||Q_n(\theta,\mu_0) - Q(\theta,\mu_0)|$, where $Q(\theta,\mu) = E_\pi[q(X_1|X_0;\theta,\mu)]$; and 4) $\theta \mapsto Q(\theta,\mu_0)$ is continuous with a unique maximum at θ_0.

To prove 1), write

$$q(x|x_0;\theta,\hat\mu) - q(x|x_0;\theta,\hat\mu_0) = \log\left(\frac{E_B[\psi(x|x_0;\theta,\hat\mu)]}{E_B[\psi(x|x_0;\theta,\hat\mu_0)]}\right).$$

Using that $(x-1)/x \le \log(x) \le x-1$, we see that

$$\frac{E_B[\psi(x|x_0;\theta,\hat\mu)] - E_B[\psi(x|x_0;\theta,\hat\mu_0)]}{E_B[\psi(x|x_0;\theta,\hat\mu)]} \le \frac{E_B[\psi(x|x_0;\theta,\hat\mu)] - E_B[\psi(x|x_0;\theta,\hat\mu_0)]}{E_B[\psi(x|x_0;\theta,\hat\mu_0)]}.$$

By Jensen's inequality and a 2nd order Taylor expansion of the exponential-function, we obtain

$$\begin{aligned}
&|E_B[\psi(x|x_0;\theta,\hat\mu)] - E_B[\psi(x|x_0;\theta,\hat\mu_0)]| \\
\le\ & E_B\left[\exp\left[2\Delta\int\limits_0^1 a\lambda_Y(Z_t(x|x_0;\theta);\theta,\hat\mu) + (1-a)\,\lambda_Y(Z_t(x|x_0;\theta);\theta,\hat\mu_0)\,dt\right]\right. \\
&\left.\times\int\limits_0^1|\lambda_Y(Z_t(x|x_0;\theta);\theta,\hat\mu) - \lambda_Y(Z_t(x|x_0;\theta);\theta,\hat\mu_0)|\,dt\right],
\end{aligned}$$

where, by Lemma 12,

$$\begin{aligned}
&\int\limits_0^1|\lambda_Y(Z_t;\theta;\theta,\hat\mu) - \lambda_Y(Z_t;\theta;\theta,\hat\mu_0)|\,dt \qquad (29)\\
\le\ & \|\lambda_Y(\cdot;\theta,\hat\mu) - \lambda_Y(\cdot;\theta,\hat\mu_0)\|_\infty \le C\sum_{i=0}^{1}||\hat\mu^{(i)} - \hat\mu_0^{(i)}||_\infty.
\end{aligned}$$

Using Jensen's inequality and (29) once more,

$$\frac{E_B[\exp[\Delta\int_0^1 a\lambda_Y(Z_t;\theta,\hat\mu) + (1-a)\,\lambda_Y(Z_t;\theta,\hat\mu_0)\,dt]]}{E_B[\psi(x|x_0;\theta,\hat\mu)]} \le \exp\left[C\sum_{i=0}^{1}||\hat\mu^{(i)} - \hat\mu_0^{(i)}||_\infty\right].$$

In total,

$$\frac{|E_B[\psi(x|x_0;\theta,\hat\mu)] - E_B[\psi(x|x_0;\theta,\hat\mu_0)]|}{E_B[\psi(x|x_0;\theta,\hat\mu)]} \le C\exp\left[C\sum_{i=0}^{1}||\hat\mu^{(i)} - \hat\mu_0^{(i)}||_\infty\right]\sum_{i=0}^{1}||\hat\mu^{(i)} - \hat\mu_0^{(i)}||_\infty, \qquad (30)$$

uniformly in θ. The above bound also holds with $\hat\mu$ and $\hat\mu_0$ interchanged. Claim 1) now follows from Lemma 11 and 39 together with the assumptions on the bandwidth and trimming parameter in (C1.6a).

To prove Claim 2), write $q(x|x_0;\theta,a) = q(x|x_0;\theta,\hat T(a)\,\mu_0)$. We then make a Taylor expansion, $q(x|x_0;\theta,a) = q(x|x_0;\theta,0) + \partial_a q(x|x_0;\theta,\bar a)\,a$, $\bar a \in [0,a]$, and claim that $|\partial_a q(x|x_0;\theta,a)| \le b(x|x_0)\,E_B[\int_0^1|\partial_a\hat T(Z_t;a)|^2dt]^{1/2}$ uniformly in $\theta\in\Theta$, where $E_\pi[b(X_1|X_0)] < \infty$. This will yield 2) since

$$\bar a E_{\pi\ \ B}\left[\int_0^1|\partial_a\hat T(Z_t;\bar a)|^2dt\right]^{1/2} \le CP_{\pi\ \ B}(a/2 \le \hat\pi(Z_t) \le a)^{1/2} \to 0,$$

where the 2nd inequality holds for n sufficiently large. We have

$$|\partial_a q\left(x|x_0;\theta,a\right)| \leq \frac{E_B\left[|\partial_a \psi\left(x|x_0;\theta,a\right)|\right]}{E_B\left[\psi\left(x|x_0;\theta,a\right)\right]} \leq E_B\left[\Delta\int_0^1 |\partial_a \lambda_Y\left(Z_t;\theta,a\right)|\, dt\right],$$

Using Lemma 12 together with (C1.2)-(C1.3),

$$|\partial_a \lambda_Y\left(z;\theta,a\right)| \leq C\left(1+|z|^q\right)|\partial_a \hat{T}\left(z;a\right)|,$$

such that by Lemma 13 $b\left(x|x_0\right) = C\left(1+|x|^q+|x_0|^q\right)$ will satisfy the desired bound. By (A0), b has a first moment.

Finally, 3) and 4) follow from standard fully parametric uniform LLN, see for example Tauchen (1985, Lemma 1): Observe that (i) Θ is compact; (ii) $\theta \mapsto q\left(x|x_0;\theta,\mu_0\right)$ is continuous; and (iii) $|q\left(x|x_0;\theta,\mu_0\right)| \leq C\left(1+|x|^q+|x_0|^q\right)$ with q given in (C1.3). The last claim follows from the fact that, by (C1.3) and Lemma 12,

$$-C\log(E_B[\exp[-C\int_0^1 |B_t|^q\, dt]])\left(1+|x|^q+|x_0|^q\right) \leq \log\left(E_B\left[\psi\left(x|x_0;\theta,\mu_0\right)\right]\right) \leq \bar{\lambda}_Y$$

where $E_B[\exp[-C\int_0^1 |B_t|^q\, dt]] < \infty$. We have now shown that the conditions of Newey and McFadden (1994, Theorem 2.1) are satisfied, and thereby that $\hat{\theta}$ is consistent. ■

Proof of Theorem 6. Define $s_0\left(x|x_0\right) \equiv s\left(x|x_0;\theta_0,\mu_0\right)$. Lemma 14-22 then establishes that

$$\sqrt{n}(\hat{\theta}-\theta_0) = \left(H_0^{-1}+o_P(1)\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{s_0\left(X_i|X_{i-1}\right)+\delta\left(X_{i-1}\right)\right\}+o_P(1).$$

Using a CLT for mixing sequences, see e.g. Doukhan et al. (1994), we are able to conclude that $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N\left(0, H_0^{-1}\Sigma_\infty H_0^{-1}\right)$, where $\Sigma_\infty = \Sigma_0 + 2\sum_{i=1}^{\infty}\Sigma_i$ and

$$\Sigma_i = E_\pi\left[\left\{s_0\left(X_1|X_0\right)+\delta\left(X_0\right)\right\}\left\{s_0\left(X_{i+1}|X_i\right)+\delta\left(X_i\right)\right\}\right].$$

The moments Σ_∞ and H_0 are well-defined by Lemma 14, 21, and 22. Using that the process $\{X_t\}$ is a time reversible stationary Markov process, c.f. Hansen and Scheinkman (1995), together with the fact that is a martingale difference, it holds for any $i \geq 1$ that

$$E_\pi\left[s_0\left(X_1|X_0\right)s_0\left(X_{i+1}|X_i\right) \quad |X_i,X_1,X_0\right] = s_0\left(X_1|X_0\right)\times 0,$$

$$E_\pi\left[s_0\left(X_1|X_0\right)\delta\left(X_i\right) \quad\right] = E_\pi\left[s_0\left(X_{i-1}|X_i\right)\delta\left(X_0\right) \quad\right] = E_\pi\left[0\times\delta\left(X_0\right) \quad\right],$$

and similar for the second cross-term. ■

Proof of Theorem 7. Define

$$\tilde{\delta}_i = n^{-1}\sum_{k=1}^{n}\nabla_\pi s\left(X_k|X_{k-1};\theta_0,\hat{\mu}_0\right)\left[K_h\left(\cdot - X_i\right)\right], \quad \bar{\delta}_i = E_\pi\left[\nabla_\pi s\left(X_1|X_0;\theta_0,\hat{\mu}_0\right)\left[K_h\left(\cdot - X_i\right)\right]\right]$$

and observe that $\hat{\delta}_i = n^{-1}\sum_{k=1}^{n}\nabla_\pi s\left(X_k|X_{k-1};\hat{\theta},\hat{\mu}\right)\left[K_h\left(\cdot - X_i\right)\right]$ by definition of the pathwise derivative of s w.r.t. π, The first part of the proof then follows the one of Newey (1994, Lemma 5.5): By

Lemma 16,

$$\begin{aligned}||\hat{\delta}_i - \tilde{\delta}_i|| &\leq C \sum_{i,j=0}^{2} ||\nabla \partial^{ij}_{x,\theta} \mu \left[K_h \left(\cdot - X_i\right)\right]|| \\ &\quad \times \left\{ \sum_{i,j=0}^{2} ||\partial^{ij}_{x,\theta}\hat{\mu} - \partial^{ij}_{x,\theta}\hat{\mu}_0||_\infty + a E_B \left[\int_0^1 |\partial_a \hat{T}\left(Z_t; a\right)|^2 dt\right]^{1/2} + ||\hat{\theta} - \theta_0|| \right\},\end{aligned}$$

where $||\nabla^{ij}_{x,\theta}\mu \left[K_h \left(\cdot - X_i\right)\right]||_\infty \leq \sum_{k=0}^{i+1} a^{-1-k} h_k^{-1-k}$. It is then easily checked that under (C1.6b) $n^{-1}\sum_{i=1}^{n} ||\hat{\delta}_i - \tilde{\delta}_i||^2 = o_P\left(n^{-1/2}\right)$. From the proof of Lemma 19 it follows that $n^{-1}\sum_{i=1}^{n} ||\tilde{\delta}_i - \bar{\delta}_i||^2 = o\left(n^{-1/2}\right)$, while by Lemma 21, $n^{-1}\sum_{i=1}^{n} ||\bar{\delta}_i - \delta_i||^2 = o\left(n^{-1/2}\right)$ with $\delta_i = \delta\left(X_{i\Delta}\right)$. In total, $n^{-1}\sum_{i=1}^{n} ||\hat{\delta}_i - \delta_i||^2 = o_P\left(n^{-1/2}\right)$. Assume that $d = \dim\left(\theta\right) = 1$ (otherwise consider $a\, \hat{\delta}_i \hat{\delta}_j a$ for any d-dim. vector a), and obtain

$$\begin{aligned}|\hat{\Omega}_i - \Omega_i| &\leq \frac{1}{n}\sum_{i=1}^{n} |\hat{\delta}_i\hat{\delta}_j - \delta_i\delta_j| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \left\{ \delta_i |\hat{\delta}_j - \delta_j| + |\delta_j(\hat{\delta}_i - \delta_i)| + |(\hat{\delta}_j - \delta_j)(\hat{\delta}_i - \delta_j)| \right\} \\ &\leq 2\left(\frac{1}{n}\sum_{i=1}^{n}\delta_i^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}|\hat{\delta}_j - \delta_j|^2\right)^{1/2} + \frac{2}{n}\sum_{i=1}^{n}|\hat{\delta}_i - \delta_i|^2 \\ &= o_P(n^{-1/4}).\end{aligned}$$

Next, very much copying the arguments of Newey and West (1987, Proof of Theorem 2), it then follows that $\hat{V} \to^P V_0$ under our conditions. Finally, $\hat{H} \to^P H_0$ by Lemma 22. ■

Proof of Theorem 8. As we did in Class 1, we modify our criterion function, and define

$$\hat{\theta} = \arg\min_{\theta \in \Theta} Q_n\left(\theta, \hat{\sigma}^2\left(\cdot; \theta\right)\right),$$

where $Q_n\left(\theta, \sigma^2\right) = \frac{1}{n}\sum_{i=1}^{n} q\left(X_i | X_{i-1}; \theta, \sigma^2\right)$ and

$$\begin{aligned}q\left(x|x_0; \theta, \sigma^2\right) &= -\frac{1}{4}\log\left[\sigma^2\left(x\right)\right] - \frac{1}{4}\log\left[\sigma^2\left(x_0\right)\right] \\ &\quad -\frac{1}{2\Delta}\left(\int_{x_0}^{x} \sigma^{-1}\left(w\right) dw\right)^2 + \log\left(E_B\left[\psi\left(x|x_0; \theta, \sigma^2\right)\right]\right).\end{aligned}$$

We now follow the same three steps as in proof of Theorem 5, and therefore do not give all details. We have

$$\frac{\hat{\sigma}_0^2\left(z;\theta\right) - \hat{\sigma}^2\left(z;\theta\right)}{\hat{\sigma}^2\left(z;\theta\right)} \leq \log\left(\frac{\hat{\sigma}^2\left(z;\theta\right)}{\hat{\sigma}_0^2\left(z;\theta\right)}\right) \leq \frac{\hat{\sigma}^2\left(z;\theta\right) - \hat{\sigma}_0^2\left(z;\theta\right)}{\hat{\sigma}_0^2\left(z;\theta\right)}$$

where $\hat{\sigma}^2\left(z;\theta\right), \hat{\sigma}_0^2\left(z;\theta\right) \geq \underline{\sigma}^2$. Thus,

$$\left|\log\left(\frac{\hat{\sigma}^2\left(z;\theta\right)}{\hat{\sigma}_0^2\left(zx;\theta\right)}\right)\right| \leq \frac{\left\|\hat{\sigma}^2\left(\cdot;\theta\right) - \hat{\sigma}_0^2\left(\cdot;\theta\right)\right\|_\infty}{\underline{\sigma}^2}$$

With $z = x$ and x_0, this establishes the desired bound for the first two terms. The third term satisfies

$$\begin{aligned} & \left(\int_{x_0}^{x} \hat{\sigma}\left(x;\theta\right)^{-1} dw \right)^2 - \left(\int_{x_0}^{x} \hat{\sigma}_0\left(x;\theta\right)^{-1} dw \right)^2 \\ = & \left(\int_{x_0}^{x} \hat{\sigma}\left(x;\theta\right)^{-1} - \hat{\sigma}_0\left(x;\theta\right)^{-1} dw \right) \left(\int_{x_0}^{x} \hat{\sigma}\left(x;\theta\right)^{-1} + \hat{\sigma}_0\left(x;\theta\right)^{-1} dw \right), \end{aligned}$$

where

$$\begin{aligned} \int_{x_0}^{x} \left| \hat{\sigma}\left(w;\theta\right)^{-1} - \hat{\sigma}_0\left(w;\theta\right)^{-1} \right| dw & \leq \underline{\sigma}^{-1}\left(|x| + |x_0|\right) \left\| \hat{\sigma}^2\left(\cdot;\theta\right) - \hat{\sigma}_0^2\left(\cdot;\theta\right) \right\|_\infty \\ \int_{x_0}^{x} \hat{\sigma}\left(w;\theta\right)^{-1} + \hat{\sigma}_0\left(w;\theta\right)^{-1} dw & \leq 2\underline{\sigma}^{-1}\left(|x| + |x_0|\right), \end{aligned}$$

Define $\hat{Z}_t = Z_t\left(\hat{\sigma}^2\right)$ and $\hat{Z}_{0t} = Z_t\left(\hat{\sigma}_0\right)$. We then have

$$\begin{aligned} \left| \log\left(\frac{E_B\left[\psi\left(x|x_0;\theta,\hat{\sigma}^2\right)\right]}{E_B\left[\psi\left(x|x_0;\theta,\hat{\sigma}_0^2\right)\right]} \right) \right| & \leq E_B[\exp[\Delta \int_0^1 |\lambda_Y(\hat{Z}_t;\theta,\hat{\sigma}^2) - \lambda_Y(\hat{Z}_{0t};\theta,\hat{\sigma}_0^2)|dt]] \\ & \quad \times E_B[\int_0^1 |\lambda_Y(\hat{Z}_t;\theta,\hat{\sigma}^2) - \lambda_Y(\hat{Z}_{0t};\theta,\hat{\sigma}_0^2)|dt], \end{aligned}$$

where, using a Taylor expansion together with Lemma 25 and 27,

$$|\lambda_Y\left(\hat{Z}_t;\theta,\hat{\sigma}_0^2\right) - \lambda_Y\left(\hat{Z}_{0t};\theta,\hat{\sigma}_0^2\right)| \leq C\left(1 + a^{-3}\right) \left\|\hat{\sigma}^2 - \hat{\sigma}_0^2\right\|_\infty.$$

This together with Lemma 26, implies that

$$\left|Q_n\left(\theta,\hat{\sigma}^2\right) - Q_n\left(\theta,\hat{\sigma}_0^2\right)\right| \leq C\left(1 + \left(|x| + |x_0|\right)\right) a^{-3} \sum_{i=0}^{2} \left\|\partial_x^i\hat{\sigma}^2 - \partial_x^i\hat{\sigma}_0^2\right\|_\infty$$

where, by Lemma 24 and 39 together with (C2.6a), $a^{-3}\left\|\partial_x^i\hat{\sigma}^2 - \partial_x^i\hat{\sigma}_0^2\right\|_\infty = o_P(1)$, $0 \leq i \leq 2$, and $E_\pi\left[|X_0|\right] < \infty$. Next, define $q\left(x|x_0;\theta,a\right) = q(x|x_0;\theta,\hat{T}\left(\cdot;a\right)\sigma_0^2 + (1 - \hat{T}\left(\cdot;a\right))\underline{\sigma}^2)$. We obtain by Lemma 25, 26, 27 and (C.1.3) that

$$\begin{aligned} \left|\partial_a q\left(x|x_0;\theta,a\right)\right| & \leq b\left(x|x_0\right)\{|\partial_a\hat{T}\left(x;a\right)| + |\partial_a\hat{T}\left(x;a\right)| \\ & \quad + \int_{x_0}^{x} |\partial_a\hat{T}\left(w;a\right)|dw + E_B[\int_0^1 |\partial_a\hat{T}\left(Z_t;a\right)dt|^2]^{1/2}\}, \end{aligned}$$

with $b\left(x|x_0\right) = C(1 + |x|^{2q} + |x_0|^{2q})$, and conclude $|Q_n(\hat{\theta},\hat{\sigma}_0^2) - Q_n(\hat{\theta},\sigma_0^2)| \to^P 0$ by the properties of $\hat{T}$. Finally, $\sup_{\theta\in\Theta}\left|Q_n\left(\theta,\sigma_0^2\right) - Q\left(\theta,\sigma_0^2\right)\right| \to^P 0$ where $Q\left(\theta,\sigma_0^2\right) = E_\pi\left[q\left(X_1|X_0;\theta,\sigma_0^2\right)\right]$ since: (i) Θ is compact; (ii) $\theta \mapsto q\left(x|x_0;\theta,\sigma_0^2\left(\cdot;\theta\right)\right)$ is continuous; (iii) by (C2.3) and Lemma 25, $\left|q\left(x|x_0;\theta,\sigma_0^2\right)\right| \leq C(1 + |x|^{2q} + |x_0|^{2q})$. ■

Proof of Theorem 9. This follows the same steps as the proof of Theorem 6, now only using Lemma 28-36. ■

Proof of Theorem 10. The claim is proved in the same fashion as Theorem 7, this time using Lemma 30, 33, 35 and 36. ■

B Lemmas

In this section we state all lemmas used in the proofs of the theorems above without any proofs. These can be obtained from the author upon request.

B.1 Class 1

Lemma 11 *Under (C1.1)-(C1.6a),*

$$||\partial_\theta^j \hat{\mu}^{(k)}(\cdot;\theta) - \partial_\theta^j \hat{\mu}_0^{(k)}(\cdot;\theta)||_\infty \leq C \sum_{i=0}^{k+1} a^{i-k-2} ||\hat{\pi}^{(i)} - \pi_0^{(i)}||_\infty,$$

$$||\partial_\theta^j \partial_x^k \hat{\mu}(\cdot;\theta) - \partial_\theta^j \partial_x^k \hat{\mu}_0(\cdot;\theta) - \partial_\theta^j \partial_x^k \nabla_\pi \hat{\mu}_0(\cdot;\theta)[\hat{\pi} - \pi_0]||_\infty \leq C \sum_{i=0}^{k+1} a^{i-k-3} ||\hat{\pi}^{(i)} - \pi_0^{(i)}||_\infty^2$$

uniformly over $\theta \in \Theta$, where $\partial_\theta^j \partial_x^k \nabla_\pi \hat{\mu}_0(\cdot;\theta)[d\pi]$ is given in (??)-(??).

Lemma 12 *Under (C1.1)-(C1.6a),*

$$\left\|\partial_{x,\theta}^{kl} \lambda_Y(\cdot;\theta,\hat{\mu}) - \partial_{x,\theta}^{kl} \lambda_Y(\cdot;\theta,\hat{\mu}_0)\right\|_\infty \leq C \sum_{i,j=0}^{k+1} a^{i-k-1} ||\partial_{x,\theta}^{ij}\hat{\mu} - \partial_{x,\theta}^{ij}\hat{\mu}_0^{(j)}||_\infty,$$

for $0 \leq k,l \leq 2$. Moreover,

$$\begin{aligned} |\lambda_Y(z;\theta,\mu)| &\leq C\left(1 + |\mu(z;\theta)|^2 + |\mu^{(1)}(z;\theta)|\right), \\ |\lambda_Y^{(1)}(z;\theta,\mu)| &\leq C\left(1 + |\mu(z;\theta)|^2 + |\mu(z)|\,\mu^{(1)}(z;\theta)| + |\mu^{(2)}(z;\theta)|\right), \\ ||\dot{\lambda}_Y(z;\theta,\mu)|| &\leq C\left(1 + |\mu(z;\theta)|^2 + |\mu(z)|\,|\dot{\mu}(z;\theta)| + ||\dot{\mu}^{(1)}(z;\theta)||\right). \end{aligned}$$

Lemma 13 *Under (C1.1)-(C1.4), $\left|\partial_\theta^i Z_t(x|x_0;\theta)\right| \leq C(1 + |x| + |x_0| + |B_t|)$, $i = 0,1,2$.*

B.1.1 The Score

Lemma 14 *Under (C1.1)-(C1.7), $S_n(\theta_0,\hat{\mu}_0) = S_n(\theta_0,\mu_0) + o_P(n^{-1/2})$, where for some $\delta > 0$, $E_\pi[\left\|s(X_1|X_0;\theta_0;\sigma_0^2)\right\|^{2+\delta}] < \infty$.*

B.1.2 The Pathwise Derivative of the Score

Lemma 15 *Assume that (C1.1)-(C1.5) hold. Then for any direction $\nabla\mu$,*

$$\|\nabla s(\theta_0,\mu_0)[\nabla\mu]\| \leq b(x|x_0) \sum_{i,j=0}^{2} E_B\left[\int_0^1 ||\partial_{x,\theta}^{i,j}\nabla\mu(Z_t)||^2 dt\right]^{1/2}, \tag{31}$$

where $E_\pi[b^{2+\delta}(X_1|X_0)] < \infty$, for some $\delta > 0$.

Lemma 16 *Under (C1.1)-(C1.5), there exist a function* b *with* $E_\pi\left[b^{2+\delta}(X_1,X_0)\right]<\infty$ *such that for all* $\theta\in\Theta$ *and with* $||\mu||_{2,\infty}\equiv\sum_{i,j=0}^{2}||\partial_{x,\theta}^{ij}\mu||_\infty$*,*

$$\begin{aligned}&||\nabla s\left(x|x_0;\theta,\hat{\mu}\right)[d\mu]-\nabla s\left(x|x_0;\theta_0,\mu_0\right)[d\mu]|| \\ \leq\;& b\left(x|x_0\right)||d\mu||_{2,\infty}\left\{||\hat{\mu}-\hat{\mu}_0||_{2,\infty}+aE_B\left[\int_0^1|\partial_a\hat{T}\left(Z_t;a\right)|^2dt\right]^{1/2}+||\theta-\theta_0||\right\}\end{aligned} \tag{32}$$

Lemma 17 *Under (C1.1)-(C1.6), there exists a function* b *with* $E_\pi\left[b\left(X_1|X_0\right)\right]<\infty$ *such that*

$$||s\left(\theta_0,\hat{\mu}\right)-s\left(\theta_0,\hat{\mu}_0\right)-\nabla s\left(\theta_0,\hat{\mu}_0\right)[\hat{\mu}-\hat{\mu}_0]||\leq b\left(x|x_0\right)\sum_{i,j=0}^{2}||\partial_{x,\theta}^{ij}\hat{\mu}-\partial_{x,\theta}^{ij}\hat{\mu}_0||_\infty.$$

We define

$$\nabla_\pi s\left(\mu\right)[d\pi]\equiv\nabla s\left(\mu\right)[\nabla_\pi\mu\left(x\right)[\hat{\pi}-\pi_0]], \tag{33}$$

where $\nabla_\pi\mu$ is given in Lemma 11. This function is the pathwise derivative w.r.t. π.

Lemma 18 *Under (C1.1)-(C1.5),*

$$||\nabla s\left(x|x_0;\theta,\hat{\mu}_0\right)[\hat{\mu}-\hat{\mu}_0]-\nabla_\pi s\left(x|x_0;\theta,\hat{\mu}_0\right)[\hat{\pi}-\pi_0]||\leq b\left(x|x_0\right)\sum_{i=0}^{3}a^{i-5}||\hat{\pi}^{(i)}-\pi_0^{(i)}||_\infty^2$$

uniformly in $\theta\in\Theta$*, where* $\nabla_\pi s$ *is given in (33) and* $E_\pi\left[b\left(X_1|X_0\right)\right]<\infty$*.*

B.1.3 The Adjustment Term

In this section we show that the pathwise derivative of the score can be written as a normalised sum and a remainder term which can be ignored.

Lemma 19 *Under (C1.1)-(C1.6),* $\nabla S_n\left(\hat{\mu}_0\right)[\hat{\mu}-\hat{\mu}_0]=\nabla S\left(\hat{\mu}_0\right)[\hat{\mu}_0-\hat{\mu}_0]+o_P\left(n^{-1/2}\right)$*.*

Lemma 20 *Under (C1.1)-(C1.6),* $\nabla S\left(\hat{\mu}_0\right)[\hat{\mu}-\hat{\mu}_0]=\nabla S\left(\mu_0\right)[\hat{\mu}-\hat{\mu}_0]+o_P\left(n^{-1/2}\right)$*.*

Lemma 21 *Under (C1.1)-(C1.7), there exists a function* δ *with* $E_\pi\left[\delta\left(X_0\right)\right]=0$ *and* $E_\pi\left[\|\delta\left(X_0\right)\|^{2+\varepsilon}\right]<\infty$ *for some* $\varepsilon>0$ *such that*

$$\nabla S\left(\mu_0\right)[\hat{\mu}-\hat{\mu}_0]=\frac{1}{n}\sum_{i=1}^{n}\delta\left(X_i\right)+o_P(n^{-1/2}).$$

B.1.4 The Hessian

Lemma 22 *Under (C1.1)-(C1.7),* $\sup_{\theta\in\Theta}\|H_n\left(\theta,\hat{\mu}\right)-H\left(\theta,\mu_0\right)\|\to^P 0$*.*

B.2 Class 2

Lemma 23 *Assume that: (i) (A0) holds: (ii) $\theta \mapsto \mu(x;\theta)$ is $k+1$ times continuously differentiable, satisfying $||\partial_\theta^i \mu(x;\theta)|| \leq C|x|^{\bar{q}/(2+\delta)}$ for $0 \leq i \leq k+1$ and some $\delta > 0$. Then*

$$\sup_{(x,\theta)\in\mathbb{R} \quad \Theta} \left| \frac{1}{n} \sum_{i=1}^n 1_{(l,x)}(X_i)\, \partial_\theta^k \mu(X_i;\theta) - \int_l^x \pi_0(y)\, \partial_\theta^k \mu(y;\theta)\, dy \right| = O_P(n^{-1/2}).$$

Lemma 24 *Under (C2.1)-(C2.4), the following holds uniformly over $(x,\theta) \in \mathbb{R} \times \Theta$, $0 \leq i \leq 4$, and $0 \leq j \leq 2$:*

$$||\partial_x^i \partial_\theta^j \hat{\sigma}^2(x;\theta) - \partial_x^i \partial_\theta^j \hat{\sigma}_0^2(x;\theta)|| \leq O_P(a^{-1-i} n^{-1/2}) + C \sum_{k=0}^i a^{-2-i+k} ||\hat{\pi}^{(k)} - \pi_0^{(k)}||_\infty,$$

$$\begin{aligned} & ||\partial_x^i \partial_\theta^j \hat{\sigma}^2(x;\theta_0) - \partial_x^i \partial_\theta^j \hat{\sigma}_0^2(x;\theta_0) - \partial_x^i \partial_\theta^j \nabla_\pi \hat{\sigma}_0^2(x;\theta_0)[\hat{\pi} - \pi_0]|| \\ \leq \;& O_P(1) \times \sum_{k=0}^i a^{-3-i+k} ||\hat{\pi}^{(k)} - \pi_0^{(k)}||_\infty^2 + O_P(n^{-1/2}) \times \sum_{k=0}^i a^{-2-i+k} ||\hat{\pi}^{(k)} - \pi_0^{(k)}||_\infty. \end{aligned}$$

Lemma 25 *Under (C2.1)-(C2.4), (i) $-(1 + C|z|^q) \leq \lambda_Y(z;\theta,\sigma_0^2) \leq \bar{\lambda}_Y$ and (ii) $|\lambda_Y^{(1)}(z;\theta,\hat{\sigma}_0^2)| \leq C + |\mu^{(2)}(z;\theta)|$.*

Lemma 26 *Under (C2.1)-(C2.4) and (C2.6a),*

$$\left| \partial_{x,\theta}^{ij} \lambda_Y(z;\theta,\hat{\sigma}^2) - \partial_{x,\theta}^{ij} \lambda_Y(z;\theta,\hat{\sigma}_0^2) \right| \leq C \sum_{k=0}^{i+2} \sum_{l=0}^{j} a^{-3-i} ||\partial_{x,\theta}^{kl} \hat{\sigma}^2 - \partial_{x,\theta}^{kl} \hat{\sigma}_0^2||_\infty$$

for $0 \leq i, j \leq 2$.

Lemma 27 *Under (C2.1)-(C2.4) and (C2.6a), for all $(x, x_0, \theta) \in \mathbb{R}^2 \times \Theta$,*

1. $|\hat{Z}_t(\theta)|, |\hat{Z}_{0t}(\theta)|, |Z_{0t}(\theta)| \leq C(|x| + |x_0| + |B_t|)$.

2.
$$\begin{aligned} |\hat{Z}_t(\theta) - \hat{Z}_{0t}(\theta)| &\leq C \left\| \hat{\sigma}^2 - \hat{\sigma}_0^2 \right\|_\infty, \\ ||\partial_\theta^k Z_t(\theta) - \partial_\theta^k Z_{0t}(\theta)|| &\leq b(x|x_0) \sum_{i=0}^k \left\| \partial_\theta^i \hat{\sigma}^2 - \partial_\theta^i \hat{\sigma}_0^2 \right\|_\infty, \end{aligned}$$

   where $E_\pi[b(X_1|X_0)] < \infty$.

B.2.1 The Score

Lemma 28 *Under (C2.1)-(C2.7),*

$$S_n(\theta_0;\hat{\sigma}_0^2) = S_n(\theta_0;\sigma_0^2) + o_P(n^{-1/2}),$$

where $E_\pi[\left\| s(X_1|X_0;\theta_0;\sigma_0^2) \right\|^{2+\delta}] < \infty$.

B.2.2 The Pathwise Derivative of the Score

Lemma 29 *Assume that (C2.1)-(C2.6) hold. Then*

$$\begin{aligned}\left\|\nabla s\left(x|x_0;\theta,\sigma_0^2\right)\left[\nabla\sigma^2\right]\right\| &\leq C\sum_{j=0}^{1}\left(||\partial_\theta^j\nabla\sigma^2\left(x\right)||+||\partial_\theta^j\nabla\sigma^2\left(x_0\right)||\right)\\ &+b\left(x|x_0\right)\sum_{j=0}^{1}\int_{\min\ x,x_0}^{\max\ x,x_0}|\partial_\theta^j\nabla\sigma^2\left(w\right)|dw\\ &+b\left(x|x_0\right)\sum_{i=0}^{2}\sum_{j=0}^{1}E_B\left[\int_0^1||\partial_x^i\partial_\theta^j\nabla\sigma^2\left(Z_t\right)||^2dt\right]^{1/2},\end{aligned}$$

where $E_\pi\left[b^{2+\delta}\left(X_1|X_0\right)\right]<\infty$*.*

Lemma 30 *Under (C2.1)-(C2.5), there exist a function b with $E_\pi\left[b^{2+\delta}\left(X_1,X_0\right)\right]<\infty$ such that for all $\theta\in\Theta$ and with $||\sigma^2||_{3,1,\infty}\equiv\sum_{i=0}^{3}\sum_{j=0}^{1}||\partial_{x,\theta}^{ij}\sigma^2||_\infty$,*

$$\begin{aligned}&\left\|\nabla s\left(x|x_0;\theta,\hat\sigma^2\right)\left[d\sigma^2\right]-\nabla s\left(x|x_0;\theta_0,\sigma_0^2\right)\left[d\sigma^2\right]\right\|\\ \leq\ &b\left(x|x_0\right)||d\sigma^2||_{3,1,\infty}\left\{||\hat\sigma^2-\sigma_0^2||_{3,1,\infty}+\|\theta-\theta_0\|+a\hat A\right\},\end{aligned}$$

where

$$\hat A=|\partial_a\hat T\left(x;a\right)|+|\partial_a\hat T\left(x;a\right)|+\int_{\min\ x,x_0}^{\max\ x,x_0}|\partial_a\hat T\left(w;a\right)|dw+E_B\left[\int_0^1|\partial_a\hat T\left(Z_t;a\right)|^2dt\right]^{1/2}.$$

Lemma 31 *Under (C2.1)-(C2.6),*

$$\left\|S_n\left(\theta_0,\hat\sigma^2\right)-S_n\left(\theta_0,\hat\sigma_0^2\right)-\nabla S_n\left(\theta_0,\hat\sigma_0^2\right)\left[\hat\sigma^2-\hat\sigma_0^2\right]\right\|\leq O_P\left(1\right)\times\left\|\hat\sigma^2-\hat\sigma_0^2\right\|_\infty^2.$$

We define $\nabla_\pi s\left(x|x_0;\theta,\sigma^2\right)[d\pi]=\nabla s\left(x|x_0;\theta,\sigma^2\right)\left[\nabla\sigma^2\left[d\pi\right]\right]$, and obtain

Lemma 32 *Under (C2.1)-(C2.6), there exists b with $E_\pi\left[b\left(X_1|X_0\right)\right]<\infty$ such that*

$$\left\|\nabla s\left(x|x_0;\theta_0,\hat\sigma_0^2\right)\left[\hat\sigma^2-\hat\sigma_0^2\right]-\nabla_\pi s\left(x|x_0;\theta_0,\hat\sigma_0^2\right)\left[\hat\pi-\pi_0\right]\right\|\leq b\left(x|x_0\right)\sum_{i=0}^{3}a^{i-5}||\hat\pi^{(i)}-\pi_0^{(i)}||_\infty^2.$$

B.2.3 The Adjustment Term

Lemma 33 *Under (C2.1)-(C2.7),*

$$\nabla S_n\left(\theta_0,\hat\sigma_0^2\right)\left[\hat\sigma^2-\sigma_0^2\right]=\nabla S\left(\theta_0,\hat\sigma_0^2\right)\left[\hat\sigma^2-\sigma_0^2\right]+o_P(n^{-1/2}).$$

Lemma 34 *Under (C2.1)-(C2.7),*

$$\nabla S\left(\theta_0,\hat\sigma_0^2\right)\left[\hat\sigma^2-\hat\sigma_0^2\right]=\nabla S\left(\theta_0,\sigma_0^2\right)\left[\hat\sigma^2-\hat\sigma_0^2\right]+o_P(n^{-1/2}).$$

Lemma 35 *Under (C2.1)-(C2.7), there exists a function δ with $E_\pi\left[\delta\left(X_0\right)\right]=0$ and $E_\pi\left[\left\|\delta\left(X_0\right)\right\|^{2+\epsilon}\right]<\infty$ such that*

$$\nabla S\left(\theta_0,\sigma_0^2\right)\left[\hat\sigma^2-\hat\sigma_0^2\right]=\frac{1}{n}\sum_{i=1}^{n}\delta\left(X_i\right)+o_P(n^{-1/2}).$$

B.2.4 The Hessian

Lemma 36 *Under (C2.1)-(C2.7),* $\mathsf{sup}_{\theta\in\Theta}\left\|H_n\left(\theta,\hat{\sigma}^2\right)-H\left(\theta,\sigma_0^2\right)\right\|\to^P \mathsf{0}$.