SCHOOL OF ECONOMICS AND MANAGEMENT
FACULTY OF SOCIAL SCIENCES
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

CREATES Research Paper 2009-41

Pseudo-Maximum Likelihood Estimation in Two Classes of Semiparametric Diffusion Models

Dennis Kristensen

School of Economics and Management
Aarhus University
Bartholins Allé 10, Building 1322, DK-8000 Aarhus C
Denmark

Pseudo-Maximum Likelihood Estimation in Two Classes of Semiparametric Diffusion Models*

Dennis Kristensen†
Columbia University and CREATES‡

September 18, 2009

Abstract

A novel estimation method for two classes of semiparametric scalar diffusion models is proposed: In the first class, the diffusion term is parameterised and the drift is left unspecified, while in the second class only the drift term is specified. Under the assumption of stationarity, the unspecified term can be identified as a functional of the parametric component and the stationary density. Given a discrete sample with a fixed time distance, the parametric component is then estimated by maximizing the associated likelihood with a preliminary estimator of the unspecified term plugged in. It is shown that this Pseudo-MLE (PMLE) is $\sqrt{n}$-consistent and asymptotically normally distributed under regularity conditions, and demonstrate how the models and estimators can be used in a two-step specification testing strategy of fully parametric models. Since the likelihood function is not available on closed form, the practical implementation of our estimator and tests will rely on simulated or approximate PMLE's. Under regularity conditions, it is verified that approximate/simulated versions of the PMLE inherits the properties of the actual but infeasible estimator. A simulation study investigates the finite-sample performance of the PMLE, and finds that it performs well and is comparable to parametric MLE both in terms of bias and variance.

Keywords: Diffusion process, fixed-time distance asymptotics, kernel estimation, pseudo-likelihood, semiparametric.

JEL-Classification: C12, C13, C14, C22.

*This paper was part of my PhD dissertation at the Department of Econonomics, LSE, and an earlier version was circulated with the title "Estimation in Two Classes of Semiparametric Diffusion Models." I am indebted to my supervisor Oliver Linton for his guidance and support. 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 Aarhus, University of Copenhagen, LSE, UCSD, Iowa State, Georgetown, Chicaco GSB, UW-Madison, Pittsburgh, Oxford, UCL and Yale for comments and suggestions. I gratefully acknowledge the financial support from the Danish Research Agency (research grant no. 2114-04-0001) and the Financial Markets Group, LSE.

†E-mail: dk2313@columbia.edu. Address: 1018 International Affairs Bldg, 420 West 118th St, New York, NY 10027.

‡Center for Research in Econometric Analysis of Time Series, funded by the Danish National Research Foundation.

1 Introduction

Continuous-time stochastic processes are widely used in finance to describe the dynamics of variables such as asset prices, interest rates and exchange rates; see Björk (2004) for an overview. The theory however puts very few restrictions on specification of the processes, and the applied researcher will therefore normally face the problem of choosing a suitable model for a given data set. In the empirical literature, there is a lack of consensus about the appropriate parametric model to use with many models having been proposed. This motivates this study where we take a semiparametric approach to the modelling, estimation and testing of diffusion processes. This allows us to obtain robust estimates with a smaller risk of misspecification, and to develop tests for fully parametric specifications.

A diffusion process is fully characterised by its so-called drift and diffusion function and if the researcher has some prior beliefs about the shape of either of these, he may be willing to impose a parametric form on one term while remaining agnostic about the other. This leads to two classes of semiparametric diffusion models: In the first class, only the diffusion is specified parametrically, while in the second class only the drift is specified.

A fully nonparametric specification is evidently more robust than the two semiparametric classes considered here, but one pays a price in terms of precision of the associated estimators: We here propose an estimator of the parametric component and show it converges with $\sqrt{n}$-rate, while in a nonparametric framework both terms will converge at a slower rate. Moreover, our theoretical results only rely on low-frequency observations, i.e. the asymptotics are derived with time distance between observations remaining fixed, and in this setting the fully nonparametric estimators proposed in the literature so far suffer from the disadvantage that their asymptotic distributions are unknown. Thus, they have limited use in drawing inference about the correct specification of the drift and diffusion term. In contrast, we here develop formal statistical tests that enable the researcher to compare fully parametric, semiparametric and nonparametric models: First, we propose a test of the null that a given semiparametric model is correct against the nonparametric alternative that the observed process is Markov. Second, we develop two tests of a fully parametric specification against either of its two semiparametric alternatives. The second set of tests complements the ones proposed in the companion paper Kristensen (2008a) where estimators of the unspecified component in the two semiparametric classes are derived and used in misspecification tests. Thus, the modelling, estimation and testing framework proposed here provides a set of tools that are useful in the search for a parsimonious parametric model when only low frequency observations are available.

The proposed estimation method relies on the assumption of stationarity of the observed diffusion process. Under this assumption, we demonstrate that the drift (diffusion) term can be expressed as a functional of the stationary marginal 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. Our estimator is then obtained through the following two steps: First, the stationary density is estimated nonparametrically and plugged into the log-likelihood of the discretely observed process together with the parametric specification of either the drift or the diffusion. We then define our semiparametric estimator as the maximizer of this pseudo-log-likelihood. Under regularity conditions, we show consistency and $\sqrt{n}$-asymptotic normality of this pseudo-maximum-likelihood estimator (PMLE).

The two sets of test statistics developed here are also based on the likelihood approach: To test the semiparametric model against the nonparametric alternative, we propose to estimate the transition density nonparametrically and then compare this with the constrained version associated with the semiparametric model; this can be done either using an L_2-distance or the Kullback-Leibler one. As such, these tests are similar in spirit to the ones developed in Aït-Sahalia, Fan and Peng (2009). To test a fully parametric specification nested within the semiparametric model, we use the score function of the semiparametric model to develop a Lagrange multiplier (LM) type test statistic as advocated in Whang and Andrews (1993). For both null hypotheses, we derive the asymptotic distribution of the relevant test statistics.

Since it is not possible to directly evaluate the likelihood function, we propose either to use approximate (e.g. Aït-Sahalia, 2002; Pedersen, 1995) or simulation-based (e.g. Kristensen and Shin, 2008) methods in order to implement the PMLE and the test statistics. We show that the estimator and tests obtained from such methods will enjoy the same properties as the actual, but infeasible, ones under weak conditions. The finite sample properties of the estimator using simulated likelihood are investigated in a small simulation study. Here, it is shown that, for even moderate sample sizes, the PMLE performs well compared to the fully parametric MLE, and that the general method of Kristensen and Shin (2008) yields a good approximation of the actual likelihood funtion for this particular application.

Our estimator fits nicely into the general class of semiparametric two-step estimators considered in e.g. Chen, Linton and van Keilegom (2003) and Newey and McFadden (1994, Section 8), where general conditions for consistency and asymptotic normality for such estimators are derived. We follow a similar proof strategy as these studies, but unfortunately the problem in consideration is not contained in their framework for two reasons: Firstly, they only consider 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 geometrically β-mixing. Secondly, we have to introduce trimming of our nonparametric estimators which is not considered in the aforementioned studies; see also Ai (1997) and Robinson (1988) for similar applications of trimming.

Nonparametric estimators of the drift and diffusion term have been widely studied, see e.g. Bandi and Phillips (2003), Chen, Hansen and Scheinkman (2000, 2009), Darolles and Gouriéroux (2001), Gobet et al. (2004). However, only few studies have considered semiparametric models: Aït-Sahalia (1996a) considers a semiparametric model which belongs to the first class of models considered here, while Conley et al. (1997) specify a flexible parametric model which can be interpreted as a semiparametric model of Class 2. Bandi and Phillips (2007) propose general semiparametric estimators using kernel methods based on in-fill asymptotics where time distance between observations shrinks to zero. In contrast, the estimators and asymptotics developed in Aït-Sahalia (1996a) and Conley et al. (1997) and this study are based on a fixed time distance.

Numerous misspecification tests of diffusion models given low-frequency data have been proposed in the literature. However, most of these test a fully parametric specification against a nonparametric Markov alternative (see e.g., Aït-Sahalia, 1996b; Aït-Sahalia et al, 2009; Corradi and Swanson, 2005; Hong and Li, 2004). Thus, they simultaneously test both the specification of the drift and diffusion term, and, as a consequence, these tests are not informative about the possible cause of a given rejection and do not give much guidance in the search for a correct specification. In contrast, we develop a two-step procedure where we test for the correct specification of one term at a time. In case of overall rejection, our test strategy is therefore able to detect whether the drift or the diffusion term is misspecified (or both), and so should be a more helpful in guiding the researcher towards a correctly specified model.

The rest of the paper is organised as follows: In Section 2, we introduce the semiparametric models, and develop the PMLE's of the parametric components. The asymptotics of the PMLE and misspecification tests are derived in Section 3 and 4 respectively. Section 5 deals with the implementation of the estimator and test statistics, and the results of the simulation study are presented in Section 6. We conclude in Section 7. Proofs and lemmas have been collected into the appendices. We will use the following notation: For a function $f(z)$, we write $f^{(i)}(z) = \partial^i f(z) / \partial z^i$. For a function $f(z;\theta)$, we write $\partial_\theta^i f(z;\theta) = \partial^i f(z;\theta) / \partial\theta^i$.

2 Models and Estimators

Let $\{X_t\} = \{X_t : t \geq 0\}$ be a univariate time-homogenous diffusion process solving the following stochastic differential equation (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 $-\infty \leq l < r \leq \infty$. The functions $\mu : I \mapsto \mathbb{R}$ and $\sigma^2 : I \mapsto \mathbb{R}_+$ are the so-called drift and diffusion term respectively. Suppose that we have observed $n+1$ observations from (1), $X_0, X_\Delta, X_{2\Delta} ..., X_{n\Delta}$, where $\Delta > 0$ is the fixed time distance between observation. We then wish to draw inference regarding the drift and diffusion term.

The process $\{X_t\}$ is Markov, and we base our estimation method on its transition density $p_t(y|x)$, $P(X_{s+t} \in A | X_s = x) = \int_A p_t(y|x)\, dy$ for $s, t \geq 0$ and any Borel set $A \subseteq I$. Under regularity conditions, the transition density solves the so-called backward Kolmogorov equation,

$$\frac{\partial p_t(y|x)}{\partial t} = \mathcal{A}\left[\mu, \sigma^2\right] p_t(y|x), \quad t > 0, \ \ (x, y) \in I, \tag{2}$$

where $\mathcal{A}$ is the infinitesimal generator,

$$\mathcal{A}\left[\mu, \sigma^2\right] p_t(y|x) = \mu(x) \frac{\partial p_t(y|x)}{\partial x} + \frac{1}{2}\sigma^2(x) \frac{\partial^2 p_t(y|x)}{\partial x^2}.$$

Since $p_t(y|x)$ is completely characterised by the generator $\mathcal{A}\left[\mu,\sigma^2\right]$, it is a functional of μ and σ and we will write $p_t(y|x) = p_t(y|x;\mu,\sigma)$. For an overview of partial differential equations (PDE's) and their connection to stochastic differential equations, we refer to Friedman (1976).

Suppose that $\{X_t\}$ is strictly stationary and ergodic, in which case it has a stationary marginal density which we denote π, $P(X_t \in A) = \int_A \pi(x)\,dx$, for $t \geq 0$ and any Borel set $A \subseteq I$. Observe that π is invariant, $\pi(y) = \int_I p_t(y|x)\,\pi(x)\,dx$, $t \geq 0$. An alternative characterization of the transition density is through the forward Kolmogorov equation,

$$\frac{\partial p_t(y|x)}{\partial t} = -\frac{\partial\left[\mu(y)\,p_t(y|x)\right]}{\partial y} + \frac{1}{2}\frac{\partial^2\left[\sigma^2(y)\,p_t(y|x)\right]}{\partial y^2}, \quad t > 0, (x,y) \in I.$$

Multiplying this PDE with $\pi(x)$ on both sides and integrating w.r.t. x, the following two equivalent expressions linking together μ, σ^2 and π are obtained:

$$\mu(x) = \frac{1}{2\pi(x)}\frac{\partial}{\partial x}\left[\sigma^2(x)\,\pi(x)\right], \tag{3}$$

$$\sigma^2(x) = \frac{2}{\pi(x)}\int_l^x \mu(y)\,\pi(y)\,dy. \tag{4}$$

From these two expressions, we see that instead of specifying the drift and diffusion term, 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 Hansen and Scheinkman (1995) and Hansen et al. (1998). Since the stationary density is nonparametrically identifiable, we only have to specify either the drift or the diffusion term; we can then identify the remainder term from the two others. This observation leads us to focus on the two following classes of semiparametric diffusion models:

Class 1: $\mu(\cdot)$ is unknown and $\sigma^2(\cdot;\theta_1)$ known up to some parameter $\theta_1 \in \Theta_1$.

Class 2: $\mu(\cdot;\theta_2)$ is known up to some parameter $\theta_2 \in \Theta_2$ and $\sigma^2(\cdot)$ unknown.

For any model in either class, we have an unknown finite-dimensional parameter, θ_k, and an infinite-dimensional one, μ (in Class 1) or σ^2 (in Class 2). Observe that any parametric model can be nested as a submodel both in Class 1 and 2. So by estimating the corresponding semiparametric model, we obtain more robust estimates of either the drift and the diffusion. Also, we will be able to test any fully parametric specification against each of its two semiparametric alternatives.

To discuss the estimation of the two classes of models, let us first consider a model from Class 1. In this case, we are given a parameterisation of the diffusion term, $\sigma^2(\cdot;\theta_1)$, which we plug into the right-hand side of (3) yielding $\mu(x;\theta_1) = \frac{\partial}{\partial x}\left[\sigma^2(x;\theta_1)\,\pi(x)\right]/\left[2\pi(x)\right]$. Since π is unknown, we substitute a nonparametric estimator for it and here choose to use a nonparametric kernel density estimator:

$$\hat{\pi}(x) = \frac{1}{n}\sum_{i=1}^n K_h(x - X_{i\Delta}), \tag{5}$$

where $K_h(z) = K(z/h)/h$ for a kernel $K : \mathbb{R} \mapsto \mathbb{R}$ and a bandwidth $h = h_n > 0$; see Silverman (1986) for an introduction to this estimator. By plugging in $\hat{\pi}$, we arrive at the estimator $\hat{\mu}(x;\theta_1) = \frac{\partial}{\partial x}\left[\sigma^2(x;\theta_1)\hat{\pi}(x)\right]/[2\hat{\pi}(x)]$. Given $\hat{\mu}(\cdot;\theta_1)$ and $\sigma^2(\cdot;\theta_1)$, we then propose to estimate θ_1 by pseudo-maximum-likelihood (PMLE), $\hat{\theta}_1 = \arg\max_{\theta_1\in\Theta_1} L_n\left(\hat{\mu}(\cdot;\theta_1), \sigma^2(\cdot;\theta_1)\right)$, where

$$L_n\left(\mu,\sigma^2\right) = \frac{1}{n}\sum_{i=1}^{n}\log p_\Delta\left(X_{i\Delta}|X_{(i-1)\Delta};\mu,\sigma^2\right). \tag{6}$$

We refer to $\hat{\theta}_1$ as a PMLE since the log-likelihood is not optimized over both unknown parameters, θ_1 and π; rather we use a fixed preliminary estimator of π, see the end of Section 4 for further discussion. Once $\hat{\theta}_1$ has been found, the obvious pointwise estimators of $\mu(x)$ and $\sigma^2(x)$ are $\hat{\mu}(x;\hat{\theta}_1)$ and $\sigma^2(x;\hat{\theta}_1)$ respectively; the asymptotic properties of these are derived in Kristensen (2008a).

The above estimation strategy is also applicable to models from Class 2. For a given parameterisation of $\mu(\cdot) = \mu(\cdot;\theta_2)$, substitute this into (4) together with $\hat{\pi}$, thereby obtaining an estimator of the unknown diffusion term, $\hat{\sigma}^2(x;\theta_2) = \frac{2}{\hat{\pi}(x)}\int_l^x \mu(y;\theta_2)\hat{\pi}(y)\,dy$.[1] This can now be plugged into $L_n\left(\mu,\sigma^2\right)$ together with $\mu(\cdot;\theta_2)$, and we again have a pseudo-likelihood function which can be maximized w.r.t. θ_2.

There is a variety of other estimating procedures in the literature for diffusion models based on other objective functions than the log-likelihood; see Aït-Sahalia and Hansen (2006) for an overview. But given that the transition density gives a complete description of the model, the PMLE is the most natural candidate.[2] In particular, we demonstrate that the only requirement that we need in order to identify θ_k for a semiparametric model in Class k, $k = 1,2$, is that the parametric submodel with π_0 known is identifiable through its log-likelihood. Since no other objective function carry more information about the parameter, it will not be able to identify θ_k unless the log-likelihood can, and so the PMLE is able to identify θ_k under the weakest possible conditions.

Aït-Sahalia (1996a) demonstrates that in the case of linear drift and unspecified diffusion, it is possible to estimate θ_2 without using information about the stationary density. For a general model in Class 2, he observes that

$$g_t(x;\theta_2,\pi) \equiv E_{\theta_2,\pi}[X_t|X_0 = x] = X_0 + \int_0^t E_{\theta_2,\pi}[\mu(X_s;\theta_2)|X_0 = x]\,ds, \tag{7}$$

where $E_{\theta_2,\pi}[\cdot|X_0 = x]$ is the conditional mean under the model specificed by $\mu(x;\theta_2)$ and π. In the linear case $\mu(x;\theta_2) = \theta_{2,1} + \theta_{2,2}x$, he obtains an analytical expression of g from (7) that does not depend on π. This enables him to estimate θ_2 by least squares of the corresponding nonlinear regression model, $X_{i\Delta} = X_{(i-1)\Delta} + g_\Delta\left(X_{(i-1)\Delta};\theta_2\right) + \varepsilon_{i\Delta}$, where $E_{\theta_2,\pi}\left[\varepsilon_{i\Delta}|X_{(i-1)\Delta}\right] = 0$. One could hope for that this estimation strategy would carry over to more general model, thereby avoiding having to rely on two-step estimators, but unfortunately this is not the case: The function g will in

[1] An alternative estimator would be $\hat{\sigma}^2(x;\theta) = \frac{2}{\hat{\pi}(x)}\sum_{i=1}^n \mathbb{I}\{X_i \le x\}\mu(X_i;\theta_2)/n$. This includes an unbiased estimator of the integral component.

[2] Alternatively, the L_2 objective function of Altissimo and Mele (2009) could be used since the resulting estimator is shown to be as efficient as the MLE.

general depend on both θ_2 and π, and the least-squares approach will also require an estimator of π. Similarly, for models in Class 1 in general, it does not seem possible to derive moment conditions that do not depend on π. In conclusion, the semiparametric estimation of the parameter in either class will in general require as input an estimator of π.

3 Asymptotic Properties of Estimators

In order to avoid repetitions, we present results for estimators in both classes joinly. Hopefully, this should create no confusion. Let $\theta_{0,k}$ and π_0 denote the true, data generating parameters for a model in Class k, $k = 1, 2$. Here and in the following we will often suppress a function's dependence on π when evaluated at $\pi = \pi_0$. For example, for the drift and diffusion term of the parametric submodel where π_0 is known, we will write

$$\text{Class 1} : \mu(x; \theta_1) = \mu(\cdot; \theta_1, \pi_0) = \frac{1}{2\pi_0(x)} \frac{\partial}{\partial x} \left[\sigma^2(x; \theta_1)\pi_0(x)\right], \tag{8}$$

$$\text{Class 2} : \sigma^2(x; \theta_2) = \sigma^2(\cdot; \theta_2, \pi_0) = \frac{2}{\pi_0(x)} \int_l^x \mu(y; \theta_2)\, \pi_0(y)\, dy. \tag{9}$$

Let $p_{t,k}(y|x; \theta_k) = p_t\left(y|x; \mu(\cdot; \theta_k), \sigma^2(\cdot; \theta_k)\right)$ denote the associated transition densities of the parametric submodel in Class k. For a given model in Class k, we impose the following conditions:

A1 The process $\{X_t\}$ is stationary and geometrically β-mixing, and has domain $I = \mathbb{R}$.

A2 The true density, π_0, is $m \geq 2$ times continuously differentiable on I with bounded derivatives, and satisfies $\int_I \pi_0(x)^{1-\delta_1}\, dx < \infty$ for some $\delta_1 > 0$.

A3 $\mu(x; \theta_k)$ and $\sigma^2(x; \theta_k)$ are twice differentiable in θ_k and satisfy for some function B,

$$\left\|\partial_\theta^i \mu(x; \theta_k)\right\| \leq B(x), \quad \left\|\partial_\theta^i \sigma^2(x; \theta_k)\right\| \leq B(x), \quad (x, \theta_k) \in I \times \Theta_k,$$

for $i = 0, 1, 2$, where $E[B^{2+\delta_2}(X_0)] < \infty$ for some $\delta_2 > 0$.

A4 (i) The transition density $p_{t,k}(y|x; \theta_k)$ exists as a solution to (2) and satisfies $\left|\partial_x^i p_{t,k}(y|x; \theta_k)\right| \leq \gamma_t(y|x)$, $(t, x, y, \theta_k) \in (0, \Delta] \times I^2 \times \Theta_k$, $i = 0, 1, 2$, where

$$\gamma_t(y|x) = c_1 \frac{|y|^{\lambda_1} + |x|^{\lambda_1}}{t^{\alpha_1}} \exp\left[-c_2 \frac{|y|^{\lambda_2} + |x|^{\lambda_2}}{t^{\alpha_2}}\right] \tag{10}$$

for constants c_j,α_j, $\lambda_j > 0$, $j = 1, 2$.

(ii) For some function q and constant $\delta_3 > 0$: $|\log(p_{\Delta,k}(y|x; \theta_k))| \leq q_k(y|x)$ for all $(x, y, \theta_k) \in I^2 \times \Theta_k$, and $E[q_k^{1+\delta_3}(X_\Delta|X_0)] < \infty$. The moment function $\theta_k \mapsto L_k(\theta_k) \equiv E[\log p_{\Delta,k}(X_\Delta|X_0; \theta_k)]$ has a unique maximum at $\theta_{k,0} \in \Theta_k$.

A5 The parameter space $\Theta_k \subset \mathbb{R}^{d_k}$ is compact.

For technical reasons, we strengthen the stationarity assumption to β-mixing in (A1), since this enables us to employ standard results for uniform convergence of kernel density estimators, and U-statistics, c.f. Hansen (2008) and Arcones (1995). Sufficient conditions for (A1) in terms of the drift and the diffusion term can be found in e.g. Meyn and Tweedie (1993) and Hansen and Scheinkman (1995); see also Karatzas and Shreve (1991, Section 5.5). Most parametric models found in the literature can be shown to satisfy (A1) under suitably restrictions on the parameters. The assumption that the domain is the whole real line is imposed to avoid any boundary issues. Since processes with domains different from $\mathbb{R}$ can be transformed to have domain $I = \mathbb{R}$, we do not find this very restrictive. If, for example, Y_t is a Markov diffusion process with domain $I_Y = (0, +\infty)$, then $X_t = \log(Y_t)$ is also a Markov diffusion by Itô's Lemma with domain $I = \mathbb{R}$.

The smoothness criteria on π_0 in (A2) together with the use of higher-order kernels defined below decrease the bias from the kernel estimation; the differentiability condition is met if the drift and the diffusion are $m-1$ and m times differentiable respectively. The integrability assumption rules out fat-tailed densities and is used to control for the effect of the trimming introduced below.

Assumption (A3) is used to ensure that relevant moments exist. These are fairly weak conditions and are satisfied by standard models used in the literature. Sufficient conditions can be found in Meyn and Tweedie (1993). The differentiability requirement can be disposed of when showing consistency, but is here maintained throughout for simplicity.

For general diffusion models, weak conditions for the existence of a transition density are unfortunately not available. A simple set of sufficient conditions for the existence of the transition density $p_t\left(y|x;\mu,\sigma^2\right)$ as the solution to the PDE in Eq. (2) can be found in e.g. Ilyin et al. (2002). These require however that the drift and diffusion functions are bounded which is very restrictive and violated by most standard models used in the literature; furthermore, the assumption of mixing in (A1) rules out bounded coefficients, c.f. Chen et al. (1999). We therefore impose the high level conditions in (A4) which for example is satisfied by the Vasicek and (the log-transformed) CIR model. Assumption (A4.i) could be replaced by alternative conditions such as the ones in Aït-Sahalia (2002) where a representation of p_t is established that does not rely on PDE's. This representation is more complicated however and proves to be very difficult to work with within our framework.[3] We therefore here use the representation of p_t as a solution to a PDE. This allows us to import and use the technology developed in Kristensen (2008) in the analysis of the estimated transition density with π_0 replaced by a kernel estimator.

Condition (A4.ii) together with (A5) imply that the MLE of the parametric submodel with π_0 known is consistent. In particular, the uniform bound q ensures that the log-likelihood converges uniformly towards $L_k(\theta_k)$ which in turn uniquely identifies $\theta_{k,0}$. Note that since θ_k enters both the drift and diffusion of the parametric submodels as defined in Eq. (8) and (9) respectively, the identification of $\theta_{k,0}$ is done jointly through the drift and the diffusion. This is in contrast to standard specifications of parametric diffusion models where the parameters entering the drift are separate from the ones entering the diffusion. Again, (A4.ii) is a high-level condition which is imposed since there does not seem to exist any general conditions in the literature in terms of

[3] See also a previous version of this paper, Kristensen (2004a).

$\mu(x;\theta_k)$ and $\sigma^2(x;\theta_k)$ for consistency to hold. We note however that (A4.ii) holds for standard parametric models such as the Vasicek (1977) and the CIR (1985) model.

Next, we impose the regularity conditions found in Hansen (2008) on the kernel K, and introduce a trimming function $\tau_a(\cdot)$:

K(m) The kernel $K:\mathbb{R}\mapsto\mathbb{R}$ is differentiable with

$$\left|K^{(i)}(z)\right|\leq C|z|^{-\eta}, \quad \left|K^{(i)}(z)-K^{(i)}(z')\right|\leq C|z-z'|, \quad i=0,1,$$

for some $\eta>0$, and is of order $m\geq 2$: $\int_{\mathbb{R}}K(z)\,dz=1$, $\int_{\mathbb{R}}z^iK(z)\,dz=0$, $i=1,...,m-1$, and $\int_{\mathbb{R}}|z|^mK(z)\,dz<\infty$.

T1 The trimming function $\tau_a:\mathbb{R}\mapsto[0,1]$, $a>0$, satisfies $\tau_a(z)=1$ for $z\geq a$ and $\tau_a(z)=0$ for $z\leq a/2$. It is twice continuously differentiable with $|\tau_a^{(i)}(z)|=O(a^i)$, $i=1,2$.

The kernel is here chosen to be of order m, where m matches up with the number of derivatives of π_0, c.f. assumption (A2). The trimming function will be used to control the tail behaviour of the estimators of the unspecified drift or diffusion term, but also the transition density itself. We here follow the idea of Andrews (1995) and Ai (1997) and use a smooth trimming function, $\tau_a(z)$. The differentiability of τ is assumed in order for the trimmed likelihood to be twice differentiable in the parameters. The speed with which the trimming parameter a goes to zero will be restricted, such that the trimming has no effect on the asymptotics. A simple way of constructing $\tau_a(z)$ is to choose a cdf F with support $[0,1]$, and define $\tau_a(z)=F((2z-a)/a)$ which then in great generality will satisfy (T1); see also Andrews (1995, p. 572).

We then use τ_a to define trimmed versions of the preliminary estimators. For a density $\pi(x)$, define

$$\text{Class 1}:\hat{\mu}(x;\theta_1,\pi)=\frac{\hat{\tau}_a(x)}{2\pi(x)}\frac{\partial}{\partial x}\left[\sigma^2(x;\theta_1)\pi(x)\right],\quad \hat{\sigma}^2(x;\theta_1)=\hat{\tau}_a(x)\sigma^2(x;\theta_1)+\underline{\sigma}^2(1-\hat{\tau}_a(x)), \tag{11}$$

$$\text{Class 2}:\hat{\mu}(x;\theta_2)=\hat{\tau}_a(x)\mu(x;\theta_2),\quad \hat{\sigma}^2(x;\theta_2,\pi)=\frac{2\hat{\tau}_a(x)}{\hat{\pi}(x)}\int_l^x\mu(y;\theta_2)\hat{\pi}(y)\,dy+\underline{\sigma}^2(1-\hat{\tau}_a(x)), \tag{12}$$

where $\hat{\tau}_a(x)=\tau_a(\hat{\pi}(x))$, $a=a_n>0$ is a trimming sequence and $\underline{\sigma}^2>0$ a constant. The inclusion of the additional term $\underline{\sigma}^2(1-\hat{\tau}_a(x))$ in the diffusion estimator guarantees that it is strictly positive for all $x\in I$ for n sufficiently large.[4] The motivation for the trimming is two-fold: First, the trimming of the nonparametric component is used to show that $\hat{\mu}(x;\theta_1,\hat{\pi})\to^P\tau_a(\pi_0(x))\mu(x;\theta_1)$ and $\hat{\sigma}^2(x;\theta_2,\hat{\pi})\to^P\tau_a(\pi_0(x))\sigma^2(x;\theta_2)$ uniformly over (x,θ_k), $k=1,2$, c.f. Lemmas 9-10. Second, the trimming of the parametric component is introduced to ensure that the associated transition density exists: Due to trimming, $\hat{\mu}$ and $\hat{\sigma}^2$ are bounded and $\hat{\sigma}^2>0$, and we can therefore apply Ilyin et al. (2002) to ensure that the associated diffusion process has a well-defined transition density.

[4] For models in Class 2 with small samples, one may want to use for example $\tilde{\sigma}^2(x;\theta)=\max\left\{\hat{\sigma}^2(x;\theta),a\right\}$ instead of $\hat{\sigma}^2(x;\theta)$ to ensure that the estimator is postive, since we can only guarantee that $\hat{\sigma}^2(x;\theta)>0$ almost surely as $n\to\infty$.

We also introduce a trimmed version of the log-likelihood $L_n\left(\mu,\sigma^2\right)$ given by

$$\hat{L}_n\left(\mu,\sigma^2\right)=\frac{1}{n}\sum_{i=1}^{n}\tau_{b,i}\left(\mu,\sigma^2\right)\log p_\Delta\left(X_{i\Delta}|X_{(i-1)\Delta};\mu,\sigma^2\right), \tag{13}$$

where $\tau_{b,i}\left(\mu,\sigma^2\right):=\tau_b\left(p_\Delta\left(X_{i\Delta}|X_{(i-1)\Delta};\mu,\sigma^2\right)\right)$ and $b=b_n\to 0$ is another trimming parameter. The trimming of $p_\Delta\left(y|x;\mu,\sigma^2\right)$ is needed for technical reasons in the proofs, and we conjecture that only trimming of the preliminary drift and diffusion estimators is required for our theoretical results. Plugging the trimmed drift and diffusion terms into the trimmed log-likelihood, we obtain a pseudo-log-likelihood function that depends on θ_k and π. Let

$$\hat{L}_{n,1}\left(\theta_1,\pi\right)=\hat{L}_n\left(\hat{\mu}\left(\cdot;\theta_1,\pi\right),\hat{\sigma}^2\left(\cdot;\theta\right)\right)\quad\text{and}\quad\hat{L}_{n,2}\left(\theta_2,\pi\right)=\hat{L}_n\left(\hat{\mu}\left(\cdot;\theta_2\right),\hat{\sigma}^2\left(\cdot;\theta_2,\pi\right)\right)$$

denote the trimmed log-likelihood for a given model in Class 1 and 2 respectively. We then define the PMLE of θ_k in class $k\in\{1,2\}$ as:

$$\hat{\theta}_k=\arg\max_{\theta_k\in\Theta_k}\ \hat{L}_{n,k}\left(\theta_k,\hat{\pi}\right). \tag{14}$$

We impose the following restrictions on the bandwidths and trimming parameters to obtain consistency and asymptotic normality of the PMLE:

B1.1 $nha^6b^2/\log(n)\to\infty$, $nh^3a^4b^2/\log(n)\to\infty$, $h^ma^{-3}b^{-1}\to 0$, $a^{\delta_1}b^{-2}\to 0$.

B2.1 $\sqrt{n}ha^6b^2/\log(n)\to\infty$, $\sqrt{n}h^3a^4b^2/\log(n)\to\infty$, $n^{1/4}h^ma^{-3}b^{-1}\to 0$, $a^{\delta_1}b^{-2}\to 0$.

B1.2 $nha^4b^2/\log(n)\to\infty$, $h^ma^{-2}b^{-1}\to 0$, $a^{\delta_1}b^{-2}\to 0$.

B2.2 $\sqrt{n}ha^4b^2/\log(n)\to\infty$, $n^{1/4}h^ma^{-2}b^{-1}\to 0$, $a^{\delta_1}b^{-2}\to 0$.

For a model in Class k, $k=1,2$, (B1.k) is imposed to show consistency, while asymptotic normality requires that (B2.k) holds; note that condition (B2.k) implies (B1.k). The rate with which h, a and b can go to zero depends on: The number of derivatives, $m\geq 2$, of π_0, and the measure of its "tail-thickness", $\delta_1>0$, as given in (A2). For values of δ_1 close to zero, the trimming parameters a and b have to go to zero at a fast rate and vice versa, while we in general have to use higher order kernels ($m>2$) to ensure that the bias and variance component of the estimator of the unspecified term disappears asymptotically. As an example, consider a model in Class 2: (B2.2) will then for example hold with $b=O\left(n^{-\varepsilon}\right)$ and $a=O\left(n^{-(3\varepsilon)/\delta_1}\right)$ for some $\varepsilon<\bar{\varepsilon}=1/\left(3+12\delta_1\right)$, while $h=O\left(n^{-(\varepsilon/\bar{\varepsilon}-1)/m}\right)$. In practice, the use of higher-order kernels does not have much of an effect on the resulting semiparametric estimator; this is confirmed in the simulation study where the Gaussian kernel was employed and yielded precise estimates of θ_k.

Theorem 1 *For a model in Class $k\in\{1,2\}$: Assume that (A1)-(A5) and (B1.k) hold. Then, the PMLE defined in Eq. (14) is consistent: $\hat{\theta}_k\to^P\theta_{k,0}$.*

Next, we analyze the asymptotic distribution of $\hat{\theta}_k$. To this end, we first introduce some additional notation: Let

$$\hat{p}_{\Delta,1}(y|x;\theta_1,\pi) = p_\Delta\left(y|x;\hat{\mu}(\cdot;\theta_1,\pi),\hat{\sigma}^2(\cdot;\theta_1)\right), \quad \hat{p}_{\Delta,2}(y|x;\theta_2,\pi) = p_\Delta\left(y|x;\hat{\mu}(\cdot;\theta_2),\hat{\sigma}^2(\cdot;\theta_2,\pi)\right) \tag{15}$$

denote the transition densities associated with the estimators in Class 1 and 2 respectively, and define the corresponding individual Score and Hessian as:

$$\hat{s}_{\Delta,k}(y|x;\theta_k,\pi) = \frac{\partial}{\partial\theta_k}\log\hat{p}_{\Delta,k}(y|x;\theta_k,\pi), \quad \hat{h}_{\Delta,k}(y|x;\theta_k,\pi) = \frac{\partial^2}{\partial\theta_k\partial\theta_k'}\log\hat{p}_{\Delta,k}(y|x;\theta_k,\pi)$$

Also, we introduce the individual Score and Hessian of the two corresponding parametric submodels:

$$s_{\Delta,k}(y|x;\theta_k) = \frac{\partial}{\partial\theta_k}\log p_{\Delta,k}(y|x;\theta_k), \quad h_{\Delta,k}(y|x;\theta_k) = \frac{\partial^2}{\partial\theta_k\partial\theta_k'}\log p_{\Delta,k}(y|x;\theta_k),$$

for $k = 1, 2$. For a given model in Class k, we can write the associated trimmed Score and Hessian as

$$\hat{S}_{n,k}(\theta_k,\pi) = \frac{1}{n}\sum_{i=1}^n \tau_{b,i}(\theta_k,\pi)\,\hat{s}_{\Delta,k}\left(X_{i\Delta}|X_{(i-1)\Delta};\theta_k,\pi\right) + o_P\left(n^{-1/2}\right),$$

$$\hat{H}_{n,k}(\theta_k,\pi) = \frac{1}{n}\sum_{i=1}^n \tau_{b,i}(\theta_k,\pi)\,\hat{h}_{\Delta,k}\left(X_{i\Delta}|X_{(i-1)\Delta};\theta_k,\pi\right) + o_P\left(n^{-1/2}\right),$$

where we have left out higher-order terms involving derivatives of $\tau_{b,i}(\theta_k,\pi)$.

To derive the asymptotic distribution, we follow the proof strategy outlined in Newey and McFadden (1994, Section 8). First, use a standard Taylor expansion argument of the first-order condition to write the PMLE in terms of the trimmed Score and Hessian,

$$0 = \hat{S}_{n,k}(\hat{\theta}_k,\hat{\pi}) = \hat{S}_{n,k}(\theta_{k,0},\hat{\pi}) + \hat{H}_{n,k}\left(\bar{\theta}_k,\hat{\pi}\right)(\hat{\theta}_k - \theta_{k,0}), \tag{16}$$

for some $\bar{\theta}_k \in [\theta_{k,0},\hat{\theta}_k]$. If π_0 was known and no trimming was employed, the first term on the right hand side would be equal to the score of the parametric submodel, $S_{n,k}(\theta_{k,0},\pi_0)$, and a standard Central Limit Theorem (CLT) could be employed to derive the asymptotic distribution. However, in general, $\hat{S}_{n,k}(\theta_{k,0},\hat{\pi})$ is not equivalent to $S_{n,k}(\theta_{k,0},\pi_0)$ and we have to take into account the use of $\hat{\pi}$ and trimming. To account for $\hat{\pi}$, we make a functional Taylor expansion of $\hat{S}_{n,k}(\theta_{k,0},\hat{\pi})$ around π_0. We define the pathwise derivative of $\hat{S}_{n,k}(\theta_{k,0},\pi)$ w.r.t. π in the directions $d\pi = \hat{\pi} - \pi_0$ as

$$\nabla\hat{S}_{n,k}[d\pi] = \frac{1}{n}\sum_{i=1}^n \tau_{b,i}(\theta_k,\pi) \bigtriangledown \hat{s}_{\Delta,k}\left(X_{i\Delta}|X_{(i-1)\Delta}\right)[d\pi], \tag{17}$$

where $\bigtriangledown\hat{s}_{\Delta,k}(y|x)[d\pi]$ is the pathwise derivative of $\hat{s}_{\Delta,k}(y|x;\theta_{0,k},\pi)$ (see Appendix B for its expression). We then show that the following first order expansion is valid:

$$\hat{S}_{n,k}(\theta_{k,0},\hat{\pi}) = \hat{S}_{n,k}(\theta_{k,0},\pi_0) + \nabla\hat{S}_{n,k}[\hat{\pi} - \pi_0] + o_P\left(n^{-1/2}\right). \tag{18}$$

Finally, with the trimming parameters vanishing sufficiently fast, the trimmed and untrimmed versions of the score and its pathwise derivative are asymptotically equivalent,

$$\hat{S}_{n,k}\left(\theta_{k,0},\pi_0\right)+\nabla\hat{S}_{n,k}\left[\hat{\pi}-\pi_0\right]=S_{n,k}\left(\theta_{k,0},\pi_0\right)+\nabla S_{n,k}\left[\hat{\pi}-\pi_0\right]+o_P\left(n^{-1/2}\right), \tag{19}$$

where $\nabla S_{n,k}\left[d\pi\right]$ denotes the untrimmed version of $\nabla\hat{S}_{n,k}\left[\hat{\pi}-\pi_0\right]$.

From the above expressions, we see that the pathwise derivative will account for additional statistical errors due to the use of $\hat{\pi}$ instead of π_0. The final part of the proof analyzes the behaviour of $\nabla S_{n,k}\left[\hat{\pi}-\pi_0\right]$. We show that $S_{n,k}\left[\hat{\pi}-\pi_0\right]$ can be written as a second order U-statistic:

$$\nabla S_{n,k}\left[\hat{\pi}-\pi_0\right]=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}s_{\Delta,k}\left(X_{i\Delta}|X_{(i-1)\Delta};\theta_{k,0}\right)\frac{d_k\left(X_{j\Delta},X_{i\Delta},X_{(i-1)\Delta}\right)}{p_{\Delta,k}\left(X_{i\Delta}|X_{(i-1)\Delta};\theta_{k,0}\right)}+o_P(n^{-1/2}), \tag{20}$$

where the functions d_k, $k=1,2$, take the following forms:

$$d_1\left(z,y,x\right)=-\frac{1}{2\pi_0\left(z\right)}\int_0^{\Delta}\frac{\partial}{\partial z}\left[\sigma^2\left(z;\theta_{1,0}\right)\frac{\partial p_{t,1}\left(y|z;\theta_{1,0}\right)}{\partial z}p_{t,1}\left(z|x;\theta_{1,0}\right)\right]dt, \tag{21}$$

$$\begin{aligned} d_2\left(z,y,x\right) &= 2\mu\left(z;\theta_{2,0}\right)\int_0^{\Delta}\int_l^z\frac{1}{\pi_0\left(w\right)}\frac{\partial^2 p_{t,2}\left(y|w;\theta_{2,0}\right)}{\partial w^2}p_{t,2}\left(w|x;\theta_{2,0}\right)dwdt \\ &\quad -2\frac{\sigma^2\left(z;\theta_{2,0}\right)}{\pi_0^2\left(z\right)}\int_0^{\Delta}\frac{\partial^2 p_{t,2}\left(y|z;\theta_{2,0}\right)}{\partial z^2}p_{t,2}\left(z|x;\theta_{2,0}\right)dt. \end{aligned} \tag{22}$$

In both cases, one can check that d_k satisfies $E\left[d_k\left(X_0,x,y\right)\right]=0$ for all (x,y). Thus, using standard U-statistics arguments,

$$\nabla S_{n,k}\left[\hat{\pi}-\pi_0\right]=\frac{1}{n}\sum_{i=1}^{n}D_k\left(X_{i\Delta}\right)+o_P(n^{-1/2}), \tag{23}$$

where

$$D_k\left(x\right)=E\left[s_{\Delta,k}\left(X_1|X_0;\theta_{k.0}\right)\frac{d_k\left(x,X_{i\Delta},X_{(i-1)\Delta}\right)}{p_{\Delta,k}\left(X_{i\Delta},X_{(i-1)\Delta};\theta_{k,0}\right)}\right] \tag{24}$$

satisfies $E\left[D_k\left(X_0\right)\right]=0$ and $E[\|D_k\left(X_0\right)\|^2]<\infty$, $k=1,2$. Combining Eqs. (16), (18) and (23) with $\hat{H}_{n,k}\left(\bar{\theta}_k,\hat{\pi}\right)\to^P-\mathcal{I}_k$, where $\mathcal{I}_k$ is the information of the parametric submodel (see Assumption A6 below),

$$\sqrt{n}(\hat{\theta}_k-\theta_{k,0})=\mathcal{I}_k^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{s_{\Delta,k}\left(X_{i\Delta}|X_{(i-1)\Delta};\theta_{k,0}\right)+D_k\left(X_{i\Delta}\right)\right\}+o_P\left(n^{-1/2}\right), \tag{25}$$

and the limiting distribution now follows by a CLT for mixing sequences.

We impose three additional assumptions to formally establish the above claims. Again, note that the following conditions cover both models in Class 1 and 2:

A6 $\theta_{k,0} \in \text{int}(\Theta_k)$. The Hessian of the parametric submodel satisfies $\|h_{\Delta,k}(y|x;\theta_k)\| \leq \bar{h}_{\Delta,k}(y|x)$ for θ_k in a neighbourhood of $\theta_{k,0}$, where $E\left[\bar{h}_{\Delta,k}(X_\Delta|X_0)\right] < \infty$. The associated information is positive definite,

$$\mathcal{I}_k \equiv E\left[s_{\Delta,k}(X_\Delta|X_0;\theta_{k,0})\, s_{\Delta,k}(X_\Delta|X_0;\theta_{k,0})'\right] > 0. \tag{26}$$

A7 Eq. (19) holds under (B2.k).

A8 The following moment exists:

$$\int_I E\left[\frac{d_k^2(x, X_\Delta|X_0)}{p_{\Delta,k}^2(X_\Delta|X_0;\theta_{k,0})}\right]\pi_0(x)\,dx < \infty.$$

Assumption (A6) ensures that the MLE of the parametric submodel is asymptotically normally distributed; a sufficient set of conditions for (A6) in terms of $\mu(x;\theta_k)$ and $\sigma^2(x;\theta_k)$ can be found in Aït-Sahalia (2002). The positive definiteness of $\mathcal{I}_k$ will normally follow under the identification condition given in (A4).

Assumption (A7) is a high-level condition implying that the trimmed score and pathwise derivative are asymptotically equivalent to their untrimmed versions. It would of course be more satisfactory if we could state primitive conditions for (A7) to hold, but due to the complexity of the model and estimator, we have not been able to do so. By the same arguments as in the proofs of Lemmas 6 and 8, we can establish a bound for the left hand side of Eq. (19) which in turn is of order $O_P\left(b^{-1}a^{\delta_1/2}\right) + O_P(|\log b|^{-\delta_3})$. Thus, Eq. (19) would hold if (*) $\sqrt{n}b^{-1}a^{\delta_1/2} \to 0$ and $\sqrt{n}\,|\log b|^{-\delta_3} \to 0$. However, these two requirements collide with the other restrictions imposed on a and b in (B.2.k): No feasible trimming sequences exist that simultaneously satisfy (B.2.k) and (*). We believe that this is due to our bound not being sharp enough. There are two main problems in deriving a sharp enough bound: First, the trimmed score is not a martingale since the "outer" trimming term, $\tau_{b,i}$, is a function of $X_{i\Delta}$. This is in contrast to most other semiparametric estimation problems where trimming can be designed such that the martingale property is maintained; see e.g. Robinson (1989) and Ai (1997). Second, the trimmed versions, $\hat{\mu}$ and $\hat{\sigma},^2$ enter $p_\Delta\left(y|x;\mu,\sigma^2\right)$ in a complex manner, c.f. Eq. (2), thereby making it difficult to analyze the impact of the "inner" trimming of $\hat{L}_{n,k}(\theta_k,\pi)$. An alternative to (A7) would be to impose parametric restrictions on the tail behaviour of the stationary density. By imposing this knowledge in the estimation, we could avoid the "inner" trimming. We give a further discussion of the case where (A7) fails to hold below.

Assumption (A8) implies that $E[\|D_k(X_0)\|^2] < \infty$, $k = 1, 2$. To give some further insight into what is required for $E\left[\|D_k(X_0)\|^2\right] < \infty$, consider a fully parametric submodel where either the diffusion (Class 1) or the drift (Class 2) is specified up to a parameter θ_k together with the following specification of the marginal density: For given $h > 0$ and $x_0 \in I$, define $\pi_{h,x_0}(x;\alpha) = (1-\alpha)\pi_0(x) + \alpha K_h(x - x_0)$, where $\alpha \in [0,1]$ is the unknown parameter. One can then show that

$D_k(x_0) = \lim_{h \to 0} \mathcal{I}_{k,h}(x_0)$ with $\mathcal{I}_{k,h}(x_0)$ being the information of this parametric submodel:

$$\mathcal{I}_{k,h}(x_0) = -E\left[\frac{\partial^2 \log p_{\Delta,k}(X_\Delta | X_0; \theta_k, \pi_{h,x_0}(\cdot; \alpha))}{\partial \theta_k \partial \alpha}\right]_{\theta_k = \theta_{k,0}, \alpha = 0}. \tag{27}$$

So $E[\|D_k(X_0)\|^2] < \infty$ translates into $\lim_{h \to 0} \int_I \|\mathcal{I}_{k,h}(x_0)\|^2 \pi_0(x_0) \, dx_0 < \infty$.

Theorem 2 *For a model in Class $k \in \{1, 2\}$: Assume that (A1)-(A8) and (B2.k) hold. Then,*

$$\sqrt{n}(\hat{\theta}_k - \theta_{k,0}) \xrightarrow{d} N\left(0, \mathcal{I}_k^{-1} \Sigma_{k,\infty} \mathcal{I}_k^{-1}\right),$$

where $\mathcal{I}_0$ is given in Eq. (26) and $\Sigma_{k,\infty} = \Sigma_{k,0} + \sum_{i=1}^{\infty} \left(\Sigma_{k,i} + \Sigma'_{k,i}\right)$ with

$$\Sigma_{k,i} = E\left[\left\{s_{\Delta,k}(X_\Delta | X_0; \theta_{k,0}) + D_k(X_0)\right\}\left\{s_{\Delta,k}\left(X_{(i+1)\Delta} | X_{i\Delta}; \theta_{k,0}\right) + D_k(X_{i\Delta})\right\}'\right], \quad i \geq 0,$$

and $D_k(x)$ as specified in Eq. (24).

The asymptotic variance of $\hat{\theta}_k$ consists of two components, $\mathcal{I}_k$ and $\Sigma_{k,\infty}$. The covariance matrix $\Sigma_{k,\infty}$ is the sum of the score of the parametric submodel and the adjustment term D_k. In particular, if $D_k = 0$, then $\Sigma_{k,\infty} = \mathcal{I}_k$. This would for example happen if π_0 was known, in which case we could estimate the parametric submodel by MLE, which, under assumptions (A1)-(A6), has asymptotic variance $\mathcal{I}_k^{-1}$. However, in general $D_k \neq 0$ due to the use of $\hat{\pi}$ instead of π_0 in the estimation. So we here pay a price in terms of variance, $\mathcal{I}_k^{-1} \Sigma_{k,\infty} \mathcal{I}_k^{-1} > \mathcal{I}_k^{-1}$, for the lack of information about π_0. In particular, the PMLE is not adaptive since it does not reach the Cramer-Rao bound of the parametric submodel (see Newey, 1989).

Not all semiparametric estimators share this property. Rather, in many cases, their asymptotic variances are equal to the ones of the infeasible estimators of the parametric submodels; see for example, Ai (1997), Robinson (1988) and Ichimura (1993). This is due to the fact that for those estimators, the parametric and nonparametric estimators are asymptotically orthogonal implying that the functional derivative of the first-order condition is $o_P(1/\sqrt{n})$. In contrast, in our case the functional derivative, $\triangledown S_{n,k}$, adds the additional term D_k to the asymptotic variance. For a further discussion of orthogonality conditions in semiparametric problems, we refer to Andrews (1994) and Newey (1989), and note that Andrews (1994) explicitly imposes such a condition in his Assumption N(c). Thus, our estimation problem is situated outside of his framework.

However, our estimator is not the only one that suffers from this drawback. A simple example is the index-model estimator of Powell, Stock and Stoker (1989) which has an additional variance component due to the presence of a first-step nonparametric estimator. Another example, which is more closely related to our problem, is semiparametric copula estimation. In this class of models, a parametric structure is imposed on the so-called copula while the marginal densities are left unspecified. If one estimates the marginal densities using kernel methods, one can show along the same lines as in Genes Ghoudi and Rivest (1995) that the asymptotic variance of the parametric estimator has the same structure as in our case.

Eventhough the PMLE is not adaptive, one could still hope for it be semiparametrically efficient. However, we conjecture that this is not the case since our nonparametric estimator of π_0 does not adapt to the specific parameterisation of the drift or diffusion term. A formal proof of this conjecture would require deriving the semiparametric efficiency bound. Unfortunately, we have not been able to do so due to the complexity of the model. However, if this conjecture is true, an obvious next step would be to develop an efficient estimator. We expect that one such estimator can be obtained by the method of sieves: Assume that π_0 belongs to some normed function space $\mathcal{F}$, and that there exists a sieve space $\mathcal{F}_n$ such that, loosely speaking, $\mathcal{F}_n \to \mathcal{F}$; see Chen (2007) for a more precise introduction to these concepts. We could then estimate (θ_k, π) in Class k by nonparametric maximum-likelihood,

$$(\hat{\theta}_{\mathrm{MLE},k}, \hat{\pi}_{\mathrm{MLE},k}) = \arg \max_{(\theta_k,\pi)\in\Theta_k\times\mathcal{F}_n} L_{n,k}(\theta_k, \pi). \tag{28}$$

The difference between the sieve MLE in Eq. (28) and the PMLE given in Eq. (14) is that the sieve MLE estimates θ_k and π jointly. Thus, we expect that $\hat{\pi}_{\mathrm{MLE},k}$ will adapt to the specification of the parametric component and therefore $\hat{\theta}_{\mathrm{MLE},k}$ will reach the semiparametric efficiency bound. Returning to the related example of semiparametric copula models discussed earlier, Chen, Fan and Tsyrennikov (2006) demonstrate that the sieve MLE indeed is semiparametrically efficient for this particular model, while the PMLE in general is not.[5]

While the sieve MLE therefore most likely enjoys better asymptotic properties compared to the PMLE, it is more difficult to implement, since it requires numerical optimization over both Θ_k and $\mathcal{F}_n$, where the dimension of $\mathcal{F}_n$ will be "large". Thus, the PMLE is a computationally attractive alternative despite its conjectured lack of efficiency.

An important assumption for Theorem 2 to hold is (A7). If the effect from trimming does not vanish, $\hat{\theta}_k$ will still be $\sqrt{n}$-asymptotically normally distributed, but will suffer from asymptotic biases. The bias can be evaluated in terms of a trimmed version of the parametric submodel associated with a model in Class k,

$$\bar{\mu}_a(x;\theta_k) = \tau_a(\pi_0(x))\,\mu(x;\theta_k), \quad \bar{\sigma}_a^2(x;\theta_k) = \tau_a(\pi_0(x))\,\sigma^2(x;\theta_k) + \underline{\sigma}^2(1-\tau_a(\pi_0(x))), \tag{29}$$

where $\mu(x;\theta_k)$ and $\sigma^2(x;\theta_k)$ are given in Eqs. (8)-(9). Let $p_{\Delta,k}^{(a)}(y|x;\theta_k) = p_{\Delta,k}\left(y|x;\bar{\mu}_a(\cdot;\theta_k), \bar{\sigma}_a^2(x;\theta_k)\right)$ be the corresponding transition density, and

$$\theta_k^{(a,b)} = \arg\max_{\theta_k\in\Theta_k} E[\tau_b(p_{\Delta,k}^{(a)}(X_\Delta|X_0;\theta_k))\log p_{\Delta,k}^{(a)}(X_\Delta|X_0;\theta_k)]$$

the parameter that maximizes the population version of the trimmed log-likelihood. The proofs of Theorems 1 and 2 still go through with $\theta_k^{(a,b)}$ replacing $\theta_{k,0}$ and we are able to conclude that $\sqrt{n}(\hat{\theta}_k - \theta_k^{(a,b)}) \to^d N(0, (\mathcal{I}_k^{(a,b)})^{-1}\Sigma_{k,\infty}^{(a,b)}(\mathcal{I}_k^{(a,b)})^{-1})$ where the asymptotic variance terms are trimmed versions of the ones appearing in Theorem 2. So if (A7) fails to hold, the PMLE will have contain

[5] In the case of Gaussian marginals, Klassen and Wellner (1997) show that the PMLE does reach the semiparametric efficiency bound.

an asymptotic bias component, $\theta_k^{(a,b)} - \theta_{k,0}$. In our simulation study, we find that the bias of the PMLE is small and comparable to the one of the parametric MLE, so we believe that Theorem 2 provides a valid asymptotic approximation.

The asymptotic variance of $\hat{\theta}_k$ can be estimated in the following manner: The component $\mathcal{I}_k$ is straightforwardly estimated by $-H_{n,k}(\hat{\theta}_k, \hat{\pi})$, c.f. Proof of Theorem 2, while the estimation of $\Sigma_{k,\infty}$ is somewhat more complicated. The functions $D_k(x)$, $k = 1, 2$, can be estimated consistently by replacing population moments with sample averages and unknown terms with estimators in Eqs. (24)-(22). However, this involves numerical evaluation of integrals of the type $\int_0^\Delta \partial^i p_t(y|z) / (\partial z^i)\, p_t(z|x)\, dt$, $i = 1, 2$, which can be rather difficult since $\lim_{t\to0} p_t(z|x) = \delta(z - x)$, where δ is the Dirac delta function. A more attractive estimator is obtained by following the idea of Newey (1994), and utilize the characterization of $D_k(x)$ given in Eq. (27). Newey (1994) state regularity conditions under which

$$\hat{D}_k(x_0) = \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^2 \log p_{\Delta,k}(X_{i\Delta}|X_{(i-1)\Delta};\theta_k, \hat{\pi}_{h,x_0}(\cdot;\alpha)))}{\partial\theta_k\partial\alpha}\right|_{\theta_k=\hat{\theta}_k,\alpha=0},$$

where $\hat{\pi}_{h,x_0}(x;\alpha) = (1-\alpha)\hat{\pi}(x) + \alpha K_h(x - x_0)$, is a consistent estimators of $D_k(x)$. One advantage of this estimator is its relative simple implementation; one can calculate it by numerical differentiation of the log-transition density w.r.t. θ_k and α.

Once an estimator of $\hat{D}_k(z)$ has been obtained, we can define for $i \geq 0$,

$$\hat{\Sigma}_{k,i} = \frac{1}{n-i}\sum_{j=1}^{n-i}\left\{\hat{s}_{k,j} + \hat{D}_{k,j}\right\}\left\{\hat{s}_{k,j} + \hat{D}_{k,j}\right\}',$$

where $\hat{s}_{k,j} = s(X_{(j+1)\Delta}|X_{j\Delta};\hat{\theta}_k,\hat{\pi})$ and $\hat{D}_{k,j} = \hat{D}_k(X_{j\Delta})$. These can in turn be used to construct an estimator of the HAC variance $\Sigma_{k,\infty}$, see e.g. Robinson and Velasco (1997) for an overview. One specific HAC estimator is the Newey and West (1987) one given by

$$\hat{\Sigma}_{k,\infty} = \hat{\Sigma}_{k,0} + \sum_{i=1}^{M} w_{M,i}(\hat{\Sigma}_{k,i} + \hat{\Sigma}_{k,i}), \quad w_{M,i} = 1 - [i/(M+1)].$$

As $M \to \infty$ and $M/m^{1/8} \to 0$, this will yield a consistent estimator of $\Sigma_{k,\infty}$.

4 Specification Testing

The semiparametric models proposed here relax some of the restrictions implied by a fully parametric diffusion model. They still impose some restrictions on the data generating process however, and it therefore desirable to be able to test the semiparametric specification against a nonparametric alternative. Once a semiparametric model is accepted by data, a natural question is what fully parametric specifications are consistent with the semiparametric model. Again, a formal test for the fully parametric specification against the semiparametric alternative is needed. We here construct two sets of test statistics which allow us to perform these tasks. The two sets of tests together supply us with a two-step strategy for testing a fully parametric model: First, one can test the correct specification of either the drift or diffusion term. If this is accepted, one can then test for the correct specification of the remaining term in the second step. This is in contrast to existing specification tests for diffusion models that only allow joint testing of the correct specification of both the drift and diffusion term.

We first consider the testing of a given semiparametric model against a nonparametric alternative: The relevant hypotheses for models in Class 1 and 2 respectively are:

$$H_{\mathrm{SP},1}: dX_t = \mu(X_t)\,dt + \sigma(X_t;\theta_{1,0})\,dW_t \text{ for some } \theta_{1,0} \in \Theta_1,$$

$$H_{\mathrm{SP},2}: dX_t = \mu(X_t;\theta_{2,0})\,dt + \sigma(X_t)\,dW_t \text{ for some } \theta_{2,0} \in \Theta_2.$$

We wish to test either of these null hypotheses against a nonparametric alternative which we here specify as:

$$H_{\mathrm{NP}}: \{X_t\} \text{ is a Markov process with transition density } p_t(y|x).$$

Implicitly, we are therefore jointly testing the assumption that (i) $\{X_t\}$ is a diffusion process and (ii) the parametric specification of either the drift or diffusion (depending on whether we are considering $H_{\mathrm{SP},1}$ or $H_{\mathrm{SP},2}$). The hypothesis (i) could be pretested using the tests developed in Kanaya (2008) and Florens et al (1998) before proceeding to the test proposed here. Under $H_{\mathrm{SP},k}$, we can estimate the transition density semiparametrically by $\hat{p}_{\mathrm{sp},k}(y|x) = \hat{p}_{\Delta,k}(y|x;\hat{\theta}_k,\pi)$ as given in Eq. (15), while under the maintained assumption H_{NP}, it can be estimated nonparametrically by

$$\hat{p}_{\mathrm{np}}(y|x) = \frac{\sum_{i=1}^{n} K_{h_{\mathrm{np}}}(X_{i\Delta} - y)\, K_{h_{\mathrm{np}}}\left(X_{(i-1)\Delta} - x\right)}{\sum_{i=1}^{n} K_{h_{\mathrm{np}}}\left(X_{(i-1)\Delta} - x\right)},$$

for a (different) bandwidth $h_{\mathrm{np}} > 0$. We then propose to test $H_{\mathrm{SP},k}$ against H_{NP} by comparing the two transition density estimates: Let $T_k = d(\hat{p}_{\mathrm{np}}, \hat{p}_{\mathrm{sp},k})$, $k = 1,2$, for some distance function $d(\cdot,\cdot)$ where three specific choices of d are:

$$\text{Kullback-Leibler (KL):} \quad d_{\mathrm{KL}}(p,p_0) = \int_{I^2} \log\left(\frac{p_0(y|x)}{p(y|x)}\right) p_0(y|x)\,\pi_0(x)\,dydx;$$

$$\text{Pearson Chi-square (PC):} \quad d_{\mathrm{PC}}(p,p_0) = \int_{I^2} \left[\frac{p_0(y|x) - p(y|x)}{p_0(y|x)}\right]^2 p_0(y|x)\,\pi_0(x)\,dydx;$$

$$L_2: \quad d_2(p,p_0) = \int_{I^2} [p_0(y|x) - p(y|x)]^2\, w(x,y)\,dydx.$$

Here, p and p_0 are two transition densities, and w is a weighting function. Tests of nonparametric hypotheses using the KL-distance are discussed in detail in Robinson (1991) where a modified version of the KL distance is used to test for independence. There are however a number of difficulties involved using the KL-distance as pointed out in Robinson (1991) so we'll here focus on the two other distances. Tests of fully parametric specification testing of (jump-)diffusion models using the modified KL and the L_2-distance have been proposed in Aït-Sahalia et al. (2009) where it is shown that $T_{\mathrm{PC}} = d_{\mathrm{P}}(\hat{p}_{\mathrm{np}}, \hat{p}_{\mathrm{fp}})$ and $T_2 = d_2(\hat{p}_{\mathrm{np}}, \hat{p}_{\mathrm{fp}})$, where $\hat{p}_{\mathrm{fp}}$ is the estimated transition density of the fully parametric model, follows a standard normal distribution when suitably normalized. The proofs of these results proceed in two steps: (i) Show T_{PC} is asymptotically equivalent to $T'_{\mathrm{PC}} = d_{\mathrm{PC}}(\hat{p}_{\mathrm{np}}, p)$ where p is the true transition density and (ii) Derive the asymptotic distribution of T'_{PC}; the same strategy is employed for T_2. Thus, the asymptotic results of Aït-Sahalia et al. (2009) carry over to our setting if we can show that

$$T_{\mathrm{PC},k} = d_{\mathrm{PC}}(\hat{p}_{\mathrm{np}}, \hat{p}_{\mathrm{sp},k}), \quad T_{2,k} = d_2(\hat{p}_{\mathrm{np}}, \hat{p}_{\mathrm{sp},k}),$$

are equivalent to T'_{PC} and $T'_2 = d_2(\hat{p}_{\mathrm{np}}, p)$ respectively. Since, as demonstrated in the Appendix, $\hat{p}_{\mathrm{sp},k}$ have faster convergence rate than $\hat{p}_{\mathrm{np}}$, this result follows along the exact same lines as in Aït-Sahalia et al. (2009) and we obtain under (A1)-(A7), (B1.k)-(B2.k) and the regularity conditions in Aït-Sahalia et al. (2009):

$$\frac{T_{\mathrm{PC},k} - \mu_{\mathrm{PC}}}{\sigma_{\mathrm{PC}}} \to^d N(0,1), \quad \frac{T_{2,k} - \mu_2}{\sigma_2} \to^d N(0,1),$$

for $k = 1, 2$, where expressions of μ_{PC}, σ_{PC}, μ_2 and σ_2 can be found in Aït-Sahalia et al. (2009).

Next, we wish to test a fully parametric specification,

$$H_{\mathrm{P}} : dX_t = \mu(X_t; \theta_{2,0})\, dt + \sigma(X_t; \theta_{1,0})\, dW_t \text{ for some } (\theta_{1,0}, \theta_{2,0}) \in \Theta_1 \times \Theta_2,$$

against either of the semiparametric alternatives. Tests for H_{P} against $H_{\mathrm{SP},k}$ based on a nonparametric estimator of the unspecified term under $H_{\mathrm{SP},k}$ have been developed in Kristensen (2008a). We here propose tests which only rely on estimators of the parametric component which has the advantage over the tests in Kristensen (2008a) that they do not rely on additional bandwidth sequences. On the other hand, this also implies that the test proposed here may have lower power and fail to reject certain alternatives as discussed below.

To compare the fully parametric model with its semiparametric alternative, we first need to estimate the model under H_{P}. A natural estimator of the parameters $\theta_0 = (\theta'_{1,0}, \theta'_{2,0})'$ under the null is the MLE, but other estimation methods are available, and we will assume that:

A9 The estimators $(\tilde{\theta}_1, \tilde{\theta}_2)$ of the fully parametric model satisfy for some $(\bar{\theta}_1, \bar{\theta}_2) \in \Theta_1 \times \Theta_2$:

$$\begin{pmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \end{pmatrix} = \begin{pmatrix} \bar{\theta}_1 \\ \bar{\theta}_1 \end{pmatrix} + \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \psi_{1,i} \\ \frac{1}{n}\sum_{i=1}^n \psi_{2,i} \end{pmatrix} + o_P(1/\sqrt{n}),$$

where $\psi_i = (\psi_{1,i}, \psi_{2,i}) = \psi(X_{i\Delta}, X_{(i)-1\Delta})$ satisfies $E[\psi_i] = 0$ and $E[\|\psi_i\|^{2+\delta}] < \infty$.

Assumption (A9) is satisfied for many different estimators in great generality. Note that we here allow for misspecification such that (A9) remains valid even if H_{P} is false. If H_{P} is true, then $(\bar{\theta}_1, \bar{\theta}_2) = (\theta_{1,0}, \theta_{2,0})$ while if H_{P} is false we expect $(\bar{\theta}_1, \bar{\theta}_2) \neq (\theta_{1,0}, \theta_{2,0})$. A natural way to test H_{P} against either $H_{\mathrm{SP},1}$ or $H_{\mathrm{SP},2}$ would be to adopt the same approach as in the test of the semiparametric models, and compare the transition densities under the two hypotheses, for example through a (pseudo-)likelihood-ratio statistic. Deriving the asymptotic distribution of the resulting test statistic is complicated though since (i) the parametric and semiparametric estimator of the transition density both converge with $\sqrt{n}$-rate and (ii) the joint distribution of the semiparametric estimator, and the fully parametric ones is difficult to derive. Instead, we follow Whang and Andrews (1993) and propose to test H_{P} against $H_{\mathrm{SP},k}$ using a Lagrange Multiplier (LM)-type test statistic:

$$T_{\mathrm{LM},k} = n\hat{S}_{n,k}(\tilde{\theta}_k, \hat{\pi})' \hat{V}_{n,k}^{-1} \hat{S}_{n,k}(\tilde{\theta}_k, \hat{\pi}), \tag{30}$$

where $\hat{S}_{n,k}(\theta_k, \pi)$ was introduced in Section 3 as the (trimmed) semiparametric score in Class k, while $\hat{V}_{n,k}$ is an estimator of its variance, $k = 1, 2$. The asymptotic variance takes the form $V_k = V_{k,0} + \sum_{i=1}^{\infty} \left(V_{k,i} + V_{k,i}' \right)$, where

$$V_{k,i} = [I_d, I_d]\, E \begin{bmatrix} (s_{k,0} + D_{k,0})(s_{k,i} + D_{k,i})' & \psi_{k,0}(s_{k,i} + D_{k,i})' \\ \psi_{k,0}(s_{k,i} + D_{k,i})' & \psi_{k,i}\psi_{k,0}' \end{bmatrix} [I_d, I_d]',$$

with $s_{k,i} = s_{\Delta,1}\left(X_{i\Delta} | X_{(i-1)\Delta}; \theta_{k,0} \right)$, $D_{k,i} = D_k\left(X_{(i-1)\Delta} \right)$, and I_d is the $(d \times d)$ identity matrix.

The motivation for $T_{\mathrm{LM},k}$ is that $E\left[\partial \log p_{\Delta,k}(X_\Delta | X_0; \theta_k, \pi_0) / \partial\theta_k \right] = 0$ if and only if $\theta_k = \theta_{k,0}$. Thus, under H_{P}, $\tilde{\theta}_k \to^P \bar{\theta}_k = \theta_{k,0}$ and it follows that $\sqrt{n}\hat{S}_{n,k}(\tilde{\theta}_k, \hat{\pi}) \to^d N(0, V_k)$. If however the null is false, we expect $\bar{\theta}_1 \neq \theta_{k,0}$ and $||\sqrt{n}\hat{S}_{n,k}(\tilde{\theta}_k, \hat{\pi})|| \to^P +\infty$. Further insight can be obtained through the following representation of $T_{\mathrm{LM},k}$ as a Hausmann-type test:

$$T_{\mathrm{LM},k} = n(\tilde{\theta}_k - \hat{\theta}_k)' \hat{W}_{n,k}^{-1} (\tilde{\theta}_k - \hat{\theta}_k) \tag{31}$$

where $\hat{W}_{n,k} = \hat{H}_{n.k}(\hat{\theta}_k, \hat{\pi}) \hat{V}_{n,k} \hat{H}_{n.k}(\hat{\theta}_k, \hat{\pi})$. Thus, we will reject if the fully parametric and semiparametric estimator of the parametric component are (statistically) significantly different from each other. This however also entails that the test may fail to reject a misspecified model in some cases: Consider for example testing H_{P} against $H_{\mathrm{SP},k}$ in which case we compare the semiparametric and fully parametric estimator of the diffusion parameter, α, using $T_{\mathrm{LM},1}$. If the drift and diffusion estimators obtained under the null are orthogonal to each other, such that $\tilde{\theta}_1$ is unaffected by the misspecification of $\mu(x; \theta_2)$, we expect that $\bar{\theta}_1 = \theta_{1,0}$ such that $\sqrt{n}(\tilde{\theta}_1 - \hat{\theta}_1) = O_P(1)$ under the alternative. If this is the case, $T_{\mathrm{LM},1}$ may fail to detect the misspecification. As such, $T_{\mathrm{LM},1}$ and $T_{\mathrm{LM},2}$ may have less power than the ones developed in Kristensen (2008a).

Theorem 3 *Under (A.1)-(A.9), (B2.k) and* H_{P}, $T_{\mathrm{LM},k} \to^d \chi^2_{d_k}$, $k \in \{1, 2\}$.

If (A7) does not hold, we will need to ensure that the fully parametric estimators are centered at the same parameter value as the semiparametric one. This can be done by choosing $(\tilde{\theta}_1, \tilde{\theta}_2) = \arg\max_{\theta \in \Theta} \hat{L}_n(\bar{\mu}_a(\cdot; \theta_2), \bar{\sigma}_a(\cdot; \theta_1))$, where $\bar{\mu}_a$ and $\bar{\sigma}_a$ are defined in Eq. (29).

5 Implementation

Two important issues are not dealt with in the theoretical sections: (i) How to evaluate the transition density $p_\Delta\left(y|x;\mu,\sigma^2\right)$, and (ii) how to choose the bandwidth h in finite sample. We discuss each in turn in the following.

The transition density p does not in general have a closed form expression, and so one cannot directly evaluate it. Instead, a number of different suggestions for how to either approximate or simulate it have been proposed in the literature. Lo (1988) suggests to use finite-difference methods to solve the PDE in Eq. (2) numerically, while a closed-form approximation of p can be found in Aït-Sahalia (2002). Simulation-based method for the evaluation of the likelihood are considered in Elerian et al. (2001), Kristensen and Shin (2008), and Pedersen (1995); see also Altissimo and Mele (2008).

Each of the above-mentioned studies leads to density approximations $p_{N,\Delta}\left(y|x;\mu,\sigma^2\right)$, $N\geq 1$, such that $p_{N,\Delta}\left(\cdot|\cdot;\mu,\sigma^2\right)\rightarrow^P p_\Delta\left(\cdot|\cdot;\mu,\sigma^2\right)$ as $N\rightarrow\infty$ for a suitable class of $\left(\mu,\sigma^2\right)$.[6] For both classes of semiparametric models, the transition density can be written as $p_\Delta\left(\cdot|\cdot;\hat{\mu},\hat{\sigma}^2\right)$ where $(\hat{\mu},\hat{\sigma}^2)$ are given in Eq. (11) and (12) respectively. The semiparametric estimator in either of the two classes can then be approximated by

$$\hat{\theta}_k^{(N)}=\arg\max_{\theta_k\in\Theta_k}\sum_{i=1}^n\tau_b(\hat{p}_{\Delta,k}^{(N)}\left(X_{i\Delta}|X_{(i-1)\Delta};\theta_k\right))\log\hat{p}_{\Delta,k}^{(N)}\left(X_{i\Delta}|X_{(i-1)\Delta};\theta_k\right), \tag{32}$$

where $\hat{p}_\Delta^{(N)}\left(y|x;\theta_k\right)=p_\Delta\left(y|x;\hat{\mu}\left(\cdot;\theta_k\right),\hat{\sigma}^2\left(\cdot;\theta_k\right)\right)$ is the approximation of the semiparametric density. The following theorem states that under weak conditions the approximate estimator is asymptotically equivalent to the actual estimator.

Theorem 4 *Assume that for a model in Class $k\in\{1,2\}$:*

(i) (A1)-(A8) and (B2.k) hold.

(ii) For any bounded and Lipschitz continuous pair, $\left(\mu\left(\cdot;\theta_k\right),\sigma^2\left(\cdot;\theta_k\right)\right)$, with $\sigma^2\left(x;\theta_k\right)\geq\underline{\sigma}$, and any $(x,y)\in I\times I$: $\theta_k\mapsto p_{N,\Delta}\left(y|x;\mu\left(\cdot;\theta_k\right),\sigma^2\left(\cdot;\theta_k\right)\right)$ is continuous, and

$$\sup_{\theta_k\in\Theta_k}\left|p_\Delta^{(N)}\left(y|x;\mu\left(\cdot;\theta_k\right),\sigma^2\left(\cdot;\theta_k\right)\right)-p_\Delta\left(y|x;\mu\left(\cdot;\theta_k\right),\sigma^2\left(\cdot;\theta_k\right)\right)\right|=o_P\left(1\right)$$

as $N\rightarrow\infty$.

Then there exists a sequence $N=N\left(n\right)\rightarrow\infty$ such that

$$\sqrt{n}(\hat{\theta}_k^{(N)}-\theta_{k,0})\xrightarrow{d}N\left(0,\mathcal{I}_k^{-1}\Sigma_{k,\infty}\mathcal{I}_k^{-1}\right).$$

[6] Note here that eventhough many of the approximation methods cited here are described within a parametric framework, they do not rely on a parametric model for the drift and diffusion.

The conditions in (i) are made to ensure that the actual estimator is consistent and asymptotically normally distributed. Condition (ii) together with the compactness of Θ_k implies that $\hat{\theta}_k^{(N)}$ is well-defined and unique, and that the approximate likelihood will converge uniformly towards the actual likelihood. Note that (ii) allows for $p_\Delta^{(N)}\left(y|x,\mu,\sigma^2\right)$ to be random as is the case when simulation-based methods are employed, and is satisfied for all the aforementioned approximation methods. While the theorem states that there exists a sequence N such that the approximate PMLE is asymptotically first-order equivalent to the actual PMLE, it is silent about how this sequence should be chosen. This is similar to the results of Aït-Sahalia (2002) and Pedersen (1995).

The dependence of the PMLE's on the smoothing parameter $h > 0$ chosen by the user is an undesirable feature, which they share with many other non- and semiparametric estimators. An obvious way of choosing the bandwidths would be cross-validation methods (Hart and Vieu, 1990) or rule-of-thumb and plug-in methods (Hall et al., 1995); see also Robinson (1983). Most existing methods however are designed to minimise the mean square error, while the conditions imposed on the set of bandwidths here require them to be of a different order. So the above methods do not appear to be directly applicable to semiparametric estimation problems as demonstrated in e.g. Powell and Stoker (1996). Ichimura and Todd (2007) contains further discussion of bandwidth selection in semiparametric estimation

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 as shown in Pritsker (1998). However, the estimation of θ_k involves smoothing of the kernel density estimator and so this problem won't be as pronounced in our semiparametric setting. This is supported by the results of our simulation study where the semiparametric estimator performs very well and is comparable to the fully parametric MLE.

6 A Simulation Study

In this section we present results from a small simulation study. The simulation study demonstrates that the estimator performs well even for moderate sample sizes for the models we consider. The two data-generating models are chosen as

$$dX_t = \{\beta_1 + \beta_2 X_t\}\, dt + \sqrt{\alpha_1 X_t^{\alpha_2}} dW_t, \tag{CKLS}$$

$$dX_t = \left\{\beta_1 + \beta_2 X_t + \beta_3 X_t^2 + \beta_4 X_t^{-1}\right\} dt + \sqrt{\alpha_1 X_t^{\alpha_2}} dW_t. \tag{AS}$$

The first is the CKLS model while the second is a restricted version of the short-term interest rate model proposed by Aït-Sahalia (1996b), and contains the CKLS models as special case. For both models, the data-generating parameters were chosen to match the estimates obtained when fitting the models by MLE to the data set of daily observations of the Eurodollar interest rate data used in Aït-Sahalia (1996a,b). We measure time in years and set the time distance $\Delta = 1/252$. The parameter estimates satisfy the β-mixing conditions in Aït-Sahalia (1996b) such that (A1) holds.

We then estimate the two following semiparametric models when CKLS and AS is the data generating process respectively,

$$
\begin{aligned}
\text{CKLS 1:} \qquad & dX_t = \mu(X_t)\,dt + \sqrt{\alpha_1 X_t^{\alpha_2}}\,dW_t, \\
\text{CKLS 2:} \qquad & dX_t = \{\beta_1 + \beta_2 X_t\}\,dt + \sigma(X_t)\,dW_t,
\end{aligned}
$$

$$
\begin{aligned}
\text{AS 1:} \qquad & dX_t = \mu(X_t)\,dt + \sqrt{\alpha_1 X_t^{\alpha_2}}\,dW_t, \\
\text{AS 2:} \qquad & dX_t = \left\{\beta_1 + \beta_2 X_t + \beta_3 X_t^2 + \beta_4 X_t^{-1}\right\}dt + \sigma(X_t)\,dW_t,
\end{aligned}
$$

For a given simulated sample, we obtain semiparametric estimates of $\theta_1 = (\alpha_i)$ and $\theta_2 = (\beta_i)$. We also estimate the fully parametric models in Eqs. (CKLS)-(AS) by MLE which allows us to compare the semiparametric and parametric estimates in finite sample. Note that our theoretical results do not offer any (asymptotic) comparison of the parametric MLE's of Eq. (CKLS) and (AS) relative to their semiparametric counterparts, only for the parametric MLE with the stationary density being known. So the semiparametric MLE might actually asymptotically be more efficient than the fully parametric MLE considered here. In order to evaluate the likelihood in both the parametric and semiparametric case, we employ the simulated likelihood method of Kristensen and Shin (2008) combined with the Euler scheme. For example, for the semiparametric model CKLS 1, this is implemented as follows: First obtain the nonparametric estimator $\hat{\pi}$ and the associated trimmed drift estimator, $\hat{\mu}(x;\theta_1) = \tau_a(x)/2\hat{\pi}(x) \times \partial\left[\alpha_1 x^{\alpha_2}\hat{\pi}(x)\right]/\partial x$, where $\theta_1 = (\alpha_1, \alpha_2)'$. We can now simulate from the model

$$
dX_t^* = \hat{\mu}(X_t^*;\alpha)\,dt + \sqrt{\alpha_1 (X_t^*)^{\alpha_2}}\,dW_t, \quad X_0^* = x.
$$

Let $X_s(x;\theta_1,\hat{\pi})$, $s = 1, ..., N$, be i.i.d. simulated values of X_Δ^* from this model for any given θ_1 and x using the Euler scheme; these will have distribution $p_{\Delta,1}(\cdot|x;\theta_1,\hat{\pi})$.[7] For any (y, x), we may then calculate $p_{\Delta,k}^{(N)}(y|x;\theta_1,\hat{\pi})$ by

$$
p_{\Delta,1}^{(N)}(y|x;\theta_1,\hat{\pi}) = \frac{1}{N}\sum_{k=1}^{N} K_{h_{\text{sim}}}\left(X_k(x;\theta_1,\hat{\pi}) - y\right),
$$

for a bandwidth $h_{\text{sim}} > 0$. By evaluating this kernel estimator at $(y, x) = \left(X_{i\Delta}, X_{(i-1)\Delta}\right)$, $i = 1, ..., n$, we obtain $\hat{L}_{n,1}^{(N)}(\theta_1,\hat{\pi}) = \sum_i \log p_{\Delta,1}^{(N)}\left(X_{i\Delta}|X_{(i-1)\Delta};\theta_1,\hat{\pi}\right)$. This can now be maximized w.r.t. θ_1.

We have a number of nuisance parameters which have to be chosen in the implementation of the semiparametric estimation procedure: The results reported here are based on a Gaussian kernel K, and a bandwidth choice of $h = n^{-\gamma}h^*$, where h^* was obtained by Silverman's Rule-of-thumb (c.f. Silverman, 1986, p. 47) and γ chosen to match the theoretical conditions in (B1.k)-(B2.k).

[7] We here ignore the discretization bias of the Euler scheme; see Kristensen and Shin (2008) for details on this.

The trimming parameters (a, b) were chosen to exclude the upper and lower 1% of the data, and we used $N = 500$ number of simulations in the evaluation of the likelihood. We experimented with other choices of kernel, bandwidth, trimming parameters and number of simulations, and within a reasonable range we obtained very similar results. In conclusion, the semiparametric estimator seems to be quite robust towards the choice of the various nuisance parameters.

In the simulation of the observations, we use the standard Euler scheme with a discretization step chosen as $\Delta/100$. We simulated 400 data sets, and for each we calculated the semiparametric estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ and the fully parametric MLE. We did this for sample sizes $n = 1000$, 2500 and 5000, roughly corresponding to 4, 10 and 20 years of data.

The results in terms of bias, standard deviation (std) and root-mean square error (root-MSE) for the CKLS model are reported in Table 1. From these we see, that the semiparametric estimator of β_2 performs better than the fully parametric MLE, and the other way around in the estimation of α_1, α_2 and β_1. The drift term is in general harder to estimate than the diffusion with higher biases and std's for all sample sizes. For moderate and small sample sizes, both the semiparametric and parametric estimates of β_2 tend to be rather imprecise. The resulting semiparametric (for Class 2) and parametric estimates of the drift are plotted in Figure 1. As can be seen, the drift estimates get more imprecise out in the tails of the support.

Qualitatively similar results are found for the AS model: As can be seen from Table 2, the semiparametric estimator outperforms the fully parametric one in the drift estimation and vice versa for the diffusion. As with the CKLS model, the drift estimates are more imprecise than the diffusion ones. For both estimators, the parameters β_1 and β_2 are particularly imprecisely estimated. We believe that this is caused by poor identification of these in finite sample: The process does not visit the upper region of its domain very often since the drift used here is practically equal to zero in a large part of the domain, and it is observations in the upper range that allows one to distinguish between β_1 and β_2. So for small and moderate sample sizes, it proves difficult to distinguish between the linear and quadratic effect. We believe this is why there is a significant gain in using the semiparametric estimator for the drift parameters since in this case the parameters are jointly identified through both the drift and diffusion term. The mean and the confidence bands of the drift and diffusion estimates estimators are plotted in Figure 2 and 3 respectively. From Figure 2, we see that the high bias and variance reported for the estimates of β_1 and β_2 respectively to a certain degree offset each other when used to calculate the resulting drift estimator. Still, the confidence bands are very wide for small sample sizes. The diffusion estimator exhibits similar behaviour with widening confidence bands as we move away from zero; however they are throughout significantly narrower than the drift ones, and the bias is negligible.

7 Concluding Remarks

We have proposed a new framework for the modelling, estimation and testing of scalar diffusion models. Estimators of the models together with associated test statistics were developed, and their asymptotic and finite sample properties investigated. The use of the proposed semiparametric models in bond and option pricing is investigated in Kristensen (2008b) where asymptotics of implied bond and option prices based on estimators from the semiparametric models are derived. Kristensen (2004b) uses the modelling and estimation procedure developed here to evaluate a semiparametric model for the US short-term interest rate.

A number of extensions would be of interest: Instead of relying on asymptotic approximations when drawing inference, an alternative would be bootstrapping or subsampling. Kristensen (2008a) proposes a Markov bootstrap method to be used in conjunction with the nonparametric estimators of the drift and diffusion term resulting from the semiparametric estimators developed here. We conjecture that this method can also be applied to obtain consistent confidence intervals for the parametric component of the model. The verification of this claim is outside of the scope of this paper however.

Our estimation procedure cannot readily be extended to general multivariate diffusion models since the identifying link between the invariant density, the drift and the diffusion term utilised here does not necessarily hold in higher dimensions. However, if one is willing to restrict one's attention to the class of multivariate models satisfying this relation, the proposed estimation and testing procedure carries can also be employed in a multivariate setting.[8]

[8] This restriction is for example imposed by Chen et al (2009) in their nonparametric study of multivariate diffusion models.

References

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.

Aït-Sahalia, Y., J. Fan and H. Peng (2009) Nonparametric Transition-Based Tests for Jump-Diffusions. Forthcoming in *Journal of the American Statistical Association*.

Aït-Sahalia, Y. and L.P. Hansen (eds.) (2006) *Handbook of Financial Econometrics* (forthcoming). Amsterdam: North-Holland.

Altissimo, F. and A. Mele (2009) Simulated Nonparametric Estimation of Dynamic Models. *Review of Economic Studies* 76, 413-450.

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

Bandi, F.M. and P.C.B. Phillips (2007) A Simple Approach to the Parametric Estimation of Potentially Nonstationary Diffusions. *Journal of Econometrics* 137, 354-395.

Björk, T. (2004) *Arbitrage Theory in Continuous Time, 2nd Edition*. Oxford: Oxford University Press.

Chen, X. (2007) Large Sample Sieve Estimation of Semi-Nonparametric Models. In *Handbook of Econometrics* , Vol. 6 (eds. J. Heckman and E. Leamer), Chapter 76.

Chen, X., Y. Fan and V. Tsyrennikov (2006) Efficient Estimation of Semiparametric Multivariate Copula Models. *Journal of the American Statistical Association* 101, 1228-1240.

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

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

Chen, X., L.P. Hansen and J. Scheinkman (2009) Nonlinear Principal Components and Long Run Implications. Forthcoming in *Annals of Statistics*.

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

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

Corradi, V. and N.R. Swanson (2005) Bootstrap Specification Tests for Diffusion Processes. *Journal of Econometrics* 124, 117-148.

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

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

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

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

Florens, J.-P., E. Renault and N. Touzi (1998) Testing Embeddability by Stationary Reversible Continuous-Time Markov Processes. *Econometric Theory* 14, 744-769.

Friedman, A. (1976) *Stochastic Differential Equations and Applications, Vol. 1.* New York: Academic Press.

Genest, C., K. Ghoudi and L.-P. Rivest (1995) A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions. *Biometrika* 82, 543-552.

Gobet, E., M. Hoffmann and M. Reiß (2004) Nonparametric Estimation of Scalar Diffusions Based on Low Frequency Data. *Annals of Statistics* 32, 2223–2253.

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

Hart, J. and P. Vieu (1990) Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statististics* 18, 873–890.

Hansen, B. (2008) Uniform Convergence Rates of Kernel Estimation with Dependent Data. *Econometric Theory* 24, 726-748.

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

Hong, Y. and Li, H. (2005) Nonparametric Specification Testing for Continuous-Time Models with Application to Spot Interest Rates. *Review of Financial Studies* 18, 37-84.

Ilyin, A.M., A.S. Kalashnikov and O.A. Oleynik (2002) Linear Second-Order Partial Differential Equations of The Parabolic Type. *Journal of Mathematical Sciences* 108, 435-542.

Kanaya, S. (2008) Non-Parametric Specification Testing for Continuous-Time Markov Processes: Do the Processes Follow Diffusions? Manuscript, University of Wisconsin.

Karatzas, I. and S.E. Shreve (1991) *Brownian Motion and Stochastic Calculus, 2nd Edition*. New York: Springer Verlag.

Klaassen, C. and J. Wellner (1997) Efficient Estimation in the Bivariate Normal Copula Model: Normal Margins are Least-Favorable. *Bernoulli* 3, 55-77.

Kristensen, D. (2004a) Estimation in Two Classes of Semiparametric Diffusion Models. FMG Discussion Papers DP500, LSE.

Kristensen, D. (2004b) A Semiparametric Single-Factor Model of the Term Structure. FMG Discussion Papers DP501, LSE.

Kristensen, D. (2008a) Nonparametric Estimation and Misspecification Testing of Diffusion Models. CREATES Research Paper 2007-01, University of Aarhus.

Kristensen, D. (2008b) Estimation of Partial Differential Equations with Applications in Finance. *Journal of Econometrics* 144, 392-408.

Kristensen, D. and A. Rahbek (2005) Asymptotics of the QMLE for a Class of ARCH(q) Models. *Econometric Theory* 21, 946-961.

Kristensen, D. and Y. Shin (2008) Estimation of Dynamic Models with Nonparametric Simulated Maximum Likelihood. Manuscript, University of Wisconsin.

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

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

Newey, W.K. (1991) Semiparametric Efficiency Bounds. *Journal of Applied Econometrics* 5, 99-135.

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

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

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

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.

Powell, J.L., J.H. Stock and T.M. Stoker (1989) Semiparametric Estimation of Index Coefficients. *Econometrica* 57, 1403-1430.

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

Pritsker, M.G. (1998) Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models. *Review of Financial Studies* 11, 449–487.

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

Robinson, P.M. (1991) Consistent Nonparametric Entropy-Based Testing. *Review of Economic Studies* 58, 437-453.

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

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

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.

Whang, Y.-J. and D.W.K. Andrews (1993) Tests of Specification for Parametric and Semiparametric Models. *Journal of Econometrics* 57, 277-318.

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.

A Proofs

Proof of Theorem 1. Consider a given model in Class 1: By Ilyin et al (2002), $(\hat{\mu}(\cdot;\theta_1,\pi),\hat{\sigma}^2(\cdot;\theta_1))$ and $(\bar{\mu}(\cdot;\theta_1,\pi),\bar{\sigma}^2(\cdot;\theta_1))$ as defined in Eq. (11) and (29) respectively are both situated in $\mathcal{H}$ as defined at the beginning of Appendix B. We,may then appeal to Lemma 6(i) yielding

$$\begin{aligned}&\sup_{\theta_1\in\Theta_1}\left|\hat{L}_n\left(\hat{\mu}(\cdot;\theta_1,\hat{\pi}),\hat{\sigma}^2(\cdot;\theta_1)\right)-L_n\left(\mu(\cdot;\theta_1),\sigma^2(\cdot;\theta_1)\right)\right|\\ =\;& O_P\left(b^{-1}\varepsilon_n\right)+O_P\left(|\log(b)|^{-\delta}\right)+O_P\left(b^{-1}a^{\delta_1/2}\right)\end{aligned}$$

where, by Lemma 9(i) together with (B1.1),

$$\varepsilon_n=\sup_{\theta_1\in\Theta_1}\max\left\{\|\hat{\mu}(\cdot;\theta_1,\hat{\pi})-\bar{\mu}(\cdot;\theta_1)\|_\infty,\left\|\hat{\sigma}^2(\cdot;\theta_1)-\bar{\sigma}^2(\cdot;\theta_1)\right\|_\infty\right\}=o_P(1).$$

Finally, (A.4.ii) together with Kristensen and Rahbek (2005, Proposition 1) imply

$$\sup_{\theta_1\in\Theta_1}\left|L_n\left(\mu(\cdot;\theta_1),\sigma^2(\cdot;\theta_1)\right)-L_1(\theta_1)\right|\to^P 0,$$

where $\theta_{1,0}=\arg\max_{\theta\in\Theta}L(\theta_1)$. Consistency now follows from Newey and McFadden (1994, Theorem 2.1).

The proof for models in Class 2 follows along the same lines: Here, the submodel and estimators are given in Eqs. (9) and (12). Appealing to Lemma 6(i) together with Lemma 10, we obtain $\sup_{\theta_2\in\Theta_2}\left|\hat{L}_n\left(\hat{\mu}(\cdot;\theta_2),\hat{\sigma}^2(\cdot;\theta_2,\hat{\pi})\right)-L_n\left(\mu(\cdot;\theta_2),\sigma^2(\cdot;\theta_2)\right)\right|=o_P(1)$ under (B2.1). The rest of the proof is identical to the one for Class 1. ■

Proof of Theorem 2. We follow the arguments outlined on p. 11-12 to prove the claimed result. First, by a standard Taylor expansion, we easily obtain that eq. (16) holds. Next, the functional Taylor expansion in Eq. (18) holds under (B.2.1) due to Lemma 7 and either Lemma 9 (Class 1) or 10 (Class2), while Eq. (19) follows from (A.7). By Lemma 8 in conjunction with either Lemma 9 or 10 imply $\left\|\hat{H}_{n,k}(\bar{\theta}_k,\hat{\pi})-H_{n,k}(\bar{\theta}_k,\pi_0)\right\|\to^P 0$. Next, Kristensen and Rahbek (2005, Proposition 1), which is applicable due to (A.6), yields $H_{n,k}(\bar{\theta}_k,\pi_0)\to^P-\mathcal{I}_k$. The long variance $\Sigma_{k,\infty}<\infty$ since, by Cauchy-Schwarz's inequality and (A8),

$$E\left[\|D_k(X_0)\|^2\right]\leq E\left[\|s_{\Delta,k}(X_\Delta|X_0;\theta_{k,0})\|^2\right]\times\int_I E\left[\frac{\|d_k(x,X_\Delta|X_0)\|^2}{p_{0,\Delta}^2(X_\Delta|X_0;\theta_{k,0})}\right]\pi_0(x)\,dx<\infty,$$

Thus, by Eq. (25), $\sqrt{n}(\hat{\theta}_k-\theta_{k,0})\xrightarrow{d}N\left(0,\mathcal{I}_k^{-1}\Sigma_{k,\infty}\mathcal{I}_k^{-1}\right)$ if indeed Eq. (23) holds.

We now verify Eq. (23) for both classes of models. In the following, we keep $\theta_k=\theta_{k,0}$ fixed, and suppress dependence on this parameter. Also, for a model in either class, let μ_0, σ_0^2 and $p_{0,t}(y|x)$ denote the true drift, diffusion and transition density respectively. Consider first a model situated in Class 1: Since $\nabla\mu[d\pi]$ as given in Lemma 9 is the pathwise derivatives w.r.t. π, it follows by the chain rule that the pathwise derivative w.r.t. π is given by

$$\nabla S_{n,1}[d\pi] = \nabla S_{n,1}^{(1)}[d\pi] + \nabla S_{n,1}^{(2)}[d\pi] \equiv \nabla S_n^{(1)}[\nabla\mu[d\pi], 0] + \nabla S_n^{(2)}[\nabla\mu[d\pi], 0],$$

where $\nabla S_n^{(i)}[d\mu, d\sigma^2]$, $i = 1, 2$, are given in Eq. (48) while $\dot{p}$, ∇p, and $\nabla\dot{p}$ are defined in Eqs. (34)-(37). We first take a closer look at $\nabla p_\Delta[\nabla\mu[\hat{\pi} - \pi_0], 0]$. By definition,

$$\begin{aligned}
& \nabla p_\Delta[\nabla\mu[\hat{\pi} - \pi_0], 0](y|x) \\
= & \int_0^\Delta \int_I \nabla\mu[\hat{\pi} - \pi_0](w) \frac{\partial p_{0,t}(y|w)}{\partial w} p_{0,t}(w|x)\, dw dt, \\
= & -\frac{1}{2}\int_0^\Delta \int_I \sigma_0^2(w) \frac{\pi_0^{(1)}(w)}{\pi_0^2(w)} \{\hat{\pi}(w) - \pi_0(w)\} \frac{\partial p_{0,t}(y|w)}{\partial w} p_{0,t}(w|x)\, dw dt \\
& + \frac{1}{2}\int_0^\Delta \int_I \sigma_0^2(w) \frac{1}{\pi_0(w)} \left\{\hat{\pi}^{(1)}(w) - \pi_0^{(1)}(w)\right\} \frac{\partial p_{0,t}(y|w)}{\partial w} p_{0,t}(w|x)\, dw dt \\
= & \frac{1}{n}\sum_{j=1}^n \{\psi_1[K_h(X_{j\Delta} - \cdot)](y, x) - \psi_1[\pi_0](y, x)\},
\end{aligned}$$

where

$$\begin{aligned}
\psi_1[f](y, x) &:= \int_I f(w)\, d_{11}(w, y, x)\, dw - \int_I f^{(1)}(w)\, d_{12}(w, y, x)\, dw, \\
d_{11}(w, y, x) &:= -\frac{1}{2}\sigma_0^2(w) \frac{\pi_0^{(1)}(w)}{\pi_0^2(w)} \int_0^\Delta \frac{\partial p_{0,t}(y|w)}{\partial w} p_{0,t}(w|x)\, dt, \\
d_{12}(w, y, x) &:= -\frac{1}{2}\frac{\sigma_0^2(w)}{\pi_0(w)} \int_0^\Delta \frac{\partial p_{0,t}(y|w)}{\partial w} p_{0,t}(w|x)\, dt.
\end{aligned}$$

Observe that $\psi_1[\pi_0](y, x) = 0$ and, using standard results for kernel smoothers,

$$\begin{aligned}
\psi_1[K_h(\cdot - z)](y, x) &= \int_I K_h(w - z)\, d_{11}(w, y, x)\, dw - \int_I K_h^{(1)}(w - z)\, d_{12}(w, y, x)\, dw \\
&= \left\{d_{11}(z, y, x) + \frac{\partial d_{12}(z, y, x)}{\partial z}\right\} + O(h^m) \\
&= d_1(z, y, x) + O(h^m),
\end{aligned}$$

uniformly in (z, y, x) where $d_1(z, y, x)$ is defined in Eq. (21). We obtain $\nabla\dot{p}_\Delta[\hat{\mu} - \mu_0, 0](y|x) = n^{-1}\sum_{j=1}^n \dot{d}_1(X_{j\Delta}, y, x) + O_P(h^m)$ by the same arguments. We have now shown that

$$\begin{aligned}
\nabla S_{n,1}^{(1)}[\hat{\pi} - \pi_0] &:= -\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \frac{\dot{p}_{\Delta,i}(\mu_0)}{p_{\Delta,i}^2(\mu_0)} d_1\left(X_{j\Delta}, X_{i\Delta}, X_{(i-1)\Delta}\right) + O_P(h^m), \\
\nabla S_{n,2}^{(2)}[\hat{\pi} - \pi_0] &:= \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \frac{1}{p_{\Delta,i}(\mu_0)} \dot{d}_1\left(X_{j\Delta}, X_{i\Delta}, X_{(i-1)\Delta}\right) + O_P(h^m),
\end{aligned}$$

where under (B1.2) $h^m = o(\sqrt{n})$. It's straightforward to check that $E[d_1(X_0, y, x)] = E[\dot{d}_1(X_0, y, x)] =$

0 and, by some simple manipulations,

$$E\left[\frac{\dot{d}_1\left(z,X_\Delta,X_0\right)}{p_{0,\Delta}\left(X_\Delta|X_0\right)}\right]=\frac{\partial}{\partial\theta_k}\int_I\int_I d_1\left(z,y,x\right)\pi_0\left(x\right)dydx=0.$$

Apply standard U-statistics results for β-mixing processes, e.g. Arcones (1995), to obtain

$$\nabla S_{n,1}^{(1)}\left[\hat{\pi}-\pi_0\right]=\frac{1}{n}\sum_{j=1}^{n}D_1\left(X_{j\Delta}\right)+o_P\left(1/\sqrt{n}\right),\quad \nabla S_{n,1}^{(2)}\left[\hat{\pi}-\pi_0\right]=o_P\left(1/\sqrt{n}\right),$$

where $D_1\left(z\right)$ is given in Eq. (24).

Next, consider a model situated in Class 2. The adjustment term can be written as

$$\nabla S_{n,2}\left[d\pi\right]=\nabla S_{n,2}^{(1)}\left[d\pi\right]+\nabla S_{n,2}^{(2)}\left[d\pi\right]\equiv\nabla S_n^{(1)}\left[0,\nabla\sigma_0^2\left[d\pi\right]\right]+\nabla S_n^{(2)}\left[0,\nabla\sigma_0^2\left[d\pi\right]\right],$$

where $\nabla\sigma_0^2\left[d\pi\right]$ is given in Lemma 10. As in Class 1, $\nabla S_n^{(2)}\left[0,\nabla\sigma_0^2\left[\hat{\pi}-\pi_0\right]\right]=o_P\left(1/\sqrt{n}\right)$ while

$$\begin{aligned}
&\nabla p_\Delta\left[0,\nabla\sigma_0^2\left[\hat{\pi}-\pi_0\right]\right](y|x)\\
=\;&\int_0^\Delta\int_I\nabla\sigma_0^2\left[\hat{\pi}-\pi_0\right](z)\,\frac{\partial^2 p_{0,t}\left(y|z\right)}{\partial z^2}p_{0,t}\left(z|x\right)dzdt\\
=\;&2\int_0^\Delta\int_I\frac{\int_l^z\hat{\pi}\left(y\right)\mu_0\left(y\right)dy}{\pi_0\left(z\right)}\frac{\partial^2 p_{0,t}\left(y|z\right)}{\partial z^2}p_{0,t}\left(z|x\right)dzdt\\
&-2\int_0^\Delta\int_I\frac{\int_l^z\pi_0\left(y\right)\mu_0\left(y\right)dy}{\pi_0^2\left(z\right)}\hat{\pi}\left(z\right)\frac{\partial^2 p_{0,t}\left(y|z\right)}{\partial z^2}p_{0,t}\left(z|x\right)dzdt+O_P\left(h^m\right)\\
=\;&\frac{1}{n}\sum_{j=1}^{n}d_2\left(X_{j\Delta},y,x\right)+O_P\left(h^m\right),
\end{aligned}$$

where $d_2\left(z,y,x\right)$ is given in Eq. (22). By the same arguments used for Class 1, this shows that $\nabla S_{n,2}^{(1)}\left[\hat{\pi}-\pi_0\right]$ can be written on the form given in Eq. (23). ■

Proof of Theorem 3. For the test against a model in Class k, we write, by a Taylor expansion,

$$\sqrt{n}\hat{S}_{n,k}(\tilde{\theta}_k,\hat{\pi})=\sqrt{n}\hat{S}_{n,k}(\theta_{k,0},\hat{\pi})+\hat{H}_{n,k}(\bar{\theta}_k,\hat{\pi})\sqrt{n}(\tilde{\theta}_k-\theta_{k,0}),$$

for some $\bar{\theta}_k\in\left[\theta_{k,0},\tilde{\theta}_k\right]$, where from the proof of Theorem 2,

$$\hat{S}_{n,k}(\theta_{k,0},\hat{\pi})=\frac{1}{n}\sum_{i=1}^{n}\left\{s\left(X_{i\Delta}|X_{(i-1)\Delta};\theta_{k,0}\right)+D_k\left(X_{(i-1)\Delta}\right)\right\}+o_P\left(n^{-1/2}\right). \tag{33}$$

Substituting in the right hand side of (33) together with $\tilde{\theta}_k-\theta_{k,0}=\sum_{i=1}^n\psi_{k,i}/n+o_P\left(1/\sqrt{n}\right)$ and appealing to the CLT for mixing processes yield the result. To show that Eqs. (30) and (31) are equivalent, we use the same expansion as above except that it's made around $\hat{\theta}_k$ instead of $\theta_{k,0}$ such that: $\sqrt{n}\hat{S}_{n,k}(\tilde{\theta}_k,\hat{\pi})=\hat{H}_{n,k}(\breve{\theta}_k,\hat{\pi})\sqrt{n}(\tilde{\theta}_k-\hat{\theta}_k)$ for some $\breve{\theta}_k\in\left[\tilde{\theta}_k,\theta_{k,0}\right]$. ■

Proof of Theorem 4. We only give a proof for models in Class 1. The proof for Class 2 follows along the exact same lines. Consider some given parametric pair $\left(\mu\left(\cdot;\theta_1\right), \sigma^2\left(\cdot;\theta_1\right)\right)$ satisfying (ii) of the theorem, and define $\varepsilon_N\left(y|x\right)$ as the associated uniform approximation error:

$$\varepsilon_N\left(y|x\right) = \sup_{\theta_1\in\Theta_1} \left|\tau_b\left(p_\Delta\left(y|x;\theta_1\right)\right)\log\left(p_\Delta\left(y|x;\theta_1\right)\right) - \tau_b\left(p_{\Delta,N}\left(y|x;\theta_1\right)\right)\log\left(p_{\Delta,N}\left(y|x;\theta_1\right)\right)\right|,$$

where $p_\Delta\left(y|x;\theta_1\right) = p_\Delta\left(y|x;\mu\left(\cdot;\theta_1\right), \sigma^2\left(\cdot;\theta_1\right)\right)$ and $p_\Delta^{(N)}\left(y|x;\theta_1\right) = p_{\Delta,N}\left(y|x;\mu\left(\cdot;\theta_1\right), \sigma^2\left(\cdot;\theta_1\right)\right)$. Due to continuity of $z \mapsto \log\left(z\right)\tau_b\left(z\right)$, $\varepsilon_N\left(y|x\right) \to^P 0$ as $N \to \infty$ according to condition (ii) of the theorem.

Choose any *fixed* $n \geq 1$, and condition on the sample $\left(X_0, X_\Delta, ..., X_{n\Delta}\right)$ so we can treat the semiparametric estimators $\left(\hat{\mu}\left(\cdot;\theta_1,\hat{\pi}\right), \hat{\sigma}^2\left(\cdot;\theta_1\right)\right)$ as non-random. We then have that the approximation error associated with this pair, which we denote $\hat{\varepsilon}_N\left(y|x\right)$, goes to zero in probability as $N \to \infty$ for any given (x, y). Thus,

$$\bar{\varepsilon}_{N,n} := \sup_{\theta_1\in\Theta_1} |\hat{L}_{n,1}\left(\theta_1\right) - \hat{L}_n^{(N)}\left(\theta_1\right)| \leq \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_N\left(X_{i\Delta}|X_{(i-1)\Delta}\right) \to^P 0, \quad N \to \infty,$$

since $\hat{\varepsilon}_N\left(X_{i\Delta}|X_{(i-1)\Delta}\right) \to^P 0$ as $N \to \infty$ for all $i = 1, ..., n$. Now, consider any given $\delta > 0$: Due to the above results, we can for any given $n \geq 1$ choose $N = N\left(n\right)$ such that $P\left(\sqrt{n}\bar{\varepsilon}_{N,n} > \delta\right) \leq e^{-n}$. This translates into $P(\sqrt{n}||\hat{\theta}_1^{(N)} - \hat{\theta}_1|| > \delta) \leq e^{-n}$ for all $n \geq 1$. Thus, $\sqrt{n}(\hat{\theta}_1^{(N)} - \theta_{1,0}) = \sqrt{n}(\hat{\theta}_1 - \theta_{1,0}) + \sqrt{n}(\hat{\theta}_1^{(N)} - \hat{\theta}_1) = \sqrt{n}(\hat{\theta}_1 - \theta_{1,0}) + o_P\left(1\right)$, and the result follows by condition (i). ■

B Lemmas

Let $\mathcal{H}$ denote the set of function pairs $\left(\mu, \sigma^2\right)$ such that an associated solution, $p_t\left(y|x;\mu,\sigma^2\right)$, to (2) exists and satisfies

$$\left|\frac{\partial^k p_t\left(y|x;\mu,\sigma^2\right)}{\partial x^k}\right| \leq C\gamma_t\left(y|x\right), \quad k = 0, 1, 2,$$

for $(x, y) \in I \times I$ and $t \in (0, \Delta)$. We also define $\bar{\gamma}\left(y|x\right) := \int_0^\Delta \gamma_t\left(y|x\right) dt$ as the integral of the upper bound. We introduce a generic diffusion model characterized by drift $\mu_\theta\left(x\right)$ and diffusion $\sigma_\theta^2\left(x\right)$; these are parameterized by some parameter $\theta \in \Theta$ (we will later choose $\theta = \theta_1$ for models in Class 1 and $\theta = \theta_2$ for models in Class 2). Let $\partial_\theta^k\mu_\theta\left(x\right)$ and $\partial_\theta^k\sigma_\theta^2\left(x\right)$ denote the kth derivatives w.r.t. θ, and $p_t\left(y|x;\mu_\theta,\sigma_\theta^2\right)$ the associated transition density.

We now derive expressions of the pathwise derivative of p_t w.r.t. $\left(\mu, \sigma^2\right)$, which we denote ∇p_t, and its first and second order derivative w.r.t. θ, which we denote $\dot{p}_t$ and $\ddot{p}_t$. Finally, let $\nabla\dot{p}_t$ denote the derivative w.r.t. θ of ∇p_t. These are given as solutions to the following PDE's:

$$\frac{\partial\dot{p}_t}{\partial t} = \mathcal{A}\left[\mu_\theta, \sigma_\theta^2\right]\dot{p}_t + \mathcal{A}\left[\partial_\theta\mu_\theta, \partial_\theta\sigma_\theta^2\right]p_t, \tag{34}$$

$$\frac{\partial \ddot{p}_t}{\partial t} = \mathcal{A}\left[\mu_\theta, \sigma_\theta^2\right] \ddot{p}_t + \left\{2\mathcal{A}\left[\partial_\theta \mu_\theta, \partial_\theta \sigma_\theta^2\right] \dot{p}_t + \mathcal{A}\left[\partial_\theta^2 \mu_\theta, \partial_\theta^2 \sigma_\theta^2\right] p_t\right\}, \tag{35}$$

$$\frac{\partial \nabla p_t}{\partial t} = \mathcal{A}\left[\mu_0, \sigma_0^2\right] \nabla p_t + \mathcal{A}\left[d\mu, d\sigma^2\right] p_{0,t}, \tag{36}$$

$$\begin{aligned} \frac{\partial \nabla \dot{p}_t}{\partial t} &= \mathcal{A}\left[\mu_0, \sigma_0^2\right] \nabla \dot{p}_t + \mathcal{A}\left[d\mu, d\sigma^2\right] \dot{p}_{0,t} \\ &\quad + \mathcal{A}\left[d\partial_\theta \mu_\theta, d\partial_\theta \sigma_\theta^2\right] p_{0,t} + \mathcal{A}\left[\partial_\theta \mu_0, \partial_\theta \sigma_0^2\right] \nabla p_t, \end{aligned} \tag{37}$$

where all have zero intial condition. Here, $d\mu$ and $d\sigma^2$ denotes the directions of the partical differentials. The following lemma states some useful properties of p_t and its derivatives:

Lemma 5 *Assume that $(\mu_{j,\theta}, \sigma_{j,\theta}^2) \in \mathcal{H}$ for all $\theta \in \Theta$, $j = 1, 2$. Then the following inequalities hold uniformly over $x, y \in I$ and for $k = 0, 1, 2$:*

$$\left|\partial_\theta^k p_\Delta\left(y|x; \mu_{1,\theta}, \sigma_{1,\theta}^2\right) - \partial_\theta^k p_\Delta\left(y|x; \mu_{2,\theta}, \sigma_{2,\theta}^2\right)\right| \leq C \sum_{i=0}^{k} B_{i,\theta}^{1/2}(x),$$

$$\left|\partial_\theta^k p_\Delta\left(y|x; \mu_{1,\theta}, \sigma_{1,\theta}^2\right) - \partial_\theta^k p_\Delta\left(y|x; \mu_{2,\theta}, \sigma_{2,\theta}^2\right) - \partial_\theta^k \nabla p_\Delta\left(y|x; \mu_{2,\theta}, \sigma_{2,\theta}^2\right)\left[d\mu_\theta, d\sigma_\theta^2\right]\right| \leq C \sum_{i=0}^{k} B_{i,\theta}(x),$$

where $d\mu_\theta = \mu_{1,\theta} - \mu_{2,\theta}$, $d\sigma_\theta^2 = \sigma_{1,\theta}^2 - \sigma_{2,\theta}^2$, and

$$B_{i,\theta}(x) := \int_I \left(\left|\partial_\theta^i \mu_{1,\theta}(w) - \partial_\theta^i \mu_{2,\theta}(w)\right|^2 + \left|\partial_\theta^i \sigma_{1,\theta}^2(w) - \partial_\theta^i \sigma_{2,\theta}^2(w)\right|^2\right) \bar{\gamma}(w|x)\, dw.$$

Proof. Let $p_{k,t}(y|x) = p_t\left(y|x; \mu_k, \sigma_k^2\right)$ where we suppress the dependence on θ. First note that for $k = 0, 1, 2$,

$$\int_0^\Delta \int_I \left|\frac{\partial^k p_{1,t}(y|w)}{\partial w^k}\right| p_{2,t}(w|x)\, dwdt \leq \int_0^\Delta \int_I \gamma_t(y|w)\, \gamma_t(w|x)\, dwdt, \tag{38}$$

where the right hand side is uniformly bounded over $x, y \in I$ and $\theta \in \Theta$. Then observe that the function $q_t := p_{1,t} - p_{2,t}$ solves the PDE,

$$\frac{\partial q_t}{\partial t} = \mathcal{A}\left[\mu_1, \sigma_1^2\right] p_{1,t} - \mathcal{A}\left[\mu_2, \sigma_2^2\right] p_{2,t} = \mathcal{A}\left[\mu_2, \sigma_2^2\right] q_t + \mathcal{A}\left[\mu_1 - \mu_2, \sigma_1^2 - \sigma_2^2\right] p_{1,t}.$$

Thus,

$$\begin{aligned} p_{1,\Delta}(y|x) - p_{2,\Delta}(y|x) &= \int_I \int_0^\Delta \mathcal{A}\left[\mu_1 - \mu_2, \sigma_1^2 - \sigma_2^2\right] p_{1,t}(y|w)\, p_{2,t}(w|x)\, dwdt \\ &= \int_I \int_0^\Delta \left\{\mu_1(w) - \mu_2(w)\right\} \frac{\partial p_{1,t}(y|w)}{\partial w} p_{2,t}(w|x)\, dtdw \\ &\quad + \frac{1}{2} \int_I \int_0^\Delta \left\{\sigma_1^2(w) - \sigma_2^2(w)\right\} \frac{\partial^2 p_{1,t}(y|w)}{\partial w^2} p_{2,t}(w|x)\, dtdw. \end{aligned}$$

Using Cauchy-Schwarz's inequality and Eq. (38), we then obtain $|p_{1,\Delta}(y|x) - p_{2,\Delta}(y|x)| \leq CB_0^{1/2}(x)$.

For later use, we also note that

$$\begin{aligned}
\frac{\partial^k \{p_{1,t} - p_{2,t}\}(z|y)}{\partial y^k} &= \int_0^t \int_{\mathbb{R}} \{\mu_1(w) - \mu_2(w)\} \frac{\partial p_{1,s}(z|w)}{\partial w} \frac{\partial^k p_{2,s}(w|y)}{\partial y^k} dwds \\
&\quad + \frac{1}{2}\int_0^t \int_{\mathbb{R}} \{\sigma_1^2(w) - \sigma_2^2(w)\} \frac{\partial^2 p_{1,s}(z|w)}{\partial w^2} \frac{\partial^k p_{2,s}(w|y)}{\partial y^k} dwds,
\end{aligned} \tag{39}$$

and

$$\begin{aligned}
\frac{\partial^k \{\dot{p}_{1,t} - \dot{p}_{2,t}\}(z|y)}{\partial y^k} &= \int_0^t \int_{\mathbb{R}} \{\partial_\theta \mu_1(w) - \partial_\theta \mu_2(w)\} \frac{\partial p_{1,s}(z|w)}{\partial w} \frac{\partial^k p_{2,s}(w|y)}{\partial y^k} dwds \\
&\quad + \int_0^t \int_{\mathbb{R}} \{\mu_1(w) - \mu_2(w)\} \frac{\partial \dot{p}_{1,s}(z|w)}{\partial w} \frac{\partial^k p_{2,s}(w|y)}{\partial y^k} dwds \\
&\quad + \frac{1}{2}\int_0^t \int_{\mathbb{R}} \{\partial_\theta \sigma_1^2(w) - \partial_\theta \sigma_2^2(w)\} \frac{\partial^2 p_{1,s}(z|w)}{\partial w^2} \frac{\partial^k p_{2,s}(w|y)}{\partial y^k} dwds, \\
&\quad + \frac{1}{2}\int_0^t \int_{\mathbb{R}} \{\sigma_1^2(w) - \sigma_2^2(w)\} \frac{\partial^2 \dot{p}_{1,s}(z|w)}{\partial w^2} \frac{\partial^k p_{2,s}(w|y)}{\partial y^k} dwds.
\end{aligned} \tag{40}$$

A similar, but more lengthy, expression can be derived for $\ddot{p}_{1,t} - \ddot{p}_{2,t}$. Using the same bounds as before,

$$\begin{aligned}
\left|\frac{\partial^k \{p_{1,t} - p_{2,t}\}(z|y)}{\partial y^k}\right| &\leq CB_{0,\theta}^{1/2}(x), \quad \left|\frac{\partial^k \{\dot{p}_{1,t} - \dot{p}_{2,t}\}(z|y)}{\partial y^k}\right| \leq C\sum_{i=0}^{1} B_i^{1/2}(x), \\
\left|\frac{\partial^k \{\ddot{p}_{1,t} - \ddot{p}_{2,t}\}(z|y)}{\partial y^k}\right| &\leq C\sum_{i=0}^{2} B_i^{1/2}(x).
\end{aligned}$$

Next, redefine $q_t := p_{1,t} - p_{2,t} - \nabla p_{2,t}\left[\mu_1 - \mu_2, \sigma_1^2 - \sigma_2^2\right]$, and we obtain

$$\begin{aligned}
\frac{\partial q_t}{\partial t} &= \mathcal{A}\left[\mu_1, \sigma_1^2\right] p_{1,t} - \mathcal{A}\left[\mu_2, \sigma_2^2\right] p_{2,t} - \mathcal{A}\left[\mu - \mu_0, \sigma^2 - \sigma_0^2\right] p_{2,t} - \mathcal{A}\left[\mu_0, \sigma_0^2\right] \nabla p_{2,t} \\
&= \mathcal{A}\left[\mu_0, \sigma_0^2\right] q_t + \mathcal{A}\left[\mu - \mu_0, \sigma^2 - \sigma_0^2\right] (p_{1,t} - p_{2,t}).
\end{aligned}$$

Applying the above bounds on the derivatives on $p_{1,t} - p_{2,t}$ and Cauchy-Schwarz's inequality,

$$\begin{aligned}
&\left|p_{1,\Delta}(y|x) - p_{2,\Delta}(y|x) - \nabla p_{2,\Delta}(y|x)\left[\mu_1 - \mu_2, \sigma_1^2 - \sigma_2^2\right]\right| \\
\leq& \int_0^\Delta \int_I \left|\mathcal{A}\left[\mu_1 - \mu_2, \sigma^2 - \sigma_2^2\right](w)\left(p_{1,t}(y|w) - p_{2,t}(y|w)\right)\right| p_{2,t}(w|x)\, dwdt \\
\leq& \int_0^\Delta \int_I |\mu_1(w) - \mu_2(w)| \left|\frac{\partial \{p_{1,t} - p_{2,t}\}(y|w)}{\partial y}\right| p_{2,t}(w|x)\, dwdt \\
&+ \int_0^\Delta \int_I \left|\sigma_1^2(w) - \sigma_2^2(w)\right| \left|\frac{\partial^2 \{p_{1,t} - p_{2,t}\}(y|w)}{\partial y^2}\right| p_{2,t}(w|x)\, dwdt \\
\leq& CB_0(x).
\end{aligned}$$

Similarly for derivatives w.r.t. θ. ■

Next, we can apply the above lemma to show that the log-likelihood, the score and the hessian are well-behaved functionals of (μ, σ^2). With $p_{\Delta,i}(\mu_\theta, \sigma_\theta^2) = p_\Delta(X_{i\Delta}|X_{(i-1)\Delta}; \mu_\theta, \sigma_\theta^2)$ and similarly for its derivatives, and $\tau_{b,i} = \tau_b(p(X_{i\Delta}|X_{(i-1)\Delta}; \mu_\theta, \sigma_\theta^2))$, the untrimmed score takes the form

$$S_n(\mu_\theta, \sigma_\theta^2) = \frac{1}{n}\sum_{i=1}^{n} \dot{p}_{\Delta,i}(\mu_\theta, \sigma_\theta^2) / p_{\Delta,i}(\mu_\theta, \sigma_\theta^2), \tag{41}$$

while the trimmed version is given by

$$\hat{S}_n(\mu_\theta, \sigma_\theta^2) = \hat{S}_n^{(1)}(\mu_\theta, \sigma_\theta^2) + \hat{S}_n^{(2)}(\mu_\theta, \sigma_\theta^2), \tag{42}$$

$$\hat{S}_n^{(1)}(\mu_\theta, \sigma_\theta^2) = \frac{1}{n}\sum_{i=1}^{n} \tau_{b,i} \frac{\dot{p}_{\Delta,i}(\mu_\theta, \sigma_\theta^2)}{p_{\Delta,i}(\mu_\theta, \sigma_\theta^2)}, \quad \hat{S}_n^{(2)}(\mu_\theta, \sigma_\theta^2) = \frac{1}{n}\sum_{i=1}^{n} \tau_{b,i}^{(1)} \log p_{\Delta,i}(\mu_\theta, \sigma_\theta^2)\, \dot{p}_{\Delta,i}(\mu_\theta, \sigma_\theta^2). \tag{43}$$

The untrimmed and trimmed Hessian are given as:

$$H_n(\mu, \sigma^2) = \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{\ddot{p}_{\Delta,i}(\mu, \sigma^2)}{p_{\Delta,i}(\mu, \sigma^2)} + \frac{\dot{p}_{\Delta,i}(\mu, \sigma^2)}{p_{\Delta,i}(\mu, \sigma^2)} \frac{\dot{p}_{\Delta,i}(\mu, \sigma^2)'}{p_{\Delta,i}(\mu, \sigma^2)} \right\}, \tag{44}$$

$$\hat{H}_n(\mu_\theta, \sigma_\theta^2) = \hat{H}_n^{(1)}(\mu_\theta, \sigma_\theta^2) + \hat{H}_n^{(2)}(\mu_\theta, \sigma_\theta^2) + \hat{H}_n^{(3)}(\mu_\theta, \sigma_\theta^2), \tag{45}$$

where

$$\begin{aligned} \hat{H}_n^{(1)}(\mu_\theta, \sigma_\theta^2) &= \frac{1}{n}\sum_{i=1}^{n} \tau_{b,i}(\mu_\theta, \sigma_\theta^2) \left\{ \frac{\ddot{p}_{\Delta,i}(\mu_\theta, \sigma_\theta^2)}{p_{\Delta,i}(\mu_\theta, \sigma_\theta^2)} + \frac{\dot{p}_{\Delta,i}(\mu_\theta, \sigma_\theta^2)}{p_{\Delta,i}(\mu_\theta, \sigma_\theta^2)} \frac{\dot{p}_{\Delta,i}(\mu_\theta, \sigma_\theta^2)'}{p_{\Delta,i}(\mu_\theta, \sigma_\theta^2)} \right\}, \\ \hat{H}_n^{(2)}(\mu_\theta, \sigma_\theta^2) &= \frac{1}{n}\sum_{i=1}^{n} \tau_{b,i}^{(1)}(\mu, \sigma^2) \left\{ \frac{\dot{p}_{\Delta,i}(\mu_\theta, \sigma_\theta^2)\, \dot{p}_{\Delta,i}(\mu_\theta, \sigma_\theta^2)'}{p_{\Delta,i}(\mu_\theta, \sigma_\theta^2)} + \log p_{\Delta,i}(\mu_\theta, \sigma_\theta^2)\, \ddot{p}_i(\mu_\theta, \sigma_\theta^2) \right\} \\ \hat{H}_n^{(3)}(\mu_\theta, \sigma_\theta^2) &= \frac{1}{n}\sum_{i=1}^{n} \tau_{ab,i}^{(2)}(\mu_\theta, \sigma_\theta^2) \log p_{\Delta,i}(\mu_\theta, \sigma_\theta^2)\, \dot{p}_{\Delta,i}(\mu_\theta, \sigma_\theta^2)\, \dot{p}_{\Delta,i}(\mu_\theta, \sigma_\theta^2)'. \end{aligned} \tag{46}$$

Finally, the pathwise derivatives of the trimmed and untrimmed score w.r.t. μ and σ^2 at (μ_0, σ_0^2) are given by:

$$\begin{aligned} \nabla \hat{S}_n[d\mu, d\sigma^2] &= \frac{1}{n}\sum_{i=1}^{n} \tau_{b,i} \left\{ \frac{\nabla \dot{p}_{\Delta,i}[d\mu, d\sigma^2]}{p_{0,\Delta,i}} - \frac{\dot{p}_{0,\Delta,i}}{p_{0,\Delta,i}} \frac{\nabla p_{\Delta,i}[d\mu, d\sigma^2]}{p_{0,\Delta,i}} \right\} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \tau_{b,i}^{(1)} \left\{ \frac{\nabla p_{\Delta,i}[d\mu, d\sigma^2]}{p_{0,\Delta,i}} \dot{p}_{0,\Delta,i} + \log p_{0,\Delta,i}(\mu_0) \nabla \dot{p}_{\Delta,i}[d\mu, d\sigma^2] \right\} \\ &=: \nabla \hat{S}_n^{(1)}[d\mu, d\sigma^2] + \nabla \hat{S}_n^{(2)}[d\mu, d\sigma^2], \end{aligned} \tag{47}$$

$$\nabla S_n\left[d\mu,d\sigma^2\right] = -\frac{1}{n}\sum_{i=1}^{n}\frac{\dot{p}_{0,\Delta,i}}{p_{0,\Delta,i}}\frac{\nabla p_{\Delta,i}\left[d\mu,d\sigma^2\right]}{p_{0,\Delta,i}}+\frac{1}{n}\sum_{i=1}^{n}\frac{\nabla\dot{p}_{\Delta,i}\left[d\mu,d\sigma^2\right]}{p_{0,\Delta,i}} \tag{48}$$

$$= :\nabla S_n^{(1)}\left[d\mu,d\sigma^2\right]+\nabla S_n^{(2)}\left[d\mu,d\sigma^2\right]. \tag{49}$$

where $p_{0,t}(y|x)=p_t\left(y|x;\mu_0,\sigma_0^2\right)$. Let in the following $\|\cdot\|_\infty$ denote the sup-norm.

Lemma 6 *Assume that the model characterized by $\left(\mu_\theta,\sigma_\theta^2\right)$ satisfies the conditions in (A1)-(A5), and that $\left(\hat{\mu}_\theta,\hat{\sigma}_\theta^2\right)\in\mathcal{H}$ for all $\theta\in\Theta$ with $\sup_{\theta\in\Theta}||\hat{\mu}_\theta-\bar{\mu}_\theta||_\infty=O_P(\varepsilon_n)$ and $\sup_{\theta\in\Theta}||\hat{\sigma}_\theta^2-\bar{\sigma}_\theta^2||_\infty=O_P(\varepsilon_n)$ where $\varepsilon_n\to0$ and $\bar{\mu}$ and $\bar{\sigma}^2$ are given in (29). Then with $L_n\left(\mu,\sigma^2\right)$ and $\hat{L}_n\left(\hat{\mu}_\theta,\hat{\sigma}_\theta^2\right)$ given in (6) and (13), the following hold:*

$$\sup_{\theta\in\Theta}\left|\hat{L}_n\left(\hat{\mu}_\theta,\hat{\sigma}_\theta^2\right)-L_n\left(\mu_\theta,\sigma_\theta^2\right)\right|=O_P\left(b^{-1}\varepsilon_n\right)+O_P\left(b^{-1}a^{\delta_1/2}\right)+O_P\left(|\log b|^{-\delta_3}\right).$$

Proof. Let $p_{\Delta,0}(y|x)=p_\Delta(y|x;\mu_0,\sigma_0^2)$ where $\mu_0=\mu_{\theta_0}$ and $\sigma_0^2=\sigma_{\theta_0}^2$. In the following we suppress the dependence on θ. We introduce two auxiliary trimming sets, $A(\eta)=\left\{p\left(y|x;\mu,\sigma^2\right)\geq\eta b_1\right\}$ and $A_0(\eta)=\{p_{\Delta,0}(y|x)\geq\eta b_1\}$, for any $\eta>0$. Defining $\bar{A}(\eta)=A(\eta)\cap A_0(\eta)$, it follows by Lemma 5, using the same arguments as in Andrews (1995, p.588), that $A_0(1/2)\subseteq A(1)\subseteq A_0(2)\subseteq\bar{A}(4)$ for $\varepsilon_n\to0$ sufficiently small. Since $|\log(p/p_0)|\leq|p-p_0|/p_0+|p-p_0|/p$, it holds that,

$$\begin{aligned}
& |\hat{L}_n\left(\hat{\mu},\hat{\sigma}^2\right)-\hat{L}_n\left(\bar{\mu},\bar{\sigma}^2\right)| \\
\leq\ & \sup_{x,y\in I}|\mathbb{I}_{\bar{A}(4)}(y,x)\left|\log p_\Delta(y|x;\hat{\mu},\hat{\sigma}^2)-\log p_\Delta(y|x;\bar{\mu},\bar{\sigma}^2)\right| \\
\leq\ & \sup_{x,y\in I}\left\{\mathbb{I}_{\bar{A}(4)}\left|\frac{p_\Delta(y|x;\hat{\mu},\hat{\sigma}^2)-p_\Delta(y|x;\bar{\mu},\bar{\sigma}^2)}{p_\Delta(y|x;\hat{\mu},\hat{\sigma}^2)}\right|\right\}+\sup_{x,y\in I}\left\{\mathbb{I}_{\bar{A}(4)}\left|\frac{p_\Delta(y|x;\hat{\mu},\hat{\sigma}^2)-p_\Delta(y|x;\bar{\mu},\bar{\sigma}^2)}{p_\Delta(y|x;\bar{\mu},\bar{\sigma}^2)}\right|\right\} \\
\leq\ & (4b)^{-1}\sup_{x,y\in I}\left|p_\Delta(y|x;\hat{\mu},\hat{\sigma}^2)-p_\Delta(y|x;\mu,\sigma^2)\right|,
\end{aligned}$$

Thus,

$$\begin{aligned}
& E\left[\sup_{\theta\in\Theta}\left|\hat{L}_n\left(\bar{\mu},\bar{\sigma}^2\right)-\hat{L}_n\left(\mu,\sigma^2\right)\right|\right] \\
\leq\ & b^{-1}\int_{I\times I}\sup_{\theta\in\Theta}\left|p_\Delta\left(y|x;\bar{\mu},\bar{\sigma}^2\right)-p_\Delta\left(y|x;\mu,\sigma^2\right)\right|p_\Delta\left(y|x;\mu,\sigma^2\right)\pi_0(x)\,dydx,
\end{aligned}$$

where, by Lemma 5, uniformly in $\theta\in\Theta$,

$$\begin{aligned}
\left|p_\Delta(y|x;\hat{\mu},\hat{\sigma}^2)-p_\Delta(y|x;\bar{\mu},\bar{\sigma}^2)\right| &\leq C\left(\int_I\left(\left|\hat{\mu}(w)-\bar{\mu}(w)\right|^2+\left|\hat{\sigma}^2(w)-\bar{\sigma}^2(w)\right|^2\right)\bar{\gamma}(w|x)\,dw\right)^{1/2} \\
&\leq C\sup_{w\in I}\left\{\left|\hat{\mu}(w)-\bar{\mu}(w)\right|+\left|\hat{\sigma}^2(w)-\bar{\sigma}^2(w)\right|\right\}.
\end{aligned}$$

Using Eq. (39) and Cauchy-Schwarz's inequality,

$$\begin{aligned}
& \left|p_\Delta\left(y|x;\bar{\mu},\bar{\sigma}^2\right)-p_\Delta\left(y|x;\mu,\sigma^2\right)\right| \qquad (50)\\
\leq & \int_0^\Delta\int_I\left|\tau_a\left(\pi_0\left(w\right)\right)-1\right|\mu\left(w\right)\left|\frac{\partial p_t\left(y|w;\bar{\mu},\bar{\sigma}^2\right)}{\partial w}\right|p_t\left(w|x;\mu,\sigma^2\right)dydt \\
& +\int_0^\Delta\int_I\left|\tau_a\left(\pi_0\left(w\right)\right)-1\right|\sigma^2\left(w\right)\left|\frac{\partial p_t\left(y|w;\bar{\mu},\bar{\sigma}^2\right)}{\partial w^2}\right|p_t\left(w|x;\mu,\sigma^2\right)dydt \\
& +\underline{\sigma}^2\int_0^\Delta\int_I\left|\tau_a\left(\pi_0\left(w\right)\right)-1\right|\left|\frac{\partial p_t\left(y|w;\bar{\mu},\bar{\sigma}^2\right)}{\partial w^2}\right|p_t\left(w|x;\mu,\sigma^2\right)dydt \\
\leq & C_a^{1/2}\left(y|x\right)\left(\int_0^\Delta\int_I\left|\tau_a\left(\pi_0\left(w\right)\right)-1\right|^2p_t\left(w|x;\mu,\sigma^2\right)dwdt\right)^{1/2},
\end{aligned}$$

where

$$\begin{aligned}
C_a\left(y|x\right) = & \int_0^\Delta\int_I\mu^2\left(w\right)\left|\frac{\partial p_t\left(y|w;\bar{\mu},\bar{\sigma}^2\right)}{\partial y}\right|^2p_t\left(w|x;\mu,\sigma^2\right)dwdt \\
& +\int_0^\Delta\int_I\sigma^4\left(w\right)\left|\frac{\partial p_t\left(y|w;\bar{\mu},\bar{\sigma}^2\right)}{\partial y^2}\right|^2p_t\left(y|x;\mu,\sigma^2\right)dwdt \\
& +\underline{\sigma}^4\int_0^\Delta\int_I\left|\frac{\partial p_t\left(y|w;\bar{\mu},\bar{\sigma}^2\right)}{\partial y^2}\right|^2p_t\left(y|x;\mu,\sigma^2\right)dwdt
\end{aligned}$$

As $a\to0$, $C_a\left(x\right)\to C_0\left(x\right)$ where

$$\begin{aligned}
E\left[\sup_{\theta\in\Theta}C_0\left(X_\Delta|X_0\right)\right] \leq & C\left(\int_0^\Delta\int_{I\times I}\left\{\sup_{\theta\in\Theta}\mu^2\left(y\right)\right\}p_t\left(y|x;\mu_0,\sigma_0^2\right)\pi_0\left(x\right)dydxdt\right)^{1/2} \\
& +C\left(\int_0^\Delta\int_I\left\{\sup_{\theta\in\Theta}\sigma^4\left(y\right)\right\}p_t\left(y|x;\mu,\sigma^2\right)\pi_0\left(x\right)dydxdt\right)^{1/2}+C\underline{\sigma}^4 \\
\leq & C\left(\int_I\sup_{\theta\in\Theta}\left\{\mu^2\left(y\right)+\sigma^4\left(y\right)\right\}\pi_0\left(x\right)dx\right)^{1/2}+C\underline{\sigma}^4 \\
\leq & C\sqrt{1+E[B^2\left(X_0\right)]}<\infty,
\end{aligned}$$

where B is given in (A3), while

$$\begin{aligned}
\int_0^\Delta\int_I\left|\tau_a\left(\pi_0\left(w\right)\right)-1\right|^2p_t\left(w|x;\mu,\sigma^2\right)dwdt \leq & \int_0^\Delta\int_I\mathbb{I}\left\{\pi_0\left(w\right)\leq a_2\right\}p_t\left(w|x;\mu,\sigma^2\right)dwdt \\
\leq & a^\delta\int_0^\Delta\int_I\pi_0\left(y\right)^{-\delta}p_t\left(y|x;\mu,\sigma^2\right)dydt.
\end{aligned}$$

In total,

$$
\begin{aligned}
& E\left[\sup_{\theta\in\Theta}\left|\hat{L}_n\left(\bar{\mu},\bar{\sigma}^2\right)-\hat{L}_n\left(\mu,\sigma^2\right)\right|\right] \\
\leq\ & b^{-1}\int\sup_{\theta\in\Theta}C_a^{1/2}(x)\left(a^{\delta_1}\int_0^{\Delta}\int_I\pi_0(y)^{-\delta_1}p_t\left(y|x;\mu,\sigma^2\right)dydt\right)^{1/2}p_0(y|x)\,\pi_0(x)\,dydx \\
\leq\ & b^{-1}a^{\delta_1/2}\left(\int_I\int_I\left\{\sup_{\theta\in\Theta}C_a(x)\right\}p_0(y|x)\,\pi_0(x)\,dydx\right)^{1/2} \\
& \times\left(\int_I\int_I\pi_0(y)^{-\delta_1}\left\{\sup_{\theta\in\Theta}p_\Delta\left(y|x;\mu,\sigma^2\right)\right\}p_{0,\Delta}(y|x)\,\pi_0(x)\,dydx\right)^{1/2} \\
\leq\ & Cb^{-1}a^{\delta_1/2}\left(\int_I\pi_0(x)^{1-\delta_1}dx\right)^{1/2}.
\end{aligned}
$$

Finally, use that as $b\to 0$,

$$
\begin{aligned}
& E\left[\sup_{\theta\in\Theta}\left|\hat{L}_n\left(\mu,\sigma^2\right)-L_n\left(\mu,\sigma^2\right)\right|\right] \\
\leq\ & \int_{I\times I}\sup_{\theta\in\Theta}\mathbb{I}\left\{\frac{\left|\log p_\Delta(y|x;\mu,\sigma^2)\right|^{\delta_3}}{\left|\log(b)\right|^{\delta_3}}>1\right\}\left|\log p_\Delta\left(y|x;\mu,\sigma^2\right)\right|p_{\Delta,0}(y|x)\,\pi_0(x)\,dydx \\
\leq\ & \left|\log(b)\right|^{-\delta_3}\int_{I\times I}\sup_{\theta\in\Theta}\left|\log p_\Delta\left(y|x;\mu,\sigma^2\right)\right|^{1+\delta_3}p_{\Delta,0}(y|x)\,\pi_0(x)\,dydx
\end{aligned}
$$

where $E\left[q^{-\delta_3}\left(X_{i\Delta}|X_{(i-1)\Delta}\right)\right]<\infty$ by (A4.ii). ■

Lemma 7 *Assume that the conditions of Lemma 6 hold. Then with* $\hat{S}_n\left(\mu_\theta,\sigma_\theta^2\right)$ *and* $\nabla\hat{S}_n$ *given in (42)-(43) and (47), the following hold:*

$$
\left\|\hat{S}_n\left(\hat{\mu}_{\theta_0},\sigma_{\theta_0}^2\right)-\hat{S}_n\left(\bar{\mu}_{\theta_0},\bar{\sigma}_{\theta_0}^2\right)-\nabla\hat{S}_n\left[\hat{\mu}_{\theta_0}-\bar{\mu}_{\theta_0},\hat{\sigma}_{\theta_0}^2-\bar{\sigma}_{\theta_0}^2\right]\right\|=O_P\left(b^{-2}\varepsilon^2\right),
$$

Proof. In the following, let $p_{\Delta,i}\left(\mu,\sigma^2\right)=p_\Delta\left(X_{i\Delta}|X_{(i-1)\Delta};\mu,\sigma^2\right)$, $(\hat{\mu}_0,\hat{\sigma}_0^2)=(\hat{\mu}_{\theta_0},\hat{\sigma}_{\theta_0}^2)$, $(\bar{\mu}_0,\bar{\sigma}_0^2)=(\bar{\mu}_{\theta_0},\bar{\sigma}_{\theta_0}^2)$, $(\mu_0,\sigma_0^2)=(\mu_{\theta_0},\sigma_{\theta_0}^2)$, $\tau_b(x,y)=\tau_b\left(p_{0,\Delta}(x,y)\right)$ and $\tau_{b,i}=\tau_b\left(X_{(i-1)\Delta},X_{(i-1)\Delta}\right)$. Given the definitions in Eqs. (42)-(43) and (47), we obtain that

$$
\left\|\hat{S}_n\left(\hat{\mu}_0,\hat{\sigma}_0^2\right)-\hat{S}_n\left(\bar{\mu}_0,\bar{\sigma}_0^2\right)-\nabla\hat{S}_n\left[\hat{\mu}_0-\bar{\mu}_0,\hat{\sigma}_0^2-\bar{\sigma}_0^2\right]\right\|\leq\|A_1\|+\|A_2\|+\|A_3\|,
$$

where

$$
\begin{aligned}
A_1 \;=\; & \frac{1}{n}\sum_{i=1}^n\bar{\tau}_{b,i}\left(\frac{\dot{p}_{\Delta,i}(\hat{\mu}_0,\hat{\sigma}_0^2)}{p_{\Delta,i}(\hat{\mu}_0,\hat{\sigma}_0^2)}-\frac{\dot{p}_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)}{p_{0,\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)}-\frac{\dot{p}_{\Delta,i}(\hat{\mu}_0,\hat{\sigma}_0^2)-\dot{p}_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)}{p_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)}\right. \\
& \left.+\frac{\dot{p}_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)}{p_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)}\frac{p_{\Delta,i}(\hat{\mu}_0,\hat{\sigma}_0^2)-p_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)}{p_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)}\right),
\end{aligned}
$$

$$
\begin{aligned}
A_2 &= \frac{1}{n}\sum_{i=1}^{n}\bar{\tau}_{b,i}\frac{1}{p_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)}\left(\dot{p}_{\Delta,i}(\hat{\mu}_0,\hat{\sigma}_0^2)-\dot{p}_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)-\nabla\dot{p}_{\Delta,i}\left[\hat{\mu}_0-\bar{\mu}_0,\hat{\sigma}_0^2-\bar{\sigma}_0^2\right]\right),\\
A_3 &= \frac{1}{n}\sum_{i=1}^{n}\bar{\tau}_{b,i}\frac{\dot{p}_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)}{p_{\Delta,i}^2(\bar{\mu}_0,\bar{\sigma}_0^2)}\left(p_{\Delta,i}(\hat{\mu}_0,\hat{\sigma}_0^2)-p_{\Delta,i}(\bar{\mu}_0,\bar{\sigma}_0^2)-\nabla p_{\Delta,i}\left[\hat{\mu}_0-\bar{\mu}_0,\hat{\sigma}_0^2-\bar{\sigma}_0^2\right]\right).
\end{aligned}
$$

Next, it is easily seen that

$$
\begin{aligned}
\|A_1\| &\leq Cb^{-1}\left\|\dot{p}_\Delta(\hat{\mu}_0,\hat{\sigma}_0^2)-\dot{p}_\Delta(\bar{\mu}_0,\bar{\sigma}_0^2)\right\|_\infty^2+b^{-2}\left\|p_\Delta(\hat{\mu}_0,\hat{\sigma}_0^2)-p_\Delta(\bar{\mu}_0,\bar{\sigma}_0^2)\right\|_\infty^2\\
&\leq Cb^{-2}\left(\|\hat{\mu}-\bar{\mu}\|_\infty^2+\left\|\hat{\sigma}^2-\bar{\sigma}^2\right\|_\infty^2\right),
\end{aligned}
$$

where the second inequality follows from Lemma 5. Similarly,

$$
\|A_2\|\leq Cb^{-1}\left(\|\hat{\mu}-\bar{\mu}\|_\infty^2+\left\|\hat{\sigma}^2-\bar{\sigma}^2\right\|_\infty^2\right),\quad \|A_3\|\leq Cb^{-2}\left(\|\hat{\mu}-\bar{\mu}\|_\infty^2+\left\|\hat{\sigma}^2-\bar{\sigma}^2\right\|_\infty^2\right),
$$

where we again appeal to Lemma 5. ■

Lemma 8 *Assume that the conditions of Lemma 6 hold. Then with $H_n\left(\mu_\theta,\sigma_\theta^2\right)$ and $\hat{H}_n\left(\hat{\mu}_\theta,\hat{\sigma}_\theta^2\right)$ given in (44) and (45), the following hold:*

$$
\sup_{\theta\in\Theta}\left\|\hat{H}_n\left(\hat{\mu}_\theta,\hat{\sigma}_\theta^2\right)-H_n\left(\mu_\theta,\sigma_\theta^2\right)\right\|=O_P\left(b^{-2}\varepsilon_n\right)+O_P\left(b^{-2}a^{\delta_1/2}\right)+O_P\left(|\log b|^{-\delta_3}\right).
$$

Proof. We use the same notation and strategy as in the proof of Lemma 7. Recall the definition of $\hat{H}_n\left(\mu,\sigma^2\right)$ in Eqs. (45)-(46). The first term satisfies

$$
\begin{aligned}
&\left\|\hat{H}_n^{(1)}\left(\hat{\mu},\hat{\sigma}^2\right)-\hat{H}_n^{(1)}\left(\bar{\mu},\bar{\sigma}^2\right)\right\|\\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\left\|\tau_{a,i}\left(\hat{\mu},\hat{\sigma}^2\right)\frac{\ddot{p}_{\Delta,i}\left(\hat{\mu},\hat{\sigma}^2\right)}{p_{\Delta,i}\left(\hat{\mu},\hat{\sigma}^2\right)}-\tau_{a,i}\left(\mu,\sigma^2\right)\frac{\ddot{p}_{\Delta,i}\left(\bar{\mu},\bar{\sigma}^2\right)}{p_{\Delta,i}\left(\bar{\mu},\bar{\sigma}^2\right)}\right\|\\
&+\frac{1}{n}\sum_{i=1}^{n}\left\|\tau_{a,i}\left(\hat{\mu},\hat{\sigma}^2\right)\frac{\dot{p}_{\Delta,i}\left(\hat{\mu},\hat{\sigma}^2\right)}{p_{\Delta,i}\left(\hat{\mu},\hat{\sigma}^2\right)}\frac{\dot{p}'_{\Delta,i}\left(\hat{\mu},\hat{\sigma}^2\right)}{p_{\Delta,i}\left(\hat{\mu},\hat{\sigma}^2\right)}-\tau_{a,i}\left(\mu,\sigma^2\right)\frac{\dot{p}_{\Delta,i}\left(\bar{\mu},\bar{\sigma}^2\right)}{p_{\Delta,i}\left(\bar{\mu},\bar{\sigma}^2\right)}\frac{\dot{p}'_{\Delta,i}\left(\bar{\mu},\bar{\sigma}^2\right)}{p_{\Delta,i}\left(\bar{\mu},\bar{\sigma}^2\right)}\right\|\\
\leq\ & b^{-2}\left\{\left\|\ddot{p}_\Delta\left(\hat{\mu},\hat{\sigma}^2\right)-\ddot{p}_\Delta\left(\bar{\mu},\bar{\sigma}^2\right)\right\|_\infty+\left\|\dot{p}_\Delta\left(\hat{\mu},\hat{\sigma}^2\right)-\dot{p}_\Delta\left(\bar{\mu},\bar{\sigma}^2\right)\right\|_\infty+\left\|p_\Delta\left(\hat{\mu},\hat{\sigma}^2\right)-p_\Delta\left(\bar{\mu},\bar{\sigma}^2\right)\right\|_\infty\right\}\\
=\ & O_P\left(b^{-2}\varepsilon_n\right),
\end{aligned}
$$

uniformly over θ. The proofs of $\sup_\theta\left\|\hat{H}_n^{(k)}\left(\hat{\mu},\hat{\sigma}^2\right)-\hat{H}_n^{(k)}\left(\bar{\mu},\bar{\sigma}^2\right)\right\|=O_P\left(b_1^{-2}\varepsilon_n\right)$, $k=2,3$, follow along the same lines. Next, by similar arguments

$$
\begin{aligned}
& E\left[\sup_{\theta}\left\|\hat{H}_n\left(\bar{\mu},\bar{\sigma}^2\right)-\hat{H}_n\left(\mu,\sigma^2\right)\right\|\right] \\
\leq\ & b^{-1}\int_{I\times I}\sup_{\theta}\left|p_{\Delta}\left(y|x;\bar{\mu}_0,\bar{\sigma}_0^2\right)-p_{0,\Delta}\left(y|x;\mu_0,\sigma_0^2\right)\right|p_{0,\Delta}\left(y|x;\mu_0,\sigma_0^2\right)\pi_0\left(x\right)dydx \\
& +b^{-1}\int_{I\times I}\sup_{\theta}\left|\dot{p}_{\Delta}\left(y|x;\bar{\mu}_0,\bar{\sigma}_0^2\right)-\dot{p}_{\Delta}\left(y|x;\mu_0,\sigma_0^2\right)\right|p_{0,\Delta}\left(y|x;\mu_0,\sigma_0^2\right)\pi_0\left(x\right)dydx \\
& +b^{-1}\int_{I\times I}\sup_{\theta}\left|\ddot{p}_{\Delta}\left(y|x;\bar{\mu}_0,\bar{\sigma}_0^2\right)-\ddot{p}_{\Delta}\left(y|x;\mu_0,\sigma_0^2\right)\right|p_{0,\Delta}\left(y|x;\mu_0,\sigma_0^2\right)\pi_0\left(x\right)dydx.
\end{aligned}
$$

Applying the inequality in Eq. (50) and the bounds established after that, we obtain that the third term is $O\left(b^{-1}a^{\delta_1/2}\right)$. For the second term, we first note that by Eq. (40),

$$
\begin{aligned}
& \left|\dot{p}_{\Delta}\left(y|x;\bar{\mu}_0,\bar{\sigma}_0^2\right)-\dot{p}_{\Delta}\left(y|x;\mu_0,\sigma_0^2\right)\right| \\
\leq\ & \int_0^{\Delta}\int_I\left|\tau_a\left(\pi_0\left(y\right)\right)-1\right|\left|\partial_{\theta}\mu_0\left(w\right)\right|\frac{\partial p_s\left(y|w;\bar{\mu}_0,\bar{\sigma}_0^2\right)}{\partial w}p_s\left(w|x;\mu_0,\sigma_0^2\right)dwds \\
& +\int_0^{\Delta}\int_I\left|\tau_a\left(\pi_0\left(y\right)\right)-1\right|\left|\mu_0\left(w\right)\right|\frac{\partial \dot{p}_s\left(y|w;\bar{\mu}_0,\bar{\sigma}_0^2\right)}{\partial w}p_s\left(w|x;\mu_0,\sigma_0^2\right)dwds \\
& +\frac{1}{2}\int_0^{\Delta}\int_I\left|\tau_a\left(\pi_0\left(y\right)\right)-1\right|\left|\partial_{\theta}\sigma_0^2\left(w\right)\right|\frac{\partial^2 p_s\left(y|w;\bar{\mu}_0,\bar{\sigma}_0^2\right)}{\partial w^2}p_s\left(w|x;\mu_0,\sigma_0^2\right)dwds, \\
& +\frac{1}{2}\int_0^{\Delta}\int_I\left|\tau_a\left(\pi_0\left(y\right)\right)-1\right|\left|\sigma_0^2\left(w\right)\right|\frac{\partial^2 \dot{p}_s\left(y|w;\bar{\mu}_0,\bar{\sigma}_0^2\right)}{\partial w^2}p_s\left(w|x;\mu_0,\sigma_0^2\right)dwds.
\end{aligned}
$$

We can now use the same arguments as those following Eq. (50) to show that the second term is of order $O\left(b^{-1}a^{\delta_1/2}\right)$. The same arguments are employed to show that the first term is also $O\left(b^{-1}a^{\delta_1/2}\right)$. Finally

$$
\begin{aligned}
& E\left[\sup_{\theta}\left\|\hat{H}_n\left(\mu,\sigma^2\right)-H_n\left(\mu,\sigma^2\right)\right\|\right] \\
\leq\ & \int_{I^2}\sup_{\theta}\left|\tau_b\left(x,y\right)-1\right|\left\{\frac{\left\|\ddot{p}_{\Delta}\left(y|x\right)\right\|}{p_{\Delta}\left(y|x\right)}+\frac{\left\|\dot{p}_{\Delta}\left(y|x\right)\right\|^2}{p_{\Delta}^2\left(y|x\right)}\right\}p_{0,\Delta}\left(y|x\right)\pi_0\left(x\right)dxdy \\
& +\int_{I^2}\sup_{\theta}\left|\tau_b^{(1)}\left(x,y\right)\right|\left\{\frac{\left\|\dot{p}_{\Delta,i}\left(\mu,\sigma^2\right)\right\|^2}{p_{\Delta,i}\left(\mu,\sigma^2\right)}+\log p_{\Delta,i}\left(\mu,\sigma^2\right)\left\|\ddot{p}_i\left(\mu,\sigma^2\right)\right\|\right\}p_{0,\Delta}\left(y|x\right)\pi_0\left(x\right)dxdy \\
& +\int_{I^2}\sup_{\theta}\tau_b^{(2)}\left(x,y\right)\left|\log p_{\Delta,i}\left(\mu,\sigma^2\right)\right|\left\|\dot{p}_{\Delta,i}\left(\mu,\sigma^2\right)\right\|^2p_{0,\Delta}\left(y|x\right)\pi_0\left(x\right)dxdy
\end{aligned}
$$

where

$$\begin{aligned}
& \int_I |\tau_b(x,y)-1| \frac{\|\dot{p}_{0,\Delta}(y|x)\|}{p_{0,\Delta}(y|x)} p_{0,\Delta}(y|x)\pi_0(x)\,dxdy \\
\leq & \left(\int_{I\times I} \mathbb{I}\left\{\frac{|\log p_{0,\Delta}(y|x)|}{|\log b|^{\delta_3}}>1\right\} p_{0,\Delta}(y|x)\pi_0(x)\,dxdy\right)^{1/2} \left(\int_{I\times I} \frac{\|\dot{p}_{0,\Delta}(y|x)\|^2}{p^2_{0,\Delta}(y|x)} p_{0,\Delta}(y|x)\pi_0(x)\,dxdy\right)^{1/2} \\
\leq & |\log b|^{-\delta_3}\sqrt{E\left[q\left(X_\Delta|X_0\right)\right]}\times\sqrt{E\left[\|s_0\left(X_\Delta|X_0\right)\|^2\right]},
\end{aligned}$$

and

$$\begin{aligned}
& \int_I \sup_\theta \left|\tau_b^{(1)}(x,y)\right| \left\{\frac{\left\|\dot{p}_\Delta\left(\mu,\sigma^2\right)\right\|^2}{p_\Delta\left(\mu,\sigma^2\right)} + \log p_\Delta\left(\mu,\sigma^2\right)\left\|\ddot{p}\left(\mu,\sigma^2\right)\right\|\right\} p_{0,\Delta}(y|x)\pi_0(x)\,dxdy \\
\leq & \; Cb\left\{E\left[|\log p_\Delta\left(X_\Delta|X_0\right)|\right] + E\left[\|s\left(X_\Delta|X_0\right)\|^2\right]\right\}.
\end{aligned}$$

Similarly, the third term can be shown to be $O_P\left(b^2\right)$. ■

Lemma 9 (Class 1) *Under (A1)-(A3), uniformly in $\theta_1\in\Theta_1$ and for $k=0,1,2$,*

$$\left\|\partial^k_{\theta_1}\hat{\mu}\left(\cdot;\theta_1,\hat{\pi}\right)-\partial^k_{\theta_q}\bar{\mu}\left(\cdot;\theta_1\right)\right\|_\infty = O_P\left(a^{-3}\sqrt{\log(n)}\,(nh)^{-1/2}+a^{-2}\sqrt{\log(n)}\left(nh^3\right)^{-1/2}+a^{-2}h^m\right),$$

$$\left\|\partial^k_{\theta_1}\hat{\mu}\left(\cdot;\theta_{1,0}\right)-\partial^k_{\theta_1}\bar{\mu}\left(\cdot;\theta_{1,0}\right)-\partial^k_{\theta_1}\nabla\bar{\mu}\left[\hat{\pi}-\pi_0\right]\right\|_\infty = O_P\left(a^{-3}\log(n)\,(nh)^{-1}+a^{-2}\log(n)\left(nh^3\right)^{-1}+a^{-2}h^{2m}\right),$$

where $\bar{\mu}(x;\theta_1)$ is defined in Eq. (29) and

$$\nabla\bar{\mu}[d\pi](x)=\frac{\tau_a\left(\pi_0(x)\right)}{2}\sigma^2\left(x;\theta_{1,0}\right)\left[\frac{d\pi^{(1)}(x)}{\pi_0(x)}-\frac{\pi_0^{(1)}(x)}{\pi_0^2(x)}d\pi(x)\right],$$

Proof. Using standard techniques, we obtain under (A1)-(A3) and $K\in\mathcal{K}(m)$, that for some $\bar{h}\in[0,h]$,

$$E[\hat{\pi}^{(s)}(x)]-\pi_0^{(s)}(x)=\frac{h^m}{m!}\int_I \pi_0^{(s)}\left(x+z\bar{h}\right)z^m K(z)\,dz=O\left(h^m\right),\quad s=0,1,$$

uniformly in $x\in I$. Next, we define

$$\hat{G}_s(x)=\frac{1}{nh}\sum_{i=1}^n G_s\left(\frac{x-X_i}{h}\right),\quad G_s(z)=K^{(s)}(z).$$

It is easily checked that with $K\in\mathcal{K}(m)$, G_s satisfies Assumption 1 in Hansen (2008) and that

(A1) implies his Assumption 2. We then obtain from Hansen (2008, Proof of Theorem 3) that

$$
\begin{aligned}
E\left[\sup_{x\in I}|\hat{\pi}^{(s)}(x)-\pi_0^{(s)}(x)|^2\right] &\leq \sup_{x\in I}|E[\hat{\pi}^{(s)}(x)]-\pi_0^{(s)}(x)|^2+h^{-2s}E\left[\sup_{x\in I}|\hat{G}(x)-E[\hat{G}(x)]|^2\right]\\
&= O\left(h^{2m}\right)+O_P\left(\log(n)\,n^{-1}h^{-(1+2s)}\right).
\end{aligned}
$$

Since $||\hat{\pi}-\pi_0|| \to^P 0$, we can as in the Proof of Lemma 9 appeal to Andrews (1995, p.588), and show uniform convergence over the set $\bar{A}=\{x:\hat{\pi}(x)\geq a/4,\pi_0(x)\geq a/4\}$. We then use the convergence rate obtained above together with $\pi_0(x)\sigma^2(x;\theta_1)\leq C$ uniformly in (x,θ_1) by (A3) to obtain uniformly in $\theta_1\in\Theta_1$,

$$
\begin{aligned}
\|\hat{\mu}(\cdot;\theta_1)-\bar{\mu}(\cdot;\theta_1)\|_\infty &\leq \frac{1}{2}\sup_{x\in I}\mathbb{I}\{x\in\bar{A}\}\pi_0^{-1}(x)\left\{\pi_0(x)\sigma^2(x;\theta_1)\right\}\left|\frac{\hat{\pi}^{(1)}(x)}{\hat{\pi}(x)}-\frac{\pi_0^{(1)}(x)}{\pi_0(x)}\right|\\
&\leq C\left(a^{-2}||\hat{\pi}^{(1)}-\pi^{(1)}||_\infty+a^{-3}\|\hat{\pi}-\pi\|_\infty\right)\\
&= O\left(a^{-2}\sqrt{\log(n)}n^{-1/2}h^{-3/2}+a^{-3}\sqrt{\log(n)}n^{-1/2}h^{-1/2}+a^{-2}h^m\right).
\end{aligned}
$$

and similarly when we take derivatives w.r.t θ_1. Next, by the same arguments as before,

$$
\begin{aligned}
|\hat{\mu}(\cdot;\theta_{1,0})-\bar{\mu}(\cdot;\theta_{1,0})-\nabla\bar{\mu}[\hat{\pi}-\pi_0]| &\leq Ca^{-1}\mathbb{I}\{\bar{A}\}\left|\frac{\hat{\pi}^{(1)}}{\hat{\pi}}-\frac{\pi_0^{(1)}}{\pi_0}-\frac{\hat{\pi}^{(1)}-\pi_0^{(1)}}{\pi_0}+\frac{\pi_0^{(1)}}{\pi_0^2}[\hat{\pi}-\pi_0]\right|\\
&\leq C\left(a^{-2}||\hat{\pi}^{(1)}-\pi^{(1)}||_\infty+a^{-3}\|\hat{\pi}-\pi\|_\infty\right),
\end{aligned}
$$

and the second result is obtained. ■

Lemma 10 (Class 2) *Under (A1)-(A3), uniformly in $\theta_2\in\Theta_2$ and for $k=0,1,2$,*

$$
\left\|\partial_{\theta_2}^k\hat{\sigma}^2(\cdot;\theta_2)-\partial_{\theta_2}^k\bar{\sigma}^2(\cdot;\theta_2)\right\|_\infty = O_P\left(a^{-2}\sqrt{\log(n)}\,(nh)^{-1/2}+a^{-2}h^m+a^q\right),
$$

$$
\left\|\partial_{\theta_2}^k\sigma^2(\cdot;\theta_{2,0})-\partial_{\theta_2}^k\bar{\sigma}^2(\cdot;\theta_{2,0})-\partial_{\theta_2}^k\nabla\bar{\sigma}^2[\hat{\pi}-\pi_0]\right\|_\infty = O_P\left(a^{-2}\log(n)\,(nh)^{-1}+a^{-2}h^{2m}\right),
$$

where $\bar{\sigma}^2(x;\theta_2)$ is defined in Eq. (29) and

$$
\nabla\sigma^2[d\pi](x)=2\tau_a(\pi_0(x))\left[\frac{\int_l^x d\pi(y)\,\mu(y;\theta_{2,0})\,dy}{\pi_0(x)}-2\frac{\int_l^x\pi_0(y)\,\mu(y;\theta_{2,0})\,dy}{\pi_0^2(x)}d\pi(x)\right].
$$

Proof. We claim that:

$$
\sup_{(x,\theta 2)\in I\times\Theta_2}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{I}\{X_i\leq x\}\,\partial_{\theta_2}^k\mu(X_i;\theta_2)-\int_l^x\pi_0(y)\,\partial_{\theta_2}^k\mu(y;\theta_2)\,dy\right| = O_P(n^{-1/2}).
$$

Observe that $\left|\mathbb{I}\{z\leq x\}\,\partial_{\theta_2}^k\mu(z;\theta_2)\right|\leq B(z)$ where $B(z)$ is given in (A3) with $E\left[B^{2+\delta}(X_0)\right]<\infty$. The result will now follow by applying Doukhan et al (1994) if: The ε-entropy with bracketing of $\mathcal{G}$ for the $L_q(\pi_0)$-metric, $H_{B,q}(\varepsilon,\mathcal{G},\pi_0)$, where $\mathcal{G}=\left\{g|g(z)=1_{(l,x)}(z)\,\partial_{\theta_2}^k\mu(z;\theta_2),\ (x,\theta_2)\in I\times\Theta_2\right\}$, satisfies $H_{B,q}(\varepsilon,\mathcal{G},\pi)\leq C\varepsilon^{-\rho}$ for some $\rho<1/2$ and $q>2$. . To prove this claim, define $\mathcal{F}=\{f|f(z)=\mathbb{I}\{z\leq x\},\ x\in I\}$, such that $\mathcal{G}=\left\{g|g(z;\theta_2,h)=f(z)\,\partial^k_{\theta_2}\mu(z;\theta_2),\ \ (\theta,h)\in\Theta_2\times\mathcal{F}\right\}$. It holds that for any $(\theta_2,f),\left(\theta_2',f'\right)\in\Theta_2\times\mathcal{F}$,

$$
\begin{aligned}
\left|g(z;\theta_2,f)-g\left(z;\theta_2',f'\right)\right| &\leq |\mu(z;\theta_2)|\left\|f-f'\right\|_q+\left\|f'\right\|_q|\partial^{k+1}_{\theta_2}\mu\left(z;\bar{\theta}_2\right)|\left\|\theta_2-\theta_2'\right\| \\
&\leq B(z)\left(\left\|f-f'\right\|_q+\left\|\theta_2-\theta_2'\right\|\right).
\end{aligned}
$$

By the same arguments as in the proof of Chen et al (2003, Theorem 3), it now follows that

$$
H_{B,q}(\varepsilon,\mathcal{G},\pi)\leq H(C\varepsilon^{1/s},\Theta_2,\|\cdot\|)+H_q(C\varepsilon^{1/s},\mathcal{F},\pi_0),
$$

for any $s\in(0,1]$, where $H(\varepsilon,\Theta_2,\|\cdot\|)$ and $H_q(\varepsilon,\mathcal{F},\pi_0)$ are the ε-entropies of Θ_2 (for the Euclidean norm) and $\mathcal{F}$ (for the $L_q(\pi_0)$-metric) respectively. By Van de Geer (2000, Lemma 2.5), $H(\varepsilon,\Theta_2,\|\cdot\|)\leq d\log\left(4C\varepsilon^{-1}+1\right)$ while, by Van de Geer (2000, Theorem 3.11 and Example 3.7.4a), $H_q(\varepsilon,\mathcal{F},\pi_0)\leq\log(C\varepsilon^{-q})$. Next, by standard bias arguments,

$$
\sup_{(x,\theta_2)\in I\times\Theta_2}\left|\int_l^x\hat{\pi}(y)\,\partial^k_{\theta_2}\mu(y;\theta_2)\,dy-\frac{1}{n}\sum_{i=1}^n\mathbb{I}\{X_i\leq x\}\,\partial^k_{\theta_2}\mu(X_i;\theta_2)\right|=O_P(h^m).
$$

We now use this together with the uniform result for the kernel estimator and the definition of the set $\bar{A}$ in the proof of Lemma 9 to obtain

$$
\begin{aligned}
\left|\hat{\sigma}^2(x;\theta_2)-\bar{\sigma}^2(x;\theta_2)\right| &\leq 2\mathbb{I}\left\{x\in\bar{A}\right\}\left|\frac{\int_l^x\hat{\pi}(y)\,\partial^k_{\theta_2}\mu(y;\theta_2)\,dy}{\hat{\pi}(x)}-\frac{\int_l^x\pi_0(y)\,\partial^k_{\theta_2}\mu(y;\theta_2)\,dy}{\pi_0(x)}\right| \\
&\leq C\sup_{\hat{\pi}(x),\pi_0(x)>a}\left|\frac{\hat{\pi}^{(1)}(x)}{\hat{\pi}(x)}-\frac{\pi_0^{(1)}(x)}{\pi_0(x)}\right|+O_P\left(a^{-1}n^{-1/2}\right)+O_P\left(a^{-1}h^m\right) \\
&\leq Ca^{-2}\left\|\hat{\pi}-\pi\right\|_\infty+O_P\left(a^{-1}n^{-1/2}\right)+O_P\left(a^{-1}h^m\right) \\
&= O\left(a^{-2}\sqrt{\log(n)}n^{-1/2}h^{-3/2}+a^{-2}h^m\right),
\end{aligned}
$$

and similar when we take derivatives w.r.t θ_2. Finally, the second result follows by:

$$
\begin{aligned}
&\left|\hat{\sigma}^2(x;\theta_{2,0})-\bar{\sigma}^2(x;\theta_{2,0})-\nabla\bar{\sigma}^2\left[\hat{\pi}-\pi_0\right](x)\right| \\
\leq\ & 2\mathbb{I}\left\{x\in\bar{A}\right\}\left|\int_l^x\pi_0(y)\,\partial^k_{\theta_2}\mu(y;\theta_{2,0})\,dy\right|\left|\frac{1}{\hat{\pi}(x)}-\frac{1}{\pi_0(x)}-\frac{\hat{\pi}(x)-\pi_0(x)}{\pi_0^2(x)}\right| \\
=\ & O\left(a^{-2}\left\|\hat{\pi}-\hat{\pi}\right\|_\infty^2\right),
\end{aligned}
$$

■

C Tables

The true parameter values are: (CKLS) $\alpha_0 = (1.8207, 2.6217)$ and $\beta_0 = (0.0344, -0.2921)$; (AS) $\alpha_0 = (1.8086, 2.6195)$ and $\beta_0 = (-0.3582, 6.6653, -35.4326, 0.0065)$.

	α_1	α_2	β_1	β_2
	$n = 5000$			
Parametric	-0.0109	-0.0044	0.0154	-0.2071
	0.2004	0.0421	0.0256	0.4004
	0.2007	0.0423	0.0298	0.4508
Semiparametric	-0.1506	0.0416	0.0160	-0.01180
	0.2890	0.0838	0.0630	0.3693
	0.3259	0.0935	0.0650	0.3877
	$n = 2500$			
Parametric	-0.0263	-0.0103	0.0331	-0.4358
	0.3197	0.0677	0.0437	0.6290
	0.3208	0.0685	0.0548	0.7652
Semiparametric	-0.1379	-0.0393	0.0396	-0.3268
	0.3402	0.0867	0.1604	0.5847
	0.3671	0.0952	0.1652	0.6698
	$n = 1000$			
Parametric	0.0690	-0.0097	0.0988	-1.1780
	0.6846	0.1391	0.1509	1.3476
	0.6881	0.1395	0.1803	1.7899
Semiparametric	-0.0277	-0.0327	0.1199	-0.9742
	0.6579	0.1423	0.5427	1.2096
	0.6584	0.1460	0.5558	1.5531

Table 1: Parametric and Semiparametric bias, std. and root-MSE in the CKLS model.

Note: The three elements in each cell are, from top to bottom, bias, std and root-MSE.

	α_1	α_2	β_1	β_2	β_3	β_4
	$n = 5000$					
	-0.0484	-0.0124	-0.2286	3.2306	-14.7289	0.0053
Parametric	0.2258	0.0489	0.3668	5.1285	23.6135	0.0090
	0.2310	0.0504	0.4322	6.0612	27.8305	0.0104
	-0.0494	-0.0133	-0.0870	1.2501	-5.906	0.0022
Semiparametric	0.2386	0.0516	0.3005	4.2194	20.5818	0.0076
	0.2436	0.0533	0.3128	4.4007	21.4248	0.0079
	$n = 2500$					
	-0.0570	-0.0177	-0.5462	7.4802	-34.4485	0.0135
Parametric	0.3455	0.0752	0.8737	11.2375	52.5934	0.0243
	0.3502	0.07773	1.0304	13.4994	62.8711	0.0278
	-0.0581	-0.0197	-0.3066	4.1607	-20.1144	0.0081
Semiparametric	0.3653	0.0810	0.6214	8.0931	41.3468	0.0174
	0.3699	0.0834	0.6930	9.1000	45.9799	0.0192
	$n = 1000$					
	0.0685	-0.0153	-2.6473	34.2092	-161.2317	0.0737
Parametric	0.7735	0.1558	3.9338	44.0910	228.4953	0.1371
	0.7766	0.1565	4.7417	55.8058	279.6530	0.1557
	0.0903	-0.0169	-1.2287	17.0347	-88.6357	0.0332
Semiparametric	0.8477	0.1636	2.0262	26.1552	148.5297	0.0635
	0.8525	0.1645	2.3697	31.2134	172.9664	0.0716

Table 2: Parametric and Semiparametric bias, std. and root-MSE in the AS model.

Note: The three elements in each cell are, from top to bottom, bias, std and root-MSE.

D Figures

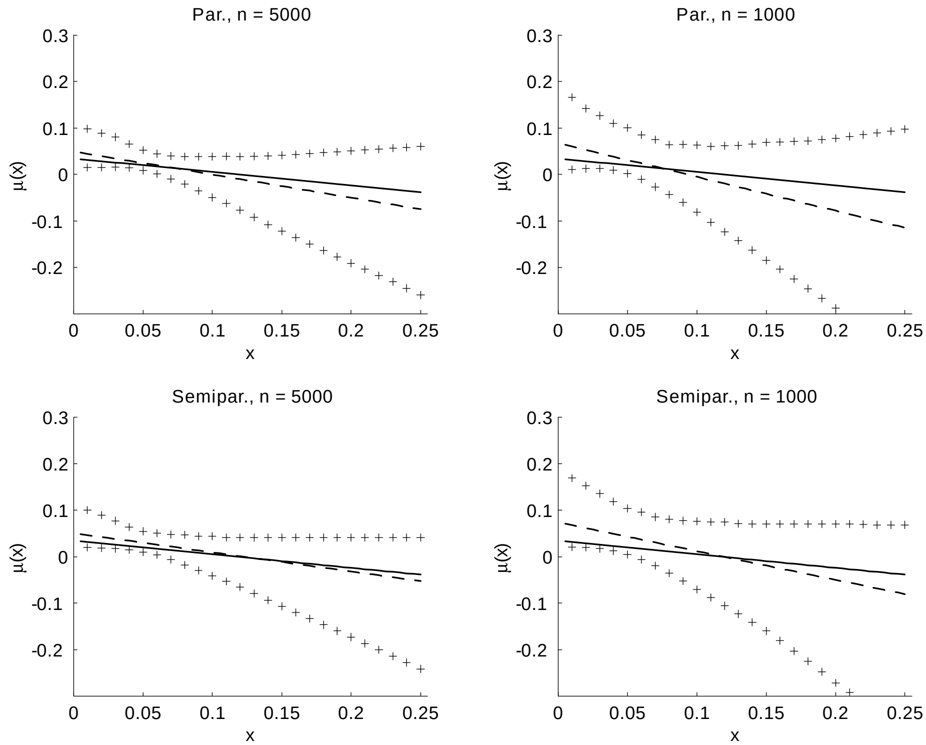

Figure 1: Parametric and semiparametric estimates of the drift in the CKLS model. Note: Full line = true drift, dashed line = mean of estimate, '+' = 95% conf. bands

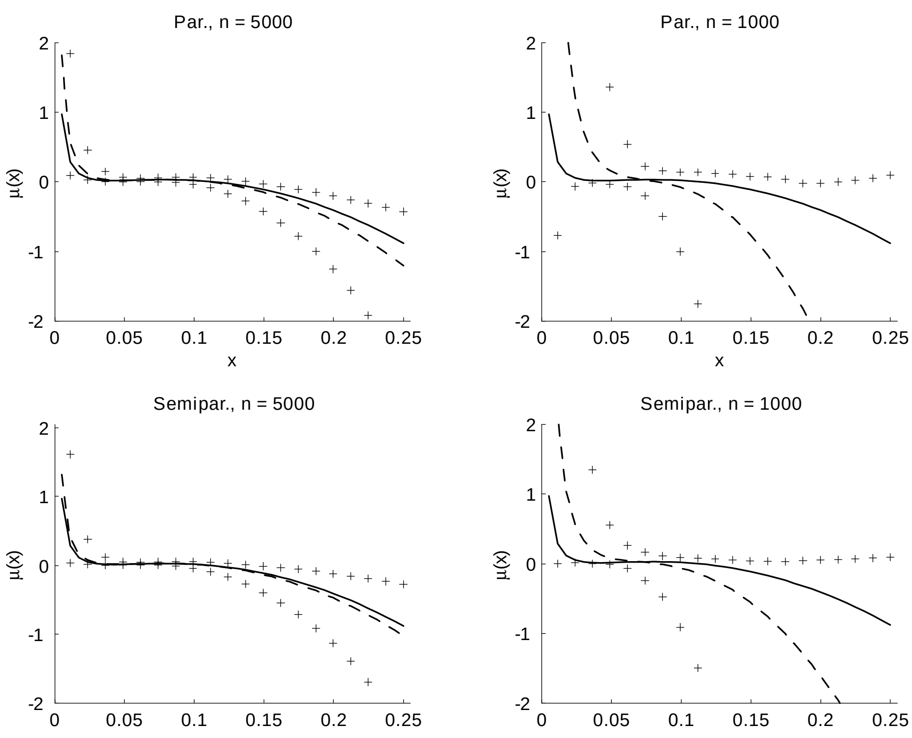

Figure 2: Parametric and semiparametric estimates of the drift in the AS model. Note: Full line

= true drift, dashed line = mean of estimate, '+' = 95% conf. bands

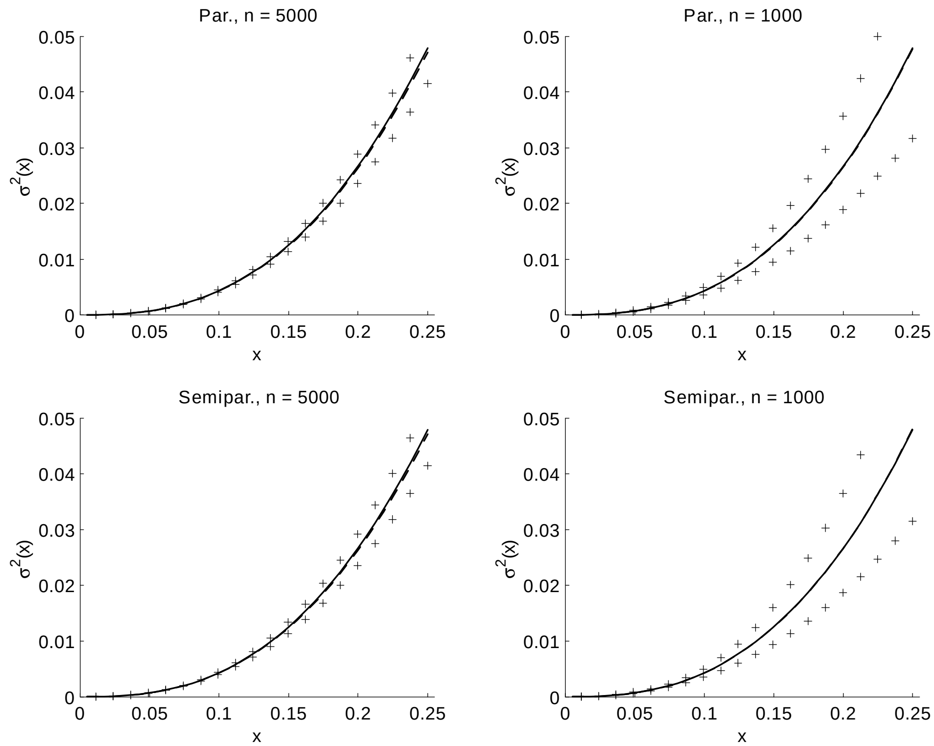

Figure 3: Parametric and semiparametric estimates of the diffusion in the AS model. Note: Full line = true diffusion, dashed line = mean of estimate, '+' = 95% conf. bands

Research Papers 2009