Simulated Nonparametric Estimation of Continuous Time Models of Asset Prices and Returns*

Filippo Altissimo
European Central Bank, and CEPR

Antonio Mele
London School of Economics

December 29, 2003

Abstract

This paper introduces a new parameter estimator of dynamic models in which the state is a multidimensional, continuous-time, partially observed Markov process. The estimator minimizes appropriate distances between nonparametric joint (and/or conditional) densities of sample data and nonparametric joint (and/or conditional) densities estimated from data simulated out of the model of interest. Sample data and model-simulated data are smoothed with the same kernel. This makes the estimator: 1) consistent independently of the amount of smoothing; and 2) asymptotically root-T normal when the smoothing parameter goes to zero at a reasonably mild rate. When the underlying state is observable, the estimator displays the same asymptotic efficiency properties as the maximum-likelihood estimator. In the partially observed case, we derive conditions under which efficient estimators can be implemented with the help of auxiliary prediction functions suggested by standard asset pricing theories. The method is flexible, fast to implement and possesses finite sample properties that are well approximated by the asymptotic theory.

JEL: C14, C15, C32, G12

Keywords: nonparametric estimation, continuous time asset pricing, continuum of moments, simulations

Corresponding author: Antonio Mele, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Email: a.mele@lse.ac.uk.

*We thank Yacine Aït-Sahalia, Torben Andersen, Marine Carrasco, Mikhail Chernov, Frank Diebold, Cristian Huse, Dennis Kristensen, Oliver Linton, Nour Meddahi, Angelo Melino, Eric Renault, Christopher Sims, Sam Thompson, seminar participants at the LSE, Princeton University, the University of Pennsylvania and, especially, Valentina Corradi for her valuable comments and Fabio Fornari who worked with us in the initial phase of this project. We remain, however, solely responsible for any omission and mistakes, as well as for the views expressed in the paper, which do not reflect those of the European Central Bank.

1 Introduction

This paper introduces a new class of parameter estimators for partially observed systems, called Simulated Nonparametric Estimator (hereafter SNE). The SNE works by making the finite dimensional distributions of the model's observables as close as possible to their empirical counterparts estimated through standard nonparametric techniques. Since the distribution of the model's observables is in general analytically intractable, we recover it through two steps. In the first step, we simulate the whole partially observed system. In the second step, we obtain model's density estimates through the application of the same nonparametric devices used to smooth the sample data. The result is a consistent and root-T asymptotically normal estimator displaying a number of attractive properties. First, our estimator is based on simulations. Consequently, it can be implemented fastly and in a straightforward manner to cope with a variety of estimation problems. Second, the SNE is purposely designed to minimize distances of densities smoothed with the same kernel. Therefore, it is consistent regardless of the smoothing parameter behavior; and it achieves the same asymptotic efficiency as the maximum-likelihood estimator in the case of fully observed Markov processes. Third, even in the presence of partially observed systems, the SNE may remain efficient when some suitable prediction functions suggested by economic theory are used in conjuction with data generated by the original unobserved systems. Finally, Monte Carlo experiments reveal that our estimator does exhibit a proper finite sample behavior.

Partially observed systems arise naturally in many areas of economics. Examples in macroeconomics include models of stochastic growth with human capital and/or sunspots, job duration models, or models of investment-specific technological changes. Examples arising in finance include latent factor models, processes with jumps, continuous time Markov chains, and even scalar diffusions. While the methods developed in this paper are well suited to address estimation issues in all such areas, we restrict our attention to the estimation of the typical diffusion models arising in financial economics.

As is well-known, the major difficulty arising from the estimation of partially observed systems is related to criterion functions that are extremely complex to evaluate. The simulated method of moments of Duffie and Singleton (1993), the indirect inference approach of Gouriéroux, Monfort and Renault (1993), or the efficient method of moments (EMM) of Gallant and Tauchen (1996) represent the first attempts at addressing this issue through extensions of the generalized method of moments; see Gouriéroux and Monfort (1996) for a survey on such methods and related approaches. The main characteristic of these approaches is that they are general-purpose. Their drawback is that they lead to inefficient estimators even in the case of fully observed systems. As an example, the EMM estimator is efficient only under the so-called "smooth embedding condition"; and as Gallant and Long (1997) demonstrated, such a condition holds when the (parameter) dimension of the auxiliary score gets higher and higher.

Estimation methods specifically designed to deal with diffusion processes include moments generating techniques (e.g., Bibby and Sørensen (1995), Hansen and Scheinkman (1995), Singleton (2001) and Chacko and Viceira (2003)), approximations to maximum likelihood (e.g., Pedersen (1995) and Aït-Sahalia (2002, 2003)) and, on a radically different perspective, Markov Chain Monte Carlo approaches (see, e.g., Elerian, Chib and Shephard (2001) and Eraker (2001)). As regards the moments generating approach, Singleton developed estimating conditions generated by the characteristic function, and identified an optimal (but generally unfeasible) instrument leading to efficient estimators. On the other hand, both Pedersen's and Aït-Sahalia's estimators constitute arbitrarily accurate approximations to the (generally unfeasible) maximum-likelihood estimator. These estimators work by approximating the transition density of arbitrary multivariate diffusion processes. Specifically, the Pedersen's estimator recovers the transition density through Monte Carlo integration; and the Aït-Sahalia's estimator recovers the transition density through closed-form expansions based on Hermite polynomials.

Our approach does not rely on the approximation of the maximum-likelihood estimator for diffusion processes. Instead, we construct criterion functions leading to a general estimation approach. In many cases of interest, these criterion functions are asymptotically equivalent to Neyman's chi-square measures of distance. It is precisely such an asymptotic equivalence which makes our resulting estimators as efficient as the maximum-likelihood estimator. However, we emphasize that our estimators are quite distinct from any possible approximation to the maximum-likelihood estimator. In the language of indirect inference theory, we rely on "auxiliary criterion functions", which generally give rise to asymptotically inefficient but consistent estimators. But as soon as model's and data transition densities are estimated through an asymptotically shrinking smoothing parameter, these criterion functions converge to Neyman's chi-squares, and our estimators become efficient. In this sense, the role played by the smoothing parameter in our context parallels the role played by the smooth embedding condition within the efficient method of moments.[1] The distinctive feature of our method is that we do not require that the (parameter) dimension of the "auxiliary" criterion goes to infinity. We only require that the smoothing parameter goes to zero at a reasonably mild rate. Furthermore, we smooth model-generated data and observations with the same kernel. Therefore, the behavior of the smoothing parameter does not affect the consistency of the estimator - as it would happen for example in the case of non-parametric simulated maximum-likelihood estimators (see, e.g., Fermanian and Salanié (2003)). An asymptotically shrinking smoothing parameter can only favorably affect the precision of our estimator.

Our methods display the attractive features of both moments generating techniques and maximum-likelihood. As we have argued, our methods are general-purpose - just as the generalized method of moments and its extensions. In this article, we demonstrate their specific ability

[1] We are grateful to Christopher Sims for bringing this point to our attention.

to address parameter estimation of arbitrary partially observed, multivariate diffusion processes. At the same time, our methods can be as efficient as maximum likelihood whenever the state is a fully observable Markov process. Finally, we demonstrate that the finite sample performance of our estimators is at least as good as maximum likelihood.

In a related paper, Carrasco, Chernov, Florens and Ghysels (2002) developed an estimation technology which also leads to asymptotic efficiency in the case of fully observed Markov processes. The authors built on previous work by Carrasco and Florens (2000), and formulated a "continuum of moment conditions" leading to match model-based (simulated) characteristic functions with data-based characteristic functions. Our estimator also relies on a "continuum of moments", but it is different. We use more classical ideas from the statistical literature, and develop estimating equations leading to match (model-based) simulated nonparametric density estimates with their empirical counterparts. Earlier estimators based on similar ideas include the ones succinctly surveyed in the next section. Two particularly important contributions related to this literature are in Aït-Sahalia (1996) and in Diebold, Ohanian and Berkowitz (1998). Aït-Sahalia developed an estimator matching marginal densities. Diebold, Ohanian and Berkowitz proposed to match spectra, thereby feeding their resulting estimator with information about the dynamic structure of a model. At the same time, matching spectra might entail loss of information about potential nonlinearities. By matching joint and/or conditional densities, we combine the relative strengths of these two approaches.

The paper is organized in the following manner. The next section motivates the design of the estimator developed in the core of the paper. Section 3 introduces basic notation and assumptions, as well as examples of models to which the estimator can be applied. Section 4 provides large sample theory. Section 5 develops conditions under which our methods can be used to implement parameter estimation of asset pricing models. Section 6 assesses the finite sample and computational properties of the estimator. Section 7 concludes. The appendix gathers proofs and regularity conditions omitted in the main text.

2 Methods: a heuristic overview

This section provides a heuristic introduction to the main ideas in this paper. Theory, extensions and computational features of the method are in sections 3 through 6. Readers willing to access directly to our results can thus proceed to section 3 without loss of continuity.

2.1 Closeness of density functions

To keep this heuristic presentation as simple as possible, we initially consider a sample $\{x_1, \cdots, x_T\}$ of independent draws from a distribution with continuous density π_0 ($x \in \mathbb{R}^d$ and $d \geq 1$). We assume that π_0 belongs to a specified parametric family $\pi(\cdot;\theta)$ indexed by a vector $\theta \in \Theta$, where

Θ denotes the parameter space. The purpose is to estimate the (supposedly) unique $\theta_0 \in \Theta$ making $\pi_0(x) = \pi(x;\theta_0)$, $x \in \mathbb{R}^d$.

Our estimation methodology is related to a classical field of the statistical literature initiated by Bickel and Rosenblatt (1973). By and large, this literature aims at testing the closeness of two arbitrary density functions f and g through the integrated squared difference:

$$I = \int_{\mathbb{R}^d} [f(x) - g(x)]^2 w(x) dx, \tag{1}$$

where w is a given weighting function. As an example, suppose that $g = \pi_0$, and consider testing the null $H_0 : \pi(x,\theta_0) \equiv f(x) = \pi_0(x)$ on $\mathbb{R}^d$, against its negation. Let π_T be a nonparametric estimator of π_0 obtained as $\pi_T(x) \equiv (T\lambda^d)^{-1} \sum_{t=1}^{T} K((x_t - x)/\lambda)$, $x \in \mathbb{R}^d$, where the bandwidth $\lambda > 0$, and K is a symmetric bounded kernel of the r-th order (see the appendix for more details on the kernels used in this paper). Consider the following empirical counterpart to (1):

$$I_T(\theta_T) = \int_{\mathbb{R}^d} [\pi(x;\theta_T) - \pi_T(x)]^2 w_T(x) dx, \tag{2}$$

where w_T is a weighting function possibly depending on data, and θ_T is a given consistent estimator of θ_0. Rescaled versions of (2) can now be used to implement tests of H_0 (see, e.g., Pagan and Ullah (1999, p. 60-71) for a comprehensive survey on those tests).[2]

The focus of this paper is on using a metric related to (2) to estimate the unknown parameter vector θ_0. Accordingly, consider endogenizing sequence θ_T in (2), and define

$$\theta_T^I = \arg\min_{\theta \in \Theta} I_T(\theta). \tag{3}$$

Notice that if $w_T \equiv \pi_T$, θ_T^I collapses to the estimator proposed by Aït-Sahalia (1996) in the context of scalar diffusion processes.

An important feature of the empirical measure of distance in (2) is that a parametric density estimate, $\pi(\cdot;\theta)$, is compared with a nonparametric one, $\pi_T(\cdot)$. Under correct model specification, $\pi_T(x) \xrightarrow{p} K * \pi(x;\theta_0) \equiv \int_{\mathbb{R}^d} \lambda^{-d} K((u-x)/\lambda)\, \pi(u;\theta)\, du$ (x-pointwise). As is well-known, the result that $\pi_T(x) \xrightarrow{p} \pi(x;\theta_0)$ (x-pointwise) only holds if the bandwidth $\lambda \equiv \lambda_T$ (say), $\lim_{T\to\infty} \lambda_T \to 0$ and $\lim_{T\to\infty} T\lambda_T^d \to \infty$. Therefore, bandwidth choice is critical for (2) and (3) to be really informative in finite samples. Furthermore, this choice becomes even more fundamentally critical in the case of dependent observations that we will deal with later in this paper. In the next subsection, we discuss how to circumvent this problem through a convenient change of the distance measure in (2).

[2] Corradi and Swanson (2003) have recently developed new specification tests for diffusion processes based on cumulative probability functions.

2.2 On "twin-smoothing"

A simple alternative to (2) is an empirical distance in which the nonparametric estimate π_T is matched asymptotically to its probability limit conditional on a given bandwidth value:

$$L_T(\theta_T) = \int_{\mathbb{R}^d} [K * \pi(x;\theta_T) - \pi_T(x)]^2 w_T(x)\, dx, \tag{4}$$

where θ_T is an arbitrarily given but consistent estimator of θ_0. Fan (1994) developed a class of bias-corrected goodness of fit tests based on $L_T(\theta_T)$ and weighting function $w_T \equiv \pi_T$.

One basic idea in this paper is to combine the attractive features of θ_T^I in (3) with the bias-corrected empirical measure in (4). To achieve this objective, we endogenize sequence θ_T in (4) rather than in (2), and consider general empirical weighting functions w_T. As we argue, our resulting estimator is free from biases related to density estimates. Specifically, define the following estimator:

$$\theta_T^L = \arg\min_{\theta\in\Theta} L_T(\theta), \tag{5}$$

where $w_T(x) \xrightarrow{p} w(x)$ (x-pointwise), and w is another positive function satisfying some basic regularity conditions (see section 4.2). As it turns out, bandwidth conditions affect the two estimators θ_T^I and θ_T^L in a quite different manner. Consistency of θ_T^L holds independently of the bandwidth behavior (i.e., λ can be any strictly positive number). Consistency of θ_T^I requires the additional conditions that $\lim_{T\to\infty}\lambda_T \to 0$ and $\lim_{T\to\infty} T\lambda_T^d \to \infty$. To illustrate one reason explaining the difference, consider the following decomposition:

$$I_T(\theta) = L_T(\theta) + M_T(\theta) + N_T(\theta),$$

where

$$M_T(\theta) \equiv \int_{\mathbb{R}^d} [\pi(x;\theta) - K * \pi(x;\theta)]^2 w_T(x)\, dx;$$

$$N_T(\theta) \equiv 2\int_{\mathbb{R}^d} [\pi(x;\theta) - K * \pi(x;\theta)]\,[K * \pi(x;\theta) - \pi_T(x)]\, w_T(x)\, dx.$$

Let $I(\theta) \equiv \int_{\mathbb{R}^d} [\pi(x;\theta) - \pi_0(x)]^2 w(x)dx$. As is well-known, conditions ensuring consistency of θ_T^I include a uniform weak law of large numbers for $I_T(\theta)$ or, equivalently, stochastic equicontinuity of $I_T(\theta)$ and the condition that $\forall\theta \in \Theta$, $I_T(\theta) \xrightarrow{p} I(\theta)$ (see appendix A.1). Now consider the simplest case $\theta = \theta_0$. Clearly, $I(\theta_0) = 0$. Furthermore, $L_T(\theta_0) \xrightarrow{p} 0$ and $N_T(\theta_0) \xrightarrow{p} 0$, both by a generalization of Glick's (1974) theorem (see appendix B.1). In contrast, $M_T(\theta_0)$ is not $o_p(1)$ unless $\lambda_T \to 0$ at an appropriate rate. To ensure consistency of θ_T^I, the bandwidth behavior must be restricted in a way to make the effect of this extra term asymptotically negligible. Results in Pritsker (1998) (for dependent observations) suggest that under these restrictions, the asymptotic theory for θ_T^I is of practical guidance only in correspondence of very large sample sizes.

As for consistency, θ_T^L and θ_T^I are asymptotically normally distributed under different bandwidth restrictions. Specifically, one has that in the i.i.d. case considered in this section,

$$\sqrt{T}\left(\theta_T^s - \theta_0\right) \xrightarrow{d} N\left(0, V\right), \quad s = I, L, \tag{6}$$

where $V = \text{var}[\Psi\left(x_1\right)]$, $\Psi\left(x\right) \equiv \left[\int_{\mathbb{R}^d} |\nabla_\theta \pi\left(x;\theta_0\right)|_2 \, w\left(x\right) dx\right]^{-1} \nabla_\theta \pi\left(x;\theta_0\right) w\left(x\right)$, and $|\cdot|_2$ denotes outer product. Such a convergence result holds under mild regularity conditions but different conditions on λ. Precisely, θ_T^L is asymptotically normal under the standard assumptions that $\lim_{T\to\infty} \lambda_T \to 0$ and $\lim_{T\to\infty} T\lambda_T^d \to \infty$. Instead, θ_T^I is asymptotically normal under the additional condition that $\lim_{T\to\infty} \sqrt{T}\lambda_T^r \to 0$. Intuitively, this condition ensures that a density bias estimate vanishes at an appropriate rate without affecting the asymptotic behavior of θ_T^I, and that a functional central limit theorem can be applied. In contrast, density bias issues are totally absent if one implements estimator θ_T^L.

Table 1 summarizes our discussion. θ_T^I is consistent if $\lambda_T \to 0$ and $T\lambda_T^d \to \infty$.[3] Furthermore, θ_T^I is asymptotic normal under the additional condition that $\sqrt{T}\lambda_T^r \to 0$. In contrast, θ_T^L is consistent without the conditions $\lambda_T \to 0$ and $T\lambda_T^d \to \infty$. These (and only these) bandwidth conditions are required in order for θ_T^L to be asymptotic normal. As we demonstrate in the Monte Carlo experiments of section 6, conditions $\lambda_T \to 0$ and $T\lambda_T^d \to \infty$ are much less restrictive for asymptotic normality than for consistency.

Table 1 - Bandwitdh assumptions and asymptotic behavior of θ_T^I in (3) and θ_T^L in (5)

	Consistency	Asymptotic normality
θ_T^I	$T\lambda_T^d \to \infty$, $\lambda_T \to 0$	$T\lambda_T^d \to \infty$, $\lambda_T \to 0$, and $\sqrt{T} \cdot \lambda_T^r \to 0$
θ_T^L	always	$T\lambda_T^d \to \infty$, $\lambda_T \to 0$

2.3 Efficiency, and robustness

Informal inspection of the variance term in (6) suggests that if the weighting function w is equal to $1/\pi_0$, both θ_T^I and θ_T^L asymptotically achieve the Cramer-Rao lower bound. Efficiency can thus be implemented with $w_T = 1/\left(\pi_T + \alpha_T\right)$, where α_T is any strictly positive sequence such that $\alpha_T \xrightarrow{p} 0$ (e.g., $\alpha_T = T^{-1}$).

[3] Other estimators related to (3) suffer from exactly the same drawback. Two examples are 1) estimators based on nonparametric density estimates of the log-likelihood function obtained through simulations; and 2) estimators based on the so-called Kullback-Leibler distance (or relative entropy) $\int_{\mathbb{R}^d} \log[\pi(x,\theta)/\pi_0(x)]\pi(x,\theta)dx$ (see Robinson (1991)). We are grateful to Oliver Linton for having suggested the latter example to us.

We emphasize that such an efficiency property coincides with the classical first-order efficiency criterion prescribed by Rao (1962). Furthermore, results by which estimators based on closeness of density functions retain efficiency properties are not a novelty in the statistical literature. In the context of independent observations with fully parametric discrete densities, Lindsay (1994) presented a class of minimum disparity estimators nesting a number of estimators such as Hellinger's distance, Pearson's chi-square, Neyman's chi-square, Kullback-Leibler distance, and maximum likelihood. Lindsay showed that while all these estimators are first-order efficient, they may differ in terms of second-order efficiency, and robustness. Basu and Lindsay (1994) extended this theory to the case of continuous densities. Such an extension can be used to illustrate some fundamental properties of our estimator.

Our estimator θ_T^L in (5) can be thought of as a member belonging to a general class of minimum disparity estimators θ_T defined by the following estimating equation:

$$0 = \int A(\delta(x)) \left[\nabla_\theta \left(K * \pi \left(x; \theta_T\right)\right)\right] dx,$$

where

$$\delta(x) \equiv \frac{K * \pi_T(x) - K * \pi \left(x; \theta_T\right)}{K * \pi \left(x; \theta_T\right)},$$

and A is an increasing continuous function in $(-1, \infty)$.[4] Under regularity conditions, function A determines how sensible an estimator is to the presence of outliers. Indeed, function δ is high exactly when a point in the sample space has been accounted much more than predicted by the model. Accordingly, a robust estimator is one able to mitigate the effect of large values of δ. As a benchmark example, the likelihood disparity sets $A(\delta) = \delta$. Estimators with the property that $A(\delta)$ ¿ δ for large δ are more robust to the presence of outliers than maximum likelihood. For instance, the Hellinger's distance sets $A(\delta) = 2[\sqrt{\delta + 1} - 1]$, and the Kullback-Leibler distance has $A(\delta) = \log(1 + \delta)$. It is easily seen that if $w_T = 1/\left(\pi_T + \alpha_T\right)$, our L_T is asymptotically a Neyman's chi-squared measure of distance, with $A(\delta) = \delta/\left(1 + \delta\right)$. These simple facts suggest that the class of estimators that we consider displays interesting robustness properties.

Naturally, this article aims at extending the above class of estimators to the case of dynamic models. However, we do not further investigate the robustness properties of our resulting estimators. Using robustness, and/or second-order efficiency criteria as discrimination devices of alternative parameter estimators of diffusion models is an interesting area that we leave for future research.

[4] When $\lambda \downarrow 0$, A and δ collapse to what Lindsay (1994) termed residual adjustment function and Pearson's residual, respectively.

2.4 Unknown density functions

A fundamental objective of this paper is to extend the previous ideas and results to more general situations. Specifically, suppose that the analytical solution for density $\pi\left(x;\theta\right)$ in (4) is unknown, but that it is still possible to simulate from that density. Consider the following estimator:

$$\theta_T = \arg\min_{\theta\in\Theta} \int_{\mathbb{R}^d} \left[\frac{1}{S} \sum_{i=1}^{S} \pi_T^i\left(x;\theta\right) - \pi_T(x) \right]^2 w_T(x)dx, \tag{7}$$

where S is a positive integer, $\pi_T^i\left(x;\theta\right) \equiv (T\lambda^d)^{-1}\sum_{t=1}^T K((x_t^{(i)}(\theta)-x)/\lambda)$, and $\{x_t^{(i)}(\theta)\}_{t=1}^T$ is the i-th sequence simulated from $\pi\left(\cdot;\theta\right)$.

The appealing feature of this estimator is that π_T^i and π_T are computed with the same) kernel and bandwidth. Such a "twin" kernel smoothing procedure operates on sample and model generated data in exactly the same manner as in (4). Consequently, the asymptotic properties of θ_T in (7) and θ_T^L in (5) are identical under the same bandw l

of biases in the density estimates, θ_T^I performs worst at all horizons in terms of MSE. This problem is absent within the SNEs. While the SNEs are implemented with simulations, the MSEs associated with the SNEs are even less than the MSEs associated with θ_T^I. Furthermore, the Opt-SNE performs better than the basic SNE. Finally, maximum likelihood performs better than the other three methods.

2.5 Extensions

This paper extends the previous ideas and examples to deal with parameter estimation of dynamic models. In the i.i.d. case covered in this heuristic section, estimators minimizing disparity measures of marginal densities have interesting properties. In sections 2.2 and 2.3, we argued that they are consistent and asymptotically efficient. And the simple Monte Carlo experiment of section 2.4 revealed that the asymptotic theory does provide practical guidance in finite samples. The use of marginal densities is no longer appropriate in the dynamic case. To exploit all the information conveyed by the probabilistic structure of a dynamic model, this paper introduces estimators based on more general measures of closeness. Accordingly, we develop three estimators that share the same "twin-smoothing" features described in section 2.2 and 2.4.

The first, basic estimator (the SNE) extends the framework of the previous sections to the case of joint densities (see section 4.2). The second estimator minimizes measures of closeness of conditional densities, and is called Conditional Density (CD)-SNE (see section 4.3). As it turns out, the CD-SNE can be made as efficient as the MLE if the state is fully observable, and Markov. The third estimator is the CD-SNE applied to "functionals of state variables" (see section 5). By "functionals of state variables", we mean that the processes of interest may have state variables, some of which not observable, that are linked to other observable variables by some nonlinear (deterministic) function(al)s. Typically, such function(al)s are suggested by standard asset pricing theories - for these theories predict that asset prices are deterministic functions of the underlying state. We derive conditions for partially observed systems to be embedded in this new (functionally interdependent) format, and then develop conditions ensuring both feasibility and efficiency of our CD-SNE. Finally, we investigate the finite sample properties of our estimators. In section 6, we show that even in the presence of persistent data and limited sample sizes, the use of joint and/or conditional densities makes the resulting estimators work as expected by the asymptotic theory.

3 The model of interest

Let $\Theta \subset \mathbb{R}^{p_\theta}$ be a compact parameter set, and for a given parameter vector $\theta \in \Theta$, consider the following data generating process $y = \{y(\tau)\}_{\tau \geq 0}$:

$$dy(\tau) = b\left(y(\tau), \theta\right) d\tau + a\left(y(\tau), \theta\right) dW(\tau),\ \tau \geq 0, \tag{8}$$

where W is a standard d-dimensional Brownian motion; b and a are vector and matrix valued functions in $\mathbb{R}^d$ and $\mathbb{R}^{d \times d}$, respectively; a is full rank almost surely; and y takes values in $Y \subseteq \mathbb{R}^d$. To simplify the presentation, we do not describe parameter estimation for jump-diffusion processes. Yet jump-diffusion processes are continuous time Markov processes. If their finite-dimensional distributions satisfy assumption 1 below, our results can be extended to cover parameter estimation in this class of models under the same assumptions developed for the (pure) multidimensional diffusion case (8).

The following regularity conditions are imposed to system (8) throughout the paper.

Maintained assumptions I. System (8) admits a strong solution and it is strictly stationary.

The purpose of this paper is to provide estimators of the true parameter vector $\theta_0 \in \Theta$. We consider a general situation in which some components of y are not observed. Accordingly, we partition vector y as:

$$y = \begin{pmatrix} y^o \\ \cdots \\ y^u \end{pmatrix},$$

where $y^o \in Y^o \subseteq \mathbb{R}^{q^*}$ is the vector of observable variables and $y^u \in Y^u \subseteq \mathbb{R}^{d-q^*}$ is the vector of unobservable variables. Data are assumed to be sampled at regular intervals, and are collected in a $q^* \times T$ matrix with elements $\{y^o_{j,t}\}_{j=1,\cdots,q^*;t=1,\cdots,T}$, where $y^o_{j,t}$ denotes the t-th observation of the j-th component of vector y^o, and T is the sample size. Since our general interest lies in the estimation of partially observed diffusion processes, we may wish to recover as much information as possible about the dependence structure of the observables in (8). We thus set $q = q^*(1+l)$, let $y^o_t = (y^o_{1,t}, \cdots, y^o_{q^*,t})$ and

$$x_t \equiv \left(y^o_t, \cdots, y^o_{t-l}\right), \quad t = t_l \equiv 1+l, \cdots, T, \tag{9}$$

and define $X \subseteq \mathbb{R}^q$ as the domain of x_t.

We now provide examples of models that can be dealt with the methods introduced in this paper, and formulate some further assumptions.

Example 1. (Affine three-factor short-term rate models with stochastic volatility and stochastic central tendency) Consider the following data generating process: $y = \{r(\tau), \sigma(\tau), \ell(\tau)\}_{\tau \geq 0}$, with

$$\begin{cases} dr(\tau) &= b_1 \left(\ell(\tau) - r(\tau)\right) d\tau + a_1 \sigma(\tau) dW^{(1)}(\tau) \\ d\sigma(\tau)^2 &= \left(b_2 - b_3 \sigma(\tau)^2\right) d\tau + a_2 \sigma(\tau) dW^{(2)}(\tau) \\ d\ell(\tau) &= \left(b_4 - b_5 \ell(\tau)\right) d\tau + a_3 dW^{(3)} \end{cases} \tag{10}$$

where $\{W_i(\tau)\}_{i=1}^3$ are independent Brownian motions, (r, σ, ℓ) denote the short-term rate, stochastic volatility and central tendency processes, and $\theta \equiv (b_1, \cdots, b_5, a_1, a_2, a_3)$ is the parameter vector. This model was introduced by Balduzzi, Das, Foresi and Sundaram (1996), and is called affine because the characteristic function associated with it is exponential-affine in y.[5] Suppose that the short-term rate r is the only observable of this system. Then, $y^o = r$ and $q^* = 1$. A possible choice for the variables of interest could then be $x_t = (r_t, r_{t-1})$ (i.e. $q = 2$). The extension to correlated Brownian motions and more elaborated affine models (as in Dai and Singleton (2000) for example) is immediate, as it is the extension to nonaffine models.

In some situations of interest, the Maintained Assumptions I can not be entirely satisfied. The celebrated geometric Brownian motion is one counterexample to those assumptions. Fortunately, our method may also work with data generated by this kind of processes. For example, if the price of a share $Q(\tau)$ follows a geometric Brownian motion, then its m-period returns $R_t^q(m) \equiv \log(Q(\tau_t)/Q(\tau_t - m))$ ($t = 1, 2, \cdots$) forms a stationary sequence (see Das and Sundaran (1999) for a detailed analysis of the moments of $R_t^q(m)$ in a fairly general stochastic volatility model such as the one in example 2 below). We then extend stationarity properties to this kind of data transformations:

Maintained assumptions II and admissible data transformation. Nonstationary components y^o of vector x are replaced with functionals of y^o and its k lags, $\varphi_t \equiv \varphi(y_t^o, ..., y_{t-k}^o)$, which are stationary.

[5] See, for instance, Duffie, Pan and Singleton (2000) for an analysis of more general affine models. Conditions for the existence of a strong solution in settings more general than (10) are established in a theorem in Duffie and Kan (1996, p. 388).

Example 2. (Stochastic volatility share price models) Consider a share price process $Q(\tau)$ solution to

$$\begin{cases} \dfrac{dQ(\tau)}{Q(\tau)} &= \mu \cdot d\tau + \sigma(\tau) dW_1(\tau) \\ \\ d\sigma(\tau)^2 &= \kappa\left(\overline{v} - \sigma(\tau)^2\right) d\tau + \psi \sigma(\tau)^{\xi} \left(\rho dW_1(\tau) + \sqrt{1-\rho^2} dW_2(\tau)\right) \end{cases} \tag{11}$$

where $\theta \equiv (\mu, \kappa, \overline{v}, \psi, \xi, \rho)$. In this example, $q^* = 1$ and $y^o = Q$, and it is easy to find the parametric restrictions that ensure a strong solution to (11) up to an explosion time.[6] Even if (11) has a strong solution, Q does not satisfy the stationarity condition of our Maintained Assumptions I. This difficulty can be circumvented by choosing $x_t = (R_t, R_{t-1})$ and $R_t = \log(Q_t/Q_{t-1})$.

By our maintained assumptions, the finite dimensional distributions associated with x are well-defined, and we let $\pi(x;\theta)$ denote the joint density induced by (8) on x when the parameter vector is $\theta \in \Theta$. Let $\pi_0(x) \equiv \pi(x;\theta_0)$ and let $|\nabla_\theta \pi(x;\theta)|_2$ denote the outer product of vector $\nabla_\theta \pi(x;\theta)$. The following assumption further characterizes the family of processes we are investigating. It contains mild regularity conditions on π as well as one standard condition that is necessary for identifiability of the diffusion:

Assumption 1. $\pi(x;\theta)$ is continuous and bounded on $X \times \Theta$. For all $x \in X$, function $\theta \mapsto \pi(x;\theta)$ is twice differentiable and its derivatives are bounded on Θ. Finally, there exists a neighborhood N of θ_0 such that matrix $E\left[|\nabla_\theta \pi(x;\theta)|_2\right]$ has full rank in $N \cap \Theta$.

The maintained assumption that (8) is stationary implies that the "observed skeleton" of the diffusion inherits the same features of the continuous time process. To ensure the feasibility of the asymptotic theory, we also need to make the following assumption on the decay of dependence in the observables in (8):

Assumption 2. Vector y is a β-mixing sequence[7] with mixing coefficients β_k satisfying $\lim_{k\to\infty} k^{\delta}\beta_k \to 0$, for some finite $\delta > 1$.

[6] As it is well-known (see Aït-Sahalia (1996, appendix)), there exists a strong solution to the volatility equation in (11), up to an explosion time. Parametric restrictions are then found through two steps. The first step makes use of the classical boundary classification analysis (see, e.g., Karatzas and Shreve (1991, p. 342-353)), and aims at finding restrictions on $\kappa, \overline{v}, \psi$ and ξ ensuring that both boundaries of σ^2 (i.e. zero and infinity) are unattainable in finite expected time. The second step aims at finding parameter restrictions ensuring that σ^2 is square-integrable against its invariant measure. Under such a square-integrability condition, Q is solution to $dQ = Q \cdot dL$, where L is a square-integrable semimartingale. Strongness of (11) then follows from Revuz and Yor (1999, theorem 2.1 p. 375). Finally, note that the characteristic function associated with this model is exponential affine if and only if $\xi \equiv 1/2$.

[7] A strictly stationary process x on a finite-dimensional Euclidean space is β-mixing (or absolutely regular) if

The mixing condition of assumption 2 is critical for the application of a functional central limit theorem due to Arcones and Yu (1994). Precisely, assumption 2 ensures convergence of suitably rescaled integrals of kernel functions to stochastic integrals involving time-changed Brownian Bridges. This kind of convergence is exactly what we need to prove asymptotic normality of our estimators (see appendixes A.2, B.2 and C.2). Chen, Hansen and Carrasco (1999) provide primitive conditions guaranteeing that assumption 2 holds in the case of scalar diffusions. A scalar diffusion is β-mixing with exponential decay if their "pull measure", defined as $\frac{b}{a} - \frac{1}{2}\frac{\partial a}{\partial y}$, is negative (positive) at the right (left) boundary (the authors also provide conditions ensuring β-mixing with polynomial decay in the case of zero pull measure at one of the boundaries (see their remark 5)). As regards multidimensional diffusions, β-mixing with exponential decay can be checked through results developed by Meyn and Tweedie (1993, section 6 p. 535-537) for exponential ergodicity, as in Carrasco, Hansen and Chen (1999). Finally, Carrasco, Hansen and Chen (1999) provide more specific results pertaining to the partially observed diffusions case (such as our model example 2).

4 Theory

4.1 Simulations

The first step of our estimation strategy requires simulated paths of the observable variables in (8). To generate such simulated paths, various discretization schemes can be used (see, e.g., Kloeden and Platen (1999)). In this paper, we consider the simple Euler-Maruyama discrete time approximation to (8):

$$ {}_hy_{h(k+1)} - {}_hy_{hk} = b\left({}_hy_{hk}, \theta\right) \cdot h + a\left({}_hy_{hk}, \theta\right) \cdot \sqrt{h} \cdot \epsilon_{k+1}, \qquad k = 0, 1, \cdots, \tag{12} $$

where h

Assumption 3. For all $\theta \in \Theta$,

i) the high frequency simulator (12) converges weakly (or in distribution)[9] to the solution of (8) i.e., for each i, $y_h^i(\theta) \Rightarrow y(\theta)$ as $h \downarrow 0$;

ii) the di*ff*usion and drift functions a and b are Lipschitz continuous in y; their components are four times continuously di*ff*erentiable in y; and a, b and their partial derivatives up to the fourth order have polynomial growth in y;

iii) as $h \downarrow 0$ and $T \to \infty$, $h \cdot \sqrt{T} \to 0$.

Since the simulation step h can not be zero in practice, assumption 2 needs to be extended to cover the "pseudo"-skeleton behavior:

Assumption 4. For all $\theta \in \Theta$, there exists a $h^0 > 0$ such that for all $h < h^0$ and i, $y_h^i(\theta)$ is a strictly stationary β-mixing sequence satisfying assumption 2.

Primitive conditions ensuring that assumption 3-i holds are well-known and can be found in Stroock and Varadhan (1979), for instance. Primitive conditions guaranteeing that assumption 4 holds are also well-known (see, e.g., Tjøstheim (1990) for conditions ensuring that the solution of (12) is exponentially ergodic). Assumptions 3-ii,iii make our estimators asymptotically free of biases arising from the imperfect simulation of model (8) (model (8) is imperfectly simulated so long as $h > 0$). Precisely, such biases arise through terms taking the form $\sqrt{T}[E(K(x_{t,h}^i(\theta_0)) - E(K(x_t))]$, where K is a symmetric bounded kernel (see, e.g., eq. (A6) in appendix A.2). But by results summarized in Kloeden and Platen (1999, chapter 14), $\sqrt{T}[E(K(x_{t,h}^i(\theta_0)) - E(K(x_t))] = O(h \cdot \sqrt{T})$ whenever assumptions 3-i,ii hold and K is as differentiable as a and b are in assumption 3-ii. The role of assumption 3-iii is then to asymptotically eliminate such bias terms. Naturally, more precise high frequency simulators would allow h to shrink to zero at an even lower rate. Finally, assumption 3-ii can considerably be weakened. For example, one may simply require that a, b be Hölder continuous, as in Kloeden and Platen (1999, theorem 14.1.5 p. 460). These extensions are not considered here to keep the presentation as simple as possible.

4.2 Simulated Nonparametric Estimators

Densities of sample data and densities of simulated data are estimated with the same nonparametric kernels. As regards sample data, define $\pi_T(x) \equiv (T\lambda_T^q)^{-1} \sum_{t=t_l}^{T} K\left((x_t - x)/\lambda_T\right)$. As

[9] Let $({}_h y_{hk})_{k=1}^{\infty}$ be a discrete time Markov process, and $(y(\tau))_{\tau \geq 0}$ be a diffusion process. When the probability laws generating the entire sample paths of $({}_h y_{hk})_{k=1}^{\infty}$ converge to the probability laws generating $(y(\tau))_{\tau \geq 0}$ as $h \downarrow 0$, $({}_h y_{hk})_{k=1}^{\infty}$ is said to converge weakly (or in distribution) to $(y(\tau))_{\tau \geq 0}$; such a kind of convergence is usually denoted as ${}_h y \Rightarrow y$; the symbol $\xrightarrow{d}$ will be used here to denote convergence in distribution of a random variable.

regards the simulations, define $\pi^i_{T,h}(x;\theta) \equiv (T\lambda_T^q)^{-1}\sum_{t=t_l}^{T} K((x^i_{t,h}(\theta)-x)/\lambda_T)$, $i=1,\cdots,S$. Our assumption concerning the bandwidth behavior is:

Assumption 5. $\lim_{T\to\infty}\lambda_T \to 0$, $\lim_{T\to\infty} T\lambda_T^q \to \infty$.

It is well-known that assumption 5 ensures that $\pi_T(x) \xrightarrow{p} \pi_0(x)$ pointwise (see, e.g., Pagan and Ullah (1999, chapter 2)). Assumption 5 is the only bandwidth condition we actually need to develop our asymptotic theory. As it will be discussed after theorem 1 below, we do not need the additional condition that:

$$\lim_{T\to\infty}\sqrt{T}\cdot\lambda_T^r \to 0 \tag{13}$$

(where r denotes the order of the kernel). As shown in appendix A.2 (see remark 3), condition (13) would be important if the theory required a functional limit theorem for $\sqrt{T}(\int \pi_T - \int \pi_0)$. We do not need such a demanding result. We only need a functional limit theorem for $\sqrt{T}(\int \pi_T - \int E(\pi_T))$. Therefore, the order of the kernel plays no role within our asymptotic theory.

We are now in a position to provide the definition of the most basic estimator considered in this paper:

Definition 1. (SNE) For each positive integer S, the Simulated Nonparametric Estimator (SNE) is the sequence $\{\theta_{T,S,h}\}_{h,T}$ given by:

$$\theta_{T,S,h} = \arg\min_{\theta\in\Theta}\int_X \left[\frac{1}{S}\sum_{i=1}^{S}\pi^i_{T,h}(x;\theta) - \pi_T(x)\right]^2 \pi_T(x)dx. \tag{14}$$

The following result provides the asymptotic properties of the SNE:

Theorem 1. In addition to assumptions 1, 2, 3-i, 4, let assumptions 7-8 in the appendix hold; then, as $h\downarrow 0$ and $T\to\infty$, the SNE is (weakly) consistent. Furthermore, let assumptions 1-5 and assumptions 7-9 in the appendix hold, and define $\Psi(x) \equiv [E|\nabla_\theta \pi(x;\theta_0)|_2]^{-1}\pi_0(x)\nabla_\theta\pi(x;\theta)$. Then, as $h\downarrow 0$ and $T\to\infty$,

$$\sqrt{T}(\theta_{T,S,h}-\theta_0) \xrightarrow{d} N\left(0, \left(1+\frac{1}{S}\right)V\right),$$

where $V \equiv \text{var}[\Psi(x_1)] + 2\sum_{j=1}^{\infty}\text{cov}[\Psi(x_1), \Psi(x_{1+j})]$.

Proof. In appendix A. ■

Three elements characterize the asymptotic theory of the SNE. First, consistency does not rely on any condition regarding the bandwidth parameter. The familiar assumption 5 is only required to make the SNE asymptotically normal.

Second, the (unscaled) variance V of theorem 1 collapses to the variance of the Aït-Sahalia (1996) estimator in the scalar case. However, we emphasize that the two estimators are radically different. The Aït-Sahalia (1996) estimator recovers the model's density by means of parametric density estimates. Consequently, assumption 5 is needed to ensure consistency of his estimator. The SNE recovers the model's finite dimensional distributions by means of simulations. In addition to be consistent independently of bandwidth issues, the SNE can then be easily applied to address estimation of multivariate models driven by partially observed state variables with unknown distribution. Also, we explicitly consider matching joint densities of data, not marginal densities. Furthermore, the SNE minimizes a measure of closeness of two nonparametric density estimates - one on true data and the second on simulated data. Under correct model's specification, the resulting biases in the two kernel estimates cancel out each other, and asymptotic normality can then be obtained without condition (13). (A more detailed description of this phenomenon is in remark 3 in appendix A.2.) These characteristics give the SNE the potential to exhibit a finite sample behavior that is well approximated by the asymptotic theory. Such a finite sample behavior is indeed documented by our Monte Carlo experiments in section 6.

Third, similarly to the familiar asymptotics of Indirect Inference estimators (Gouriéroux, Monfort and Renault (1993)), a scaling term $(1+S^{-1})$ appears in the variance of the estimator. Such a scaling term emerges because the (unknown) density of the model is recovered by means of simulations.

The basic SNE in definition 1 overweights discrepancies occurring where observed data have more mass. Theorem 1 can be extended to accommodate any well-behaved weighting function. Consider the following estimator:

$$\theta_{T,S,h} = \arg\min_{\theta\in\Theta} \int_X \left[\frac{1}{S}\sum_{i=1}^{S} \pi^i_{T,h}(x;\theta) - \pi_T(x)\right]^2 w_T(x)dx, \tag{15}$$

where w_T is a sequence of weighting functions possibly dependent on data. In addition to the basic regularity conditions of appendix B (assumptions 10 and 11), we assume that sequence w_T satisfies:

Assumption 6. For all $x \in X$ and almost all points in the sample space, there exist two bounded functions $w^j_T(x,\lambda)$, $j=1,2$, such that $w_T(x) \equiv w_T(x,\lambda) = w^1_T(x,\lambda)/\, w^2_T(x,\lambda)$ and, for *fixed* λ, $w^j_T(x,\lambda) \xrightarrow{p} w^j(x,\lambda)$, $j=1,2$, x-pointwise, where $w^j(x,\lambda)$, $j=1,2$, are bounded functions such that for all $x \in X$, $w^1(x,\lambda) = w^2(x,\lambda)\cdot w(x,\lambda)$ for some function $w(x,\lambda)$. Finally, there exist three functions $w(x)$ and $w^j(x)$, $j=1,2$, such that $\lim_{\lambda\downarrow 0} w^j(x,\lambda) = w^j(x)$, $j=1,2$ and $w^1(x) = w^2(x)\cdot w(x)$ all $x \in X$.

Asymptotically unbounded weighting functions are not ruled out by assumption 6. Naturally,

a given unbounded weighting function does not jeopardize per se finiteness of the objective function in (15). Additional regularity conditions ensuring consistency of all of our resulting estimators are spelled out in assumption 10 in appendix B. In appendix B, we also show that under assumptions 6, 10 and 11, the asymptotic behavior of $\theta_{T,S,h}$ in (15) is as in theorem 1, with the exception that function Ψ is now replaced by:

$$\Psi(x) \equiv \left[\int_X |\nabla_\theta \pi(x;\theta_0)|_2 \, w(x)\, dx\right]^{-1} w(x)\, \nabla_\theta \pi(x;\theta). \tag{16}$$

The previous formula reveals that the asymptotic variance of the estimator depends indeed on the weighting function w at hand. However, a weighting function minimizing such an asymptotic variance is unknown, even in the case of fully observable diffusions.[10] In the next section, we show that in the case of fully observable diffusions, this problem can considerably be simplified through an appropriate change of the objective function in (15).

4.3 Conditional Density SNE, and Efficiency

This section introduces a modification of the SNE, and addresses efficiency issues within the case of fully observable diffusions. We show that by casting the estimation problem as a matching of conditional densities (instead of joint ones), our resulting estimator is asymptotically (first-order) efficient whenever the state y in (8) is fully observable.

To prepare the analysis, consider again vector $x \in X \subseteq \mathbb{R}^q$ in (9). For each t, partition x_t as $x_t = (z_t, v_t)$, where $z_t \equiv y_t^o \in Z \subseteq \mathbb{R}^{q^*}$ is the vector of observable variables, and $v_t \in V \subseteq \mathbb{R}^{q-q^*}$, is the vector of predetermined variables:

$$v_t \equiv \left(y_{t-1}^o, \cdots, y_{t-l}^o\right), \quad t = t_l \equiv 1+l, \cdots, T.$$

Consider the following conditional density matching estimator:

Definition 2. (CD-SNE) For each positive integer S, the Conditional Density SNE (CD-SNE) is the sequence $\{\theta_{T,S,h}\}_{h,T}$ given by:

$$\theta_{T,S,h} = \arg\min_{\theta\in\Theta} \int_Z \int_V \left[\frac{1}{S}\sum_{i=1}^{S} \pi_{T,h}^i(z|v;\theta) - \pi_T(z|v)\right]^2 w_T(z,v)dzdv, \tag{17}$$

where $\pi_T(z|v) \equiv \pi_T(z,v)/\pi_T(v)$, $\pi_{T,h}^i(z|v;\theta) \equiv \pi_{T,h}^i(z,v,\theta)/\pi_{T,h}^i(v,\theta)$ ($i = 1, \cdots, S$), and w_T is a sequence of weighting functions satisfying assumption 6.

[10] An exception arises exactly in an hypothetical i.i.d. case. Under the regularity conditions given in the general case of corollary 1 below, the optimal weighting function in the i.i.d. case is given by $w_T(x) = 1/(\pi_T(x) + \alpha_T)$, where α_T is as in section 2.2.

Under our assumptions, $\pi_T(z|v)$ is bounded (see, e.g. Chen, Linton and Robinson (2001, property 2)). Under the standard bandwidth assumption 5, $\pi_T(z|v) \xrightarrow{p} \pi(z|v) \equiv \pi(z,v)/\pi(v)$, (z,v)-pointwise. Chen, Linton and Robinson (2001) provide further discussion on bandwidth selection strategies for the estimation of conditional densities and dependent observations. Our practical implementation of the CD-SNE relies on some of their suggestions, which we briefly illustrate in Appendix E.

The following result provides the asymptotic properties of the CD-SNE.

Theorem 2. In addition to assumptions 1, 2, 3-i, 4, let assumptions 7, 12 and 13 in the appendix hold; then, as $h \downarrow 0$ and $T \to \infty$, the CD-SNE is (weakly) consistent. Furthermore, let assumptions 1-5 and assumptions 7, 9, 12 and 13 in the appendix hold; then, as $h \downarrow 0$ and $T \to \infty$,

$$\sqrt{T}(\theta_{T,S,h} - \theta_0) \xrightarrow{d} N(0, V),$$

where $V \equiv \mathsf{var}(\Psi_1) + 2\sum_{j=1}^{\infty}\mathsf{cov}(\Psi_1, \Psi_{1+j})$, $\Psi \equiv D_4^{-1}[\frac{1}{S}\sum_{i=1}^{S}(D_1^i + D_3^i) + D_2]$, and the terms $\{D_1^i\}_{i=1}^S$, D_2, $\{D_3^i\}_{i=1}^S$ and D_4 are given in appendix C.2.

Proof. In appendixes C.1 and C.2. ■

The previous theorem contains a general statement about the asymptotic behavior of the CD-SNE. It holds for any weighting function satisfying our regularity conditions, and even when the state vector y is partially observed. Furthermore, the variance structure of the CD-SNE differs from the one characterizing the asymptotic distribution of the SNE in section 4.2. In the CD-SNE case, one has to cope with additional terms arising because conditional densities are estimated as ratios of two densities (joints over marginals). Such additional terms are represented by $\{D_3^i\}_{i=1}^S$. As we show in appendix C.3, there exist weighting functions making these terms identically zero. In those cases, the variance terms in theorem 2 have the same representation as the variance terms in section 4.2. Proposition 3 in appendix C.3 summarizes our results on these issues.

We now heuristically demonstrate that when vector y is fully observable, there exists a simple choice of w_T that makes the CD-SNE asymptotically attain the Cramer-Rao lower bound. (As it turns out, such a function belongs to the class of functions considered in proposition 3 (see appendix C.3).) Specifically, consider the following weighting function:

$$w_T(z,v) = \frac{\pi_T(v)^2}{\pi_T(z,v) + \alpha_T}, \tag{18}$$

where α_T is any strictly positive sequence satisfying $\alpha_T \to 0$. The simple role played by sequence α_T is to ensure that w_T does not blow up in finite samples. And asymptotically, our objective function in (17) is finite under additional regularity conditions given in the appendix (see

assumption 12 in appendix C.1). If w_T is as in (18), the criterion in (17) reduces to:

$$\int_Z \int_V \left[\frac{1}{S} \sum_{i=1}^{S} \left(\frac{\pi^i_{T,h}(z|v;\theta)}{\pi_T(z|v)} - 1 \right) \right]^2 \frac{\pi_T(z,v)^2}{\pi_T(z,v) + \alpha_T} dzdv,$$

which asymptotically becomes a Neyman's chi-squared measure of distance.

The first order conditions satisfied by the CD-SNE are:

$$0 = \int_Z \int_V \frac{1}{S} \sum_{i=1}^{S} \left[\frac{\pi^i_{T,h}(z|v;\theta_{T,S,h})}{\pi_T(z|v)} - 1 \right] \frac{\pi_T(z,v)^2}{\pi_T(z,v) + \alpha_T} \frac{\nabla_\theta \pi^i_{T,h}(z|v;\theta_{T,S,h})}{\pi_T(z|v)} dzdv,$$

and a Taylor's expansion about θ_0 then yields that in large samples,

$$\sqrt{T}(\theta_{T,S,h} - \theta_0) \sim -J_{T,S,h}(\theta_0)^{-1} \left[H^0_{T,S,h}(\theta_0) + \frac{1}{S} \sum_{i=1}^{S} H^i_{T,S,h}(\theta_0) \right], \tag{19}$$

where $\sim$ stands for asymptotic equivalence (in distribution), and

$$\begin{aligned}
J_{T,S,h}(\theta_0) &\equiv \frac{1}{S} \sum_{i=1}^{S} \int_Z \int_V \left| \nabla_\theta \ln \pi^i_{T,h}(z|v;\theta_0) \right|_2 \pi_T(z,v)\, dzdv, \\
H^0_{T,S,h}(\theta_0) &\equiv \int_Z \int_V \sqrt{T} \left[\pi_T(z,v) - E(\pi_T(z,v)) \right] \left[\frac{1}{S} \sum_{i=1}^{S} \nabla_\theta \ln \pi^i_{T,h}(z|v;\theta_0) \right] dzdv, \\
H^i_{T,S,h}(\theta_0) &\equiv \int_Z \int_V \left[\frac{1}{S} \sum_{i=1}^{S} \sqrt{T} \left(\pi^i_{T,h}(z,v;\theta_0) - E(\pi^i_{T,h}(z,v;\theta_0)) \right) \right] \left[\nabla_\theta \ln \pi^i_{T,h}(z|v;\theta_0) \right] dzdv.
\end{aligned}$$

By a law of large numbers and a central limit theorem developed in appendix C.2, $J_{T,S,h}(\theta_0) \xrightarrow{p} E\left[|\nabla_\theta \ln \pi(z|v;\theta_0)|_2 \right]$ and

$$H^i_{T,S,h}(\theta_0) \xrightarrow{d} N\left(0,\ \text{var}\left(\nabla_\theta \ln \pi(z|v;\theta_0)\right)\right), \qquad i = 0, 1, \cdots, S, \tag{20}$$

as $h \downarrow 0$ and $T \to \infty$.

By the Markov property of a diffusion (see, e.g., Arnold (1992), theorem 9.2.3 p. 146),

$$\frac{1}{T} \log \pi\left(\{y_t\}_{t=1}^T; \theta_0 \right) = \frac{1}{T} \log \pi(y_1;\theta_0) + \frac{1}{T} \sum_{t=2}^{T} \log \pi(y_t|y_{t-1};\theta_0). \tag{21}$$

Since the system is fully observable, $z_t = y_t$, and $\nabla_\theta \ln \pi(y_t|y_{t-1};\theta_0)$ is a martingale difference with respect to the sigma-fields generated by y. Therefore, by taking $v_t = y_{t-1}$, we have that the variance of the CD-SNE (rescaled by $(1 + S^{-1})$) does attain the Cramer-Rao lower bound

$$E\left[|\nabla_\theta \ln \pi(y_t|y_{t-1};\theta_0)|_2 \right]^{-1}.$$

Naturally, the previous arguments are heuristic. Nevertheless, the efficiency result can be made rigorous, as in the following corollary:

Corollary 1. (Cramer-Rao lower bound) Suppose that the state is fully observable (i.e., $q^* = d$). Let the CD-SNE match one-step ahead conditional densities (i.e., $(z, v) \equiv (z_t, z_{t-1})$ in (17)) and let $w_T(z, v) = \pi_T(v)^2 / [\pi_T(z, v) + \alpha_T]$, where α_T is any strictly positive sequence satisfying $\alpha_T \xrightarrow{p} 0$. Then, under the assumptions in theorem 2, the CD-SNE is as in theorem 2, and it attains the Cramer-Rao lower bound as $S \to \infty$.

Proof. In appendix C.3. ■

Put differently, the previous efficiency result follows because the weighting function in (18) makes the resulting estimator asymptotically equivalent to the score function when the system is fully observable (see eqs. (19) and (20)). We now turn to analyze how our CD-SNE can be used to implement parameter estismation of partially observed systems coupled with new information provided by asset pricing theories. We study both feasibility and asymptotic efficiency of our resulting estimators.

5 Asset pricing, prediction functions and statistical efficiency

This section analyzes situations in which the original partially observed system (8) can be estimated by augmenting it with a number of observable deterministic functions of the state. In many situations of interest, such deterministic functions are suggested by asset pricing theories in a natural way. Typical examples include derivative asset price functions or any deterministic function(als) of asset prices (e.g., asset returns, bond yields, implied volatility, etc.). The idea to use predictions of asset pricing theories to improve the fit of models with unobservable factors is not new (see, e.g., Christensen (1992), Pastorello, Renault and Touzi (2000), Chernov and Ghysels (2000) and Singleton (2001, sections 3.2 and 3.3)). In this section, we provide a theoretical description of the mechanism leading to efficiency within the class of our estimators.

We consider a standard Markov pricing setting. For fixed $t \geq 0$, we let M be the expiration date of a contingent claim with rational price process $c = \{c(y(\tau), M - \tau)\}_{\tau \in [t, M)}$, and let $\{\zeta(y(\tau))\}_{\tau \in [t, M]}$ and $\Pi(y)$ be the associated intermediate payoff process and final payoff function, respectively. Let $\partial / \partial \tau + L$ be the usual infinitesimal generator of (8) taken under the risk-neutral measure.[11] In a frictionless economy without arbitrage opportunities, c is the solution to

[11] See, e.g., Duffie (1996) for details on the change of measure for diffusion models in financial applications.

the following partial differential equation:

$$\begin{cases} 0 = \left(\dfrac{\partial}{\partial \tau} + L - R\right) c(y, M-\tau) + \zeta(y),\ \forall (y,\tau) \in Y \times [t, M) \\ \\ c(y,0) = \Pi(y),\ \forall y \in Y \end{cases} \tag{22}$$

where $R \equiv R(y)$ is the short-term rate. We call prediction function any continuous and twice differentiable function $c\left(y; M-\tau\right)$ solution to the partial differential equation (22).

We now augment system (8) with $d - q^*$ prediction functions. Precisely, we let:

$$C(\tau) \equiv \left(c\left(y(\tau), M_1 - \tau\right), \cdots, c\left(y(\tau), M_{d-q^*} - \tau\right)\right), \quad \tau \in [t, M_1]$$

where $\{M_i\}_{i=1}^{d-q^*}$ is an increasing sequence of fixed maturity dates. Furthermore, we define the measurable vector valued function:

$$\phi\left(y(\tau); \theta, \gamma\right) \equiv \left(y^o(\tau), C\left(y(\tau)\right)\right), \quad \tau \in [t, M_1], \quad (\theta, \gamma) \in \Theta \times \Gamma,$$

where $\Gamma \subset \mathbb{R}^{p_\gamma}$ is a compact parameter set containing additional parameters. These new parameters arise from the change of measure leading to the pricing model (22) (see, e.g., example 3 below), and are now part of our estimation problem.

We assume that the pricing model (22) is correctly specified. That is, all contingent claim prices in the economy are taken to be generated by the prediction function $c(y, M-\tau)$ for some $(\theta_0, \gamma_0) \in \Theta \times \Gamma$. For simplicity, we also consider a stylized situation in which all contingent claims have the same contractual characteristics specified by $\mathsf{C} \equiv (\zeta, \Pi)$. More generally, one may define a series of classes of contingent claims $\{\mathsf{C}_j\}_{j=1}^J$, where class of contingent claims j has contractual characteristics specified by $\mathsf{C}_j \equiv \left(\zeta_j, \Pi_j\right)$.[12] The number of prediction functions that we would introduce in this case would be equal to $d - q^* = \sum_{j=1}^J M^j$, where M^j is the number of prediction functions within class of assets j. To keep the presentation simple, we do not consider such a more general situation here.

Example 3. (Example 2 continued) In the setting of model (11), $y = (y^o, y^u) \equiv (Q, \sigma)$. If in addition $\xi \equiv 1/2$, model (11) collapses to the Heston's (1993) affine stochastic volatility model, with $\theta \equiv (\mu, w, \varphi, \psi, \rho)$. The price of a European option is given by the prediction function $c\left(Q(\tau), \sigma(\tau), M-\tau; \theta, \gamma\right)$, where γ is a parameter related to the price of volatility risk.[13] The augmented price system is thus $(Q(\tau), c(Q(\tau), \sigma(\tau), M-\tau; \theta, \gamma))$.

[12] As an example, assets belonging to class $\mathcal{C}_1$ can be European options; assets belonging to class $\mathcal{C}_1$ can be bonds; and so on.

[13] See eq. (6) (p. 329) and formulae # (10)-(18) (p. 330-331) in Heston (1993).

Our objective is to provide estimators of the parameter vector (θ_0, γ_0) under which observations were generated. In exactly the same spirit as the previous sections, we want our CD-SN estimator of (θ_0, γ_0) to make the finite dimensional distributions of ϕ implied by model (8) and (22) as close as possible to their sample counterparts. Let $\Phi \subseteq \mathbb{R}^d$ be the domain on which ϕ takes values. As illustrated in Figure 2, our program is to move from the "unfeasible" domain Y of the original state variables in y (observables and not) to the domain Φ on which all observable variables take value. Ideally, we would like to implement such a change in domain in order to recover as much information as possible on the original unobserved process in (8). Clearly, ϕ is fully revealing whenever it is globally invertible. However, we will show that our methods can be implemented even when ϕ is only locally one-to-one. Further intuition on this distinction will be provided after the statement of theorem 3 below.

An important feature of the theory in this section is that it does not hinge upon the availability of contingent prices data covering the same sample period covered by the observables in (8). First, the price of a given contingent claim is typically not available for a long sample period. As an example, available option data often include option prices with a life span smaller than the usual sample span of the underlying asset prices; in contrast, it is common to observe long time series of option prices having the same maturity. Second, the price of a single contingent claim depends on time-to-maturity of the claim; therefore, it does not satisfy the stationarity assumptions maintained in this paper. To address these issues, we deal with data on assets having the same characteristics at each point in time. Precisely, consider the data generated by the following random processes:

Definition 3. (Intertertemporal (ℓ, N)-cohort of contingent claim prices) Given a prediction function $c(y; M - \tau)$ and a N-dimensional vector $\ell \equiv (\ell_1, \cdots, \ell_N)$ of *fi*xed maturities, an intertemporal (ℓ, N)-cohort of contingent claim prices is any collection of contingent claim price processes $c(\tau, \ell) \equiv (c(y(\tau), \ell_1), \cdots, c(y(\tau), \ell_N))$ ($\tau \geq 0$) generated by the pricing model (22).

Consider for example a sample realization of three-months at-the-money option prices, or a sample realization of six-months zero-coupon bond prices. Long sequences such as the ones in these examples are common to observe. If these sequences were generated by (22), as in definition 3, they would be deterministic functions of y, and hence stationary. We now develop conditions ensuring both feasibility and first-order efficiency of the CD-SNE procedure as applied to this kind of data. Let $\overline{a}$ denote the matrix having the first q^* rows of a, and let ∇C denote the Jacobian of C with respect to y. We have:

Theorem 3. (Asset pricing and Cramer-Rao lower bound) Suppose to observe an intertemporal $(\ell, d-q^*)$-cohort of contingent claim prices $c(\tau,\ell)$, and that there exist prediction functions C in $\mathbb{R}^{d-q^*}$ with the property that for $\theta=\theta_0$ and $\gamma=\gamma_0$,

$$\begin{pmatrix} \overline{a}(\tau)\cdot a(\tau)^{-1} \\ \nabla C(\tau) \end{pmatrix} \neq 0, \quad P\otimes d\tau\text{-a.s.} \quad \text{all } \tau\in[t,t+1], \tag{23}$$

where C satisfies the initial condition $C(t)=c(t,\ell)\equiv(c(y(t),\ell_1),\cdots,c(y(t),\ell_{d-q^*}))$. Let $(z,v)\equiv\left(\phi_t^c,\phi_{t-1}^c\right)$, where $\phi_t^c=(y^o(t),c(y(t),\ell_1),\cdots,c(y(t),\ell_{d-q^*}))$. Then, under the assumptions in theorem 2, the CD-SNE has the same properties as in theorem 2, with the variance terms being taken with respect to the fields generated by ϕ_t^c. Finally, suppose that ϕ_t^c is Markov, and set $w_T(z,v)=\pi_T(z)^2/\left[\pi_T(z,v)+\alpha_T\right]$, where α_T is as in corollary 1. Then, the CD-SNE attains the Cramer-Rao lower bound (with respect to the fields generated by ϕ_t^c) as $S\to\infty$.

Proof. In appendix D. ■

According to theorem 3, our CD-SNE is feasible whenever ϕ is locally invertible for a time span equal to the sampling interval. As figure 2 illustrates, condition (23) is satisfied whenever ϕ is locally one-to-one and onto.[14] If ϕ is also globally invertible for the same time span, ϕ^c is Markov. The last part of this theorem then says that in this case, the CD-SNE is asymptotically efficient. We emphasize that such an efficiency result is simply about first-order efficiency in the joint estimation of θ and γ given the observations on ϕ^c. We are not claiming that our estimator is first-order efficient in the estimation of θ in the case in which y is fully observable.

Naturally, condition (23) does not ensure that ϕ is globally one-to-one and onto. In other terms, ϕ might have many locally invertible restrictions.[15] In practice, ϕ might fail to be globally invertible because monotonicity properties of ϕ may break down in multidimensional diffusion models. In models with stochastic volatility, for example, option prices can be decreasing in the underlying asset price (see Bergman, Grundy and Wiener (1996)); and in the corresponding stochastic volatility yield curve models, medium-long term bond prices can be increasing in the short-term rate (see Mele (2003)). Intuitively, these pathological situations may occur because there is no guarantee that the solution to a stochastic differential system is nondecreasing in the initial condition of one if its components - as it is instead the case in the scalar case.

When all components of vector y^o represent the prices of assets actively traded in frictionless markets, (23) corresponds to a condition ensuring market completeness in the sense of Harrison and Pliska (1983). As an example, condition (23) for model (11) is $\partial c/\partial\sigma\neq 0$ $P\otimes d\tau$-a.s. This

[14] Local invertibility of ϕ means that for every $y\in Y$, there exists an open set Y_* containing y such that the restriction of ϕ to Y_* is invertible. And ϕ is locally invertible on Y_* if $\det J\phi\neq 0$ (where $J\phi$ is the Jacobian of ϕ), which is condition (23).

[15] As an example, consider the mapping $\mathbb{R}^2\mapsto\mathbb{R}^2$ defined as $\phi(y_1,y_2)=(e^{y_1}\cos y_2, e^{y_1}\sin y_2)$. The Jacobian satisfies $\det J\phi(y_1,y_2)=e^{2y_1}$, yet ϕ is 2π-periodic with respect to y_2. For example, $\phi(0,2\pi)=\phi(0,0)$.

condition is satisfied by the Heston's model in example 3. In fact, Romano and Touzi (1997) showed that within a fairly general class of stochastic volatility models, option prices are always strictly increasing in σ whenever they are convex in Q.[16] This suggests that many two-factor stochastic volatility models may be efficiently estimated along the lines indicated in theorem 3.[17]

Theorem 3 can be used to implement efficient estimators in other complex multidimensional models. This is the case of the short-term rate model (10) of example 1. Let $u^{(i)} = u\left(r(\tau), \sigma(\tau), \ell(\tau); M_i - \tau\right)$ denote the time τ rational price of a pure discount bond expiring at $M_i \geq \tau$, $i = 1, 2$ and take $M_1 < M_2$. Let $\phi \equiv (r, u^{(1)}, u^{(2)})$. Condition (23) for model (10) is then:

$$u_{\sigma}^{(1)} u_{\ell}^{(2)} - u_{\ell}^{(1)} u_{\sigma}^{(2)} \neq 0, \quad P \otimes dt\text{-a.s.} \quad \tau \in [t, t+1] \tag{24}$$

where subscripts denote partial derivatives. It is easily checked that this same condition must be satisfied by models with correlated Brownian motions and by yet more general models. Classes of models of the short-term rate for which condition (24) holds are more intricate to identify than in the European option pricing literature mentioned above (see Mele (2003) for an analysis regarding general qualitative properties of bond price functions up to three-factor models). Finally, it is well-known that bond prices can be computed fastly within the class of the exponential-affine models (see, e.g., Dai and Singleton (2000) and Duffie, Pan and Singleton (2000)). Approximate solutions for nonlinear models can also be obtained by truncation of the formula: $u(y; M - \tau) = \sum_{n=0}^{\infty} \left(f_n(y) \cdot (M - \tau)^n / n! \right)$, where $f_{n+1}(y) = L f_n(y) - r f_n(y)$, $f_0 \equiv 1$.[18]

6 Monte Carlo experiments

Our estimation methodology relies on both simulations and nonparametric techniques. Therefore, it is important to investigate whether the finite sample behavior of our estimators is well approximated by the asymptotic theory. In this section, we conduct Monte Carlo experiments on both one-factor and two-factor models. As regards the one-factor case, we consider the celebrated Vasicek (1977) model of the short-term rate:

$$dr(\tau) = b_2 \times (b_1 - r(\tau))\, d\tau + a_1 \times dW(\tau), \tag{25}$$

[16] Convexity with respect to Q is not a general property of option prices in models with stochastic volatility. While concavity of option prices should not be a property of real data, the only conditions that are currently known to ensure convexity of option prices are that the risk-neutral drift of the volatility process is independent of Q (see Romano and Touzi (1997)).

[17] While the model of example 3 can be implemented very easily by using the closed-form solution provided by Heston (1993), numerical solutions of other models can be obtained very fastly for short-term (say, three months) at-the-money options.

[18] This formula represents a straightforward multidimensional generalization of the one given by Chapman et al. (1999, proposition 3 p. 781).

where b_1, b_2 and a_1 are parameters. As regards the two-factor case, we consider a simple extension of (25) in which volatility is a process with constant elasticity of variance:

$$\begin{cases} dr(\tau) &= b_2 \times (b_1 - r(\tau))\, d\tau + a_1 \times \sigma(\tau) dW_1(\tau) \\ \\ d\sigma(\tau) &= b_3 \times (1 - \sigma(\tau))\, d\tau + a_2 \times \sigma(\tau) dW_2(\tau) \end{cases} \tag{26}$$

where W_1 and W_2 are uncorrelated, and b_3, a_2 are further parameters. Models (25) and (26) have both been chosen to keep the computational burden of the experiments as low as possible.

6.1 The common setup

In all experiments, data are assumed to be sampled weekly. Typically, we consider sample sizes of 500 and 1000 observations. The high-frequency generator is the Euler-Maruyama scheme (12), with $h = 1/(5 \times 52)$. We use $S = 5$ path simulations, and every piece of the experiment is made up of 1000 replications. Nonparametric density estimates are implemented through Gaussian kernels. We consider highly persistent data generating processes. Therefore, we initially pay special attention to bandwidth choice. Bandwidth choice in the case of conditional distributions and dependent data has been addressed by Chen, Linton and Robinson (2001). We closely follow the suggestions in their paper, and select the bandwidth by searching over values minimizing the average asymptotic mean integrated squared error. In appendix E, we provide further details on the factual implementation of this procedure. We evaluate the various objective functions at sample points. That is, we consider the empirical counterparts of (14) and (17). As an example, the empirical counterpart of the criterion in (14) is $T^{-1} \sum_{t=1}^{T} [S^{-1} \sum_{i=1}^{S} \pi_{T,h}^{i}(x_t; \theta) - \pi_T(x_t)]^2$. The choice between the two alternatives (i.e. (14) and (17) versus their empirical counterparts) is mainly driven by computational issues because the empirical counterparts to (14) and (17) do not involve any integration issue. Under assumption 5, the two alternatives lead to asymptotically equivalent criteria.[19]

6.2 Vasicek

The baseline parametrization of the Vasicek model (25) is given by $b_1 = 0.06$, $b_2 = 0.5$ and $a_1 = 0.03$. These parameter values imply that the resulting model-generated data have approximately the same mean, variance, and autocorrelations as US short-term interest rate. Table 2A reports estimation results obtained with this baseline parametrization. We provide mean, median, standard deviation, and the root mean square error of both the CD-SNE (with optimal weighting function and optimal bandwidth) and the MLE. When the size of the simulated

[19] Our estimators are implemented with Fortran90. The objective functions are optimized through a DFP algorithm, with a convergence criterion of the order of 10^{-5}.

samples is 1000, the performance of the two methods is comparable in terms of precision of the estimates (see Panel A). Specifically, the CD-SNE is more precise than the MLE as regards the estimation of the parameter b_2 tuning the persistence of r; and the MLE is more precise than the CD-SNE as regards the estimation of the diffusion parameter a_1. As regards bias issues, the MLE tends to underestimate the dependence of the data. Such a finite sample property of the MLE is very well-known. Interestingly, this phenomenon disappears when the model is estimated with the CD-SNE. When the simulated samples have smaller size, the variability of the estimates increases with both the CD-SNE and the MLE (see Panel B). As regards b_2, the mean bias of the MLE doubles. The mean bias of the CD-SNE increases as well, but it remains small if compared to the MLE mean bias.

As is well-known, the practical performance of nonparametric methods hinges on the proper choice of the bandwidth parameter. We then analyze the effects of the bandwidth selection on the performance of our CD-SNE. We implement two experiments: in one, we double the size of the optimal bandwidth (see Table 2B, panel A); in another, we halve the size of the optimal bandwidth (see Table 2B, panel B). On average, the experiments reported in Table 2A produced a bandwidth choice of $\lambda = 1.65 \times 10^{-2}$. The results in Table 2B now suggest that while not-optimal bandwidth choice produces some effects on the estimates, these effects are only marginal (with the CD-SNE being slightly more sensitive to oversmoothing than to undersmoothing). If any, these effects are visible more in terms of precision than in terms of bias of the estimates. This is perfectly in line with our analytical results.

Next, we tackle the choice between the CD-SNE and the SNE. The results are reported in Table 2C. We consider two experiments. In the first one, we compare the CD-SNE with the SNE in definition 1 (see eq. (14)). That is, we match the joint density of any two adjacent observations $\pi_T(r_t, r_{t-1})$, and use $w_T(r_t, r_{t-1}) = \pi_T(r_t, r_{t-1})$ as a weighting function. In the second experiment, we modify the definition of the SNE and replace simulated nonparametric estimates $S^{-1}\sum_{i=1}^{S}\pi^i_{T,h}(r_t, r_{t-1}; \theta)$ of the joint density $\pi^{\text{vas}}(r_t, r_{t-1}; \theta)$ (say) with $\pi^{\text{vas}}(r_t, r_{t-1}; \theta)$.[20] More precisely, the objective function in the second experiment takes the form:

$$\int_{(r_t, r_{t-1}) \in \mathbb{R}^2} \left[\pi^{\text{vas}}(r_t, r_{t-1}; \theta) - \pi_T(r_t, r_{t-1})\right]^2 \pi_T(r_t, r_{t-1}) dr_t dr_{t-1}. \tag{27}$$

As we demonstrated in section 4.3, the CD-SNE can be made first order efficient in the case of fully observed systems (see corollary 1). Both experiments then aim at investigating the effects of suboptimal choice of the objective function on the estimates produced by our class of estimators. At the same time, the second experiment also allows us to gauge the effects arising from dismissing the "twin-smoothing" idea discussed in section 2.2. The results reported in Panel

[20] As is well known, the transition density $\pi^{\text{vas}}(r_s|r_t; \theta)$ from date t to date s ($s > t$) is Gaussian with expectation equal to $b_1/b_2 + [r(t) - (b_1/b_2)]\exp(-b_2(s-t))$ and variance equal to $\left[b_3^2/(2b_2)\right]\left[1 - \exp(-2b_2(s-t))\right]$. The marginal density is obtained by letting $s \to \infty$.

A clearly demonstrate that moving from CD-SNE to SNE causes an increase in the variability of the estimates; this result is pronounced as regards the diffusion parameter a_1. As regards the second experiment, the results reported in panel B reveal a much larger variability of the estimates of b_2 and a_1. More interestingly, these two parameters are estimated with large biases. In particular, minimizing (27) underestimates the dependence of data (the mean bias of b_2 is 0.07) and overestimates the diffusion coefficient by 20% (the mean bias of a_1 is 0.006). Again, these results are perfectly consistent with our theoretical explanation of possible second order bias effects on parameter estimates (see section 2.2). They also extend to the dependent case results of our simple Monte Carlo experiments in section 2.4.

Finally, we address the issue arising from data generating process as characterized by different levels of persistence. We alter the baseline case of Table 2A and consider two cases with lower dependence, one with $b_2 = 1$ and another with $b_2 = 5$. The corresponding results are in Table 2D. Both the MLE and the CD-SNE produce relatively better results than in the baseline case. In particular, the MLE bias of the b_2 parameter seems to decrease. As regards the CD-SNE, we have observed the interesting phenomenon that on average, the optimal bandwidth increases with the dependence of data. As we mentioned earlier, the optimal average bandwidth was equal to 1.65×10^{-2} in the baseline case ($b_2 = 0.5$). In this set of experiments, the optimal bandwidth was equal to 1.51×10^{-2} on average when $b_2 = 1$; and it was equal to 1.20×10^{-2} on average when $b_2 = 5$.

6.3 Stochastic Volatility

The baseline parametrization of the stochastic volatility model (26) is $b_1 = 0.06$, $b_2 = 0.5$, $a_1 = 0.03$, $b_3 = 1.0$ and $a_2 = 0.3$. It implies that the unobservable volatility process is strongly dependent, but not as strongly as the r process itself. Such a difference in persistence that we are imposing is consistent with some empirical evidence that we have gathered on US short-term rate data (results are available upon request). However, we will also consider the reverse case in which $b_2 = 0.5$ and $b_3 = 0.4$.

Initially, we implement the CD-SNE with the weighting function in (18) of section 4.3 (see corollary 1), and we match the transition density of any two adjacent observations of r. While the joint process (r, σ) is Markov in (26), r is not if taken by itself. Therefore, the conditions of corollary 1 do not apply anymore, and the weighting function in (18) does not make the resulting CD-SNE asymptotically efficient in the case under consideration. (In the remainder we keep on referring to this weighting function as "optimal" due to a lack of better terminology.)

Table 3A reports the results obtained with this baseline set-up. We consider simulated samples of both 1000 and 500 observations. As regards the larger simple size case, the precision and the bias associated with b_1, b_2 and a_1 are of the same order of magnitude as in the observable case (see Table 2A, Panel A). While larger standard deviations are associated to the parameters b_3

and a_2, we do not observe any sizeable bias in the estimates of these parameters. Apart from a somehow larger sample variability, results are similar in the case of smaller sample sizes (see Panel B). Next, we compare the performance of the CD-SNE with the SNE. We implement the SNE (see eq. (14)) using $\pi_T(r_t, r_{t-1})$ as a weighting function. Table 3B reports the results. These results improve upon the ones obtained with the CD-SNE implemented through the weighting function in (18) (particularly, in terms of the variability of the estimates). Finally, we investigate the impact of persistence in volatility on the estimates. Specifically, we set $b_3 = 0.4$. Table 3C reports the results. We do not observe major differences between this case and the case with lower persistence in volatility.

7 Conclusions

This paper has introduced new methods to estimate the parameters of the typical partially observed diffusion models arising in financial applications. The building block of these methods is indeed very simple. It consists in simulating the model of interest for the purpose of recovering the corresponding density function. Our estimators are the ones which make densities on simulated data as close as possible to their empirical counterparts. We made use of ideas in the statistical literature to build up convenient measures of closeness of densities. Our estimators are easy to implement, fast to compute and in the special case of fully observed Markov systems, they can attain the same asymptotic efficiency as the maximum likelihood estimator. Furthermore, Monte Carlo experiments revealed that their finite sample performance is very satisfactory, even in comparison to the maximum likelihood benchmark.

Using simulations to recover model-implied density is not only convenient "just" because it allows one to recover estimates of densities unknown in closed-form. We demonstrated that this feature of our methods stands as a great improvement upon alternative techniques matching "closed-form" model-implied densities to data-implied densities. Consistenly with our asymptotic theory, finite sample results suggest that a careful choice of both the measures of closeness of density functions and the bandwidth functions does enhance the performance of our estimators, but only in terms of their precision. Our trick to use simulations to recover model-implied densities makes our estimators attain a high degree of accuracy in terms of unbiasedness, even in cases of unsophisticated objective functions and/or bandwidth selection procedures.

This paper has illustrated how to implement our methods to estimate the typical continuous time models arising in finance. These methods are flexible and can be adapted to address related estimation problems. As an example, the typical (discrete time) Markov models arising in applied macroeconomics may also be estimated with our methods. In these cases, too, the previous asymptotic efficiency and encouraging finite sample properties make our methods stand as a promising alternative to previous simulation-based inference methods.

Appendix

Preliminaries

Appendix A through D present regularity conditions omitted in the main text, and all proofs. To facilitate the presentation, proofs regarding asymptotic normality are organized with an hypothetical i.i.d. case being proven in the first place. Given these preliminary results and the mixing conditions of the main text, we will show that the extension to time series can be made with a mere change in notation. While it is conjectured that our theory can be developed with the functional differentiation methods of Aït-Sahalia (1994), here we adopt standard tools of analysis. A final appendix E provides a succinct description of our bandwidth selection procedure.

First, we remind some basic definitions pertaining to kernels that are of interest in this paper. A symmetric kernel K is a symmetric function around zero that integrates to one. Kernels considered in this paper are symmetric bounded kernels which are continuously differentiable with bounded derivatives up to the fourth order. A kernel K is said to be of the r-th order if: 1) $\forall \mu \in \mathbb{N}^q : |\mu| \in \{1, \cdots, r-1\}$ ($|\mu| \equiv \sum_{j=1}^{q} \mu_j$), $\int u_1^{\mu_1} \cdots u_q^{\mu_q} K(u)du = 0$; 2) $\exists \mu \in \mathbb{N}^q : |\mu| = r$ and $\int u_1^{\mu_1} \cdots u_q^{\mu_q} K(u)du \neq 0$; and 3) $\int \|u\|^r K(u)du < \infty$.

The following pieces of notation are then employed throughout the appendix. First, we write $x(\theta) \equiv \{x_t(\theta)\}_{t=t_l}^{T}$ to denote one hypothetical sequence that it would be possible to observe if the true parameter in (8) were θ. The real positive number h^0 denotes the same critical number introduced in assumption 4. Weak convergence as $h \downarrow 0$ (see footnote 8) is denoted with $\Rightarrow$ and convergence in probability is denoted with $\xrightarrow{p}$. If b is a column vector, $|b|_2$ denotes the outer product $b \cdot b^\top$. For any real number a, $|a|$ is the absolute value of a, and $|A|_{i,j}$ is the absolute value of the (i,j)-entry of a matrix A. Also, $\mathbf{0}_n$ is a column vector of n zeros. To keep notation as simple as possible, we omit to mention that all statements hold for almost all $\omega, \widetilde{\omega} \in \Omega$, where ω and $\widetilde{\omega}$ denote a sample point and S simulated points in the sample space Ω. We let:

$$\widetilde{\pi}_T(x;\theta) \equiv \frac{1}{S}\sum_{i=1}^{S} \pi_{T,h}^i(x;\theta) \quad \text{and} \quad \widetilde{\pi}_T(z|v;\theta) \equiv \frac{1}{S}\sum_{i=1}^{S} \frac{\pi_{T,h}^i(z,v;\theta)}{\pi_{T,h}^i(v;\theta)},$$

where $x \in X \subseteq \mathbb{R}^q$, $z \in Z \subseteq \mathbb{R}^{q^*}$ and $v \in V \subseteq \mathbb{R}^{q-q^*}$, as in the main text. The expectation of the kernel for a given bandwidth value λ is denoted as:

$$m(x,\theta) \equiv K * \pi(x;\theta) = \frac{1}{\lambda^q} \cdot \int K\left(\frac{x-u}{\lambda}\right)\pi(u;\theta)du.$$

We set:

$$\begin{aligned} L_T(\theta) &\equiv \int \left[\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right]^2 w_T(x,\lambda)dx; \\ L(\theta) &\equiv \int \left[m(x;\theta) - m(x;\theta_0)\right]^2 w(x,\lambda)dx. \end{aligned}$$

In appendix A, $w_T(x,\lambda) \equiv \pi_T(x)$ and $w(x,\lambda) \equiv m(x;\theta_0)$. In the remaining appendixes, both $w_T(x,\lambda)$ and $w(x,\lambda)$ are as in assumption 6. In appendix C,

$$\begin{aligned} L_T(\theta) &\equiv \int \left[\widetilde{\pi}_T(z|v;\theta) - \pi_T(z|v)\right]^2 w_T(x,\lambda)dx; \\ L(\theta) &\equiv \int \left[n(z,v,\theta) - n(z,v,\theta_0)\right]^2 w(x,\lambda)dx. \end{aligned}$$

where $n(z,v,\theta) \equiv \lambda^{-q^*} \cdot m(z,v,\theta)/m(v,\theta)$. In all appendixes, the previously defined asymptotic criteria are required to satisfy the following regularity and identifiability conditions:

Assumption 7. For all $\theta \in \Theta$, $L(\theta)$ is continuous, and the equation $L(\theta) = 0$ has exactly one solution in the interior of Θ.

A. Proof of theorem 1

A.1 Consistency

Consistency of the SNE is ensured by the following additional assumption:

Assumption 8. There exists a $\alpha > 0$ and a sequence κ_T bounded in probability as T becomes large such that for all $h < h^0$ and all $\theta^+, \theta \in \Theta$

$$\left|L_T(\theta^+) - L_T(\theta)\right| \leq \kappa_T \cdot \left\|\theta^+ - \theta\right\|_2^\alpha. \tag{A1}$$

Assumptions 7 and 8 are high level assumptions. Their role is quite standard and it will be further elucidated after the statement of proposition 1 below.

Remark 1. If data are smoothed with the popular Gaussian kernels, assumption 8 holds as $h \downarrow 0$ if a and b satisfy local Lipschitz and growth conditions with respect to θ. Indeed, according to Pedersen (1994, thm. 5 p. 23-24) and Friedman (1975, section 5 p. 117-123), $x_t(\theta)$ is differentiable with respect to θ in the L^2-sense under the previous Lipschitz and growth conditions. And again by the previous results of Pedersen and Friedman, as $h \downarrow 0$, $\nabla_\theta \widetilde{\pi}_T$ and $\nabla_{\theta\theta} \widetilde{\pi}_T$ stay bounded in the case of Gaussian kernels. As we emphasize in the next remark, boundedness of these derivatives then implies that assumption 8 holds for $h \downarrow 0$.

Remark 2. An example of conditions under which assumption 8 does hold is a global modulus of continuity condition on $\widetilde{\pi}_T(x;\cdot)$ similar to the standard one used by Duffie and Singleton (1993,

p. 938) in a related problem:

$$\forall x \in X, \quad \forall \theta^{+}, \theta \in \Theta, \quad \left|\widetilde{\pi}_T\left(x;\theta^{+}\right) - \widetilde{\pi}_T\left(x;\theta\right)\right| \leq k_T(x) \cdot \left\|\theta^{+} - \theta\right\|_2^{\alpha}, \quad \alpha > 0, \qquad \text{(A2)}$$

where $k_T(x)$ is a sequence of functions such that

$$\beta_{pT} \equiv \int k_T(x)^p \pi_T(x) dx < \infty, \quad \text{all } T \text{ and } p = 1, 2.$$

By the mean-value theorem, Cauchy-Schwartz inequality and compacteness of Θ, (A2) holds for $\alpha = 1$ whenever $\nabla_\theta \widetilde{\pi}_T(x;\theta)$ is continuous and bounded (see, also, related results by Andrews (1992, p. 248-249)). The claim that (A2) implies (A1) will be demonstrated after the proof of the next proposition.

Proposition 1. Let assumptions 1-4 hold. Then, as $h \downarrow 0$ and $T \to \infty$, $L_T(\theta) \xrightarrow{p} L(\theta)$, $\forall \theta \in \Theta$.

According to a well-known result (see Newey (1991, thm. 2.1 p. 1162); and Davidson (1994, p. 337-340) for a discussion of this and related results), the following conditions are equivalent:

C1: $\lim_{T\to\infty} P\left(\sup_{\theta\in\Theta} |L_T(\theta) - L(\theta)| > \epsilon\right) = 0.$

C2: $\forall \theta \in \Theta$, $L_T(\theta) \xrightarrow{p} L(\theta)$, and $L_T(\theta)$ is stochastically equicontinuous.

By Newey and McFadden (1994, lemma 2.9 p. 2138), assumption 8 guarantees that $L_T(\theta)$ is stochastically equicontinuous, and so weak consistency follows from the equivalence of C1 and C2 above, assumptions 7-8, lemma 1, compacteness of Θ and the classical strategy of proof in Amemiya (1985, thm. 4.1.1 pp. 106-107).

To prove proposition 1, we need the following preliminary result:

Lemma 1. (Glick's (1974) theorem) Let f_T be a density estimate on $\mathbb{R}^q$, and let f be a density on $\mathbb{R}^q$. If $f_T \xrightarrow{p} f$ pointwise, then $\int_{\mathbb{R}^q} |f_T(x) - f(x)|\, dx \xrightarrow{p} 0$.

Proof of proposition 1. We have:

$$\begin{aligned} & |L_T(\theta) - L(\theta)| \\ \leq & \int \Big| \left[\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right]^2 \pi_T(x) - \left[m(x;\theta) - m(x;\theta_0)\right]^2 \cdot m(x;\theta_0) \\ & + \left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m(x;\theta) - m(x;\theta_0)\right| \cdot \left[\left|\pi_T(x) - m(x;\theta_0)\right| - \left(\pi_T(x) - m(x;\theta_0)\right)\right] \Big| \, dx \end{aligned}$$

$$
\begin{aligned}
= & \int |\widetilde{\pi}_T(x;\theta) - \pi_T(x)| \cdot \pi_T(x) \cdot [|\widetilde{\pi}_T(x;\theta) - \pi_T(x)| - |m(x;\theta) - m(x;\theta_0)|] \\
& + |m(x;\theta) - m(x;\theta_0)| \cdot m(x;\theta_0) \cdot [|\widetilde{\pi}_T(x;\theta) - \pi_T(x)| - |m(x;\theta) - m(x;\theta_0)|] \\
& + |\widetilde{\pi}_T(x;\theta) - \pi_T(x)| \cdot |m(x;\theta) - m(x;\theta_0)| \cdot |\pi_T(x) - m(x;\theta_0)|| \, dx \\
\\
\leq & \int \{ \, |[\widetilde{\pi}_T(x;\theta) - m(x;\theta)] - [\pi_T(x) - m(x;\theta_0)]| \cdot |\widetilde{\pi}_T(x;\theta) - \pi_T(x)| \cdot \pi_T(x) \\
& + |[\widetilde{\pi}_T(x;\theta) - m(x;\theta)] - [\pi_T(x) - m(x;\theta_0)]| \cdot |m(x;\theta) - m(x;\theta_0)| \cdot m(x;\theta_0) \\
& + |m(x;\theta) - m(x;\theta_0)| \cdot |\pi_T(x) - m(x;\theta_0)| \cdot |\widetilde{\pi}_T(x;\theta) - \pi_T(x)| \, \} dx \\
\\
\equiv & \int g_{T,S,h}(x,\theta) dx.
\end{aligned}
$$

By boundedness of $\widetilde{\pi}_T(x;\theta)$, $\pi_T(x)$ and $m(x;\theta)$, there exists a sequence ζ_T bounded in probability as T becomes large such that

$$
\int g_{T,S,h}(x;\theta) \, dx \leq \zeta_T \cdot \left[\int |\widetilde{\pi}_T(x;\theta) - m(x;\theta)| \, dx + \int |\pi_T(x) - m(x;\theta_0)| \, dx \right],
$$

where all integrals are finite for all $\theta \in \Theta$.

By assumptions 1, 5 and 9, $\pi_T(x) \xrightarrow{p} m(x;\theta_0)$ pointwise. By assumption 3, $x_h^i(\theta) \Rightarrow x(\theta)$ as $h \downarrow 0$, $i = 1, \cdots, S$ (for fixed T). By continuity of $\pi_{T,h}^i(x;\theta)$ with respect to the simulated points $\{x_{t,h}^i(\theta)\}_{t=t_l}^T$ and independence of the simulated strings $\left\{x_h^i(\theta)\right\}_{i=1}^S$, for all $i = 1, \cdots, S$, $\pi_{T,h}^i(x;\theta) \Rightarrow \pi_T^i(x;\theta) \equiv \sum_{t=t_l}^T K\left(\left(x_t(\theta) - x\right)/\lambda_T\right) \Big/ \left(T\lambda_T^q\right)$ as $h \downarrow 0$ (all $\theta \in \Theta$), and $\pi_T^i(x;\theta) \xrightarrow{p} m(x;\theta)$ as $T \to \infty$ (all $\theta \in \Theta$) both x-pointwise. Since $\int_{\mathbb{R}^q} \pi_T(x) dx = 1$ and, for all $\theta \in \Theta$, $\int_{\mathbb{R}^q} \widetilde{\pi}_T(x;\theta) dx = 1$ and $\int_{\mathbb{R}^q} m(x;\theta) \, dx = 1$,

$$
\forall \theta \in \Theta, \; \int_X |\widetilde{\pi}_T(x;\theta) - m(x;\theta)| \, dx \leq \int_{\mathbb{R}^q} |\widetilde{\pi}_T(x;\theta) - m(x;\theta)| \, dx \xrightarrow{p} 0,
$$

and

$$
\int_X |\pi_T(x) - m(x;\theta_0)| \leq \int_{\mathbb{R}^q} |\pi_T(x) - m(x;\theta_0)| \xrightarrow{p} 0,
$$

by Glick's (1974) theorem in lemma 1, and the proof is complete.[21] The case in which $\lambda_T \downarrow 0$ is identical. ■

[21] The previous results (obtained with any nonzero λ_T) do not contradict lemma 5 (p. 900) in Devroye (1983). Devroye's lemma 5 refers to data drawn from a density f_∞ and convergence issues of $|f_T - f_\infty|_{L_1}$, where f_T is a nonparametric density estimate of f_∞. Here we were simply concerned with convergence issues of $|\overline{f}_T - \overline{f}_\infty|_{L_1}$ (say), where both $\overline{f}_T$ and $\overline{f}_\infty$ integrate to one in $\mathbb{R}^q$ and $\overline{f}_T \xrightarrow{p} \overline{f}_\infty$ pointwise.

Modulus of continuity example. (Ineq. (A2) implies ineq. (A1)) We have:

$$\begin{aligned} & L_T\left(\theta^+\right) - L_T\left(\theta\right) \\ = & \int \left[\widetilde{\pi}_T(x;\theta^+) - \widetilde{\pi}_T(x;\theta)\right]^2 \pi_T(x)dx \\ & +2\int \left[\widetilde{\pi}_T(x;\theta^+) - \widetilde{\pi}_T(x;\theta)\right]\left[\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right]\pi_T(x)dx. \end{aligned}$$

Next, let $B \equiv \max_{\theta,\theta' \in \Theta} \left\|\theta' - \theta\right\|_2^\alpha$. By Θ compact, $B < \infty$. By using (A2),

$$\begin{aligned} & \left|L_T\left(\theta^+\right) - L_T\left(\theta\right)\right| \\ \leq & \left\|\theta^+ - \theta\right\|_2^{2\alpha} \cdot \beta_{2T} + 2\left\|\theta^+ - \theta\right\|_2^{\alpha} \cdot \int k_T(x)\left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right|\pi_T(x)dx \\ \leq & \left\|\theta^+ - \theta\right\|_2^{2\alpha} \cdot \beta_{2T} + \left\|\theta^+ - \theta\right\|_2^{\alpha} \cdot \xi_T \cdot \beta_{1T} \\ \leq & \left\|\theta^+ - \theta\right\|_2^{\alpha} \cdot \left(B \cdot \beta_{2T} + \xi_T \cdot \beta_{1T}\right), \end{aligned}$$

where

$$\xi_T \equiv 2 \cdot \sup_{x \in X,\, \theta \in \Theta} |\widetilde{\pi}_T(x;\theta) - \pi_T(x)| < \infty \text{ all } T,$$

Since β_{1T}, β_{2T} and ξ_T are bounded in probability as T becomes large, so is $B \cdot \beta_{2T} + \xi_T \cdot \beta_{1T}$. Set then $\kappa_T \equiv B \cdot \beta_{2T} + \xi_T \cdot \beta_{1T}$ to conclude the example. ■

A.2 Asymptotic normality

As discussed in appendix A1, all of our assumptions ensure that $\nabla_\theta \widetilde{\pi}_T$ and $\nabla_{\theta\theta} \widetilde{\pi}_T$ exist in the case of Gaussian kernels and perfect simulations (as $h \downarrow 0$) if a and b satisfy local Lipschitz and growth conditions with respect to θ. In the case of a fixed h and general kernels, we formulate the following condition:

Assumption 9. For all $x \in X$ and $h < h^0$, function $\theta \mapsto \widetilde{\pi}_T(x;\theta)$ is twice differentiable, and its derivatives are continuous and bounded on Θ. Furthermore, for all $h < h^0$, $\int [\sup_{\theta\in\Theta} |\nabla_\theta f(x,\theta)| + \sup_{\theta\in\Theta} |\nabla_{\theta\theta} f(x,\theta)|]\pi_T(x)dx < \infty$, where $f(x,\theta) \equiv [\widetilde{\pi}_T(x;\theta) - \pi_T(x)]^2$.

By assumption 9, the order of derivation and integration in $\nabla_\theta L_T(\theta)$ may be interchanged (see Newey and McFadden (1994, lemma 3.6 p. 2152-2153)), and the first order conditions satisfied

by the SNE are:

$$\mathbf{0}_{p_\theta} = \int \left[\widetilde{\pi}_T(x;\theta_{T,S,h}) - \pi_T(x)\right]\, \pi_T(x)\, \nabla_\theta \widetilde{\pi}_T(x;\theta_{T,S,h}) dx. \tag{A3}$$

Next, consider the c-parametrized curves $\theta(c) = c \circ (\theta_0 - \theta_{T,S,h}) + \theta_{T,S,h}$, where, for any $c \in (0,1)^p$ and $\theta \in \Theta$, $c \circ \theta$ denotes the vector in Θ whose i-th element is $c^{(i)}\theta^{(i)}$. By assumption 9 and the intermediate value theorem, there exists a c^* in $(0,1)^p$ such that:

$$\begin{aligned}\mathbf{0}_p &= \sqrt{T}\int \left[\widetilde{\pi}_T(x;\theta_0) - \pi_T(x)\right]\, \pi_T(x)\, \nabla_\theta \widetilde{\pi}_T(x;\theta_0) dx \\ &\quad + \left\{\int \left[\left|\nabla_\theta \widetilde{\pi}_T(x;\overline{\theta})\right|_2 + \left(\widetilde{\pi}_T(x;\overline{\theta}) - \pi_T(x)\right) \nabla_{\theta\theta} \widetilde{\pi}_T(x;\overline{\theta})\right] \pi_T(x) dx\right\} \cdot \sqrt{T}(\theta_{T,S,h} - \theta_0),\end{aligned} \tag{A4}$$

where $\overline{\theta} \equiv \theta(c^*)$.

Next,

$$\begin{aligned}&\left|\int \left(\widetilde{\pi}_T(x;\overline{\theta}) - \pi_T(x)\right)\, \pi_T(x)\, \nabla_{\theta\theta} \widetilde{\pi}_T(x;\overline{\theta}) dx\right|_{i,j} \\ \leq\ & \int \left|\widetilde{\pi}_T(x;\overline{\theta}) - \pi_T(x)\right| \left|\nabla_{\theta\theta} \widetilde{\pi}_T(x;\overline{\theta})\right|_{i,j}\, \pi_T(x) dx \\ \leq\ & \sup_{x\in X}\left[\left|\nabla_{\theta\theta} \widetilde{\pi}_T(x;\overline{\theta})\right|_{i,j}\right] \cdot \int \left|\widetilde{\pi}_T(x;\overline{\theta}) - \pi_T(x)\right| dx, \text{ all } i,j.\end{aligned}$$

By assumption 5, $\pi_T(x) \xrightarrow{p} \pi_0(x)$ pointwise. By assumption 3, $x_h^i(\theta) \Rightarrow x(\theta)$ as $h \downarrow 0$, $i = 1, \cdots, S$. By continuity of $\pi_{T,h}^i(x;\theta)$ with respect to simulated points $\{x_{t,h}^i(\theta)\}_{t=t_l}^T$ and independence of the simulated strings $(x_h^i(\theta))_{i=1}^S$, $\pi_{T,h}^i(x;\theta_0) \Rightarrow \pi_T(x)$ ($i = 1, \cdots, S$) for fixed T. By assumption 9,

$$\forall \epsilon > 0,\ \exists M_\epsilon : P\left\{\sup_{\theta\in\Theta} |\nabla_\theta \widetilde{\pi}_T(x;\theta)|_i + \sup_{\theta\in\Theta} |\nabla_{\theta\theta} \widetilde{\pi}_T(x;\theta)|_{i,j} < M_\epsilon\right\} \geq 1 - \epsilon, \quad \text{all } i,j.$$

By the previous inequality, the mean value theorem, assumptions 4-5 and 9, consistency of $\theta_{T,S,h}$, and the definition of $\overline{\theta}$, as $h \downarrow 0$ and $T \to \infty$, $\widetilde{\pi}_T(x;\overline{\theta}) \xrightarrow{p} \pi(x;\theta_0)$ pointwise and $\left|\nabla_\theta \widetilde{\pi}_T(x;\overline{\theta})\right|_2 \xrightarrow{p} |\nabla_\theta \pi(x;\theta_0)|_2$ both componentwise and pointwise and so, by Glick's (1974) theorem in lemma 1,

$$\int \left[\left|\nabla_\theta \widetilde{\pi}_T(x;\overline{\theta})\right|_2 + \left(\widetilde{\pi}_T(x;\overline{\theta}) - \pi_T(x)\right)\, \nabla_{\theta\theta} \widetilde{\pi}_T(x;\overline{\theta})\right]\, \pi_T(x) dx \xrightarrow{p} E\left[|\nabla_\theta \pi(x;\theta_0)|_2\right]. \tag{A5}$$

Next, consider the first term in (A4):

$$\begin{aligned}
&\sqrt{T}\int \left[\widetilde{\pi}_T(x;\theta_0)-\pi_T(x)\right]\,\pi_T(x)\,\nabla_\theta\widetilde{\pi}_T(x;\theta_0)dx \\
= &\int \sqrt{T}\left[\widetilde{\pi}_T(x;\theta_0)-E(\widetilde{\pi}_T(x;\theta_0))\right]\,\pi_T(x)\,\nabla_\theta\widetilde{\pi}_T(x;\theta_0)dx \\
&-\int \sqrt{T}\left[\pi_T(x)-E(\pi_T(x))\right]\,\pi_T(x)\,\nabla_\theta\widetilde{\pi}_T(x;\theta_0)dx \\
&+\int \sqrt{T}\left[E(\widetilde{\pi}_T(x;\theta_0))-E(\pi_T(x))\right]\,\pi_T(x)\,\nabla_\theta\widetilde{\pi}_T(x;\theta_0)dx. \qquad \text{(A6)}
\end{aligned}$$

Under assumption 3-ii and the assumption that the kernel is four times continuously differentiable (see the preliminaries of the appendix), the last term in (A6) is $O(\sqrt{T}h)$ by Kloeden and Platen (1999, thm. 14.5.1 p. 473), and it asymptotically vanishes by assumption 3-iii.

As regards the first two terms in (A6), let $F(x;\theta)=\int_0^x \pi(v;\theta)dv$, $F_T(x)=\int_0^x \pi_T(v)dv$ and $F(x)=\int_0^x \pi_0(v)dv$. Let $Q(t)$ be a local martingale with quadratic variation $\ell \equiv \langle Q\rangle(t) = F(t)$. By the Dambis-Dubins-Schwarz theorem (see, e.g., Karatzas and Shreve (1991, thm. 4.6 p. 174)), we can define a time-changed process $B(F(t))=Q(t),\ t\in[0,\infty)$, where $B(\ell)=Q(F^{-1}(\ell))$ is a standard Brownian motion in $[0,1]$. Let $G(F)$ denote a centered Gaussian process with variance $F(x)(1-F(x))$. By Arcones and Yu (1994, corollary 2.1 p. 59-60), the mixing conditions of assumption 2 (which are trivially satisfied in the i.i.d. case) and assumption 5, $A_T \equiv \sqrt{T}(F_T(x)-E(F_T(x))) \Rightarrow G(F)$. By construction, $G(F)$ is a Brownian Bridge in $[0,1]$ and its continuous version can be written as:

$$B^0(F(x)) \equiv B(F(x)) - F(x)B(1) = Q(x) - F(x)Q(F^{-1}(1)). \qquad \text{(A7)}$$

We have,

$$\begin{aligned}
\sqrt{T}\int \left[\pi_T(x)-E\left(\pi_T(x)\right)\right]\,\nabla_\theta\widetilde{\pi}_T(x;\theta_0)\,\pi_T(x)dx \;&\equiv\; \int \nabla_\theta\widetilde{\pi}_T(x;\theta_0)\pi_T(x)dA_T(x) \\
&\xrightarrow{d}\; \int \nabla_\theta\pi(x;\theta_0))\pi_0(x)dB^0(F(x))
\end{aligned}$$

by assumptions 2 and 5, and an argument similar to the one utilized to show (A5).

This stochastic integral is

$$\int \nabla_\theta\pi(x;\theta_0)\pi_0(x)dB^0(F(x)) = \int \nabla_\theta\pi(x;\theta_0)\pi_0(x)dQ(x) - Q(F^{-1}(1))\cdot\int \nabla_\theta\pi(x;\theta_0)\pi_0(x)dF(x),$$

and is centered Gaussian with variance:

$$
\begin{aligned}
& \text{var}\left[\int \nabla_\theta \pi(x;\theta_0)\pi_0(x) dB^0(F(x))\right] \\
= & \int |\nabla_\theta \pi(x;\theta_0)\pi_0(x)|_2 \, F'(x)dx + E\left[\left|Q(F^{-1}(1))\left(\int \nabla_\theta \pi(x;\theta_0)\pi_0(x)F'(x)dx\right)\right|_2\right] \\
& -2E\left[Q(F^{-1}(1))\left(\int \nabla_\theta \pi(x;\theta_0)\pi_0(x)dQ(x)\right)\left(\int \nabla_\theta \pi(x;\theta_0)\pi_0(x)F'(x)dx\right)^\top\right] \\
= & E\left[|\nabla_\theta \pi(x;\theta_0)\pi_0(x)|_2\right] + E\left[\left|Q(F^{-1}(1))E(\nabla_\theta \pi(x;\theta_0)\pi_0(x))\right|_2\right] \\
& -2E\left[\left(\int \nabla_\theta \pi(x;\theta_0)\pi_0(x)Q(F^{-1}(1))dQ(x)\right)E(\nabla_\theta \pi(x;\theta_0)\pi_0(x))^\top\right] \\
= & E\left[|\nabla_\theta \pi(x;\theta_0)\pi_0(x)|_2\right] + |E\left[\nabla_\theta \pi(x;\theta_0)\pi_0(x)\right]|_2 \\
& -2E\left[\left(\int \nabla_\theta \pi(x;\theta_0)\pi_0(x)F'(x)dx\right)E\left(\nabla_\theta \pi(x;\theta_0)\pi_0(x)\right)^\top\right] \\
= & E\left[|\nabla_\theta \pi(x;\theta_0)\pi_0(x)|_2\right] - |E\left[\nabla_\theta \pi(x;\theta_0)\pi_0(x)\right]|_2 \\
= & \text{var}\left[\nabla_\theta \pi(x;\theta_0)\pi_0(x)\right], \qquad \text{(A8)}
\end{aligned}
$$

where we made use of (A7), the definition of Q as a local martingale with quadratic variation F, and the fact that $E[Q(F^{-1}(1))]^2 = E[B(1)^2]^2 = 1$ and $E[Q(F^{-1}(1)) \cdot dQ(x)] = E[B(1)dB(\ell)] = d\ell = F'(x)dx$.

Finally, let $F^i_{T,h}(x;\theta) \equiv \int_0^x \pi^i_{T,h}(v,\theta)dv$, $i = 1, \cdots, S$. By assumptions 3-5, the independence of the S simulations, and again by Arcones and Yu (1994, corollary 2.1 p. 59-60), $A^i_{T,h}(x;\theta_0) \equiv \sqrt{T}[F^i_{T,h}(x;\theta_0) - E(F^i_{T,h}(x;\theta_0))] \Rightarrow \mathsf{G}_i(F)$ as $h \downarrow 0$ and $T \to \infty$, where $\mathsf{G}_i(F)$ are independent Brownian Bridges in $[0,1]$. Hence,

$$\sqrt{T}\sum_{i=1}^{S}\left[F^i_{T,h}(x;\theta_0) - E(F^i_{T,h}(x;\theta_0))\right] \Rightarrow \sum_{i=1}^{S}\mathsf{G}_i(F).$$

Since $E(F^i_{T,h}(x;\theta_0)) = E(F^j_{T,h}(x;\theta_0))$ for all $i, j = 1, \cdots, S$, we use the same arguments leading to (A8) and conclude that $\int[\sqrt{T}\left(\widetilde{\pi}_T(x;\theta_0) - E(\widetilde{\pi}_T(x;\theta_0))\right)]\nabla_\theta \widetilde{\pi}_T(x;\theta_0)\ \pi_T(x)dx \xrightarrow{d} N(0,$ $\text{var}(\nabla_\theta \pi(x;\theta_0)\pi_0(x)/S))$. Finally, A and $A^i_{T,h}$, $i = 1, \cdots, S$, are all independent. Therefore, by the decomposition in (A6),

$$\sqrt{T}\int\left[\widetilde{\pi}_T(x;\theta_0) - \pi_T(x)\right]\nabla_\theta \widetilde{\pi}_T(x;\theta_0)\,\pi_T(x)dx \xrightarrow{d} N\left(0, \left(1+\frac{1}{S}\right)\text{var}\left(\nabla_\theta \pi(x;\theta_0)\pi_0(x)\right)\right). \qquad \text{(A9)}$$

The desired result now follows from (A4), (A5), (A9) and the Slutzky's theorem: we have

$$\sqrt{T}\left(\theta_{T,S,h}-\theta_0\right) \xrightarrow{d} N\left(0,\left(1+\frac{1}{S}\right)V\right),$$

where

$$V \equiv \left[E\left|\nabla_\theta \pi(x;\theta_0)\right|_2\right]^{-1} \cdot \mathrm{var}\left(\nabla_\theta \pi(x;\theta_0)\pi_0(x)\right) \cdot \left[E\left|\nabla_\theta \pi(x;\theta_0)\right|_2\right]^{\top -1},$$

which is theorem 1 in the i.i.d. case.

As regards the general dependent case, let $\mathbb{G}$ be a measurable V-C subgraph class of uniformly bounded functions (see, e.g., Arcones and Yu (1994, definition 2.2 p. 51)). Again by Arcones and Yu (1994, corollary 2.1 p. 59-60), if y^o satisfies assumption 2, then, for each $G \in \mathbb{G}$, $T^{-1/2}\sum_{t=t_l}^{T}\left[G(x_t)-EG\right]$ converges in law to a Gaussian process. Now $\lambda_T^{-q}K\left(\left(x_t-x\right)/\lambda_T\right) \in \mathbb{G}$, and the variance terms that are reported in the theorem follow.

Remark 3. A crucial step of the previous proof is given by the weak convergence $\sqrt{T}[F_T(x) - E(F_T(x))] \Rightarrow \mathsf{G}(F)$. Because $\sqrt{T}(F_T - F) = \sqrt{T}[F_T - E(F_T)] + \sqrt{T}[E(F_T) - F]$, we see that $\sqrt{T}[F_T(x) - F(x)] \Rightarrow \mathsf{G}(F)$ under the more stringent condition (13). This condition is needed to asymptotically zero the bias term $\sqrt{T}[E(F_T) - F]$, and is exactly assumption A4(r,0) in Aït-Sahalia (1994, lemma 1 p. 20). As we noted in the main text, we do not need such a more severe condition because bias effects cancel out each other through the decomposition in eq. (A6).

B. Asymptotic behavior of the SNE for general weighting functions

B.1 Consistency

We set

$$\rho_T\left(x;\theta\right) \equiv \left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| w_T(x;\lambda) \quad \text{and} \quad \rho(x,\theta) \equiv \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| w(x,\lambda),$$

The following assumption further characterizes the class of weighting functions that we consider in this paper:

Assumption 10. For all $\theta \in \Theta$,

10.1: $\left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \xrightarrow{p} 0$ as $h \downarrow 0$ and $T \to \infty$, x-pointwise;

10.2: for all $h < h^0$, and every S and T, functions $\rho_T\left(x;\theta\right)\pi_T(x)$, $\rho_T\left(x;\theta\right)\widetilde{\pi}_T(x;\theta)$, $\rho\left(x;\theta\right)m\left(x;\theta\right)$ and $\rho\left(x;\theta\right)m\left(x;\theta_0\right)$ are bounded for all $x \in X$, and integrable.

Assumption 10.1 always holds for weighting functions which make $w^2(x,\lambda) > 0$ (see assumption 6). This is trivially the case in appendix A.1, where $w^1(\cdot,\lambda) \equiv m$ and $w^2(\cdot,\lambda) \equiv 1$. Assumption 10.1 also covers the case of weighting functions implying $w^2(x,\lambda) = 0$ for some x. Finally, assumption 10.2 guarantees that $\forall\theta \in \Theta$, $L(\theta) < \infty$ and that for sufficiently small h, $L_T(\theta) < \infty$, for all S and T.

We have:

Proposition 2. Let assumptions 1-4 and 10 hold. Then, as $h \downarrow 0$ and $T \to \infty$, $L_T(\theta) \xrightarrow{p} L(\theta)$, $\forall\theta \in \Theta$.

As in appendix A.1, consistency now follows as soon as $L_T(\theta)$ also fulfils condition (A1). An example of conditions guaranteeing that (A1) holds in the setting of this appendix is: for all $x \in X$ and $\theta, \theta^+ \in \Theta$,

$$\left|\widetilde{\pi}_T(x;\theta^+) - \widetilde{\pi}_T(x;\theta)\right| \cdot w_T(x,\lambda) \leq k_T(x) \cdot \left\|\theta^+ - \theta\right\|_2^\alpha,$$

where $k_T(x)$ is a sequence of functions such that for all T,

$$\begin{aligned}
\beta_{1T} &\equiv \sup_{\theta,\theta'\in\Theta} \int k_T(x) \cdot \left|\widetilde{\pi}_T(x;\theta') - \widetilde{\pi}_T(x;\theta)\right| dx < \infty; \\
\beta_{2T} &\equiv \sup_{\theta\in\Theta} \int k_T(x) \cdot \left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| dx < \infty.
\end{aligned}$$

Indeed, simple algebra reveals that under the previous condition,

$$\left|L_T\left(\theta^+\right) - L_T\left(\theta\right)\right| \leq \left\|\theta^+ - \theta\right\|_2^\alpha \cdot \left(\beta_{1T} + 2\beta_{2T}\right).$$

As regards the proof of proposition 2, the following result will be useful.

Lemma 2. (Generalization of Glick's (1974) theorem) Let assumption 10 hold. Then, as $h \downarrow 0$ and $T \to \infty$,

$$\int \left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right| dx \xrightarrow{p} 0, \quad \text{for all } \theta \in \Theta.$$

Proof of lemma 2. For all $\theta \in \Theta$, $\left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right|$ is bounded and integrable by assumption 10.2. By Lebesgue's dominated convergence theorem

and by assumption 10.1, as $h \downarrow 0$,

$$\lim_{T\to\infty} E\left[\left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right|\right]$$

$$= E\left[\lim_{T\to\infty} \left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right|\right]$$

$$= 0 \quad \forall (x,\theta) \in X \times \Theta.$$

By Fubini's theorem, $\forall \theta \in \Theta$

$$E\left[\int \left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right| dx\right]$$

$$= \int E\left[\left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right|\right] dx.$$

Again by Lebesgue's dominated convergence theorem, as $h \downarrow 0$,

$$\lim_{T\to\infty} E\left[\int \left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right| dx\right]$$

$$= \lim_{T\to\infty} \int E\left[\left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right|\right] dx$$

$$= \int \lim_{T\to\infty} E\left[\left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right|\right] dx$$

$$= 0, \quad \text{all } \theta \in \Theta.$$

The result follows by taking limits in the Markov's inequality:

$$\forall \epsilon > 0, \; P\left\{\int \left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right| dx > \epsilon\right\}$$

$$\leq \frac{E\left[\int \left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \cdot \left|m\left(x,\theta\right) - m\left(x;\theta_0\right)\right| \cdot \left|w_T(x,\lambda) - w(x,\lambda)\right| dx\right]}{\epsilon}, \quad \text{all } \theta \in \Theta.$$

■

Proof of proposition 2. We have:

$$\left|L_T\left(\theta\right) - L(\theta)\right|$$

$$\leq \int \Big| \left[\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right]^2 w_T(x,\lambda) - \left[m\left(x;\theta\right) - m\left(x;\theta_0\right)\right]^2 w(x,\lambda)$$
$$+ \left|\widetilde{\pi}_T(x;\theta) - \pi_T(x)\right| \left|m\left(x;\theta\right) - m\left(x;\theta_0\right)\right| \left[\left|w_T(x,\lambda) - w(x,\lambda)\right| - \left(w_T(x,\lambda) - w(x,\lambda)\right)\right] \Big| \, dx$$

$$
\begin{aligned}
= & \int \left| \left| \widetilde{\pi}_T(x;\theta) - \pi_T(x) \right| \cdot w_T(x,\lambda) \cdot \left[\left| \widetilde{\pi}_T(x,\theta) - \pi_T(x) \right| - \left| m\left(x;\theta\right) - m\left(x;\theta_0\right) \right| \right] \right. \\
& + \left| m\left(x;\theta\right) - m\left(x;\theta_0\right) \right| \cdot w(x,\lambda) \cdot \left[\left| \widetilde{\pi}_T(x;\theta) - \pi_T(x) \right| - \left| m\left(x;\theta\right) - m\left(x;\theta_0\right) \right| \right] \\
& \left. + \left| \widetilde{\pi}_T(x;\theta) - \pi_T(x) \right| \cdot \left| m\left(x;\theta\right) - m\left(x;\theta_0\right) \right| \cdot \left| w_T(x,\lambda) - w(x,\lambda) \right| \right| dx \\
\leq & \int g_{T,S,h}\left(x;\theta\right) dx,
\end{aligned}
$$

where

$$
\begin{aligned}
g_{T,S,h}\left(x;\theta\right) \equiv & \left| \widetilde{\pi}_T(x;\theta) - \pi_T(x) \right| \cdot w_T(x;\lambda) \cdot \left| \left[\widetilde{\pi}_T(x;\theta) - m\left(x;\theta\right) \right] - \left[\pi_T(x) - m\left(x;\theta_0\right) \right] \right| \\
& + \left| m\left(x;\theta\right) - m\left(x;\theta_0\right) \right| \cdot w(x,\lambda) \cdot \left| \left[\widetilde{\pi}_T(x;\theta) - m\left(x;\theta\right) \right] - \left[\pi_T(x) - m\left(x;\theta_0\right) \right] \right| \\
& + \left| \widetilde{\pi}_T(x;\theta) - \pi_T(x) \right| \cdot \left| m\left(x;\theta\right) - m\left(x;\theta_0\right) \right| \cdot \left| w_T(x,\lambda) - w(x,\lambda) \right| \\
\equiv & \; G_1 + G_2 + G_3.
\end{aligned}
$$

By lemma 2, $\int G_3 \xrightarrow{p} 0$ for all $\theta \in \Theta$. Furthermore, for all $\theta \in \Theta$,

$$
\begin{aligned}
G_1 + G_2 \leq & \left[\rho\left(x;\theta\right) + \rho_T\left(x;\theta\right) \right] \\
& \times \left[2\left(\pi_T(x) + m\left(x;\theta\right) \right) + \left| \widetilde{\pi}_T(x;\theta) - \pi_T(x) \right| + \left| m\left(x;\theta\right) - m\left(x;\theta_0\right) \right| \right],
\end{aligned}
$$

which is bounded and integrable by assumption 10.2. A repeated use of Lebesgue's dominated convergence theorem and Fubini's theorem then reveals that $\int \left(G_1 + G_2 \right) \xrightarrow{p} 0$, as in the proof of lemma 2, and the proof of proposition 2 is complete. The case $\lambda_T \downarrow 0$ is identical. ■

B.2 Asymptotic normality

We show that the SNE in (15) behaves as the SNE in theorem 1, but with function Ψ given by (16). The proof is sketchy because its steps only generalize steps of proofs produced in the previous appendixes. The sequence of the weighting functions is required to satisfy an additional set of regularity conditions:

Assumption 11. There exists a neighborhood N of θ_0 such that matrix $\int_X \left| \nabla_\theta \pi\left(x;\theta\right) \right|_2 w(x) dx$ has full rank in $N \cap \Theta$. Furthermore, assumption 5 holds, and as $h \downarrow 0$ and $T \to \infty$,

11.1: $\nabla_\theta \widetilde{\pi}_T(x;\theta_0) \cdot w_T(x,\lambda) \xrightarrow{p} \nabla_\theta \pi\left(x;\theta\right) \cdot w(x)$, x-pointwise;

11.2: var$[\nabla_\theta \pi\left(x;\theta\right) \cdot w(x)] < \infty$.

11.3: For all $h < h^0$, $\int [\sup_{\theta \in \Theta} \left| \nabla_\theta f(x,\theta) \right| + \sup_{\theta \in \Theta} \left| \nabla_{\theta\theta} f(x,\theta) \right|] w_T(x,\lambda) dx < \infty$, where f is as in assumption 9.

Assumption 11.1 regards pointwise convergence to a deterministic function $\zeta(x) \equiv \nabla_\theta \pi(x;\theta_0) \cdot w(x)$. Assumption 11.2 is a condition on second order moments finiteness of function ζ when x is taken to be a random variate. Finally, assumption 11.3 allows to interchange the order of derivation and integration in the first order conditions.

We now state a simple but powerful result that will also be used to show other results in appendix C.2:

Lemma 3. Let B^0 be a Brownian Bridge. For any $\mathbb{R}^l$-valued function ζ with *finite* moments, $\int \zeta(x) dB^0(F(x))$ is centered Gaussian with

$$\text{var}\left[\int \zeta(x) dB^0(F(x))\right] = \text{var}\left[\zeta(x)\right],$$

where the variance in the last term is taken with respect to F.

Proof. The result follows from a straightforward generalization of (A8) in appendix A.2. ■

The optimality conditions for the SNE with general weighting function lead to:

$$\begin{aligned}\mathbf{0}_{p_\theta} &= \sqrt{T}\int \left[\widetilde{\pi}_T(x;\theta_0) - \pi_T(x)\right] \nabla_\theta \widetilde{\pi}_T(x;\theta_0)\, w_T(x,\lambda) dx \\ &+ \left\{\int \left[\left|\nabla_\theta \widetilde{\pi}_T(x;\overline{\theta})\right|_2 + \left(\widetilde{\pi}_T(x;\overline{\theta}) - \pi_T(x)\right) \nabla_{\theta\theta} \widetilde{\pi}_T(x;\overline{\theta})\right] w_T(x,\lambda) dx\right\} \cdot \sqrt{T}\left(\theta_{T,S,h} - \theta_0\right),\end{aligned}$$

where $\overline{\theta}$ is defined similarly as in eq. (A4) in appendix A.2. Use the arguments of appendix A.2 and lemma 3 to conclude: function Ψ generating the variance/covariance matrix is exactly the one given in (16).

C. Proof of theorem 2 and corollary 1

Appendixes C.1 and C.2 provide the general theory underlying theorem 2. The efficiency implication of corollary 1 is developed in appendix C.3.

C.1 Consistency

The consistency proof is nearly identical to the proof of consistency given in appendix B.1. We only need a change in notation. We let:

$$\begin{aligned} r_T(z,v,\theta) &\equiv \left|\widetilde{\pi}_T(z|v,\theta) - \pi_T(z|v)\right| w_T(z,v,\lambda), \\ r(z,v,\theta) &\equiv \left|n(z,v,\theta) - n(z,v,\theta_0)\right| w(z,v,\lambda), \end{aligned}$$

and formulate the following assumption:

Assumption 12. For all $\theta \in \Theta$,

12.1: $|\widetilde{\pi}_T(z|v,\theta) - \pi_T(z|v)| \cdot |w_T(z,v,\lambda) - w(z,v,\lambda)| \cdot |n(z,v,\theta) - n(z,v,\theta_0)| \xrightarrow{p} 0$ as $h \downarrow 0$ and $T \to \infty$, (z,v)-pointwise;

12.2: for all $h < h^0$, and every S and T, functions $r_T(z,v,\theta)\,\pi_T(z|v)$, $r_T(z,v,\theta)\,\widetilde{\pi}_T(z|v,\theta)$, $r(z,v,\theta)\,n(z,v,\theta)$ and $r(z,v,\theta)\,n(z,v,\theta_0)$ are bounded for all $(z,v) \in Z \times V$, and integrable.

Consistency now follows by the same arguments produced in appendix B.1.

C.2 Asymptotic normality

We consider weighting functions satisfying regularity conditions mirroring the ones in assumption 11:

Assumption 13. There exists a neighborhood N of θ_0 such that matrix $\int_Z \int_V |\nabla_\theta \pi(z|v;\theta)|_2\, w(z,v)dzdv$ has full rank in $N \cap \Theta$. Furthermore, assumption 5 holds, and as $h \downarrow 0$ and $T \to \infty$,

13.1: $\nabla_\theta \widetilde{\pi}_T(z|v,\theta_0) w_T(z,v,\lambda)/\pi_T(v) \xrightarrow{p} \nabla_\theta \pi(z|v,\theta_0) w(z,v)/\pi(v,\theta_0)$ and, for all $i = 1,\cdots,S$, $\nabla_\theta \widetilde{\pi}_T(z|v,\theta_0) w_T(z,v,\lambda)/\pi^i_{T,h}(v;\theta_0) \xrightarrow{p} \nabla_\theta \pi(z|v,\theta_0) w(z,v)/\pi(v,\theta_0)$, both (z,v)-pointwise;

13.2: $\text{var}\{\nabla_\theta \pi(z|v,\theta_0) w(z,v)/\pi(v,\theta_0)\} < \infty$.

13.3: For all $h < h^0$, $\int\int [\sup_{\theta\in\Theta} |\nabla_\theta g(z,v;\theta)| + \sup_{\theta\in\Theta} |\nabla_{\theta\theta} g(z,v;\theta)|] w_T(z,v,\lambda)dzdv < \infty$, where $g(z,v;\theta) \equiv [\widetilde{\pi}_T(z|v;\theta) - \pi_T(z|v)]^2$.

The usual expansion of the first order conditions satisfied by the CD-SNE (definition 2) leaves:

$$\begin{aligned} \mathbf{0}_{p_\theta} &= \frac{1}{S}\sum_{i=1}^{S} \sqrt{T} \int\int \left[\frac{\pi^i_{T,h}(z,v,\theta_0)}{\pi^i_{T,h}(v,\theta_0)} - \frac{\pi_T(z,v)}{\pi_T(v)}\right] \nabla_\theta \widetilde{\pi}_T(z|v,\theta_0)\; w_T(z,v,\lambda)dzdv \\ &\quad + \left[\int\int \left|\nabla_\theta \widetilde{\pi}_T(z|v,\overline{\theta})\right|_2 w_T(z,v,\lambda)dzdv\right] \cdot \sqrt{T}(\theta_{T,S,h} - \theta_0) + o_p(1), \end{aligned}$$

where $\overline{\theta}$ is defined similarly as in appendix A.2 (see eq. (A4)), and the $o_p(1)$ term in the last line emerges as a result of an argument similar to the one used to show (A5) in appendix A.2.

If assumption 3-ii holds and the kernel is four times continuously differentiable, then $\sqrt{T}[E(\pi^i_{T,h}(z,v;\theta_0)) - E(\pi_T(z,v))] \approx O(h\sqrt{T})$ and $\sqrt{T}[E(\pi^i_{T,h}(v;\theta_0)) - E(\pi_T(v))] \approx O(h\sqrt{T})$

($i = 1, \cdots, S$), as in appendix A.2 (see eq. (A6)). Therefore, as $h \downarrow 0$ as prescribed by assumption 3-iii,

$$\mathbf{0}_{p_\theta} = \frac{1}{S}\sum_{i=1}^{S}\left(D^i_{T,h,1} + D^i_{T,h,3}\right) - D_{T,h,2} + D_{T,h,4}\cdot\sqrt{T}\left(\theta_{T,S,h} - \theta_0\right) + o_p(1),$$

where

$$\begin{aligned}
D^i_{T,h,1} &\equiv \int\int \frac{\nabla_\theta \widetilde{\pi}_T(z|\, v;\theta_0) w_T(z,v,\lambda)}{\pi^i_{T,h}(v;\theta_0)} dA^i_{T,h}(z,v,\theta_0);\\
D_{T,h,2} &\equiv \int\int \frac{\nabla_\theta \widetilde{\pi}_T(z|\, v,\theta_0) w_T(z,v,\lambda)}{\pi_T(v)} dA_T(z,v);\\
D^i_{T,h,3} &\equiv \int\int \frac{\nabla_\theta \widetilde{\pi}_T(z|\, v,\theta_0) E\left[\pi_T(z,v)\right] w_T(z,v,\lambda)}{\pi^i_{T,h}(v,\theta_0)\cdot\pi_T(v)} dz\left[dA_T(v) - dA^i_{T,h}(v,\theta_0)\right];\\
D_{T,h,4} &\equiv \int\int \left|\nabla_\theta \widetilde{\pi}_T(z|\, v,\theta_0)\right|_2 \; w_T(z,v,\lambda) dz dv;
\end{aligned}$$

and $A^i_{T,h}(z,v,\theta_0)$, $A_T(z,v)$, $A_T(v)$ and $A^i_{T,h}(v,\theta_0)$ are defined similarly as in appendix A.2.

By exactly the same arguments of appendix A.2, we have that as $h \downarrow 0$ and $T \to \infty$,

$$\begin{aligned}
D^i_{T,h,1} &\xrightarrow{d} D^i_1 \equiv \int\int \frac{\nabla_\theta \pi(z|\, v;\theta_0)\, w(z,v)}{\pi(v,\theta_0)} dB^0_i\left(F(z,v,\theta_0)\right), \qquad i = 1,\cdots,S;\\
D_{T,h,2} &\xrightarrow{d} D_2 \equiv \int\int \frac{\nabla_\theta \pi(z|\, v;\theta_0)\, w(z,v)}{\pi(v,\theta_0)} dB^0\left(F(z,v)\right);\\
D^i_{T,h,3} &\xrightarrow{d} D^i_3 \equiv \int\int \frac{\nabla_\theta \pi(z|\, v;\theta_0)\, \pi(z,v)\, w(z,v)}{\pi(v,\theta_0)^2} dz\left[dB^0_v\left(F(v)\right) - dB^0_{v,i}\left(F(v,\theta_0)\right)\right],\\
&\qquad\qquad i = 1,\cdots,S;\\
D_{T,h,4} &\xrightarrow{p} D_4 \equiv \int\int \left|\nabla_\theta \pi(z|\, v;\theta_0)\right|_2 \; w(z,v) dx dv;
\end{aligned}$$

where B^0 and $B^0_i,\ i = 1,\cdots,S$, are independent Brownian Bridges; and B^0_v and $B^0_{v,i},\ i = 1,\cdots,S$, are also independent Brownian Bridges.

By lemma 3 in appendix B.2, D^i_1, $i = 1,\cdots,S$, and D_2 are all independent and asymptotically centered Gaussian with variance

$$\text{var}\left[\frac{\nabla_\theta \pi(z|\, v;\theta_0)\cdot w(z,v)}{\pi(v;\theta_0)}\right],$$

and we have the following result:

$$\sqrt{T}\left(\theta_{T,S,h} - \theta_0\right) \xrightarrow{d} N(0,V),$$

where $V = D_4^{-1} \cdot \Sigma \cdot {D_4}^{\top -1}$ and $\Sigma = \text{var}[\frac{1}{S}\sum_{i=1}^{S}(D_1^i + D_3^i) + D_2]$. Finally, the same result holds in the dependent case, with variance terms given by $\text{var}(\Psi_t) + 2\sum_{j=1}^{\infty}\text{cov}(\Psi_t, \Psi_{t+j})$, where $\Psi \equiv D_4^{-1}[\frac{1}{S}\sum_{i=1}^{S}(D_1^i + D_3^i) + D_2]$.

C.3 Proof of corollary 1 (Cramer-Rao lower bound)

Define:

$$\xi(z,v) \equiv \frac{\pi(z,v)w(z,v)}{\pi(v)^2}. \tag{C1}$$

We have:

$$\begin{aligned} D_3^i &= \int_Z \int_V \nabla_\theta \pi\left(z|\, v;\theta_0\right) \xi(z,v) dz \left[dB_v^0\left(F(v)\right) - dB_{v,i}^0\left(F(v,\theta_0)\right)\right] \\ &= \int_V \gamma(v,\theta_0)\left[dB_v^0\left(F(v)\right) - dB_{v,i}^0\left(F(v,\theta_0)\right)\right], \end{aligned}$$

where

$$\gamma(v,\theta_0) \equiv \int_Z \nabla_\theta \pi\left(z|\, v;\theta_0\right) \xi(z,v) dz. \tag{C2}$$

Again by lemma 3, and the independence of the Brownian Bridges B_v^0 and $B_{v,i}^0$, $i = 1, \cdots, S$, D_3^i, $i = 1, \cdots, S$, is also centered Gaussian with variance equal to $2 \cdot \text{var}(\gamma(v,\theta_0))$. Next, consider the following class of weighting functions:

$$W_T^\xi \equiv \left\{ w_T(z,v,\lambda) : w_T(z,v,\lambda) = \xi_T(v) \cdot \frac{\pi_T(v)^2}{\pi_T(z,v) + \alpha_T} \quad \text{all } (v,z) \in V \times Z \right\},$$

where for each T, $\alpha_T > 0$, and $\alpha_T \approx o_p(1)$, $v \not\mapsto \xi_T(v)$ is a continuous function possibly dependent on data, with $\xi_T(v) \xrightarrow{p} \xi(v)$ pointwise, and ξ is another continuous function. For any $w \in W_T^\xi$, the corresponding limiting function $\xi(z,v)$ in (C1) must necessarily take the form $\xi(z,v) = \xi(v)$. Now for all $v \in V$, $\int_Z \nabla_\theta \pi\left(z|\, v;\theta_0\right) dz = 0$. By replacing $\xi(v)$ into (C2), we conclude that

$$\text{for all } v,\ \gamma(v,\theta_0) = 0.$$

Therefore, $D_3^i \equiv 0$, and by a trivial extension of the arguments to the dependent case, we have:

Proposition 3. Under the assumptions of theorem 2, CD-SNEs with weighting functions $w \in W_T^\xi$ are consistent and asymptotically normal with variance/covariance matrix V given by $\left(1 + S^{-1}\right) \cdot [\text{var}(\Psi_1) + 2\sum_{j=1}^{\infty}\text{cov}(\Psi_1, \Psi_{1+j})]$, where

$$\Psi \equiv \left[\int_Z \int_V \left|\nabla_\theta \pi\left(z|\, v;\theta_0\right)\right|_2 \cdot w\left(z,v\right) dz dv\right]^{-1} \frac{\nabla_\theta \pi\left(z|\, v;\theta_0\right) \cdot w(z,v)}{\pi(v;\theta_0)}. \tag{C3}$$

Finally, we claim that when the state x is fully observable and Markov, variance V is minimized with $w \in W_T^{\xi}$, $\xi(t) = 1$ all t, $\xi_T(t) = 1$, all t and T. Indeed, by plugging the limiting function $w(z,v) \equiv \pi(v)^2 / \pi(z,v)$ into (C3),

$$\Psi = \left[\int_Z \int_V \left| \frac{\nabla_\theta \pi\left(z \middle| v; \theta_0\right)}{\pi\left(z \middle| v; \theta_0\right)} \right|_2 \cdot \pi\left(z, v; \theta_0\right) dz dv \right]^{-1} \frac{\nabla_\theta \pi\left(z \middle| v; \theta_0\right)}{\pi\left(z \middle| v; \theta_0\right)},$$

and the claim immediately follows by the usual score martingale difference argument.

D. Proof of theorem 3

Let $\pi_t \equiv \pi_t\left(\phi\left(y(t+1), \mathbf{M} - (t+1)\mathbf{1}_{d-q^*}\right) \middle| \phi\left(y(t), \mathbf{M} - t\mathbf{1}_{d-q^*}\right)\right)$ denote the transition density of

$$\phi\left(y(t), \mathbf{M} - t\mathbf{1}_{d-q^*}\right) \equiv \phi\left(y(t)\right) \equiv \left(y^o(t), c(y(t), M_1 - t), \cdots, c(y(t), M_{d-q^*} - t)\right),$$

where we have emphasized the dependence of ϕ on the time-to-expiration vector:

$$\mathbf{M} - t\mathbf{1}_{d-q^*} \equiv (M_1 - t, \cdots, M_{d-q^*} - t).$$

By $a(\tau)$ full rank $P \otimes d\tau$-a.s., and Itô's lemma, ϕ satisfies, for $\tau \in [t, t+1]$,

$$\begin{cases} dy^o(\tau) &= b^o(\tau)d\tau + F(\tau)a(\tau)dW(\tau) \\ \\ dc(\tau) &= b^c(\tau)d\tau + \nabla c(\tau)a(\tau)dW(\tau) \end{cases}$$

where b^o and b^c are, respectively, q^*-dimensional and $(d-q^*)$-dimensional measurable functions, and $F(\tau) \equiv \overline{a}(\tau) \cdot a(\tau)^{-1}$ $P \otimes d\tau$-a.s. Under condition (23), π_t is not degenerate. Furthermore, $C\left(y(t); \ell\right) \equiv C(t)$ is deterministic in $\ell \equiv (\ell_1, \cdots, \ell_{d-q^*})$. That is, for all $(\overline{c}, \overline{c}^+) \in \mathbb{R}^d \times \mathbb{R}^d$, there exists a function μ such that for any neighbourhood $N(\overline{c}^+)$ of $\overline{c}^+$, there exists another neighborhood $N(\mu(\overline{c}^+))$ of $\mu(\overline{c}^+)$ such that,

$$\left\{\omega \in \Omega : \phi\left(y(t+1), \mathbf{M} - (t+1)\mathbf{1}_{d-q^*}\right) \in N(\overline{c}^+) \middle| \quad \phi\left(y(t), \mathbf{M} - t\mathbf{1}_{d-q^*}\right) = \overline{c}\right\}$$

$$= \left\{\omega \in \Omega : (y^o(t+1), c(y(t+1), M_1 - t)), \cdots, c(y(t+1), M_{d-q^*} - t)) \in N(\mu(\overline{c}^+)) \right.$$
$$\left. \middle| \phi\left(y(t), \mathbf{M} - t\mathbf{1}_{d-q^*}\right) = \overline{c}\right\}$$

$$= \left\{\omega \in \Omega : (y^o(t+1), c(y(t+1), M_1 - t)), \cdots, c(y(t+1), M_{d-q^*} - t)) \in N(\mu(\overline{c}^+)) \right.$$
$$\left. \middle| (y^o(t), c(y(t), M_1 - t), \cdots, c(y(t), M_{d-q^*} - t)) = \overline{c}\right\}$$

where the last equality follows by the definition of ϕ. In particular, the transition laws of ϕ_t^c given ϕ_{t-1}^c are not degenerate; and ϕ_t^c is stationary. The feasibility of the CD-SNE is proved. The final (efficiency) claim follows by the Markov property of ϕ, eq. (21), and the usual score martingale difference argument advocated in proving corollary 1.

E. Practical bandwidth choice with SNE

In this paper, our bandwidth choice relied on results in Chen, Linton and Robinson (2001) (CLR, henceforth), which we now succinctly describe. Let $AB(z,v)$ be the asymptotic bias, defined as the leading term in the deviation of the ratio of expectations of numerator and denominator of the estimated density $f_{aac}(z|v)$ from the actual density $f(z|v)$. Here the subscripts aac identify the bandwidth used to estimate a bivariate density $f(z,v)$ through product kernels (aa) and the marginal $f(v)$ (c). Let $AV(z,v)$ be the asymptotic variance of kernel density $f_{aac}(z|v)$.

By CLR (lemma 1),[22]

$$AB(z,v) \;=\; a^2 B_1(z,v) - c^2 B_2(z,v) + o(\max(a^2,c^2)),$$

where $B_1(z,v) = \left[\partial^2 f(z,v)/\,\partial z^2 + \partial^2 f(z,v)/\,\partial v^2\right] \int u^2 K(u)du/\,[2f(v)]$ and $B_2(z,v) = f(z|v) \times \left[d^2 f(v)/\,dv^2\right] \int u^2 K(u)du/\,[2f(v)]$.

By CLR (lemma 2),

$$AV(z,v) = \frac{V_1(z,v)}{Tc} + \frac{V_2(z,v)}{Ta^2} - \frac{V_3(z,v,a,c)}{Tc} + o\left(\frac{1}{T\cdot\min(a^2,c)}\right),$$

where $V_1(z,v) = f(z|v)^2 \int K(u)^2 du/\,f(v)$, $V_2(z,v) = f(z|v)\left[\int K(u)^2 du\right]^2/f(v)$, and $V_3(z,v,a,c)$ $= 2\, f(z|v)^2 \int K(u)K(\frac{au}{c})du/\,f(v)$.

The previous results provide practical guidance to bandwidth selection. The asymptotic mean squared error of $f_{aac}(z|v)$ is $AMSE(z,v) \equiv AV(z,v) + AB(z,v)^2$.

If $a = c$ (which is our choice),

$$AMSE(z,v) = \frac{V_2(z,v)}{Ta^2} + a^4 \left[B_1(z,v) - B_2(z,v)\right]^2.$$

For any fixed (z,v), the $AMSE$ is mimimized by $a(z,v) = \{V_2(z,v)/[2T(B_1(z,v) - B_2(z,v))^2]\}^{\frac{1}{6}}$. In the practical implementation of our estimators, we searched for those bandwidths minimizing the $AMSE$ averaged over the sample points. The computation of the $AMSE$ required an initial choice of the bandwidth in order to have an initial estimate of the densities entering in the $AMSE$ formula. Towards this end, we used the optimal bandwidth under the assumption that $f(z,v)$ is Gaussian.

[22] All of our assumptions imply that the regularity conditions in CLR are met.

References

Aït-Sahalia, Y., 1994, "The Delta Method for Nonparametric Kernel Functionals," working paper, Princeton University.

Aït-Sahalia, Y., 1996, "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., 2003, "Closed-Form Likelihood Expansions for Multivariate Diffusions," working paper, Princeton University.

Amemiya, T., 1985, Advanced Econometrics, Cambridge, Mass.: Harvard University Press.

Andrews, D.W.K., 1992, "Generic Uniform Convergence," Econometric Theory, 8, 241-257.

Arcones, M.A. and B. Yu, 1994, "Central Limit Theorems for Empirical and U-Processes of Stationary Mixing Sequences," Journal of Theoretical Probability, 7, 47-71.

Arnold, L., 1992, Stochastic Differential Equations: Theory and Applications. Malabar, Florida: Krieger Publishing Company.

Balduzzi, P., S. R. Das, S. Foresi, and R. K. Sundaram, 1996, "A Simple Approach to Three Factor Affine Term Structure Models," Journal of Fixed Income, 6, 43-53.

Basu, A. and B. G. Lindsay, 1994, "Minimum Disparity Estimation for Continuous Models: Efficiency, Distributions and Robustness," Annals of the Institute of Statistical Mathematics, 46, 683-705.

Bergman, Y. Z., B. D. Grundy, and Z. Wiener, 1996, "General Properties of Option Prices," Journal of Finance, 51, 1573-1610.

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

Bickel, P.J. and M. Rosenblatt, 1973, "On Some Global Measures of the Deviations of Density Function Estimates," Annals of Statistics, 1, 1071-1095.

Carrasco, M. and J.-P. Florens, 2000, "Generalization of GMM to a Continuum of Moment Conditions," Econometric Theory, 16, 797-834.

Carrasco, M., L. P. Hansen and X. Chen, 1999, "Time Deformation and Dependence," working paper, University of Rochester.

Carrasco, M., M. Chernov, J.-P. Florens and E. Ghysels, 2002, "Efficient Estimation of Jump-Diffusions and General Dynamic Models with a Continuum of Moment Conditions," working paper, University of Rochester.

Chacko, G. and L. Viceira, 2003, "Spectral GMM Estimation of Continuous-Time Processes," Journal of Econometrics, 116, 259-292.

Chapman, D. A., J. B. Long, and N. D. Pearson, 1999, "Using Proxies for the Short Rate: When Are Three Months like an Months like an Instant?," Review of Financial Studies, 12, 763-806.

Chen, X., L. P. Hansen and M. Carrasco, 1999, "Nonlinearity and Temporal Dependence," working paper, University of Rochester.

Chen, X, O. Linton, and P. M. Robinson, 2001, "The Estimation of Conditional Densities," The Journal of Statistical Planning and Inference Special Issue in Honor of George Roussas, 71-84.

Chernov, M. and E. Ghysels, 2000, "A Study towards a Unified Approach to the Joint Estimation of Objective and Risk-Neutral Measures for the Purpose of Options Valuation," Journal of Financial Economics, 56, 407-458.

Christensen, B. J., 1992, "Asset Prices and the Empirical Martingale Model," working paper, New York University.

Corradi, V. and N. R. Swanson, 2003, "Bootstrap Specification Tests for Diffusion Processes," working paper, Queen Mary, University of London.

Dai, Q. and K. J. Singleton, 2000, "Specification Analysis of Affine Term Structure Models," Journal of Finance, 55, 1943–1978.

Das, S. R. and R. K. Sundaran, 1999, "Of Smiles and Smirks: a Term Structure Perspective," Journal of Financial and Quantitative Analysis, 34, 211-239.

Davidson, J., 1994, Stochastic Limit Theory, Oxford: Oxford University Press.

Devroye, L., 1983, "The Equivalence of Weak, Strong and Complete Convergence in L_1 for Kernel Density Estimates," Annals of Statistics, 11, 896-904.

Diebold, F. X., L. E. Ohanian and J. Berkowitz, 1998, "Dynamic Equilbrium Economies: A Framework for Comparing Models and Data," Review of Economic Studies, 65, 433-451.

Duffie, D. and R. Kan, 1996, "A Yield-Factor Model of Interest Rates," Mathematical Finance, 6, 379-406.

Duffie, D., 1996, Dynamic Asset Pricing Theory, 2nd ed., Princeton University Press, Princeton, NJ.

Duffie, D. and K.J. Singleton, 1993, "Simulated Moments Estimation of Markov Models of Asset Prices," Econometrica, 61, 929-952.

Duffie, D., J. Pan and K. Singleton, 2000, "Transform Analysis and Asset Pricing for Affine Jump-Diffusions," Econometrica, 68, 1343-1376.

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

Eraker, B., 2001, "MCMC Analysis of Diffusion Models with Applications to Finance," Journal of Business and Economic Statistics, 19, 177-191.

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

Fermanian, J.-D. and B. Salanié, 2003, "A Nonparametric Simulated Maximum Likelihood Estimation Method," forthcoming in Econometric Theory.

Friedman, A., 1975, Stochastic Differential Equations and Applications (Vol. I), New York: Academic Press.

Gallant, A. R., and J. R. Long, 1997, "Estimating Stochastic Differential Equations Efficiently by Minimum Chi-Squared," Biometrika, 84, 125-141.

Gallant, A. R. and G. Tauchen, 1996, "Which Moments to Match ?," Econometric Theory, 12, 657-681.

Glick, N., 1974, "Consistency Conditions for Probability Estimators and Integrals of Density Estimators," Utilitas Mathematica, 6, 61-74.

Gouriéroux, C. and A. Monfort, 1996, Simulation-Based Econometric Methods, Oxford: Oxford University Press.

Gouriéroux, C., A. Monfort and E. Renault, 1993, "Indirect Inference," Journal of Applied Econometrics, 8, S85-S118.

Hansen, L. and J. A. Scheinkman, 1995, "Back to the Future: Generating Moment Implications for Continuous-Time Markov Processes," Econometrica, 63, 767-804.

Harrison, J. M. and S. R. Pliska, 1983, "A Stochastic Calculus Model of Continuous Trading: Complete Markets," Stochastic Processes and their Applications 15, 313-316.

Heston, S., 1993, "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options," Review of Financial Studies, 6, 327-343.

Karatzas, I. and S. Shreve, 1991, Brownian Motion and Stochastic Calculus, Berlin: Springer Verlag.

Kloeden, P.E. and E. Platen, 1999, Numerical Solutions of Stochastic Differential Equations, Berlin: Springer Verlag.

Lindsay, B. G., 1994, "Efficiency versus Robustness: The Case for Minimum Hellinger Distance and Related Methods," Annals of Statistics, 22, 1081-1114.

Mele, A., 2003, "Fundamental Properties of Bond Prices in Models of the Short-Term Rate," Review of Financial Studies, 16, 679-716.

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, "Uniform Convergence in Probability and Stochastic Equicontinuity," Econometrica, 59, 1161-1167.

Newey, W. K. and D. L. McFadden, 1994, "Large Sample Estimation and Hypothesis Testing," in Engle, R.F. and D. L. McFadden (eds.), Handbook of Econometrics, vol. 4, chapter 36, 2111-2245. Amsterdam: Elsevier.

Pagan, A. and A. Ullah, 1999, Nonparametric Econometrics, Cambridge: Cambridge University Press.

Pastorello, S., E. Renault and N. Touzi, 2000, "Statistical Inference for Random-Variance Option Pricing," Journal of Business and Economic Statistics 18, 358-367.

Pedersen, A.R., 1994, "Quasi-Likelihood Inference for Discretely Observed Diffusion Processes," Research report no. 295, Dpt of Theoretical Statistics, University of Aarhus.

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.

Pritsker, M., 1998, "Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models," Review of Financial Studies, 11, 449-487.

Rao, C.R., 1962, "Efficient Estimates and Optimum Inference Procedures in Large Samples," Journal of The Royal Statistical Society, Series B, 24, 46-63.

Revuz, D. and M. Yor, 1999, Continuous Martingales and Brownian Motion, Berlin: Springer Verlag.

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

Romano, M. and N. Touzi, 1997, "Contingent Claims and Market Completeness in a Stochastic Volatility Model," Mathematical Finance, 7, 399-412.

Singleton, K.J., 2001, "Estimation of Affine Asset Pricing Models Using the Empirical Characteristic Function," Journal of Econometrics, 102, 111-141.

Stroock, D.W. and S.R.S. Varadhan, 1979, Multidimensional Diffusion Processes. Berlin: Springer-Verlag.

Tjøstheim D., 1990, "Non-Linear Time Series and Markov Chains," Advances in Applied Probability, 22, 587-611.

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

Table 2A - Baseline Monte Carlo study of the Vasicek model. Finite sample properties of the CD-SNE with optimal weighting function and optimal bandwidth as compared to the MLE for the Vasicek model. Parameter values in the experiment are: $b_1 = 0.06$, $b_2 = 0.5$ and $a_1 = 0.03$. Results are obtained through 1000 replications of samples with 1000 and 500 observations.

Panel A: Sample size $=$ 1000

	CD-SNE (Opt band; Opt weight)			**MLE**		
	b_1	b_2	a_1	b_1	b_2	a_1
mean	0.0638	0.4987	0.0306	0.0587	0.5490	0.0301
median	0.0612	0.4929	0.0305	0.0580	0.5136	0.0300
mean bias	0.0038	-0.0012	0.0007	-0.0013	0.0490	0.0001
std	0.0154	0.1160	0.0011	0.0138	0.1754	0.0007
Rmse	0.0158	0.1163	0.0013	0.0138	0.1822	0.0007

Panel B: Sample size $=$ 500

	CD-SNE (Opt band; Opt weight)			**MLE**		
	b_1	b_2	a_1	b_1	b_2	a_1
mean	0.0595	0.5562	0.0303	0.0567	0.5903	0.0301
median	0.0579	0.5332	0.0302	0.0181	0.2598	0.0009
mean bias	0.0005	0.0562	0.0004	-0.0032	0.0903	0.0001
std	0.0220	0.1818	0.0012	0.0182	0.2598	0.0009
Rmse	0.0197	0.1556	0.0013	0.0184	0.2751	0.0009

Table 2B - Monte Carlo study of bandwidth sensitivity (Vasicek model). Finite sample properties of the CD-SNE with optimal weighting function and optimal bandwidth as compared to the CD-SNE with doubled and halved bandwidth. Parameter values in the experiment are: $b_1 = 0.06$, $b_2 = 0.5$ and $a_1 = 0.03$. Results are obtained through 1000 replications of samples with 1000 observations.

Panel A: Sample size $= 1000$

	CD-SNE (Opt band; Opt weight)			**CD-SNE** (doubled bandwidth)		
	b_1	b_2	a_1	b_1	b_2	a_1
mean	0.0638	0.4987	0.0306	0.0657	0.4812	0.0298
median	0.0612	0.4929	0.0305	0.0611	0.4864	0.0030
mean bias	0.0038	-0.0012	0.0007	0.0057	-0.0188	-0.0002
std	0.0154	0.1160	0.0011	0.0249	0.1405	0.0013
Rmse	0.0158	0.1163	0.0013	0.0256	0.1418	0.0014

Panel B: Sample size $= 1000$

	CD-SNE (Opt band; Opt weight)			**CD-SNE** (halved bandwidth)		
	b_1	b_2	a_1	b_1	b_2	a_1
mean	0.0638	0.4987	0.0306	0.0627	0.5062	0.0307
median	0.0612	0.4929	0.0305	0.0600	0.4997	0.0304
mean bias	0.0038	-0.0012	0.0007	0.0027	0.0062	0.0007
std	0.0154	0.1160	0.0011	0.0157	0.1310	0.0010
Rmse	0.0158	0.1163	0.0013	0.0160	0.1312	0.0012

Table 2C - Monte Carlo study of weighting function choice, and twin smoothing (Vasicek model). Panel A compares the finite sample properties of the CD-SNE with optimal weighting function and optimal bandwidth with the finite sample properties of the SNE with weighting function equal to $\pi(r_t, r_{t-1})$. Panel B compares the finite sample properties of the CD-SNE (with optimal weighting function and bandwidth) with the finite sample properties of estimators replacing model-simulated nonparametric density estimates with model-implied densities expressed in analytical form. Parameter values in the experiment are: $b_1 = 0.06$, $b_2 = 0.5$ and $a_1 = 0.03$. Results are obtained through 1000 replications of samples with 1000 observations.

Panel A: Sample size $= 1000$

	CD-SNE (Opt band; Opt weight)			**SNE** (weight = $\pi(r_t, r_{t-1})$)		
	b_1	b_2	a_1	b_1	b_2	a_1
mean	0.0638	0.4987	0.0306	0.0591	0.5479	0.0288
median	0.0612	0.4929	0.0305	0.0587	0.5146	0.0299
mean bias	0.0038	-0.0012	0.0007	-0.0009	0.0479	-0.0012
std	0.0154	0.1160	0.0011	0.0142	0.1898	0.0035
Rmse	0.0158	0.1163	0.0013	0.0142	0.1948	0.0037

Panel B: Sample size $= 1000$

	CD-SNE (Opt band; Opt weight)			Analytical form		
	b_1	b_2	a_1	b_1	b_2	a_1
mean	0.0638	0.4987	0.0306	0.0596	0.5736	0.0355
median	0.0612	0.4929	0.0305	0 0607	0.4831	0.0346
mean bias	0.0038	-0.0012	0.0007	-0.0003	0.0736	0.0055
std	0.0154	0.1160	0.0011	0.0247	0.3079	0.0062
Rmse	0.0158	0.1163	0.0013	0.0247	0.3157	0.0083

Table 2D - Monte Carlo study of persistence effects (Vasicek model). Finite sample properties of the CD-SNE with optimal weighting function and optimal bandwidth versus the MLE in the case of different levels of persistence. Parameter values of the experiment are: $b_1 = 0.03$, $b_2 = 1.0$ and $a_1 = 0.03$ (Panel A); and $b_1 = 0.006$, $b_2 = 5.0$ and $a_1 = 0.03$ (Panel B). Results are obtained through 1000 replications of samples with 1000 observations.

Panel A: Sample size $= 1000$

	CD-SNE (Opt band; Opt weight)			**MLE**		
	b_1	b_2	a_1	b_1	b_2	a_1
mean	0.0620	0.9803	0.0308	0.0392	1.0396	0.0301
median	0.0599	0.9989	0.0306	0.0334	1.0003	0.0301
mean bias	0.0020	-0.0197	0.0008	-0.0208	0.0396	0.0001
std	0.0093	0.1399	0.0011	0.0166	0.2174	0.0007
Rmse	0.0095	0.1413	0.0013	0.0266	0.2210	0.0007

Panel B: Sample size $= 1000$

	CD-SNE (Opt band; Opt weight)			**MLE**		
	b_1	b_2	a_1	b_1	b_2	a_1
mean	0.0612	4.9390	0.0308	0.0613	5.0042	0.0303
median	0.0602	4.9497	0.0308	0.0602	4.9997	0.0302
mean bias	0.0011	-0.0610	0.0008	-0.0013	-0.0042	0.0003
std	0.0162	0.7964	0.0010	0.0106	0.4423	0.0007
Rmse	0.0162	0.7987	0.0013	0.0108	0.4423	0.0007

Table 3A - Baseline Monte Carlo study of the stochastic volatility model. Finite sample properties of the CD-SNE with "optimal" weighting function and optimal bandwidth for the stochastic volatility model (26). Parameter values in the experiment are: $b_1 = 0.06$, $b_2 = 0.5$, $a_1 = 0.03$, $b_3 = 1.0$ and $a_2 = 0.3$. Results are obtained through 1000 replications of samples with 1000 and 500 observations.

Panel A: Sample size $= 1000$

	CD-SNE (Opt band; Opt weight)				
	b_1	b_2	a_1	b_3	a_2
mean	0.0615	0.5228	0.0309	1.0624	0.3185
median	0.0603	0.5002	0.0307	0.9988	0.3034
mean bias	0.0015	0.0228	0.0009	0.0624	0.0185
std	0.0160	0.1352	0.0016	0.5923	0.1610
Rmse	0.0161	0.1371	0.0018	0.5955	0.1620

Panel B: Sample size $= 500$

	CD-SNE (Opt band; Opt weight)				
	b_1	b_2	a_1	b_3	a_2
mean	0.0649	0.5190	0.0309	1.0905	0.3266
median	0.0601	0.5017	0.0306	1.0028	0.3011
bias	0.0049	0.0190	0.0009	0.0905	0.0266
std	0.0281	0.1722	0.0021	0.5636	0.1787
Rmse	0.0286	0.1733	0.0023	0.5708	0.1807

Table 3B - Monte Carlo study of weighting function choice. Finite sample properties of the CD-SNE with "optimal" weighting function and optimal bandwidth as compared to the SNE implemented with weighting function equal to $\pi(r_t, r_{t-1})$. Parameter values in the experiment are: $b_1 = 0.06$, $b_2 = 0.5$, $a_1 = 0.03$, $b_3 = 1.0$ and $a_2 = 0.3$. Results are obtained through 1000 replications of samples with 1000 observations.

Panel A: Sample size $= 1000$

	CD-SNE (Opt band; Opt weight)				
	b_1	b_2	a_1	b_3	a_2
mean	0.0615	0.5228	0.0309	1.0624	0.3185
median	0.0603	0.5002	0.0307	0.9988	0.3034
mean bias	0.0015	0.0228	0.0009	0.0624	0.0185
std	0.0160	0.1352	0.0016	0.5923	0.1610
Rmse	0.0161	0.1371	0.0018	0.5955	0.1620

Panel B: Sample size $= 1000$

	SNE (weight = $\pi(r_t, r_{t-1})$)				
	b_1	b_2	a_1	b_3	a_2
mean	0.0629	0.4970	0.0292	1.0227	0.2196
median	0.0608	0.4834	0.0292	1.0058	0.2334
mean bias	0.0029	-0.0029	-0.0008	0.0227	-0.0804
std	0.0159	0.1185	0.0012	0.2426	0.1115
Rmse	0.0163	0.1185	0.0014	0.2434	0.1372

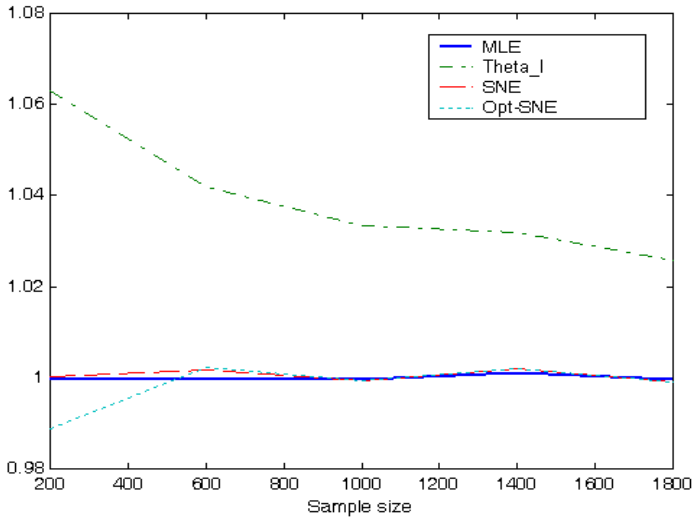
Table 3C - Monte Carlo study of persistence in volatility effects (stochastic volatility model). Finite sample properties of the CD-SNE with "optimal" weighting function and optimal bandwidth for the stochastic volatility model (26) in the case of different levels of persistence in volatility. Parameter values in the experiment are: $b_1 = 0.06$, $b_2 = 0.5$, $a_1 = 0.03$, $b_3 = 1.0$ and $a_2 = 0.3$ (panel A); and $b_1 = 0.06$, $b_2 = 0.5$, $a_1 = 0.03$, $b_3 = 0.4$ and $a_2 = 0.3$ (panel B). Results are obtained through 1000 replications of samples with 1000 observations.

Panel A: Sample size $= 1000$

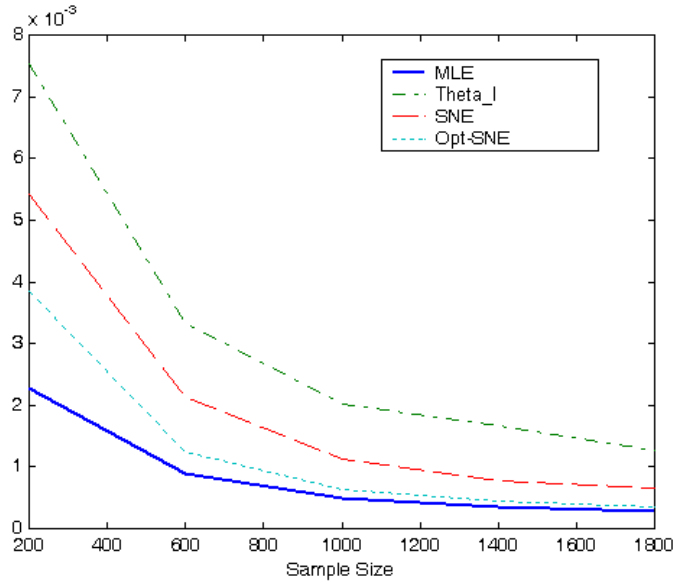
	CD-SNE (Opt band; Opt weight)				
	b_1	b_2	a_1	b_3	a_2
mean	0.0615	0.5228	0.0309	1.0624	0.3185
median	0.0603	0.5002	0.0307	0.9988	0.3034
mean bias	0.0015	0.0228	0.0009	0.0624	0.0185
std	0.0160	0.1352	0.0016	0.5923	0.1610
Rmse	0.0161	0.1371	0.0018	0.5955	0.1620

Panel B: Sample size $= 1000$

	CD-SNE (Opt band; Opt weight)				
	b_1	b_2	a_1	b_3	a_2
mean	0.0634	0.5071	0.0310	0.4490	0.3002
median	0.0612	0.4953	0.0306	0.4035	0.2950
mean bias	0.0034	0.0071	-0.0011	0.0490	0.0002
std	0.0165	0.1469	0.0024	0.2704	0.1265
Rmse	0.0168	0.1470	0.0026	0.2748	0.1266



Panel A - Average against sample size



Panel B - Mean squared error against sample size

Figure 1 - This figure reports the results of a Monte Carlo experiment in which the standard deviation of a standard normal distribution is estimated with four estimators - the MLE, estimator θ^I in (3) with weighting function equal to π_T, the SNE in (7) with $w_T = \pi_T$ and, finally, the SNE in (7) with optimal w_T (opt-SNE). Panel A reports the average of the estimates. Panel B reports the MSE of the estimates.

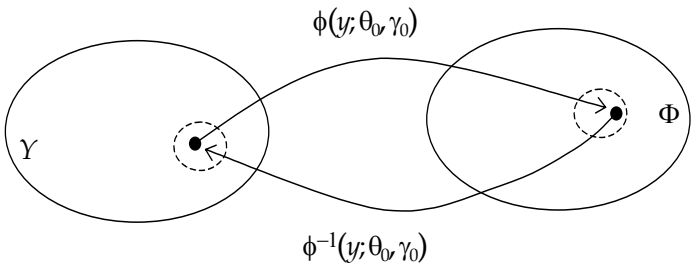


Figure 2 - Asset pricing, the Markov property, and statistical efficiency. Y is the domain on which the partially observed primitive state process $y \equiv (y^o \;\; y^u)^\top$ takes values, Φ is the domain on which the observed system $\phi \equiv (y^o \; C(y))^\top$ takes values in Markovian economies, and $C(y)$ is a contingent claim price process in $\mathbb{R}^{d-q^*}$. Let $\phi^c = (y^o, c(y,\ell_1), \cdots, c(y,\ell_{d-q^*}))$, where $\{c(y,\ell_j)\}_{j=1}^{d-q^*}$ forms an intertemporal cohort of contingent claim prices, as in definition 3. If local restrictions of ϕ are one-to-one and onto, the CD-SNE applied to ϕ^c is feasible. If ϕ is also globally invertible, the CD-SNE applied to ϕ^c achieves the maximum likelihood first-order asymptotic efficiency.