THE REVIEW OF ECONOMIC STUDIES

Review of Economic Studies (2009) **76**, 413–450 0034-6527/09/00150413$02.00

Simulated Non-Parametric Estimation of Dynamic Models

FILIPPO ALTISSIMO
Brevan Howard Asset Management LLP and CEPR

and

ANTONIO MELE
London School of Economics

First version received December 2003; *final version accepted July* 2008 (*Eds.*)

This paper introduces a new class of parameter estimators for dynamic models, called simulated non-parametric estimators (SNEs). The SNE minimizes appropriate distances between non-parametric conditional (or joint) densities estimated from sample data and non-parametric conditional (or joint) densities estimated from data simulated out of the model of interest. Sample data and model-simulated data are smoothed with the same kernel, which considerably simplifies bandwidth selection for the purpose of implementing the estimator. Furthermore, the SNE displays the same asymptotic efficiency properties as the maximum-likelihood estimator as soon as the model is Markov in the observable variables. The methods introduced in this paper are fairly simple to implement, and possess finite sample properties that are well approximated by the asymptotic theory. We illustrate these features within typical estimation problems that arise in financial economics.

1. INTRODUCTION

This paper introduces a new class of parameter estimators for dynamic models with possibly unobserved components, called simulated non-parametric estimators (hereafter SNEs). The SNE aims to minimize measures of distance between the finite-dimensional distributions of the model's observables and their empirical counterparts estimated through standard non-parametric 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 model of interest. In the second step, we obtain the model's density estimates through the application of the same non-parametric 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; thus, it can be employed to cope with a large variety of estimation problems. Second, the SNE minimizes distances of densities smoothed with the same kernel; therefore, up to identifiability, it is consistent, regardless of the smoothing parameter behaviour. Third, if the SNE is taken to match conditional densities and the model is Markov in the observables, it achieves the same asymptotic efficiency as the maximum-likelihood estimator (MLE). Finally, Monte Carlo experiments reveal that our estimator exhibits a proper finite sample behaviour.

Models with unobserved components 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 in finance include latent factor models, continuous-time Markov chains, and even scalar diffusions.

As is well known, the major difficulty in estimating dynamic models with unobserved components relates to the complexity of evaluating the criterion functions. A natural remedy

to this difficulty is to make use of simulation-based methods. The simulated method of moments (McFadden, 1989; Pakes and Pollard, 1989; Lee and Ingram, 1991; Duffie and Singleton, 1993), the simulated pseudo-maximum likelihood method of Laroque and Salanié (1989, 1993, 1994), the indirect inference approach of Gouriéroux, Monfort and Renault (1993) and Smith (1993) and the efficient method of moments of Gallant and Tauchen (1996) are the first attempts to address this problem through extensions of the generalized method of moments. The main characteristic of these approaches is that they are general purpose. Their drawback is that unless the true score belongs to the span of the moment conditions, they lead to asymptotically inefficient estimators, even in the case of fully observed systems (see Carrasco and Florens, 2004; and the early discussion in Tauchen, 1997). There exist alternative simulation-based econometric methods, which directly approximate the likelihood function through simulations (*e.g.* Lee, 1995; or Hajivassiliou and McFadden, 1998). These methods lead to asymptotic efficiency, but they are typically designed to address specific estimation problems.

This article belongs to a new strand of the literature in search for estimators combining the attractive features of both moment matching approaches and the MLE. Fermanian and Salanié (2004) and Carrasco, Chernov, Florens and Ghysels (2006) are two particularly important contributions in this area. Fermanian and Salanié (2004) introduce a general-purpose method in which the likelihood function is approximated by kernel estimates obtained through simulations of the model of interest. The resulting estimator, called the non-parametric simulated maximum likelihood (NPSML) estimator, is then both consistent and asymptotically efficient as the smoothing parameter goes to 0 at some typical convergence rate. Carrasco *et al.* (2006) develop a general estimation technology, which also leads to asymptotic efficiency in the case of fully observed Markov processes. Their method leads to a "continuum of moment conditions" matching model-based (simulated) characteristic functions to data-based characteristic functions. In the following, we illustrate the distinctive features of our approach within a scalar model and contrast it to the other approaches in the literature.

1.1. *Density matching, "twin-smoothing", and the simulated non-parametric estimators*

Let $\{y_t\}_{t=1}^T$ be a sample of data of size T, generated by some conditional law with continuous support. For the purpose of introducing the basic ideas underlying our estimator, we initially discuss a simple case, arising when the goal of the estimator is to calibrate model-based marginal densities to their empirical counterparts. Accordingly, let $\pi(\cdot;\theta_0)$ be the marginal density of y_t, where θ_0 denotes the true parameter value, a point in some parameter set Θ. Let $\pi_T(y)$ be a non-parametric estimator of $\pi(y;\theta_0)$, obtained as $\pi_T(y) \equiv (T\lambda_T)^{-1}\sum_{t=1}^T K((y_t - y)/\lambda_T)$, where λ_T is a bandwidth sequence and K is a symmetric kernel.

Our estimation methodology is related to the classical literature on minimum disparity estimators and goodness-of-fit tests. Consider the following general measure of distance between the two densities $\pi_T(\cdot;\theta)$ and $\pi_T(\cdot)$,

$$M_T(\theta) = \int D(\pi(y;\theta), \pi_T(y))w_T(y)dy, \tag{1}$$

where for each y and θ, the function $D(\pi(y;\theta), \pi_T(y))$ determines how close the model's density $\pi(y;\theta)$ is to $\pi_T(y)$, $w_T(y)$ is a weighting function possibly dependent on the data, and, finally, the integral is taken over the domain of y_t.

The criterion $M_T(\theta)$ in (1) forms the basis for a variety of estimation and testing approaches. Perhaps one of the best known approaches hinges on the goodness-of-fit tests initiated by Bickel and Rosenblatt (1973). These tests typically rely on the empirical distance $M_T(\theta)$ obtained with

$D(\pi, \pi_T) = (\pi - \pi_T)^2$, that is,

$$I_T(\theta) = \int [\pi(y;\theta) - \pi_T(y)]^2 w_T(y) dy, \tag{2}$$

where a common choice for the weighting function is $w_T(y) = \pi_T(y)$.[1] The tests, then, are based on $I_T(\hat{\theta})$, where $\hat{\theta}$ is some consistent estimator of θ_0. Alternatively, it has been noted that the empirical distance in (2) can be utilized to *estimate* the unknown parameter vector θ_0. Notably, Aït-Sahalia (1996) defined an estimator minimizing (2) in the context of scalar diffusions. He used a weighting function $w_T(y) = \pi_T(y)$ to compute the following disparity estimator,

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

Other examples of distance functions $D(\pi, \pi_T)$ in the literature relate to the family of power divergence measures introduced by Cressie and Read (1984) in the context of discrete distributions. In the continuous framework we consider, the Cressie–Read criteria are obtained by using $D(\pi, \pi_T) = [(\pi_T/\pi)^\phi - 1]/(\phi^2 + \phi)$ and $w_T(y) = \pi_T(y)$, where ϕ is a constant.[2]

In this paper, we consider estimators, which are related to the minimization of the quadratic functional $I_T(\theta)$ in (2). A remarkable feature of $I_T(\theta)$ is that a *parametric* density, $\pi(\cdot;\theta)$, is matched to a *non-parametric* density estimate, $\pi_T(\cdot)$. For a fixed bandwidth value $\lambda_T \equiv \bar{\lambda} > 0$ (say), $\pi_T(y)$ converges pointwise in probability to,

$$\pi^*(y;\theta_0,\bar{\lambda}) \equiv \bar{\lambda}^{-1} E[K((y_t - y)/\bar{\lambda})] = \bar{\lambda}^{-1} \int K((u-y)/\bar{\lambda}) \pi(u;\theta_0) du.$$

As is well known, $\pi_T(y)$ converges pointwise in probability to $\pi(y;\theta_0)$ if the bandwidth satisfies: (i) $\lambda_T \to 0$ and (ii) $T\lambda_T \to \infty$. Therefore, bandwidth choice is critical for both consistency and the finite sample behaviour of θ_T^I in (3).

A natural alternative to (2) is an empirical measure of distance in which the non-parametric estimate $\pi_T(\cdot)$ is matched by the model's density smoothed with the same kernel and bandwidth,

$$L_T(\theta) = \int [\pi^*(y;\theta,\lambda_T) - \pi_T(y)]^2 w_T(y) dy. \tag{4}$$

In terms of the empirical measure of distance (1), this alternative criterion replaces the distance function $D(\pi(y;\theta), \pi_T(y))$ in (1) with $D(\pi^*(y;\theta,\lambda_T), \pi_T(y))$, where $D(\pi^*, \pi_T) = (\pi^* - \pi_T)^2$. Fan (1994) developed bias-corrected goodness-of-fit tests based on the previous empirical distance and weighting function $w_T(y) \equiv 1$. Härdle and Mammen (1993) devised a similar bias-correction procedure for testing the closeness of a parametric regression function to a non-parametric one.

Our main idea is to combine the appealing features of the estimator θ_T^I in (3) with the bias-corrected empirical measure $L_T(\theta)$ in (4). Precisely, consider an estimator minimizing the distance in (4) rather than in (2), namely

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

1. Other choices in the literature include $w_T(y) = 1$ (see, for example, Pagan and Ullah, 1999 for a survey), or refer to discrete weightings, which lead to criteria cast in terms of sums, not integrals as in (1) or (2). See, for example, Imbens, Spady and Johnson (1998) or, for more recent work, Antoine, Bonnal and Renault (2007).

2. For example, the criterion $M_T(\theta)$ in (1) collapses to (i) Neyman's χ^2, for $\phi = -2$; (ii) the Kullback–Leibler distance, for $\phi = -1$; (iii) Hellinger's distance, for $\phi = -\frac{1}{2}$; (iv) the likelihood disparity, for $\phi = 0$; and (v) the Pearson's χ^2 distance, for $\phi = 1$.

In (4), kernel smoothing acts in the same manner on both the model-implied density and the data-based density estimate. Therefore, bandwidth conditions affect the two estimators θ_T^I and θ_T^L in a quite different manner. In particular, the criterion in (4) is bias corrected by construction; hence, consistency of θ_T^L holds independently of the bandwidth behaviour, up to identifiability and regularity conditions.

We extend these basic insights to more general settings. We make two additional innovations. First, we consider *conditional* densities, not simply marginal densities as in (4). Second, we accommodate situations in which the analytical solution for such conditional densities is unknown or difficult to compute.

To illustrate our approach, consider again our introductory example, and suppose that y_t is generated by $y_{t+1} = f(y_t, \varepsilon_{t+1}; \theta_0)$, where the transition function f is known, and ε_t is a sequence of independent and identically distributed random variables with known distribution. Assume that for each θ, it is possible to simulate S paths of length T of y_t, by iterating the equation $y_{t+1}^i(\theta) = f(y_t^i(\theta), \tilde{\varepsilon}_{t+1}^i; \theta)$, where for a given simulation $i \in \{1, \ldots, S\}$, and some initial value y_0, $\{\tilde{\varepsilon}_t^i\}_{t=1}^T$ is a sequence of draws from the distribution of ε_t. Consider the joint density estimate on *sample* data, y_t, $\pi_T(y', y) \equiv (T\lambda_T^2)^{-1} \sum_{t=2}^T K_2((y_t - y')/\lambda_T, (y_{t-1} - y)/\lambda_T)$, where K_2 is a symmetric bivariate kernel. Likewise, let $\pi_T^i(y', y; \theta)$ and $\pi_T^i(y; \theta)$ be the joint and marginal density estimates on the *simulated* data, $y_t^i(\theta)$, computed with the same kernels and bandwidth used to estimate the joint and marginal densities on sample data.

Our estimator aims to match the conditional density obtained with simulated data to their sample counterpart, as follows:

$$\theta_{T,S} = \arg\min_{\theta \in \Theta} \iint \left[\frac{1}{S} \sum_{i=1}^{S} \pi_T^i(y' \mid y; \theta) - \pi_T(y' \mid y) \right]^2 w_T(y', y) dy' dy, \tag{6}$$

where $w_T(y', y)$ is a weighting function, and where the conditional densities are estimated as ratios of joint over marginal density estimates, that is, $\pi_T(y' \mid y) \equiv \pi_T(y', y)/\pi_T(y)$ and $\pi_T^i(y' \mid y; \theta) \equiv \pi_T^i(y', y; \theta)/\pi_T^i(y; \theta)$.

The important property of the estimator $\theta_{T,S}$ in (6) is that the conditional density estimates $\pi_T(y' \mid y)$ and $\pi_T^i(y' \mid y; \theta)$ are computed with the *same kernels* and *bandwidth*. This "twin-smoothing" property is the conditional density counterpart to the kernel-smoothing device in (4), and eliminates asymptotic biases affecting non-parametric density estimates. Indeed, we show that the bandwidth behaviour is not critical for consistency of $\theta_{T,S}$, and that $\theta_{T,S}$ does not suffer from finite sample bias, even for simple choices of the bandwidth.

Instead, the bandwidth behaviour affects the *precision* of our estimator. We show that by an appropriate choice of the weighting function $w_T(y', y)$ in (6), $\theta_{T,S}$ can be asymptotically as efficient as the MLE, as soon as the bandwidth goes to 0 at some appropriate rate. Intuitively, we require the bandwidth to go to 0 in order to match, asymptotically, the model's conditional density to the true conditional density of y_t. This matching, and the particular weighting function we shall use, make the estimator asymptotically equivalent to (a linear function of) the true score, as we shall show.

The computational aspects of our estimator are useful to mention. First, the criterion in (6) can be evaluated through Monte Carlo integration, which avoids using integration quadratures. This attractive feature of the estimator extends to the general case we consider in this paper (estimation of multidimensional models with possibly unobservable variables), in which the criterion functions involve integrals of dimension larger than that in (6). Second, we shall explain, and our Monte Carlo experiments confirm, that the computational burden arising in multivariate settings can be further reduced, with a loss in efficiency, when the estimator is taken to match the conditional densities of the single elements of the observable variables.

1.2. *Related literature*

By construction, our estimation approach is not meant to approximate the likelihood function. Rather, we develop a class of simulation-based criterion functions (that in equation (6)), which generalizes a classical measure of distance between marginal densities (the quadratic functionals in equations (2) and (4)) to a setting of conditional densities. Thus, albeit based on simulations, our estimation strategy is distinct from the NPSML methodology introduced by Fermanian and Salanié (2004) and also considered by Kristensen and Shin (2006). Our approach also differs from that of Carrasco *et al.* (2006). Indeed, we also rely on a "continuum of moments" in (6); at the same time, we match model-based density estimates (not characteristic functions) to their empirical counterparts. Finally, the twin-smoothing procedure in (6) further differentiates our approach from those in these papers.

The twin-smoothing procedure is intimately related to the general indirect inference strategy put forward in the seminal papers of Gouriéroux *et al.* (1993) and Smith (1993). In the language of indirect inference, we are calibrating the parameter of interest θ by matching a model-implied (infinite-dimensional) "auxiliary" parameter (*i.e.* $\frac{1}{S}\sum_{i=1}^{S}\pi_T^i(y' \mid y;\theta)$) to the corresponding parameter computed on sample data (*i.e.* $\pi_T(y' \mid y)$). In principle, both of these "auxiliary" parameters might be estimated with an arbitrary bandwidth choice. Indeed, the important point is that the two "auxiliary" parameters be estimated with the same kernel and bandwidth. In this case, and up to identifiability, our estimator $\theta_{T,S}$ in (6) is still consistent, as we would expect it to be by the logic underlying indirect inference.

Our basic ideas are also related to the kernel-based indirect inference approach developed by Billio and Monfort (2003). The Billio–Monfort estimator matches conditional expectations of arbitrary test functions estimated through the same kernel method—one conditional expectation computed on sample data and one conditional expectation computed on simulated data. Like our estimator, their estimator is not affected by any asymptotic bias. One important difference between our estimator and the Billio–Monfort estimator is that ours is asymptotic normal at the usual parametric rate, and it can be asymptotically as efficient as the MLE. Instead, the rate of convergence of the Billio–Monfort estimator is contaminated by the rate of convergence of their bandwidth sequence to 0, although this rate can be made arbitrarily slow. Intuitively, the Billio–Monfort estimator matches a finite number of test functions. Instead, our estimator $\theta_{T,S}$ in (6), like Aït-Sahalia's (1996) estimator (3), can be understood as a device to match a continuum of moment conditions. The integration step over such a continuum of moment conditions eliminates the effect of the bandwidth on the rate of convergence of $\theta_{T,S}$ in (6).

Aït-Sahalia (1996) is one additional fundamental contribution, which this article is clearly related to. Aït-Sahalia developed a minimum distance estimator, that in equation (3), for which the measure of distance is, asymptotically, a special case of the general class of measures of distance we consider here. Our estimator, however, is different for three additional important reasons. First, Aït-Sahalia's estimator is affected by an asymptotic bias, which, instead, does not arise within the class of our estimators, due to our twin-smoothing device. Second, Aït-Sahalia's estimator only matches marginal densities. Third, our conditional-density based estimator in (6) can lead to asymptotic efficiency.

Our focus on matching density functions is also related to the "effective calibration" strategy of Gallant (2001). The main difference is that Gallant's estimator matches cumulative distribution functions, and it does not lead to asymptotic efficiency. The advantage of focusing on density functions is that it allows us to address efficiency issues.

Finally, we note that Hong and White (2005) have recently made use of a twin smoothing trick similar to ours to estimate joint and marginal densities in the context of non-parametric entropy measures of serial dependence. This trick leads to fairly weak bandwidth conditions and

a faster convergence rate for their entropy estimator. Aït-Sahalia, Fan and Peng (2005) have also recently used a device similar to ours to reduce the bias of non-parametric goodness-of-fit tests for scalar diffusion models. Naturally, the focus of these two papers is radically different from our focus to provide parameter estimators for dynamic models.

The rest of the paper is organized in the following manner. Section 2 introduces our simulated non-parametric estimators in detail. Section 3 provides the large sample theory. Section 4 assesses finite sample properties. Section 5 concludes.

A word on notation: For any $\mathbb{R}^d$-valued variable X, we use $\|X\|$ to denote the Euclidean norm and $|X|_2$ to denote the outer product. All the integrals are taken on the real coordinate space. We use the double integral notation $\iint$ in the context of conditional density matching, as in (6). We use the notation $\int$ in the context of joint density matching, and in all remaining contexts.

2. SIMULATED NON-PARAMETRIC ESTIMATORS

2.1. *The model of interest*

Let $\Theta \subset \mathbb{R}^n$ be a compact parameter set and, for a given parameter vector θ_0 in the interior of Θ, consider the following data generating process:

$$y_{t+1} = f(y_t, \varepsilon_{t+1}; \theta_0), \quad t = 0, 1, \ldots, \tag{7}$$

where $y_t \in \mathbb{R}^d$, f is known and ε_t is a sequence of $\mathbb{R}^d$-valued independent and identically distributed random variables with known distribution. The purpose of this paper is to provide estimators of the true parameter vector θ_0.

We consider a general situation in which some components of y_t are not observed. Accordingly, we partition the vector y_t as $y_t \equiv [y_t^{\mathrm{o}}, y_t^{\mathrm{u}}]$, where $y_t^{\mathrm{o}} \in \mathbb{R}^{q^*}$ is the vector of the observable variables, and $y_t^{\mathrm{u}} \in \mathbb{R}^{d-q^*}$ is the vector of the unobservable variables. Since our general interest lies in the estimation of partially observed processes, we may wish to recover as much information as possible about the dependence structure of the observables in (7). Thus, we stack y_t^{o} and l lagged values of y_t^{o} into a vector $x_t \in \mathbb{R}^q$, with $q = q^*(1+l)$, where

$$x_t \equiv [y_t^{\mathrm{o}}, \cdots, y_{t-l}^{\mathrm{o}}], \quad t = 1+l, \ldots, T. \tag{8}$$

In practice, there is a clear trade-off between increasing the lag length l and both speed of computations *and* the curse of dimensionality. In Section 2.3, we succinctly present a few practical devices on how to cope with the curse of dimensionality. Finally, for each t, we partition x_t as $x_t = [z_t, \upsilon_t]$, where $z_t \equiv y_t^{\mathrm{o}} \in \mathbb{R}^{q^*}$ is the vector of the observable variables at time t, and $\upsilon_t \in \mathbb{R}^{q-q^*}$ is the vector of predetermined variables,

$$\upsilon_t \equiv [y_{t-1}^{\mathrm{o}}, \cdots, y_{t-l}^{\mathrm{o}}], \quad t = 1+l, \ldots, T. \tag{9}$$

Throughout the paper, we let $\pi_2(x;\theta_0) \equiv \pi_2(z, \upsilon; \theta_0)$ denote the joint density induced by (7) on (8); $\pi_1(\upsilon;\theta_0)$ the joint density of the predetermined variables in (9); and $\pi(z \mid \upsilon;\theta_0)$ the conditional density of z_t given υ_t.

2.2. *Conditional density SNE*

Consider a non-parametric estimator of the joint density $\pi_2(x;\theta_0)$, obtained as $\pi_{2T}(x) \equiv (T\lambda_T^q)^{-1} \sum_{t=1+l}^{T} K_q((x_t - x)/\lambda_T)$ where K_q is a q-dimensional, r-th order, symmetric kernel,[3]

3. A symmetric kernel K is a symmetric function around 0 that integrates to 1. It is said to be of the r-th order if (i) $\forall \mu \in \mathbb{N}^q : |\mu| \in \{1, \cdots, r-1\}$, $|\mu| \equiv \sum_{j=1}^{q} \mu_j$, $\int u_1^{\mu_1} \cdots u_q^{\mu_q} K(u)du = 0$; (ii) $\exists \mu \in \mathbb{N}^q : |\mu| = r$ and $\int u_1^{\mu_1} \cdots u_q^{\mu_q} K(u)du \neq 0$; and (iii) $\int \|u\|^r K(u)du < \infty$.

and λ_T is the bandwidth function. Similarly, estimate the joint density of the predetermined variables $\pi_1(v;\theta_0)$ as $\pi_{1T}(v) \equiv (T\lambda_T^{q-q^*})^{-1}\sum_{t=1+l}^{T} K_{q-q^*}((v_t - v)/\lambda_T)$. Let

$$\pi_T(z \mid v) \equiv \frac{\pi_{2T}(z,v)}{\pi_{1T}(v)} \tag{10}$$

be an estimate of the conditional density of the observed variables z_t given the predetermined variables v_t.

Our estimator aims to match the model-implied conditional density to the conditional density $\pi_T(z \mid v)$ estimated from sample data. The first step of our estimation strategy requires simulated paths of the observable variables in (7). To generate S simulated paths for a given parameter value θ and some initial point y_0, we compute recursively,

$$y_{t+1}^i(\theta) = f(y_t^i(\theta), \tilde{\varepsilon}_{t+1}^i; \theta), \quad t = 0, 1, \ldots, T, \quad \text{for } i = 1, \ldots, S,$$

where for each simulation i, $\{\tilde{\varepsilon}_t^i\}_{t=1}^T$ is a sequence of random numbers drawn from the distribution of ε_t. Let $x_t^i(\theta)$ and $v_t^i(\theta)$ be the simulated counterparts to x_t and v_t in (8) and (9), when the parameter vector is θ. Define $\pi_{2T}^i(x;\theta) \equiv (T\lambda_T^q)^{-1}\sum_{t=1+l}^{T} K_q((x_t^i(\theta) - x)/\lambda_T)$ and $\pi_{1T}^i(v;\theta) \equiv (T\lambda_T^{q-q^*})^{-1}\sum_{t=1+l}^{T} K_{q-q^*}((v_t^i(\theta) - v)/\lambda_T)$, where K_q, K_{q-q^*} and λ_T are the same kernels and bandwidth used to compute $\pi_T(z \mid v)$ in (10). We estimate the conditional density on simulated data for a given parameter value θ as an average of the simulated ratios of joint over marginal densities,

$$\pi_{T,S}(z \mid v;\theta) \equiv \frac{1}{S}\sum_{i=1}^{S} \frac{\pi_{2T}^i(z,v;\theta)}{\pi_{1T}^i(v;\theta)}, \quad i = 1, \ldots, S. \tag{11}$$

We are now in a position to provide the definition of our estimator:

Definition 1. (CD-SNE) For each fixed S, the Conditional Density SNE (CD-SNE) is the sequence $\{\theta_{T,S}\}_T$ given by:

$$\begin{aligned}\theta_{T,S} &= \arg\min_{\theta\in\Theta} L_{T,S}^{\mathrm{CD}}(\theta) \\ &\equiv \arg\min_{\theta\in\Theta} \iint [\pi_{T,S}(z \mid v;\theta) - \pi_T(z \mid v)]^2 w_T(z,v)\mathbb{T}_{T,S}^2(v;\theta)dzdv,\end{aligned} \tag{12}$$

where $\{w_T(z,v)\}_T$ is a sequence of positive weighting functions; and for each $\theta \in \Theta$, $\{\mathbb{T}_{T,S}(v;\theta)\}_T$ is a sequence of positive trimming functions depending on the simulations.

The objective function in (12) involves a *weighting* function $w_T(z,v)$ and also a *trimming* function $\mathbb{T}_{T,S}(v;\theta)$. The role of the weighting function is to literally weight the distance of the conditional densities $\pi_{T,S}(z \mid v;\theta)$ and $\pi_T(z \mid v)$ at all the points (z,v). For instance, if $w_T(z,v) = \pi_{2T}(z,v)$, the CD-SNE overweights discrepancies occurring where observed data have more mass. In Section 3, we will indicate when and how the CD-SNE can be made asymptotically efficient with a proper choice of the weighting function $w_T(z,v)$.

The trimming function $\mathbb{T}_{T,S}(v;\theta)$ plays instead a merely technical role. The CD-SNE relies on non-parametric conditional density estimates obtained as ratios between joint over marginal density estimates. Small values of the denominators in (10)–(11) may hinder the numerical stability of the estimator and the asymptotic theory. Therefore, we need to control the tail behaviour of the marginal density estimates $\pi_{1T}(v)$ and $\pi_{1T}^i(v;\theta)$. The role of $\mathbb{T}_{T,S}(v;\theta)$ is to trim small values of these marginal density estimates. A similar "denominator" problem arises in many related

contexts, and is addressed by means of trimming functions (see, for example, Andrews, 1995; Ai, 1997; Fermanian and Salanié, 2004; Kitamura, Tripathi and Ahn, 2004). We defer to Section 3 a careful and complete description of the regularity conditions on $\mathbb{T}_{T,S}(\upsilon;\theta)$ (see Assumptions 8–10). At this juncture, we simply note that $\mathbb{T}_{T,S}(\upsilon;\theta)$ also depends on the parameter θ being used to evaluate the criterion in (12). This is because the tails of the marginal density estimates on simulated data $\{\pi^i_{1T}(\upsilon;\theta)\}_{i=1}^S$ obviously depend on the parameter vector θ used to produce the simulations.

2.3. *The curse of dimensionality*

Estimators relying on non-parametric methods are subject to two well-known critiques. A first critique relates to the difficulty to properly deal with density estimation in high dimension when the sample size is small, as in the famous "empty space phenomenon" described by Silverman (1986, Section 4.5). A second critique stems from the mere computational burden of using kernel methods in high dimension. For example, the numerical integration underlying the CD-SNE can be computationally costly when q, the dimension of x_t in (8), is large.

As regards the first critique, note that by design, and consistently with the indirect inference principle, the CD-SNE aims to make a model mimic the sample properties of an "auxiliary" infinite dimensional parameter, that is, a conditional density. Hence, the "twin-smoothing" device underlying the criterion in (12) is such that the biases in the model-implied density estimate and its empirical counterpart cancel out each other. As we shall show, this makes the CD-SNE consistent independently of how well we are able to estimate the two densities.

A standard but convenient way to address the second critique is to use Monte Carlo integration, and compute the objective functions through sample averages, rather than through integration quadratures. In this manner, a q-dimensional Riemann integral is approximated by a one-dimensional sum over kernel density estimates evaluated at the sample points. Even with integration quadratures, one can still reduce the computational burden, with a loss in efficiency. For example, the computational burden related to the lag l in (8) can be reduced by matching the conditional densities of the individual elements of x_t, as follows:

$$\hat{\theta}_{T,S} = \arg\min_{\theta\in\Theta} \sum_{k=1}^{l} \int_{\mathbb{R}^{2q^*}} [\pi_{T,S}(y^{\mathrm{o}} \mid y^{\mathrm{o}}_{-k};\theta) - \pi_T(y^{\mathrm{o}} \mid y^{\mathrm{o}}_{-k})]^2 \mathbb{T}^2_{T,S}(y^{\mathrm{o}}_{-k};\theta) w_T(y^{\mathrm{o}}, y^{\mathrm{o}}_{-k}) dy^{\mathrm{o}} dy^{\mathrm{o}}_{-k}, \tag{13}$$

where $\pi_T(y^{\mathrm{o}} \mid y^{\mathrm{o}}_{-k})$ is the estimate of the conditional density of two observations that are k lags apart in (8), $\pi_{T,S}(y^{\mathrm{o}} \mid y^{\mathrm{o}}_{-k};\theta)$ is its simulated counterpart, and $\mathbb{T}_{T,S}(y^{\mathrm{o}}_{-k};\theta)$ is a trimming function.

This estimator is similar to the CD-SNE in Definition 1, but can be implemented with a number l of $2q^*$-dimensional integrations, instead of a single q-dimensional integration, $q = q^*(1+l)$. In proposing the above estimator, we imitated Fermanian and Salanié (2004, section 4) and Singleton (2006, section 5.7), who suggested "splits" similar to those in (13) to address dimensionality issues in the context of their estimators. In tests involving stochastic volatility models, we found that the CD-SNE computed with $l = 1$ (*i.e.* the CD-SNE matching the conditional density of two adjacent observations) has a proper finite sample behaviour (see Section 4). In the remainder, we develop the asymptotic properties of the CD-SNE in Definition 1, and leave the details of the asymptotics of the estimator $\hat{\theta}_{T,S}$ in (13) in Appendix C. Dimensionality issues related to the spatial dimension q^* can be mitigated in the same vein.[4]

4. For example, a possible estimator extending the CD-SNE could match the estimates of two-dimensional conditional densities of every single pair of observables, rather than the estimates of $2q^*$-dimensional densities.

3. LARGE SAMPLE THEORY

3.1. *Regularity conditions*

This section collects the regularity conditions we need to develop the asymptotic theory for the CD-SNE. Our first assumptions further characterize the family of models underlying the data generating process in (7).

Assumption 1. (a) *For all* $z, v, \theta \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} \times \Theta$, $\pi(z \mid v; \theta)$, $\pi_2(z, v; \theta)$ *and* $\pi_1(v; \theta)$ *are bounded, and continuous in all their arguments.* (b) *For all* $z, v, \theta \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} \times \Theta$, $\pi(z \mid v; \theta)$, $\pi_2(z, v; \theta)$ *and* $\pi_1(v; \theta)$ *are twice continuously differentiable with respect to* θ*, and their derivatives up to the second order with respect to* θ *are bounded. Furthermore, for all* $y, \varepsilon, \theta \in \mathbb{R}^d \times \mathbb{R}^d \times \Theta$, $f(y, \varepsilon; \theta)$ *is continuous and twice continuously differentiable with respect to* θ*.*

Assumption 2. *The vector-valued process* y_t *in* (7) *is a Markov* β*-mixing sequence with mixing coefficients* β_k *satisfying* $\lim_{k\to\infty} k^{\mu}\beta_k \to 0$*, for some* $\mu > 1$*.*

Assumption 1(a) (resp. 1(b)) is needed to prove consistency (resp. asymptotic normality) of our estimators. Assumption 2 imposes restrictions on the data dependence and is needed for the application of an empirical process central limit theorem (Arcones and Yu, 1994) to density kernel estimates. The next assumption lists the basic regularity conditions on the kernel functions.

Assumption 3. *The kernels* K_q *and* K_{q-q^*} *are of the same order* r*, bounded, symmetric, and continuously differentiable with bounded derivatives up to the second order. Moreover, they are absolutely integrable with an absolutely integrable Fourier transform.*

Assumption 3 is needed to use and extend Andrews' (1995) results on the uniform convergence of density estimates and their derivatives. These results are of critical importance for both consistency and asymptotic normality, as we shall highlight below.[5]

We now introduce notation and regularity conditions related to the asymptotic behaviour of the criterion function in (12). First, we define the limiting bandwidth value $\bar{\lambda} : \lambda_T \to \bar{\lambda}$, and the pointwise probability limits, for fixed S and θ, and $T\lambda_T^q \to \infty$,

$$\pi_2^*(z, v; \theta, \bar{\lambda}) \equiv \operatorname*{plim}_{T\to\infty} \pi_{2T,S}(z, v; \theta), \quad \pi_1^*(v; \theta, \bar{\lambda}) \equiv \operatorname*{plim}_{T\to\infty} \pi_{1T,S}(v; \theta). \tag{14}$$

It is well known (*e.g.* Pagan and Ullah, 1999) that given Assumptions 1–3, these probability limits collapse to the joint densities $\pi_2(z, v; \theta)$ and $\pi_1(v; \theta)$, once we let $\lambda_T \to 0$.

Second, we define,

$$L^{\mathrm{CD}}(\theta, \bar{\lambda}) \equiv \iint [\pi^*(z \mid v; \theta, \bar{\lambda}) - \pi^*(z \mid v; \theta_0, \bar{\lambda})]^2 w(z, v) dz dv, \quad \pi^*(z \mid v; \theta, \bar{\lambda}) \equiv \frac{\pi_2^*(z, v; \theta, \bar{\lambda})}{\pi_1^*(v; \theta, \bar{\lambda})}, \tag{15}$$

where $w(z, v)$ is the probability limit of the weighting function $w_T(z, v)$ in (12), as formalized by the following assumption.

Assumption 4. (a) *The sequence of functions* $\{w_T(z, v)\}_T$ *is bounded and integrable, and converges in probability pointwise to some function* $w(z, v)$ *as* $T \to \infty$*, where the limiting*

5. Precisely, Assumption 3 is needed to prove Lemmas C1–C3 and Lemmas N1–N5 in the appendices through Andrews' (1995) strategy of proof.

function $w(z,\upsilon)$ is bounded and integrable. (b) *We have,* $\sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\mathbb{R}^{q-q^*}}|w_T(z,\upsilon)-w(z,\upsilon)|=O_p(T^{-\frac{1}{2}}\lambda_T^{-q})+O_p(\lambda_T^r)$.

Assumption 4(a), combined with Assumptions 1(a), 2, 3, and additional regularity conditions on the trimming function $\mathbb{T}_{T,S}(\upsilon;\theta)$ (stated below), ensures that $L^{\mathrm{CD}}(\theta,\bar{\lambda})$ is the probability limit of $L^{\mathrm{CD}}_{T,S}(\theta)$, for fixed θ, a key ingredient for the consistency of the CD-SNE. Assumption 4(b), instead, contains a high level condition we use to show asymptotic normality of the CD-SNE. This condition is satisfied, for example, by $w_T(z,\upsilon)=\pi_T(z,\upsilon)$.

We assume that the objective function $L^{\mathrm{CD}}_{T,S}(\theta)$ in (12) and the limiting objective function $L^{\mathrm{CD}}(\theta,\bar{\lambda})$ in (15) satisfy the following regularity and identifiability conditions:

Assumption 5. *For all $\theta\in\Theta$, $L^{\mathrm{CD}}_{T,S}(\theta)$ is measurable and continuous on Θ. Moreover, the function $\pi^*(z\mid\upsilon;\theta,\bar{\lambda})$ in* (15) *is bounded, $L^{\mathrm{CD}}(\theta,\bar{\lambda})$ is bounded and continuous on Θ, and there exists a unique θ_0 in the interior of Θ such that $L^{\mathrm{CD}}(\theta,\bar{\lambda})=0$ implies that $\theta=\theta_0$.*

The first part of Assumption 5 contains standard regularity conditions that ensure the existence of the CD-SNE. The identification condition in the second part of this assumption is critical. It requires that the "auxiliary" parameter $\pi^*(z\mid\upsilon;\theta,\bar{\lambda})$ in (15) has information content on the "structural" parameter θ, possibly for all the points (z,υ), and can equivalently be stated in terms of the limiting bandwidth $\bar{\lambda}$, as follows: $\bar{\lambda}:\sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\mathbb{R}^{q-q^*}}|\pi^*(z\mid\upsilon;\theta,\bar{\lambda})-\pi^*(z\mid\upsilon;\theta_0,\bar{\lambda})|=0\Longrightarrow\theta=\theta_0$. For example, if the probability limits in (14) are such that $\pi_2^*(z,\upsilon;\theta,\bar{\lambda})=\pi_2(z,\upsilon;\theta)$ and $\pi_1^*(\upsilon;\theta_0,\bar{\lambda})=\pi_1(\upsilon;\theta)$, then, this condition collapses to a standard identifiability condition for the conditional density of z given υ.

More generally, the previous condition needs not to be satisfied in some special cases, arising when the limiting bandwidth is larger than the support of the data. Consider the following counterexample. Suppose that y_t is independent and identically distributed and that $y\in Y\subset\mathbb{R}$, where Y is a compact set. Consider the uniform kernel with support $[-1,1]$, $K(y)\equiv\frac{1}{2}\mathbb{I}_{|y|\leq 1}$, where $\mathbb{I}$ is the indicator function. In this case, the identification condition is $L_Y(\theta,\bar{\lambda})=0\Longrightarrow\theta=\theta_0$, where, for some weighting function $w(y)$,

$$L_Y(\theta,\bar{\lambda})\equiv\int_{y\in Y}[\pi_1^*(y;\theta,\bar{\lambda})-\pi_1^*(y;\theta_0,\bar{\lambda})]^2w(y)dy,$$

and, by simple computations,

$$\pi_1^*(y;\theta,\bar{\lambda})-\pi_1^*(y;\theta_0,\bar{\lambda})=\frac{1}{2\bar{\lambda}}\int_{\xi\in Y}\mathbb{I}_{|y-\xi|\leq\bar{\lambda}}[\pi_1(\xi;\theta)-\pi_1(\xi;\theta_0)]d\xi. \tag{16}$$

With $\bar{\lambda}$ large enough, we have that $\mathbb{I}_{|y-\xi|\leq\bar{\lambda}}=1$ for all $(y,\xi)\in Y\times Y$, which implies that $L_Y(\theta)=0$ *for all* $\theta\in\Theta$.

The previous situation does not necessarily arise if the support of the data is $\mathbb{R}$. Indeed, if the data have unbounded support and the kernel is uniform, then

$$\pi_1^*(y;\theta,\bar{\lambda})-\pi_1^*(y;\theta_0,\bar{\lambda})=\frac{1}{2\bar{\lambda}}\int_{y-\bar{\lambda}}^{y+\bar{\lambda}}[\pi_1(\xi;\theta)-\pi_1(\xi;\theta_0)]d\xi. \tag{17}$$

Note the role that unbounded support plays here. Indeed, and unless $\bar{\lambda}=\infty$, the R.H.S. in (17) is not identically 0, as it is instead the case in (16), when $\bar{\lambda}$ is large enough. In this example, identification occurs if $\bar{\lambda}:\sup_{y\in\mathbb{R}}|\pi_1^*(y;\theta,\bar{\lambda})-\pi_1^*(y;\theta_0,\bar{\lambda})|=0\Longrightarrow\theta=\theta_0$.

In Appendix A.2, we develop one example in which identification occurs, when y_t is Gaussian, the kernel is still uniform, and the limiting bandwidth $\bar{\lambda}$ is non-zero. Appendix A.2 provides one additional example in which identifiability occurs with sufficiently small values of $\bar{\lambda}$, even when data have bounded support.

To prove consistency of the CD-SNE, we need additional regularity conditions:

Assumption 6. *There exists a $c > 0$ and a sequence $\{\kappa_{T,S}\}_T$ bounded in probability as T becomes large such that for all $(\varphi, \theta) \in \Theta \times \Theta$, $|L^{\mathrm{CD}}_{T,S}(\varphi) - L^{\mathrm{CD}}_{T,S}(\theta)| \leq \kappa_{T,S}\|\varphi - \theta\|^c$.*

Assumption 6 is a standard high level assumption we need to prove that $L^{\mathrm{CD}}_{T,S}(\theta)$ converges to $L^{\mathrm{CD}}(\theta, \bar{\lambda})$, uniformly in θ. The following conditions are needed to Taylor-expand the first order conditions satisfied by the CD-SNE and to ensure the uniform convergence of non-parametric estimates of score functions to their asymptotic counterparts.

Assumption 7. (a) *For all $(x, \theta) \in \mathbb{R}^q \times \Theta$, each element of $|\nabla_\theta K_q((x^i_t(\theta) - x)/\lambda_T)|$ and $|\nabla_\theta K_{q-q^*}((v^i_t(\theta) - v)/\lambda_T)|$ exists, is continuous in θ, bounded with bounded gradient and satisfies Assumption* 2. (b) *$\partial^{\rho+1}\pi_2(x;\theta)/\partial\theta\partial x^\rho$ and $\partial^{\rho+1}\pi_1(v;\theta)/\partial\theta\partial v^\rho$ are uniformly bounded for some $\rho \geq r$.*

The next assumption specifies the trimming function we use to address the "denominator" problems in (10) and (11):

Assumption 8. *Let g be a bounded, twice continuously differentiable density function with support $[0, 1]$, $g(0) = g(1) = 0$, and let $g_\delta(u) \equiv \frac{1}{\delta} g(\frac{u}{\delta} - 1)$.*[6] *We set, in* (12), *$\mathbb{T}_{T,S}(v;\theta) \equiv G_{\delta_T}(\pi_{1T}(v)) \prod_{i=1}^S G_{\delta_T}(\pi^i_{1T}(v;\theta))$, where $G_{\delta_T}(\ell) \equiv \int_0^\ell g_{\delta_T}(u)du$, for some sequence $\delta_T : \delta_T \to 0$.*

Our formulation of the trimming function $\mathbb{T}_{T,S}(v;\theta)$ is related to previous work by Andrews (1995) and Ai (1997). By construction, this function has the following two fundamental properties, holding for all θ: (i) $\mathbb{T}_{T,S}(v;\theta) = 0$ for all v such that $\pi_{1T}(v) < \delta_T$ or $\pi^i_{1T}(v;\theta) < \delta_T$ (for at least one simulation i); and (ii) $\mathbb{T}_{T,S}(v;\theta) = 1$ for all v such that $\pi_{1T}(v) > 2\delta_T$ and $\pi^i_{1T}(v;\theta) > 2\delta_T$ (for all the simulations $i \in \{1, \ldots, S\}$). Hence, the function $\mathbb{T}_{T,S}(v;\theta)$ trims small values of the density estimates $\pi_{1T}(v)$ and $\pi^i_{1T}(v;\theta)$ on both sample and simulated data. Finally, the condition that $\delta_T \to 0$ ensures that $\mathbb{T}_{T,S}(v;\theta) \xrightarrow{p} 1$ for all v and θ, thereby making the trimming effects asymptotically negligible.[7]

In Assumption 9 below, we gather all the regularity conditions on the asymptotic behaviour of the trimming sequence δ_T and the bandwidth sequence λ_T:

Assumption 9. *As $T \to \infty$,*

(a) *$\lambda_T \to \bar{\lambda}$, where $0 \leq \bar{\lambda} < \infty$, and $T^{\frac{1}{2}}\lambda_T^q\delta_T \to \infty$;*
(b) *$\lambda_T \to 0$, $T^{\frac{1}{2}}\lambda_T^{q+1}\delta_T^4 \to \infty$, and $\delta_T^{-1}\lambda_T^\psi \to 0$, where $\psi \equiv \min\{q^* + 1, \frac{1}{5}r\}$.*

6. Let $g^{(d)}$ be the d-th order derivative of g, with $g^{(0)} \equiv g$. By convention, the derivatives at the end points 0 and 1, $g^{(d)}(0) \equiv \lim_{x\to 0^+}[g^{(d-1)}(x) - g^{(d-1)}(0)]/x$ and $g^{(d)}(1) \equiv \lim_{x\to 0^-}[g^{(d-1)}(1+x) - g^{(d-1)}(1)]/x$, for $d = 1, 2$.

7. Another key property of this function is that it is twice continuously differentiable with respect to θ, which allows expansion of the CD-SNE through Taylor series arguments. Linton and Xiao (2000) suggested the following example of trimming function with a closed-form solution satisfying Assumption 8. Let the Beta density $g(u) \propto z^k(1 - z)^k$, for some integer k; then $G_\delta(\ell)$ is a $(2k+1)$-polynomial in $(\ell - \delta)/\delta$.

Assumption 9(a) contains conditions on the joint asymptotic behaviour of λ_T and δ_T that ensure consistency of the CD-SNE. We require that δ_T do not decay too rapidly, that is, $\delta_T : T^{\frac{1}{2}}\lambda_T^q\delta_T \to \infty$. We need this condition to ensure that the kernel estimate $\pi_T(z \mid v) = \pi_{2T}(z, v)/\pi_{1T}(v)$ in (10) converges uniformly to its population counterpart $\pi^*(z \mid v; \theta_0, \bar{\lambda})$ in (15), and over sets of v on which the "denominator" $\pi_{1T}(v)$ is bounded away from 0. Moreover, we need a "square-root-T" condition to ensure such a uniform convergence.[8]

Assumption 9(b) is needed to prove asymptotic normality of the CD-SNE. We require that the bandwidth sequence λ_T converges to 0 at the "square-root-T" rate, in the sense that $T^{\frac{1}{2}}\lambda_T^{q+1}\delta_T^4 \to \infty$, and that it enters the square-root-T condition with a power of $q+1$. These two conditions are needed to ensure uniform convergence of non-parametric estimates of gradient functions and in the case of Markov models, make the CD-SNE asymptotically as efficient as the MLE, as we shall explain in Section 3.3. The second part of Assumption 9(b) imposes that the rate of decay for the trimming sequence δ_T be slower than in Assumption 9(a). We need this condition as non-parametric estimates of gradient functions, $\nabla_\theta \pi_{T,S}(z \mid v; \theta)$, involve denominator problems that are more severe than those arising for consistency. Finally, note that the order of the kernel r also plays a role. This is because we need to ensure that estimates of density functions and score functions converge uniformly to true densities and scores. The third part of Assumption 9(b) on the order of the kernel ensures that asymptotic biases affecting non-parametric estimates of density and gradient functions are asymptotically eliminated.

3.2. *Consistency and asymptotic normality*

This section establishes consistency and asymptotic normality of the CD-SNE. To develop intuition about these asymptotic properties, let us consider the first order conditions satisfied by the CD-SNE (see Appendix A.3, equation (A10)). Under the regularity conditions of Section 3.1, these conditions can be Taylor-expanded around θ_0 to yield,

$$\iint \sqrt{T}[\pi_{T,S}(z \mid v; \theta_0) - \pi_T(z \mid v)]\nabla_\theta \pi_{T,S}(z \mid v; \theta_0) w_T(z, v)\mathbb{T}_{T,S}^2(v; \theta_0) dz dv$$

$$+ \left[\iint |\nabla_\theta \pi_{T,S}(z \mid v; \theta_0)\mathbb{T}_{T,S}(v; \theta_0)|_2 w_T(z, v) dz dv\right] \sqrt{T}(\theta_{T,S} - \theta_0) = o_p(1). \tag{18}$$

Next, add and subtract the expectation $E(\pi_{2T}(z, v)) = E(\pi_{2T}^i(z, v; \theta_0))$ in the numerators of $\pi_{T,S}(z \mid v; \theta_0)$ and $\pi_T(z \mid v)$. Then, by lengthy but straightforward computations, the expansion in (18) is, asymptotically,

$$-\mathcal{J}_{T,S}\sqrt{T}(\theta_{T,S} - \theta_0) = \frac{1}{S}\sum_{i=1}^{S}(\mathcal{I}_{1T}^i + \mathcal{I}_{2T}^i) - (\mathcal{I}_{1T}^0 + \mathcal{I}_{2T}^0) + o_p(1), \tag{19}$$

where

$$\mathcal{J}_{T,S} \equiv \iint |\nabla_\theta \pi_{T,S}(z \mid v; \theta_0)\mathbb{T}_{T,S}(v; \theta_0)|_2 w_T(z, v) dz dv \tag{20}$$

8. See Appendix A, Remarks on the proof of Lemma C1, for the basic computations needed to establish these uniform convergence results. Note that Bierens (1983, Section 5) originally developed the condition that $\sqrt{T}\lambda_T^q \to \infty$ in the context of uniform consistency for kernel estimators of density and regressor functions. Our proofs use and extend more general results developed by Andrews (1995, Theorem 1), and address uniform consistency for kernel estimators of gradient functions as well as "denominator" problems.

$$\mathcal{I}^i_{1T} \equiv \iint \eta(z,\upsilon) dA^i_T(z,\upsilon) \quad \text{and} \quad \mathcal{I}^i_{2T} \equiv \int \gamma(\upsilon) dA^i_T(\upsilon), \ \ i = 0, 1, \ldots, S \tag{21}$$

$$\eta(z,\upsilon) = \frac{\nabla_\theta \pi(z \mid \upsilon; \theta_0) w(z,\upsilon)}{\pi_1(\upsilon;\theta_0)}, \quad \gamma(\upsilon) = \int \frac{\nabla_\theta \pi(z \mid \upsilon;\theta_0)\pi_2(z,\upsilon;\theta_0) w(z,\upsilon)}{\pi_1(\upsilon;\theta_0)^2} dz \tag{22}$$

and the integrands in (21) are defined as, $dA^i_T(z,\upsilon) \equiv \sqrt{T}[\pi^i_{2T}(z,\upsilon;\theta_0) - E(\pi^i_{2T}(z,\upsilon;\theta_0))]dzd\upsilon$ and $dA^i_T(\upsilon) \equiv \sqrt{T}[\pi^i_{1T}(\upsilon;\theta_0) - E(\pi^i_{1T}(\upsilon;\theta_0))]d\upsilon$, with $\pi^0_{2T}(z,\upsilon;\theta_0)$ denoting the joint density estimate $\pi_{2T}(z,\upsilon)$, and $\pi^0_{1T}(\upsilon;\theta_0)$ denoting the marginal density estimate $\pi_{1T}(\upsilon)$.

By equation (19), then, the CD-SNE is root-T asymptotically normal if (i) the probability limit of $\mathcal{J}_{T,S}$ equals some positive definite constant matrix, and (ii) $\mathcal{I}^i_{1T}$ and $\mathcal{I}^i_{2T}$ are asymptotically normal. Precisely, we have:

Theorem 1. *Let Assumptions* 1(a), 2–3, 4(a), 5–6, 8, *and* 9(a) *hold. Then, the CD-SNE is (weakly) consistent. Furthermore, let* $\Upsilon(z,\upsilon) \equiv \eta(z,\upsilon) + \gamma(\upsilon)$, *where* $\eta(z,\upsilon)$ *and* $\gamma(\upsilon)$ *are defined in* (22), *and let* $E[\|\Upsilon(z_t,\upsilon_t)\|^\vartheta]^{1/\vartheta} < \infty$, *for some* $\vartheta > 2$. *Suppose that the* $n \times n$ *matrix* $\mathcal{J} = \iint |\nabla_\theta \pi(z \mid \upsilon;\theta_0)|_2 w(z,\upsilon) dzd\upsilon$ *is invertible. Then, under the additional Assumptions* 1(b), 4(b), 7, *and* 9(b),

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

where $V = \text{var}[\Upsilon(z_t,\upsilon_t)] + \sum_{k=1}^\infty \{\text{cov}[\Upsilon(z_t,\upsilon_t), \Upsilon(z_{t+k},\upsilon_{t+k})] + \text{cov}[\Upsilon(z_{t+k},\upsilon_{t+k}), \Upsilon(z_t,\upsilon_t)]\}$.

To develop intuition about the asymptotic variance of the CD-SNE, note, first, that the matrix $\mathcal{J}$ is the probability limit of $\mathcal{J}_{T,S}$ in (20). To understand the expression for the matrix V, consider the terms $\mathcal{I}^0_{1T}$ and $\mathcal{I}^0_{2T}$ in (21), and approximate $A^0_T(z,\upsilon) = \sqrt{T}[F_{2T}(z,\upsilon) - E(F_{2T}(z,\upsilon))]$, where $F_{2T}(z,\upsilon) = \int_{-\infty}^z \int_{-\infty}^\upsilon \pi_{2T}(s',s) ds'ds$, with $\hat{A}^0_T(z,\upsilon) \equiv \sqrt{T}[\hat{F}_{2T}(x) - E(\hat{F}_{2T}(x))]$, where $\hat{F}_{2T}(x) = \frac{1}{T}\sum_{t=1+l}^T \mathbb{I}_{x_t \le x}$ is the empirical cumulative distribution function of $x_t = [z_t, \upsilon_t]$, and $\mathbb{I}$ is the indicator function. Similarly, let us approximate $A^0_T(\upsilon)$ with $\hat{A}^0_T(\upsilon) \equiv \sqrt{T}[\hat{F}_{1T}(\upsilon) - E(\hat{F}_{1T}(\upsilon))]$, where $\hat{F}_{1T}(\upsilon)$ is the empirical cumulative distribution function of υ_t. Using these approximations, we can replace the integrals in (21) with finite sums, obtaining:

$$\begin{aligned}\mathcal{I}^0_{1T} + \mathcal{I}^0_{2T} &\approx \frac{1}{\sqrt{T}} \sum_{t=1+l}^T [\eta(z_t,\upsilon_t) - E(\eta(z_t,\upsilon_t))] + \frac{1}{\sqrt{T}} \sum_{t=1+l}^T [\gamma(\upsilon_t) - E(\gamma(\upsilon_t))] \\ &\equiv \frac{1}{\sqrt{T}} \sum_{t=1+l}^T [\Upsilon(z_t,\upsilon_t) - E(\Upsilon(z_t,\upsilon_t))],\end{aligned} \tag{23}$$

where $\Upsilon(z,\upsilon)$ is as in Theorem 1. The same approximation can be made for the terms $(\mathcal{I}^i_{1T} + \mathcal{I}^i_{2T})$ arising from the S simulations in the asymptotic expansion (19). Therefore, given (19) and the fact that the S simulation-based terms $(\mathcal{I}^i_{1T} + \mathcal{I}^i_{2T})$ have the same distribution as $(\mathcal{I}^0_{1T} + \mathcal{I}^0_{2T})$, asymptotic normality and the variance terms in Theorem 1 follow, heuristically, by the

independence of the simulations, by applying the central limit theorem to the R.H.S. of (23), and by $\mathcal{J}_{T,S} \xrightarrow{P} \mathcal{J}$.[9]

Finally, as for the indirect inference estimators, Theorem 1 requires that ST go to infinity in such a way that for a fixed number of simulations S, the size T of *every* simulated sample goes to infinity. Hence, the variance of the CD-SNE depends on the scaling term $(1 + S^{-1})$, as in the familiar asymptotics of the indirect inference estimators (*e.g.* Gouriéroux *et al.*, 1993).

3.3. *Efficiency*

By focusing on conditional densities, the CD-SNE provides a simple and appealing means to match the statistical properties of a dynamic model to those of the data, even in the presence of latent variables. The purpose of this section is to analyze when and how the CD-SNE can be asymptotically as efficient as the MLE.

It is well known that in the context of independent observations, certain minimum disparity estimators retain efficiency properties. For example, all the estimators encompassed by the Cressie and Read (1984) divergence measures mentioned in the Introduction are first-order efficient, although they may differ in terms of second-order efficiency, and robustness (see, for example, Beran, 1977; Lindsay, 1994). In fact, there is an interesting connection between the estimators minimizing the Cressie–Read divergence measures and the CD-SNE in (12). Consider the limiting criterion L^{CD} in (15), with: (i) the probability limit $\pi^*(z \mid \upsilon; \theta, \bar{\lambda}) = \pi(z \mid \upsilon; \theta)$, (ii) the following limiting weighting function,

$$w(z, \upsilon) = \frac{\pi_1(\upsilon; \theta_0)^2}{\pi_2(z, \upsilon; \theta_0)}, \tag{24}$$

and, finally, (iii) y_t observable, or $(z_t, \upsilon_t) = (y_t, y_{t-1})$. Then, by a straight forward computation, the limiting criterion of the CD-SNE can be written as:

$$E\left[\frac{\pi(y_t|y_{t-1}; \theta)}{\pi(y_t|y_{t-1}; \theta_0)} - 1\right]^2. \tag{25}$$

This criterion is, asymptotically, the *conditional density* counterpart to the Neyman's χ^2 measure of distance for marginal densities, which is a special case of the Cressie–Read divergence measures.[10]

Armed with this intuition about Neyman's χ^2, we now develop additional heuristic details about the asymptotic properties of the CD-SNE, when the asymptotic criterion is as in (25). Consider the expansion in (19). We are looking for a weighting function $w_T(z, \upsilon)$ such that: (i) asymptotically, the term $T^{-\frac{1}{2}}(\mathcal{I}^i_{1T} + \mathcal{I}^i_{2T})$ behaves as the score, and (ii) the matrix $\mathcal{J}_{T,S}$ converges in probability to the Fisher's information matrix. Let us consider, then, the limiting weighting function (24). By replacing (24) into the definition of $\eta(z, \upsilon)$ and $\gamma(\upsilon)$ in (22), we obtain

$$\eta(z, \upsilon) = \nabla_\theta \log \pi(z \mid \upsilon; \theta_0), \quad \text{and} \quad \gamma(\upsilon) = \int \nabla_\theta \pi(z \mid \upsilon; \theta_0) dz = 0.$$

9. In principle, asymptotic normality should also obtain without the first part of Assumption 9(b), although the asymptotic variance should then depend on the limiting bandwidth value, $\bar{\lambda}$. In our context, the first part of Assumption 9(b) is needed to address the efficiency issues we deal with in Section 3.3.

10. Indeed, by replacing the distance and weighting functions for marginal densities $D(\pi(y; \theta), \pi_T(y)) = [(\pi(y; \theta)/\pi_T(y))^2 - 1]$ and $w_T(y) = \pi_T(y)$ into the criterion $M_T(\theta)$ in (1), we obtain the Neyman's χ^2 for *marginal* densities, which, as noted in footnote 2, is a special case of the Cressie–Read divergence measures.

Therefore, equation (19) simplifies to,

$$-\mathcal{J}_{T,S}\sqrt{T}(\theta_{T,S}-\theta_0)=\frac{1}{S}\sum_{i=1}^{S}(\mathcal{I}_{1T}^{i}-\mathcal{I}_{1T}^{0})+o_p(1), \tag{26}$$

where, by the same heuristic arguments leading to equation (23),

$$\mathcal{I}_{1T}^{i}\approx\frac{1}{\sqrt{T}}\sum_{t=1+l}^{T}\nabla_\theta\log\pi(z_t^i(\theta_0)|v_t^i(\theta_0);\theta_0),\quad i=0,1,\ldots,S,$$

and where $[z_t^0(\theta_0)\ v_t^0(\theta_0)]\equiv[z_t\ v_t]$, the sample data.

Next, replace the limiting weighting function (24) into the matrix $\mathcal{J}$ in Theorem 1, obtaining $\mathcal{J}=\mathcal{J}_*\equiv E[|\nabla_\theta\log\pi(z_t\mid v_t;\theta_0)|_2]$. Finally, suppose that all the components of y_t are observable, and let $z_t=y_t$ and $v_t=y_{t-1}$, such that $\pi(z_t|v_t;\theta_0)=\pi(y_t|y_{t-1};\theta_0)$. Since the model (7) is first-order Markov, then, by a standard argument made in the appendices, $\nabla_\theta\log\pi(y_t|y_{t-1};\theta_0)$ is a martingale difference. In this case, (i) the terms $T^{-\frac{1}{2}}\mathcal{I}_{1T}^{i}$ are asymptotically equivalent to the score function, and (ii) $\mathcal{J}_*$ collapses to the Fisher's information matrix. Therefore, the variance of the CD-SNE (rescaled by $(1+S^{-1})$) attains asymptotically the Cramer–Rao lower bound, $E[|\nabla_\theta\log\pi(y'|y;\theta_0)|_2]^{-1}$.

The previous arguments are obviously heuristic. One critical issue is that the limiting weighting function (24) can be unbounded at the tails of the joint density $\pi_2(z,v;\theta_0)$ and, hence, does not satisfy Assumption 4. To implement the CD-SNE in this case, we use a trimming procedure similar to that which we used to cope with the denominator problems discussed earlier. Consider the following weighting function:

$$w_T(z,v)=\frac{\pi_{1T}(v)^2}{\pi_{2T}(z,v)}\mathbb{T}_{2T}(z,v),\quad \mathbb{T}_{2T}(z,v)\equiv G_{\alpha_T}(\pi_{2T}(z,v)), \tag{27}$$

where $G_{\alpha_T}(\ell)\equiv\int_0^\ell g_{\alpha_T}(u)du$, and the function g_{α_T} is as in Assumption 8, for some sequence $\alpha_T:\alpha_T\to 0$. Similarly as for the trimming function $\mathbb{T}_{T,S}(v;\theta)$ in Assumption 8, the trimming function $\mathbb{T}_{2T}(z,v)$ converges pointwise in probability to one and satisfies $\mathbb{T}_{2T}(z,v)=0$ on $\{(z,v)\in\mathbb{R}^{q^*}\times\mathbb{R}^{q-q^*}:\pi_{2T}(z,v)<\alpha_T\}$. Thus, it trims small values of the denominator $\pi_{2T}(z,v)$ and ensures that the weighting function in (27) is bounded. Under regularity conditions, (27) converges uniformly to (24). Assumption 10 below collects the regularity conditions on the asymptotic behaviour of the bandwidth sequence λ_T, the sequence δ_T in the trimming function $\mathbb{T}_{T,S}(v;\theta)$ in Assumption 8, and the sequence α_T in the trimming function $\mathbb{T}_{2T}(z,v)$.

Assumption 10. *As $T\to\infty$, $\alpha_T\to 0$, $\delta_T\to 0$, and $\delta_T/\alpha_T\to\kappa$, where κ is a constant. Moreover*

(a) *$\lambda_T\to\bar{\lambda}$, where $0\le\bar{\lambda}<\infty$, and $T^{\frac{1}{2}}\lambda_T^q\alpha_T^4\to\infty$.*
(b) *$\lambda_T\to 0$, $T^{\frac{1}{2}}\lambda_T^{q+1}\alpha_T^4\to\infty$, and $\alpha_T^{-1}\lambda_T^{\psi}\to 0$, where $\psi\equiv\min\{q^*+1,\frac{1}{5}r\}$.*

The condition that α_T and δ_T go to 0 at the same rate can be relaxed, at the cost of making the presentation more cumbersome (see Appendix B, Remarks on Assumption 10(a) and 10(b)). Assumption 10(a) ensures that the objective function of the CD-SNE with weighting function (27) converges uniformly in probability to a well-defined limit and is needed to show consistency of the CD-SNE. Assumption 10(b), instead, contains regularity conditions needed to prove asymptotic normality. The intuition about the "square-root-T" conditions in both parts of Assumption 10 is the same as that provided in Section 3.1, and relates to the need to obtain uniform convergence results for kernel density and score estimates.

We have:

Theorem 2 (Cramer–Rao lower bound). *Let $w_T(z,v)$ be as in* (27)*, and let Assumptions* 1(a)*,* 2–3*,* 5*,* 6*,* 8*,* 9(a)*, and* 10(a) *hold; then, the CD-SNE is* (*weakly*) *consistent. Moreover, suppose that $E[\|\nabla_\theta \log \pi(z_t \mid v_t;\theta_0)\|^\vartheta]^{1/\vartheta} < \infty$, for some $\vartheta > 2$, and that the $n \times n$ matrix $\mathcal{J}_* = \int\int |\nabla_\theta \log \pi(z \mid v;\theta_0)|_2 \pi_2(z,v;\theta_0) dz dv$ is invertible. Then, under the additional Assumptions* 1(b)*,* 7*,* 9(b)*, and* 10(b)*,*

$$\sqrt{T}(\theta_{T,S} - \theta_0) \xrightarrow{d} N\left(0, \left(1 + \frac{1}{S}\right)\mathcal{J}_*^{-1} V_* \mathcal{J}_*^{\top -1}\right),$$

where $V_* = \mathrm{var}[\nabla_\theta \log \pi(z_t \mid v_t;\theta_0)] + \sum_{k=1}^{\infty}\{\mathrm{cov}[\nabla_\theta \log \pi(z_t \mid v_t;\theta_0), \nabla_\theta \log \pi(z_{t+k} \mid v_{t+k};\theta_0)] + \mathrm{cov}[\nabla_\theta \log \pi(z_{t+k} \mid v_{t+k};\theta_0), \nabla_\theta \log \pi(z_t \mid v_t;\theta_0)]\}$. Finally, suppose that the state is fully observable, and let the CD-SNE match one-step ahead conditional densities, that is, $(z_t, v_t) \equiv (y_t, y_{t-1})$. Then, the CD-SNE attains the Cramer–Rao lower bound as $S \to \infty$.

In words, our CD-SNE is asymptotically as efficient as the MLE, when the number of simulations is large and the weighting function is as in (27). In this case, the criterion the CD-SNE minimizes, asymptotically, is (25), which is the conditional counterpart to the Neyman χ^2.

3.4. *Joint density SNE*

The CD-SNE hinges on matching conditional density estimates. In this subsection, we present an alternative estimator obtained by matching *joint* density estimates. This estimator is inspired by Aït-Sahalia's (1996) estimator in equation (3). Its distinctive feature is that the analytical expression for the joint density of data is replaced with an average of the joint densities computed from simulated data, as formalized by the following definition.

Definition 2. (J-SNE) For each fixed S, the Joint Density SNE (J-SNE) is the sequence $\{\theta_{T,S}\}_T$ given by

$$\begin{aligned}\theta_{T,S}^{\mathrm{J}} &= \arg\min_{\theta\in\Theta} L_{T,S}^{\mathrm{J}}(\theta) \\ &\equiv \arg\min_{\theta\in\Theta} \int [\pi_{2T,S}(x;\theta) - \pi_{2T}(x)]^2 w_T(x) dx, \end{aligned} \tag{28}$$

where $\pi_{2T,S}(x;\theta) \equiv S^{-1}\sum_{i=1}^{S} \pi_{2T}^i(x;\theta)$ and $\{w_T(x)\}_T$ is a sequence of positive weighting functions satisfying Assumption 4.

Let us define $L^{\mathrm{J}}(\theta,\bar\lambda) \equiv \int [\pi_2^*(x;\theta,\bar\lambda) - \pi_2^*(x;\theta_0,\bar\lambda)]^2 w(x) dx$, where $\pi_2^*(x;\theta,\bar\lambda)$ is as in (14) and, as usual, $\bar\lambda$ denotes the limiting bandwidth, that is, $\bar\lambda : \lambda_T \to \bar\lambda$. To prove consistency of the J-SNE, we need two sets of conditions paralleling those in Assumptions 5 and 6:

Assumption 11. *For all $\theta \in \Theta$, $L_{T,S}^J(\theta)$ is measurable and continuous on Θ. Moreover, $L^J(\theta,\bar\lambda)$ is continuous on Θ, and there exists a unique θ_0 in the interior of Θ such that $L^J(\theta,\bar\lambda) = 0$ implies that $\theta = \theta_0$.*

Assumption 12. *There exists an $\alpha > 0$ and a sequence $\{\kappa_{T,S}\}_T$ bounded in probability as T becomes large such that for all $(\varphi,\theta) \in \Theta \times \Theta$ and some $\alpha > 0$, $|L_{T,S}^J(\varphi) - L_{T,S}^J(\theta)| \le \kappa_{T,S}\|\varphi - \theta\|^\alpha$.*

We have:

Theorem 3. *Let Assumptions* 1(a), 2–3, 4(a), *and* 11–12 *hold; then, the J-SNE is* (*weakly*) *consistent. Furthermore, let* $\Psi(x) \equiv \nabla_\theta \pi_2(x;\theta_0)w(x)$, *suppose that* $E[\|\Psi(x_t)\|^\vartheta]^{1/\vartheta} < \infty$, *for some* $\vartheta > 2$, *and that the* $n \times n$ *matrix* $\mathcal{D} = \int |\nabla_\theta \pi_2(x;\theta_0)|_2 w(x)dx$ *is invertible. Then, under the additional Assumptions* 1(b), 4(b), *and* 7, *and the conditions that* $\lambda_T \to 0$ *and* $T^{\frac{1}{2}}\lambda_T^{q+1} \to \infty$ *as* $T \to \infty$,

$$\sqrt{T}(\theta^{\mathrm{J}}_{T,S} - \theta_0) \xrightarrow{d} N\left(0, \left(1 + \frac{1}{S}\right)\mathcal{D}^{-1}W\mathcal{D}^{\top -1}\right),$$

where $W \equiv \mathrm{var}[\Psi(x_t)] + \sum_{k=1}^{\infty}\{\mathrm{cov}[\Psi(x_t), \Psi(x_{t+k})] + \mathrm{cov}[\Psi(x_{t+k}), \Psi(x_t)]\}$.

A few remarks are warranted. First, the J-SNE allows one to estimate multivariate models driven by partially observed variables with unknown distribution by matching *joint* densities of observed data.

Second, and up to identifiability, bandwidth choice does not affect consistency of the J-SNE. This property parallels a similar property of the CD-SNE. It originates from the "twin-smoothing" device, by which we now smooth joint densities (on simulated data and on sample data) with the same kernel and bandwidth.

Third, the J-SNE is based on matching joint densities and, hence, does not lead to the denominator problems discussed in Sections 2.2 and 3. Thus, the J-SNE is not affected by any of the trimming issues underlying Assumptions 9 and 10. Note, also, that the order of the kernel plays no role within the asymptotic theory for the J-SNE. In the case of the CD-SNE, the order of the kernel plays a role, notably through Assumptions 9(b) and 10(b). Moreover, in the context of partially observed systems, there is no clear ranking between the asymptotic variances of Theorems 1–3. For all these reasons, the J-SNE is a reasonable alternative to the CD-SNE in the presence of partially observed systems.

Finally, the (unscaled) variance $\mathcal{D}^{-1}W\mathcal{D}^{\top -1}$ of Theorem 3 collapses to the variance of Aït-Sahalia's (1996) estimator in the scalar case and when $w_T(x) = \pi_{2T}(x)$. The J-SNE, however, is different from Aït-Sahalia's, as it relies on a "twin-smoothing" device, as we explained earlier. Finally, to show asymptotic normality, we need the square-root condition ($T^{\frac{1}{2}}\lambda_T^{q+1} \to \infty$), to ensure uniform convergence of the simulation-based kernel estimate $\nabla_\theta \pi_{2T,S}(x;\theta_0)$ to $\nabla_\theta \pi_2(x;\theta_0)$ in the function $\Psi(x)$ of Theorem 3.

4. MONTE CARLO EXPERIMENTS

In this section we perform Monte Carlo experiments to investigate the finite sample properties of our estimators. We wish to address four points. First, we wish to ascertain whether the finite sample properties of our estimators are accurately approximated by the asymptotic theory. Second, we study how our simulated non-parametric estimators compare with alternative estimators such as the Fermanian and Salanié (2004) NPSML estimator, and even the MLE. Third, we examine how the CD-SNE and the J-SNE compare with each other. Fourth, we investigate how bandwidth choice and the possible curse of dimensionality impart on our estimators' finite sample performance.

To address these points, we consider four distinct models: two discrete-time stochastic volatility models (one univariate and one bivariate) (in Section 4.1), and two continuous-time models commonly utilized in finance (namely, the standard Vasicek model and one extension of the Vasicek model with stochastic volatility) (in Section 4.2).

Our experiments on all these models share some common features. First, non-parametric density estimates are implemented through Gaussian kernels. Second, our bandwidth choice closely follows the suggestions made by Chen, Linton and Robinson (2001) in the context of conditional density estimation with dependent data; precisely, for each Monte Carlo replication, we select the bandwidth by searching over values minimizing the asymptotic mean integrated squared error of the conditional density estimated on *sample* data. Third, we trim 2% of the observations. Fourth, we set the number of path simulations equal to 5 in all experiments (*i.e.* $S = 5$). Fifth, in cases in which our estimators can not be efficient, asymptotic standard deviations are approximated through Newey–West windows of ± 12. Finally, we run 1000 Monte Carlo replications in each experiment.[11]

4.1. *Discrete-time models*

Discrete-time stochastic volatility models are very often utilized in financial applications. In this section, we gauge the finite sample performance of our simulated non-parametric estimators applied to the following stochastic volatility model,

$$\begin{cases} y_t = \sigma_b \exp(y_t^*/2)\varepsilon_{1t} \\ y_t^* = \phi y_{t-1}^* + \sigma_e \varepsilon_{2t} \end{cases} \tag{29}$$

where y_t is the observable variable, y_t^* is the latent volatility process, ε_{1t} and ε_{2t} are two innovations independent and identically distributed as standard normal and, finally, ϕ, σ_b and σ_e are the parameters of interest. The interpretation of the observable variable y_t is that of the unpredictable part of some asset return. The important reason we focus on this model is that it has become a workhorse in previous Monte Carlo studies—for example, Fermanian and Salanié (2004) tested their NPSML estimator on this model.

We consider two estimators. The first estimator is the CD-SNE in (12), which we implement by matching the model's conditional density to the conditional density $\pi_T(y_t \mid y_{t-1})$ of two adjacent observations, through the weighting function $w_T(y_t, y_{t-1}) = \frac{\pi_{1T}(y_{t-1})^2}{\pi_{2T}(y_t, y_{t-1})}$ in (27). By Theorem 2, the resulting estimator is not as efficient as the MLE, as the observable variable y_t is not first-order Markov—only the joint process (y_t, y_t^*) is first-order Markov. The second estimator we consider is the J-SNE in (28), implemented by matching the joint density of two adjacent observations, $\pi_{2T}(r_t, r_{t-1})$, and using the weighting function $w_T(y_t, y_{t-1}) = \pi_{2T}(y_t, y_{t-1})$.

The parametrization of the discrete-time model (29) is $\phi = 0{\cdot}95$, $\sigma_b = 0{\cdot}025$ and $\sigma_e = 0{\cdot}260$. We consider a sample size of 500 observations. Table 1 reports the results of our Monte Carlo experiments.[12] We report the mean, median, and sample standard deviation of the estimates over the Monte Carlo replications. As regards the CD-SNE and the J-SNE, Table 1 also reports: (i) asymptotic standard deviations, obtained through the relevant theory developed in Section 3, and (ii) coverage rates for 90% confidence intervals, computed through the usual asymptotic approximation to the distribution of the estimator—that is, the estimate plus or minus 1·645 times the asymptotic standard deviation. Finally, Table 1 reports the finite sample properties of three alternative estimation methods available in the literature, and summarized by Fermanian and Salanié (2004, table 4).

11. In the most demanding applications (diffusion processes and sample sizes of 1000 observations), computation time on a Pentium 4 with 1·7GHz is between 3 and 6 minutes. In the Monte Carlo experiments of this section, our estimators are implemented with Fortran-90. The objective functions are optimized through a Quasi-Newton algorithm, with a convergence criterion of the order of 10^{-5}.

12. Initial values of the parameters are drawn from a uniform distribution on $[0{\cdot}15, 0{\cdot}35]$ (for σ_e); on $[0{\cdot}9, 0{\cdot}99]$ (for ϕ); and on $[0{\cdot}015, 0{\cdot}035]$ (for σ_b). The correlation (over the Monte Carlo replications) between initial values and final estimates are 0·11 (for the CD-SNE) and 0·09 (for the J-SNE) on average over the parameters.

TABLE 1

Monte Carlo experiments (univariate discrete-time stochastic volatility model (29))

Estimator		ϕ	σ_b	σ_e
CD-SNE	Mean	0·909	0·024	0·229
	Median	0·939	0·023	0·210
	Sample S.D.	0·102	0·003	0·131
	Asymptotic S.D.	0·115	0·004	0·089
	Coverage rate 90% confidence interval	0·92	0·93	0·74
J-SNE	Mean	0·942	0·027	0·297
	Median	0·960	0·026	0·274
	Sample S.D.	0·095	0·005	0·144
	Asymptotic S.D.	0·121	0·005	0·093
	Coverage rate 90% confidence interval	0·94	0·89	0·72
QML	Mean	0·906	–	0·302
	Sample S.D.	0·18	–	0·17
MCL	Mean	0·930	–	0·233
	Sample S.D.	0·10	–	0·07
NPSML	Mean	0·913	0·022	0·318
	Sample S.D.	0·10	0·003	0·17

Notes: True parameter values are: $\phi = 0{\cdot}95$, $\sigma_b = 0{\cdot}025$ and $\sigma_e = 0{\cdot}260$. Sample size: $T = 500$. QML, quasi maximum likelihood; MCL, Monte Carlo maximum likelihood; NPSML, non-parametric simulated maximum likelihood.

The results in Table 1 reveal that the CD-SNE and the J-SNE exhibit a proper finite sample behaviour, also in comparison with alternative estimation methods. In particular, the sample variability of the estimates of ϕ and σ_b obtained with our methods is in line with its asymptotic counterpart. As it turns out, it is relatively more difficult to estimate the volatility parameter σ_e of the latent process y_t^*, which results in a sample standard deviation larger than its asymptotic counterpart for both the CD-SNE and the J-SNE. Finally, note that since y_t in (29) is not first-order Markov, we do not expect, and do not find, a clear ranking between the two estimators, in terms of the precision of the estimates.

Next, we explore how our methods are affected by the dimensionality of non-parametric density estimates. We consider a simple model in which *two* (unpredictable parts of) asset returns exhibit stochastic volatility. To isolate the effects of the curse of dimensionality and keep the Monte Carlo design as simple as possible, we make the simplifying assumption that the two asset return volatilities are driven by a common volatility factor,

$$\begin{cases} y_{1t} = \sigma_{b1} \exp(y_t^*/2)\varepsilon_{1t} \\ y_{2t} = \sigma_{b2} \exp(y_t^*/2)\varepsilon_{2t} \\ y_t^* = \phi y_{t-1}^* + \sigma_e \varepsilon_{3t} \end{cases} \tag{30}$$

where y_{it} $(i = 1, 2)$ are the observable variables, y_t^* is the latent volatility process, ε_{1t}, ε_{2t} and ε_{3t} are three innovations independent and identically distributed as standard normal and, finally, σ_{bi} $(i = 1, 2)$, ϕ and σ_e are the parameters of interest.

As in the previous experiments, we consider sample sizes of 500 observations, and parametrize model (30) as follows: $\phi = 0{\cdot}95$, $\sigma_{b1} = \sigma_{b2} = 0{\cdot}025$ and $\sigma_e = 0{\cdot}260$. We examine the finite sample properties of both the CD-SNE and the J-SNE. The CD-SNE is implemented by matching the conditional density of two adjacent pairs of observations, $\pi_T(y_{1t}, y_{2t} \mid y_{1t-1}, y_{2t-1})$, and using the weighting function (27). The J-SNE is implemented by matching the joint density of

TABLE 2

Monte Carlo experiments (bivariate discrete-time stochastic volatility model (30)).

Estimator		ϕ	σ_{b1}	σ_{b2}	σ_e
CD-SNE	Mean	0·916	0·025	0·026	0·289
	Median	0·919	0·026	0·027	0·287
	Sample S.D.	0·072	0·004	0·004	0·101
	Asymptotic S.D.	0·080	0·004	0·004	0·113
	Coverage rate 90% confidence interval	0·92	0·83	0·88	0·91
J-SNE	Mean	0·913	0·027	0·027	0·365
	Median	0·938	0·026	0·027	0·331
	Sample S.D.	0·084	0·004	0·004	0·164
	Asymptotic S.D.	0·085	0·005	0·005	0·154
	Coverage rate 90% confidence interval	0·88	0·92	0·93	0·88

Notes: True parameter values are $\phi = 0{\cdot}95$, $\sigma_{b1} = 0{\cdot}025$, $\sigma_{b2} = 0{\cdot}025$, and $\sigma_e = 0{\cdot}260$. Sample size: $T = 500$.

two adjacent pairs of observations, $\pi_{2T}(y_{1t}, y_{2t}, y_{1t-1}, y_{2t-1})$, and using the weighting function $\pi_{2T}(y_{1t}, y_{2t}, y_{1t-1}, y_{2t-1})$. The results are displayed in Table 2.[13]

The increase in dimensionality may produce two effects on the estimates. On the one hand, the observation of two asset returns may facilitate our understanding of the dynamic properties of the common unobserved volatility process. On the other hand, the larger dimension of the non-parametric density estimates may impinge upon the precision of the estimates. The results in Table 2 suggest that these effects arise in our experiments. Overall, an increase in dimensionality does not seem to have jeopardized the performance of our estimators in this experiment.

4.2. *Continuous-time models*

All available simulation-based techniques (and the methods developed in this article) rest on the obvious assumption that the model of interest can be simulated. However, continuous-time models can not even be simulated, except in the trivial case in which the transition density is known.[14] Rather, continuous-time models can only be *imperfectly* simulated by means of some discretization device. To deal with this complication goes well beyond the purpose of this illustrative section. In the unpublished appendix to this paper, we derive conditions under which our theory works once the discretization shrinks to 0 at some rate (see Altissimo and Mele, 2008, section E). Here, we provide the essential guidelines to estimate continuous time models with the SNE. Consider the following data generating process,

$$dy(\tau) = b(y(\tau), \theta_0)d\tau + a(y(\tau), \theta_0)dW(\tau),\ \tau \geq 0, \tag{31}$$

where $W(\tau)$ is a standard d-dimensional Brownian motion, b and a are vector and matrix valued functions in $\mathbb{R}^d$ and $\mathbb{R}^{d\times d}$, a is full rank, $y(\tau) \in \mathbb{R}^d$ and, finally, $\theta_0 \in \Theta$, where Θ is compact. Similarly as in Section 2, we partition $y(\tau)$ as $y(\tau) \equiv [y^{\mathrm{o}}(\tau)\ \ y^{\mathrm{u}}(\tau)]$, where $y^{\mathrm{o}}(\tau) \in \mathbb{R}^{q^*}$ is the vector of the observable variables. We assume that the data are sampled at regular intervals; accordingly, we still let $q \equiv q^*(1+l)$ and $x_t \equiv (y_t^{\mathrm{o}}, \cdots, y_{t-l}^{\mathrm{o}})$ $(t = 1+l, \cdots, T)$, where y_t^{o} are

13. Initial values of the parameters are drawn as in the previous footnote. Correlations between initial guesses and final estimates are also of the same order of magnitude as in the previous footnote.

14. To date, estimation methods specifically designed to deal with diffusion processes include moments generating techniques (*e.g.* Hansen and Scheinkman, 1995; Singleton, 2001), approximations to maximum likelihood (*e.g.* Pedersen, 1995; Santa-Clara, 1995; Aït-Sahalia, 2002, 2003) and, on a radically different perspective, Markov Chain Monte Carlo approaches (*e.g.* Elerian, Chib and Shephard, 2001).

the discretely sampled data. Finally, we assume that (31) is strictly stationary and that $\{y_t^o\}_{t=1}^T$ satisfy the same regularity conditions in Section 3.

To generate simulated paths of the observable variables in (31), various discretization schemes can be used (see, for example, Kloeden and Platen, 1999). In this paper, we consider the simple Euler–Maruyama discrete time approximation to (31),

$$ {}_h y_{h(k+1)} - {}_h y_{hk} = b({}_h y_{hk}, \theta)h + a({}_h y_{hk}, \theta)\sqrt{h}\varepsilon_{k+1}, \quad k = 0, 1, \ldots, \tag{32}$$

where h is the discretization step and ε_k is a sequence of independent and identically distributed $\mathbb{R}^d$-valued random variables. Let $x_{t,h}^i(\theta)$ be the i-th simulation of the t-th observation when the parameter vector is θ, and the discretization step is h. We compute joint and conditional density estimates from the simulated data $x_{t,h}^i(\theta)$ as we described in Section 2. The SNE now makes density estimates computed from simulated data as close as possible to those computed from the discretely sampled diffusion, according to the measures of distance in Definitions 1 and 2. In Altissimo and Mele (2008, theorem E.1), we provide a rate condition on h (namely that as $T \to \infty$, $h \downarrow 0$ in such a way that $\sqrt{T}h \to 0$) and additional regularity conditions under which our SNEs behave as in Theorems 1–3.

As for the experiments, we simulate the models through the Euler–Maruyama scheme in (32), taking ε_k to be normally distributed, using a stepsize $h = 1/(5 \times 52)$, and sampling the simulated data at a weekly frequency. We start by considering the celebrated Vasicek model of the short-term interest rate,

$$di(\tau) = (b_1 - b_2 i(\tau))d\tau + a_1 dW(\tau), \quad \tau \geq 0, \tag{33}$$

where $W(\tau)$ is a scalar Brownian motion, and b_1, b_2 and a_1 are the parameters of interest. This model is a useful benchmark because it is the continuous-time counterpart of a discrete-time AR(1) model, and it can be easily estimated by maximum likelihood. The parametrization we choose for this model is $b_1 = 3{\cdot}00$, $b_2 = 0{\cdot}50$, and $a_1 = 3{\cdot}00$. These parameter values imply that the model-implied mean, variance and autocorrelations are roughly the same as the 3-month U.S. interest rate in post-war data.

Let i_t be the discretely sampled data. We consider four estimators. First, we implement the CD-SNE by matching the model's conditional density to the conditional density $\pi_T(i_t \mid i_{t-1})$ of any two adjacent observations, and using $\frac{\pi_{1T}(i_{t-1})^2}{\pi_{2T}(i_t, i_{t-1})}$ as a weighting function. By Theorem 2, this estimator is asymptotically as efficient as the MLE. Second, we implement the J-SNE by matching the joint density $\pi_{2T}(i_t, i_{t-1})$ of two adjacent observations, and using $\pi_{2T}(i_t, i_{t-1})$ as a weighting function. Third, we consider an estimator we label Analytical-NE. The Analytical-NE is a modification of the J-SNE in that the simulated non-parametric estimate $\pi_{2T,S}(i_t, i_{t-1}; \theta)$ in (28) is replaced with its analytical counterpart $\pi_2^{\text{vas}}(i_t, i_{t-1}; \theta)$.[15] Thus, the objective function of the Analytical-NE is:

$$\int [\pi_2^{\text{vas}}(i_t, i_{t-1}; \theta) - \pi_{2T}(i_t, i_{t-1})]^2 \pi_{2T}(i_t, i_{t-1}) di_t di_{t-1}. \tag{34}$$

Naturally, the Analytical-NE is unfeasible in most models of interest. We consider this estimator because it allows us to gauge the practical importance of the "twin-smoothing" device underlying the SNE. The fourth estimator we consider is the MLE.

15. As is well known, the transition density $\pi^{\text{vas}}(i_s \mid i_t; \theta)$ from date t to date s ($s > t$) is Gaussian with expectation equal to $b_1/b_2 + [i_t - (b_1/b_2)]\exp(-b_2(s-t))$ and variance equal to $[a_1^2/(2b_2)][1 - \exp(-2b_2(s-t))]$. The marginal density is obtained by letting $s \to \infty$.

TABLE 3

Monte Carlo experiments (Vasicek model (33))

Sample	Estimators		b_1	b_2	a_1
$T = 1000$	CD-SNE	Mean	2·87	0·49	3·08
		Median	2·89	0·47	3·10
		Sample S.D.	0·97	0·17	0·29
		Asymptotic S.D.	1·10	0·19	0·23
		Coverage rate 90% confidence interval	0·95	0·92	0·82
	CD-SNE – Double bandwidth	Mean	2·65	0·43	3·23
		Median	2·56	0·44	3·16
		Sample S.D.	0·84	0·17	0·28
	CD-SNE – Half bandwidth	Mean	2·98	0·54	2·97
		Median	2·93	0·56	3·04
		Sample S.D.	1·06	0·23	0·40
	J-SNE	Mean	3·20	0·55	2·89
		Median	3·07	0·52	2·76
		Sample S.D.	1·11	0·25	0·41
		Asymptotic S.D.	1·24	0·22	0·31
		Coverage rate 90% confidence interval	0·95	0·85	0·81
	Analytical-NE	Mean	3·47	0·57	3·55
		Median	3·20	0·47	3·46
		Sample S.D.	2·09	0·64	0·62
	MLE	Mean	3·74	0·62	3·01
		Median	3·93	0·63	2·99
		Sample S.D.	1·21	0·20	0·07
$T = 500$	CD-SNE	Mean	2·95	0·48	3·14
		Median	2·95	0·48	3·12
		Sample S.D.	1·03	0·24	0·42
		Asymptotic S.D.	1·36	0·26	0·32
		Coverage rate 90% confidence interval	0·94	0·94	0·83
	J-SNE	Mean	3·06	0·58	2·58
		Median	3·03	0·51	2·51
		Sample S.D.	1·41	0·35	0·71
		Asymptotic S.D.	1·65	0·31	0·57
		Coverage rate 90% confidence interval	0·97	0·84	0·76
	MLE	Mean	3·99	0·70	2·99
		Median	4·01	0·69	3·00
		Sample S.D.	1·36	0·27	0·10

Notes: True parameter values are $b_1 = 3{\cdot}00$, $b_2 = 0{\cdot}50$, and $a_1 = 3{\cdot}00$.

The performance of the four estimators is tested in samples of 1000 and 500 observations. We report the results in Table 3.[16] When the size of the simulated samples is 1000, the performances of the CD-SNE and MLE are comparable in terms of variability of the estimates. Specifically, the CD-SNE has a lower standard deviation than the MLE as regards the estimation of the parameter b_2 affecting the persistence of i_t. The MLE, however, is more precise than the CD-SNE as regards the estimation of the volatility parameter a_1. As it turns out, the sample standard deviation of the CD-SNE estimates of a_1 is larger than its asymptotic counterpart, and this is reflected in a coverage rate below the nominal one. As regards biases, the MLE tends to under-estimate the persistence of the data and largely over-estimate the constant b_1 in the drift term. This phenomenon does not emerge when the model is estimated with the CD-SNE.

16. Initial values of the parameters are drawn from a uniform distribution on $[1{\cdot}5, 4{\cdot}5]$ (for b_1 and a_1); and on $[0{\cdot}1, 0{\cdot}9]$ (for b_2). The correlations (over the Monte Carlo replications) between initial values and final estimates are 0·08 (for the CD-SNE) and 0·07 (for the J-SNE) on average over the parameters.

As expected, the results in Table 3 clearly reveal that moving from the CD-SNE to the J-SNE causes an increase in the variability of the estimates; this result is pronounced for the volatility parameter a_1. Furthermore, the Analytical-NE produces a much larger variability of the estimates. Finally, the Analytical-NE produces parameters estimates with large biases: it over-estimates the volatility coefficient a_1 by 0·55 and the constant b_1 in the drift term by 0·47. These results are perfectly consistent with our theoretical explanations of biases arising when the model density and the sample density are not smoothed with the same kernel.

As is well known, the practical performance of non-parametric methods hinges on the proper choice of the bandwidth parameter. Table 3 also shows the effects of bandwidth selection on the small samples performance of the CD-SNE. We have implemented two experiments. In the first one, the CD-SNE is implemented with a bandwidth level, which is twice the size suggested by Chen *et al.* (2001) (which we utilized earlier). In the second experiment, the bandwidth size is half the size we utilized earlier. The results in Table 3 suggest that while these bandwidth choices produce some effects on the estimates, those effects are marginal. In particular, we note that: (i) under-smoothing the data increases somehow the variability of the density estimates, which in turn leads to a higher standard deviation of the parameter estimates; and (ii) over-smoothing the data tends to increase the mean bias of the parameter estimates.

Finally, Table 3 documents the performance of the CD-SNE, the J-SNE and the MLE in shorter samples of 500 observations. As expected, the performance of all these methods worsens as regards the variability of the estimates. As regards mean biases for the parameters b_1 and b_2, we note that: (i) the mean bias of the MLE almost doubles with respect to the longer sample; and (ii) the mean biases of the CD-SNE remain small, compared to the corresponding mean biases of the MLE.

A simple extension of the model in (33) is one in which the instantaneous volatility of the short-term rate $i(\tau)$ is driven by an unobservable process $\sigma(\tau)$ with constant elasticity of variance,

$$\begin{cases} di(\tau) = (b_1 - b_2 i(\tau))d\tau + a_1\sigma(\tau)dW_1(\tau) \\ d\sigma(\tau) = b_3(1-\sigma(\tau))d\tau + a_2\sigma(\tau)dW_2(\tau) \end{cases} \tag{35}$$

where $W_1(\tau)$ and $W_2(\tau)$ are two independent Brownian motions, and b_3 and a_2 are additional parameters related to the volatility dynamics. Naturally, the presence of the unobservable volatility component in model (35) now makes the MLE an unfeasible estimation alternative.

To implement Monte Carlo experiments, we choose as parameter values $b_1 = 3{\cdot}00$, $b_2 = 0{\cdot}5$, $a_1 = 3{\cdot}00$, $b_3 = 1{\cdot}0$ and $a_2 = 0{\cdot}3$. These parameter values are consistent with the estimates of similar models on U.S. short-term interest rates data. We implement the CD-SNE by matching the model's conditional density to the conditional density $\pi_T(i_t \mid i_{t-1})$ of any two adjacent observations, and using the weighting function (27). We implement the J-SNE by matching the joint density $\pi_{2T}(i_t, i_{t-1})$ of two adjacent observations, and using $\pi_{2T}(i_t, i_{t-1})$ as a weighting function. The performance of both estimators is gauged in samples of 1000 and 500 observations. We report the results in Table 4.[17]

As regards the larger simple size case and the CD-SNE, the standard deviation and the bias associated with the parameters b_1 and b_2 of the observable variable $i(\tau)$ are of the same order of magnitude as in Table 3. The presence of the unobservable volatility component makes the estimate of a_1 more imprecise than the corresponding estimates in Table 3. As regards the bias terms, the CD-SNE has a tendency to over-estimate the parameter b_3; this phenomenon becomes more pronounced in the smaller sample size.

17. Initial values of the parameters are drawn from a uniform distribution on $[1{\cdot}5, 4{\cdot}5]$ (for b_1 and a_1); on $[0{\cdot}1, 0{\cdot}9]$ (for b_2); on $[0{\cdot}5, 1{\cdot}5]$ (for b_3); and on $[0{\cdot}1, 0{\cdot}5]$ (for a_2). The correlation (over the Monte Carlo replications) between initial values and final estimates are 0·11 (for the CD-SNE) and 0·12 (for the J-SNE) on average over the parameters.

TABLE 4

Monte Carlo experiments (continuous-time stochastic volatility model (35))

Sample	Estimator		b_1	b_2	a_1	b_3	a_2
$T = 1000$	CD-SNE	Mean	3·03	0·48	3·05	1·11	0·34
		Median	3·07	0·49	3·04	0·98	0·32
		Sample S.D.	0·93	0·22	0·40	0·59	0·20
		Asymptotic S.D.	1·17	0·22	0·32	0·45	0·16
		Coverage rate 90% confidence interval	0·95	0·88	0·83	0·82	0·83
	J-SNE	Mean	2·91	0·48	2·97	1·10	0·38
		Median	2·95	0·49	2·91	1·05	0·33
		Sample S.D.	1·15	0·22	0·50	0·52	0·20
		Asymptotic S.D.	1·20	0·23	0·31	0·50	0·18
		Coverage rate 90% confidence interval	0·91	0·91	0·78	0·84	0·88
$T = 500$	CD-SNE	Mean	2·94	0·49	3·12	1·30	0·34
		Median	2·99	0·49	3·07	1·11	0·31
		Sample S.D.	1·41	0·30	0·62	0·77	0·27
		Asymptotic S.D.	1·69	0·31	0·44	0·63	0·22
		Coverage rate 90% confidence interval	0·95	0·89	0·80	0·83	0·85
	J-SNE	Mean	2·96	0·46	2·92	1·29	0·33
		Median	3·01	0·47	2·87	1·12	0·29
		Sample S.D.	1·52	0·29	0·61	0·75	0·25
		Asymptotic S.D.	1·75	0·32	0·43	0·70	0·25
		Coverage rate 90% confidence interval	0·94	0·92	0·81	0·87	0·89

Notes: True parameter values are $b_1 = 3{\cdot}00$, $b_2 = 0{\cdot}50$, $a_1 = 3{\cdot}00$, $b_3 = 1{\cdot}00$, and $a_2 = 0{\cdot}30$.

In contrast with our previous results obtained with the Vasicek model (33), we do not observe a clear ranking of the properties of the CD-SNE and the J-SNE. This phenomenon is particularly clear when the two estimators are compared in terms of the variability of the estimates. Intuitively, the unobservable volatility process $\sigma(\tau)$ destroys the Markov property of the short-term interest rate $i(\tau)$ in (33). Precisely, the joint process $(i(\tau), \sigma(\tau))$ in (35) is Markov, but the "marginal" process $i(\tau)$ is not. Therefore, the conditions in Theorem 2 for asymptotic efficiency of the CD-SNE are not met. As a result, the CD-SNE does not necessarily outperform the J-SNE, which makes the J-SNE an interesting alternative to look at in practical applications such as those considered in this section.

5. CONCLUSION

This paper has introduced new methods to estimate the parameters of partially observed dynamic models. The building block of these methods is simple. It consists of simulating the model of interest for the purpose of recovering the corresponding density function. Our estimators are those that make conditional (or joint) densities on simulated data as close as possible to their empirical counterparts. We made use of classical ideas in the statistical literature to build up convenient measures of closeness for such densities. Our estimators are easy to implement and in the special case of observable Markov systems, they can attain the same asymptotic efficiency as the maximum likelihood estimator. Furthermore, Monte Carlo experiments reveal that their finite sample performance is very satisfactory, even in comparison to maximum likelihood.

Using simulations to recover the model-implied density is not only convenient because it allows to estimate densities unknown in closed-form. We demonstrated that such a "twin-smoothing" procedure makes our methods improve upon alternative techniques matching closed-form model-implied densities to data-implied densities. Consistent with the asymptotic theory, finite sample results suggest that a careful choice of both the measures of closeness for density

functions and the bandwidth functions can enhance the performance of the estimators, but mainly in terms of their precision. Furthermore, the "twin-smoothing" device makes the estimators accurate in terms of unbiasedness, even in cases of simple bandwidth selection procedures.

In the numerical experiments, we emphasized the kind of applications arising in financial economics. But we also demonstrated that our approach is quite general, and can be used to address related estimation problems. As an example, the typical Markov models arising in applied macroeconomics can also be estimated with our methods. Extensions of our methods can be made to allow estimation of models in which some of the endogenous variables are tied up by general equilibrium or no-arbitrage conditions; see, for example, Singleton's (2006) surveys on these estimation problems, and Pastorello, Patilea and Renault (2003) for a "latent backfitting" approach to the estimation of partially observed equilibrium models. In these cases, too, the previous asymptotic efficiency and encouraging finite sample properties would make our methods stand as a promising advance into the literature of simulation-based inference methods.

APPENDICES

An unpublished appendix (Altissimo and Mele, 2008; hereafter Al-M08) includes extensive details and all the proofs of the lemmas stated in these appendices.

To simplify the notation, we shall denote the probability limits in equations (14) and (15) of the main text as follows:

$$m_2(z,\upsilon;\theta) \equiv \pi_2^*(z,\upsilon;\theta,\bar{\lambda}),\quad m_1(z\mid\upsilon;\theta) \equiv \pi_1^*(\upsilon;\theta,\bar{\lambda}),\quad m(z\mid\upsilon;\theta) \equiv \pi^*(z\mid\upsilon;\theta,\bar{\lambda}),$$

and set, in equation (15), $L^{\mathrm{CD}}(\theta) \equiv L^{\mathrm{CD}}(\theta,\bar{\lambda})$.

APPENDIX A. PROOF OF THEOREM 1

A.1. *Consistency*

Let $L_{T,S}^{\mathrm{CD}}(\theta)$ be the criterion function in (12) in the main text. We have:

Proposition 1. *Let Assumptions* 1(a), 2, 3, 4(a), *and* 5 *hold. Then* $\forall\theta\in\Theta$, $L_{T,S}^{\mathrm{CD}}(\theta)\xrightarrow{p} L^{\mathrm{CD}}(\theta)$ *as* $T\to\infty$.

According to a well-known result (see Newey, 1991, thm. 2.1, p. 1162), the following conditions are equivalent:

(C1) $\lim_{T\to\infty} P(\sup_{\theta\in\Theta}|L_{T,S}^{\mathrm{CD}}(\theta)-L^{\mathrm{CD}}(\theta)|>\varepsilon)=0$, $\varepsilon>0$.

(C2) $\forall\theta\in\Theta$, $L_{T,S}^{\mathrm{CD}}(\theta)\xrightarrow{p} L^{\mathrm{CD}}(\theta)$, and $L_T^{\mathrm{CD}}(\theta)$ is stochastically equicontinuous.

By Newey and McFadden (1994, lemma 2.9, p. 2138), Assumption 6 guarantees that $L_{T,S}^{\mathrm{CD}}(\theta)$ is stochastically equicontinuous, and so weak consistency follows from the equivalence of C1 and C2 above, Assumptions 5 and 6, compactness of Θ, and a classical argument (*e.g.* White, 1994, theorem 3.4). So we are only left to prove Proposition 1.

We need the following preliminary result.

Lemma C1. *Let Assumptions* 1(a), 2, *and* 3 *hold, and for given* $\theta\in\Theta$, *set* $\mathcal{B}_T \equiv \{\upsilon\in\mathbb{R}^{q-q^*} : \pi_{1T}(\upsilon)>\delta_T$ *and* $\pi_{1T}^i(\upsilon;\theta)>\delta_T,\ i=1,\ldots,S\}$, *where* $\delta_T\to 0$ *and* $T^{\frac{1}{2}}\lambda_T^q\delta_T\to\infty$. *We have*

(a) *Let* $\lambda_T\to\bar{\lambda}$, *where* $0\le\bar{\lambda}<\infty$. *Then,* $\sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\mathcal{B}_T}\left|\frac{\pi_{2T}(z,\upsilon)}{\pi_{1T}(\upsilon)}-m(z\mid\upsilon;\theta_0)\right|\xrightarrow{p}0$; *and for all* $\theta\in\Theta$, $\sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\mathcal{B}_T}\left|\frac{\pi_{2T}^i(z,\upsilon;\theta)}{\pi_{1T}^i(\upsilon;\theta)}-m(z\mid\upsilon;\theta)\right|\xrightarrow{p}0,\quad i=1,\ldots,S$.

(b) *Let* $\lambda_T\to 0$ *and* $\delta_T^{-1}\lambda_T^r\to 0$. *Then,* $\sup_{(z,\upsilon)\in\mathcal{B}_T}\left|\frac{\pi_{2T}(z,\upsilon)}{\pi_{1T}(\upsilon)}-\pi(z\mid\upsilon;\theta_0)\right|\xrightarrow{p}0$; *and for all* $\theta\in\Theta$, $\sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\mathcal{B}_T}\left|\frac{\pi_{2T}^i(z,\upsilon;\theta)}{\pi_{1T}^i(\upsilon;\theta)}-\pi(z\mid\upsilon;\theta)\right|\xrightarrow{p}0,\quad i=1,\ldots,S$.

Remarks on the proof of Lemma C1. The basic intuition about the proof of Lemma C1 carries over the proofs of the remaining lemmas in this appendix, which are provided in sections A.1 and A.2 of Al-M08. We now develop the basic lines of the arguments, which also help understand the Lemma's conditions on the trimming parameter δ_T. Consider the first result in Part (b) of Lemma C1 (the intuition about Part (a) is entirely similar). For some $\varepsilon > 0$, let $\mathcal{B}_{1T}(\varepsilon) \equiv \{\upsilon \in \mathbb{R}^{q-q^*} : \pi_1(\upsilon;\theta_0) \geq \varepsilon\delta_T\}$ and $\mathcal{B}_{2T}(\varepsilon) \equiv \{\upsilon \in \mathbb{R}^{q-q^*} : \pi_{1T}(\upsilon) > \varepsilon\delta_T\}$. Finally, let $\hat{\mathcal{B}}_T \equiv \hat{\mathcal{B}}_T(\varepsilon) \equiv \mathcal{B}_{1T}(\varepsilon) \cap \mathcal{B}_{2T}(\varepsilon)$. We have,

$$\begin{aligned}
\sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\hat{\mathcal{B}}_T} |\pi_T(z \mid \upsilon) - \pi(z \mid \upsilon;\theta_0)| &\leq \sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\hat{\mathcal{B}}_T} \left[\frac{1}{\pi_{1T}(\upsilon)}|\pi_{2T}(z,\upsilon) - \pi_2(z,\upsilon;\theta_0)|\right] \\
&\quad + \sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\hat{\mathcal{B}}_T} \left[\frac{\pi_2(z,\upsilon;\theta_0)}{\pi_1(\upsilon;\theta_0)\pi_{1T}(\upsilon)}|\pi_{1T}(\upsilon) - \pi_1(\upsilon;\theta_0)|\right] \\
&\leq \varepsilon^{-1}\delta_T^{-1} \sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\hat{\mathcal{B}}_T} [|\pi_{2T}(z,\upsilon) - \pi_2(z,\upsilon;\theta_0)|] \\
&\quad + c_0\varepsilon^{-1}\delta_T^{-1} \sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\hat{\mathcal{B}}_T} [|\pi_{1T}(\upsilon) - \pi_1(\upsilon;\theta_0)|] \\
&= O_p(T^{-\frac{1}{2}}\lambda_T^{-q}\delta_T^{-1}) + O_p(\delta_T^{-1}\lambda_T^r) \\
&\quad + O_p\left(T^{-\frac{1}{2}}\lambda_T^{-(q-q^*)}\delta_T^{-1}\right) + O_p(\delta_T^{-1}\lambda_T^r),
\end{aligned}$$

where $c_0 \equiv \sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\mathbb{R}^{q-q^*}} [\pi_2(z,\upsilon;\theta_0)/\pi_1(\upsilon;\theta_0)] \equiv \sup_{(z,\upsilon)\in\mathbb{R}^{q^*}\times\mathbb{R}^{q-q^*}} \pi(z \mid \upsilon;\theta_0)$. The second line follows by the identity $\frac{a}{b} - \frac{\bar{a}}{\bar{b}} \equiv \frac{1}{b}(a-\bar{a}) - \frac{\bar{a}}{b\bar{b}}(b-\bar{b})$, which holds for any four strictly positive functions $a, b, \bar{a}$, and $\bar{b}$; the third line follows by the definition of the trimming set $\hat{\mathcal{B}}_T$; the fourth line holds as $\sup_{x\in\mathbb{R}^q} |\pi_{2T}(x) - \pi_2(x;\theta_0)| = O_p(T^{-\frac{1}{2}}\lambda_T^{-q}) + O_p(\lambda_T^r)$, by theorem 1 (p. 568) of Andrews (1995). Then, Part (b) of Lemma C1 holds true as it can be shown that with probability approaching one as $T \to \infty$, the trimming set $\hat{\mathcal{B}}_T$ is the same as $\mathcal{B}_{2T}(1)$.

Remarks on the proof of Proposition 1. Before stating our proof of Proposition 1, it is useful to describe the main ideas underlying this proof. Our concern is to show that the integrand of $[L_{T,S}^{\mathrm{CD}}(\theta) - L^{\mathrm{CD}}(\theta)]$ is bounded by integrable functions independent of the sample size T, and that it converges in probability pointwise to 0 as T goes to infinity. To establish these facts, we shall rely on an inequality, which is a standard component of the consistency proof for methods of moments estimators (see, for example, Duffie and Singleton, 1993, equation (A5), p. 949). Let M_T and W_T be two sequences converging in probability to M_0 and W_0, respectively. (In our proof, M_T and W_T will be two estimated functions.) Then, the following inequality holds true,

$$|M_T^2 W_T - M_0^2 W_0| \leq |M_0||W_T - W_0||M_T| + |M_T - M_0|(W_T|M_T| + W_0|M_0|). \tag{A1}$$

We shall also use the inequality (A1) to establish consistency of the CD-SNE in Theorem 2 (Appendix B).

Proof of Proposition 1. We produce the arguments that apply to the case in which the bandwidth sequence λ_T satisfies Assumption 9(a). Accordingly, we will make a repeated use of Lemma C1(a). The case of a bandwidth sequence that satisfies Assumption 9(b) is dealt with similarly, by replacing Lemma C1(a) with Lemma C1(b). Below, we will denote $\bar{w}(z,\upsilon) \equiv E[w_T(z,\upsilon)]$.

We claim that

$$|L_{T,S}^{\mathrm{CD}}(\theta) - L^{\mathrm{CD}}(\theta)| \leq \iint (a_{1T,S}(z,\upsilon;\theta) + a_{2T,S}(z,\upsilon;\theta))dzd\upsilon, \tag{A2}$$

where

$$\begin{aligned}
a_{1T,S}(z,\upsilon;\theta) &\equiv |\pi_{T,S}(z \mid \upsilon;\theta) - \pi_T(z \mid \upsilon)|\mathbb{T}_{T,S}(\upsilon;\theta)|m(z \mid \upsilon;\theta) - m(z \mid \upsilon;\theta_0)||w_T(z,\upsilon) - \bar{w}(z,\upsilon)| \\
a_{2T,S}(z,\upsilon;\theta) &\equiv |[\pi_{T,S}(z \mid \upsilon;\theta)\mathbb{T}_{T,S}(\upsilon;\theta) - m(z \mid \upsilon;\theta)] - [\pi_T(z \mid \upsilon)\mathbb{T}_{T,S}(\upsilon;\theta) - m(z \mid \upsilon;\theta_0)]| \\
&\quad \times [\phi_{T,S}(z,\upsilon;\theta) + \phi(z,\upsilon;\theta)]
\end{aligned}$$

$$\phi_{T,S}(z,v;\theta) \equiv |\pi_{T,S}(z \mid v;\theta) - \pi_T(z \mid v)|\mathbb{T}_{T,S}(v;\theta)w_T(z,v)$$

$$\phi(z,v;\theta) \equiv |m(z \mid v;\theta) - m(z \mid v;\theta_0)|\bar{w}(z,v) \tag{A3}$$

provided the R.H.S. of (A2) is finite. Indeed, (A2) follows by applying the inequality (A1) to the integrand of $[L^{\mathrm{CD}}_{T,S}(\theta) - L^{\mathrm{CD}}(\theta)]$, after setting $M_T \equiv [\pi_{T,S}(z \mid v;\theta) - \pi_T(z \mid v)]\mathbb{T}_{T,S}(v;\theta)$, $M_0 \equiv [m(z \mid v;\theta) - m(z \mid v;\theta_0)]$, $W_T = w_T(z,v)$ and $W_0 = \bar{w}(z,v)$.

Next, we show that $\iint (a_{1T,S} + a_{2T,S}) \xrightarrow{p} 0$ for all $\theta \in \Theta$. We study the two integrals separately.

– For all $(z,v,\theta) \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} \times \Theta$, $a_{1T,S}(z,v;\theta) \leq \ell_T(z,v;\theta)\phi_{2T,S}(z,v;\theta)$, where

$$\ell_T(z,v;\theta) \equiv |m(z \mid v;\theta) - m(z \mid v;\theta_0)||w_T(z,v) - \bar{w}(z,v)|$$

$$\phi_{2T,S}(z,v;\theta) \equiv \frac{1}{S}\sum_{i=1}^{S} |\pi^i_T(z \mid v;\theta) - m(z \mid v;\theta)|\mathbb{T}_{T,S}(v;\theta) + |\pi_T(z \mid v) - m(z \mid v;\theta_0)|\mathbb{T}_{T,S}(v;\theta)$$

$$+ |m(z \mid v;\theta) - m(z \mid v;\theta_0)|\mathbb{T}_{T,S}(v;\theta). \tag{A4}$$

By Assumptions 1(a), 3, and 4, we have that for each $\theta \in \Theta$, the function ℓ_T is bounded by integrable functions independent of T. Moreover, by Assumption 4, for all $\theta \in \Theta$, $\ell_T(z,v;\theta) \xrightarrow{p} 0$ (z,v) pointwise. Finally, by Lemma C1(a),

$$\sup_{(z,v)\in\mathbb{R}^q} |\pi^i_T(z \mid v;\theta) - m(z \mid v;\theta)|\mathbb{T}_{T,S}(v;\theta) \xrightarrow{p} 0, \quad i = 1,\ldots,S.$$

This result clearly holds for the first $S+1$ terms of $\phi_{2T,S}$ in (A4) as well. Finally, $|m(z \mid v;\theta) - m(z \mid v;\theta_0)|$ is bounded. Therefore, for all $\theta \in \Theta$,

$$\iint a_{1T,S}(z,v;\theta)dzdv \xrightarrow{p} 0. \tag{A5}$$

– For all $(z,v,\theta) \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} \times \Theta$,

$$a_{2T,S}(z,v;\theta) \leq \frac{1}{S}\sum_{i=1}^{S} |\pi^i_T(z \mid v;\theta)\mathbb{T}_{T,S}(v;\theta) - m(z \mid v;\theta)|\phi_{3T,S}(z,v;\theta)$$

$$+ |\pi_T(z \mid v)\mathbb{T}_{T,S}(v;\theta) - m(z \mid v;\theta_0)|\phi_{3T,S}(z,v;\theta), \tag{A6}$$

where $\phi_{3T,S}(z,v;\theta) \equiv \phi(z,v;\theta) + \phi_{T,S}(z,v;\theta) \leq \phi(z,v;\theta) + \phi_{2T,S}(z,v;\theta)w_T(z,v)$. For each $i = 1,\ldots,S$, and $(z,v,\theta) \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} \times \Theta$,

$$|\pi^i_T(z \mid v;\theta)\mathbb{T}_{T,S}(v;\theta) - m(z \mid v;\theta)|\phi_{3T}(z,v;\theta)$$

$$\leq m(z \mid v;\theta)[1 - \mathbb{T}_{T,S}(v;\theta)][\phi(z,v;\theta) + \phi_{2T,S}(z,v;\theta)w_T(z,v)]$$

$$+ |\pi^i_T(z \mid v;\theta) - m(z \mid v;\theta)|\mathbb{T}_{T,S}(v;\theta)[\phi(z,v;\theta) + \phi_{2T,S}(z,v;\theta)w_T(z,v)]$$

$$\equiv a_{21T,S}(z,v;\theta) + a_{22T,S}(z,v;\theta),$$

where the inequality holds by the triangle inequality. Since w_T, ϕ and m are bounded, and w_T and ϕ are also integrable, $\iint a_{22T,S}(z,v;\theta) \xrightarrow{p} 0$ for all $\theta \in \Theta$ by Lemma C1(a). As for the $a_{21T,S}$ term, $|1 - \mathbb{T}_{T,S}(v;\theta)| \leq 1$. Moreover, $[1 - \mathbb{T}_{T,S}(v;\theta)] \xrightarrow{p} 0$ pointwise. Hence, by the previous results on $\phi_{2T,S}$ and Lemma C1(a), $\iint a_{21T,S}$ $(z,v;\theta) \xrightarrow{p} 0$ for all $\theta \in \Theta$. By reiterating the previous arguments, one shows that the same result holds for the second term in (A6) and, hence, for all $\theta \in \Theta$,

$$\iint a_{2T,S}(z,v;\theta)dzdv \xrightarrow{p} 0. \tag{A7}$$

Hence, the proof of the proposition is complete, by equations (A2), (A5), and (A7).

A.2. *Identifiability and bandwidth choice*

We provide two examples of kernels and data-generating processes with both bounded and unbounded support for which the identifiability condition in Assumption 5 holds.

Example 1. *Let y_t in* (7) *be independent and identically distributed as a Gaussian with unit variance and mean parameter $\theta_0 = 0$. Let the kernel be uniform, as in the example of Section* 3.1 (*see equation* (16)), *that is, $K(y) = \frac{1}{2}\mathbb{I}_{|y|\leq 1}$, where $\mathbb{I}$ is the indicator function. In this case, the asymptotic criterion of the CD-SNE is given by*

$$\mathbb{L}(\theta, \bar{\lambda}) \equiv \int_{\mathbb{R}} [\pi_1^*(y;\theta,\bar{\lambda}) - \pi_1^*(y;\theta_0,\bar{\lambda})]^2 w(y)dy, \tag{A8}$$

where,

$$\pi_1^*(y;\theta,\bar{\lambda}) - \pi_1^*(y;\theta_0,\bar{\lambda}) = \frac{1}{2\bar{\lambda}}\frac{1}{\sqrt{2\pi}}\int_{y-\bar{\lambda}}^{y+\bar{\lambda}}\left(e^{-\frac{1}{2}(\xi-\theta)^2} - e^{-\frac{1}{2}(\xi-\theta_0)^2}\right)d\xi, \quad \theta_0 = 0,$$

and where we take $w(y) = \frac{1}{\sqrt{2\pi}}e^{-(1/2)y^2}$. Identification occurs if $\mathbb{L}(\theta,\bar{\lambda}) = 0 \Longrightarrow \theta = \theta_0$, or if

$$\bar{\lambda} : \sup_{y\in\mathbb{R}} |\pi_1^*(y;\theta,\bar{\lambda}) - \pi_1^*(y;\theta_0,\bar{\lambda})| = 0. \tag{A9}$$

Let the limiting bandwidth value $\bar{\lambda} = \frac{1}{2}$. Figure A1 *below illustrates that $\mathbb{L}\left(\theta, \frac{1}{2}\right) = 0$ only with $\theta = \theta_0$. In other terms, $\theta_0 = 0$ is the only parameter value for θ that makes $\pi_1^*\left(y;\theta,\frac{1}{2}\right) - \pi_1^*\left(y;\theta_0,\frac{1}{2}\right) = 0$ for each y, which is what is formally required by (A9).*

Next, we develop one example in which data have *bounded* support, and identification occurs even when the limiting bandwidth value $\bar{\lambda}$ is non-zero.

Example 2. *Let us assume that y_t is independent and identically distributed, generated by a Beta distribution with parameters θ, β, where β_0 is known and equal to* 2. *Therefore, the support is $Y = (0,1)$, and the marginal density for y_t is,*

$$\pi_1(y;\theta) = \frac{\Gamma(\theta+2)}{\Gamma(\theta)}y^{\theta-1}(1-y).$$

Let $\theta_0 = 2$. For all $\theta \in (1,\infty)$,

$$\Delta\pi_1^*(y;\theta,\bar{\lambda}) \equiv \pi_1^*(y;\theta,\bar{\lambda}) - \pi_1^*(y;\theta_0,\bar{\lambda}) = \frac{1}{\bar{\lambda}}\int_0^1 K\left(\frac{y-\xi}{\bar{\lambda}}\right)\left[\frac{\Gamma(\theta+2)}{\Gamma(\theta)}\xi^{\theta-1} - 6\xi\right](1-\xi)d\xi.$$

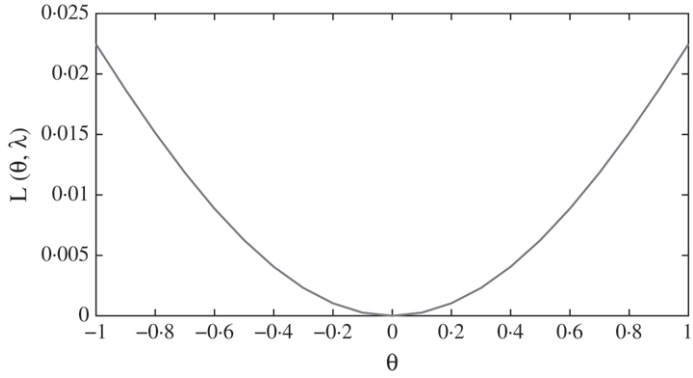

FIGURE A1

Identifiability with Normal distributions and uniform kernels. This picture depicts the SNE asymptotic criterion $\mathbb{L}(\theta,\bar{\lambda})$ in equation (A8), evaluated at $\bar{\lambda} = \frac{1}{2}$

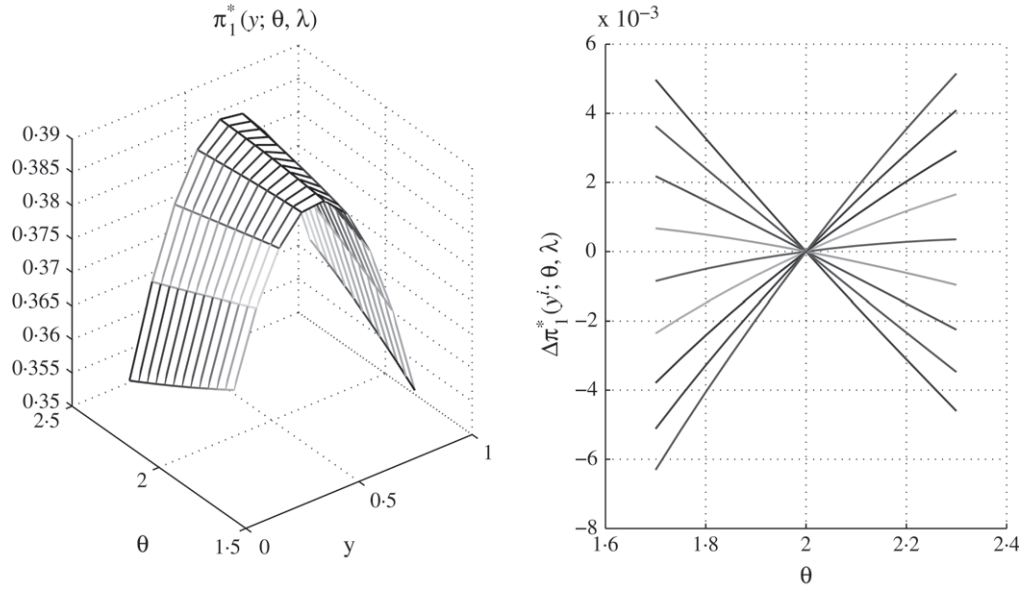

FIGURE A2

Identifiability with Beta distributions. The L.H.S. panel depicts the function $\pi_1^*(y;\theta,\bar{\lambda}) = \bar{\lambda}^{-1}\int K\left(\frac{y-\xi}{\bar{\lambda}}\right)\pi_1(\xi;\theta)d\xi$ evaluated at the points $y^i = 0{\cdot}10,\ 0{\cdot}20,\ \ldots,\ 0{\cdot}90$ and $\theta^i = 1{\cdot}70,\ 1{\cdot}75,\ldots,\ 2{\cdot}30$. The R.H.S. panel depicts the lines $\theta \mapsto \Delta\pi_1^*(y;\theta,\bar{\lambda}) = \pi_1^*(y;\theta,\bar{\lambda}) - \Delta\pi_1^*(y;\theta_0,\bar{\lambda})$ evaluated at the same points y^i, θ^i. In both cases, $\pi_1(y;\theta) = \frac{\Gamma(\theta+2)}{\Gamma(\theta)}y^{\theta-1}(1-y)$, $\theta_0 = 2$, K is the Gaussian kernel and $\bar{\lambda} = 1$

Consider the kernel $K(x) = \frac{1}{\sqrt{2\pi}}e^{-(1/2)x^2}$, *and let* $\bar{\lambda} = 1$. *Figure* A2 *plots both* $\pi_1^*(y;\theta,1)$ *and* $\Delta\pi_1^*(y;\theta,1)$. *Note, now, that for all* y, $\Delta\pi_1^*(y;\theta,1) = 0 \Longrightarrow \theta = \theta_0 = 2$. *Thus, in this example, the model and the limiting bandwidth* $\bar{\lambda}$ *satisfy an identification condition, which is even stronger than required by* (A9), *i.e.* $\sup_{y\in(0,1)}|\Delta\pi_1^*(y;\theta,1)| = 0 \Longrightarrow \theta = \theta_0$.

A.3. *Asymptotic normality*

In Lemmas N1 through N3 below, $\mathcal{B}_T$ is the same set introduced in Lemma C1, and δ_T is the same trimming sequence introduced in Assumptions 8 and 9.

Lemma N1. *Let Assumptions* 1–3 *hold, let* $x \equiv [z\ \ v]$, *as in the main text, and let Assumption* 7(b) *hold. Then, for all* $\theta \in \Theta$ *and* $j = 1,\ldots,n$,

(i) $\sup_{x\in\mathbb{R}^q}\left|\nabla_{\theta_j}\pi_{2T,S}(x;\theta) - \nabla_{\theta_j}\pi_2(x;\theta)\right| = O_p\left(T^{-\frac{1}{2}}\lambda_T^{-q-1}\right) + O_p(\lambda_T^r)$.

(ii) $\sup_{(z,v)\in\mathbb{R}^{q^*}\times\mathcal{B}_T}\left|\nabla_{\theta_j}\pi_{T,S}(z\mid v;\theta) - \nabla_{\theta_j}\pi(z\mid v;\theta)\right| = O_p\left(T^{-\frac{1}{2}}\lambda_T^{-q-1}\delta_T^{-2}\right) + O_p\left(T^{-\frac{1}{2}}\lambda_T^{-(q-q^*)-1}\delta_T^{-2}\right) + O_p\left(T^{-\frac{1}{2}}\lambda_T^{-(q-q^*)}\delta_T^{-3}\right) + O_p(\lambda_T^r\delta_T^{-3})$.

Lemma N2. *Let the assumptions in Lemma* N1 *and Assumption* 7(b) *hold. Then, for all* $j = 1,\ldots,n$,

$$\sup_{(z,v)\in\mathbb{R}^{q^*}\times\mathcal{B}_T}\left|\frac{\nabla_{\theta_j}\pi_{T,S}(z\mid v;\theta_0)w_T(z,v)}{\pi_{1T}(v)} - \frac{\nabla_{\theta_j}\pi(z\mid v;\theta_0)w(z,v)}{\pi_1(v;\theta_0)}\right|$$

$$= O_p\left(T^{-\frac{1}{2}}\lambda_T^{-q-1}\delta_T^{-3}\right) + O_p\left(T^{-\frac{1}{2}}\lambda_T^{-(q-q^*)-1}\delta_T^{-3}\right) + O_p\left(T^{-\frac{1}{2}}\lambda_T^{-(q-q^*)}\delta_T^{-4}\right) + O_p(\lambda_T^r\delta_T^{-4}).$$

Lemma N3. *Let the assumptions in Lemma* N2 *hold. Then, for all* $i = 1, \ldots, S$ *and* $j = 1, \ldots, n$,

$$\sup_{(z,v)\in\mathbb{R}^{q^*}\times\mathcal{B}_T} \left| \frac{\nabla_{\theta_j}\pi_{T,S}(z \mid v;\theta_0)E(\pi_{2T}(z,v))w_T(z,v)}{\pi^i_{1T}(v;\theta_0)\pi_{1T}(v)} - \frac{\nabla_{\theta_j}\pi(z \mid v;\theta_0)\pi_2(z,v;\theta_0)w(z,v)}{\pi_1(v;\theta_0)^2} \right|$$

$$= O_p\left(T^{-(1/2)}\lambda_T^{-q-1}\delta_T^{-4}\right) + O_p\left(T^{-(1/2)}\lambda_T^{-(q-q^*)-1}\delta_T^{-4}\right) + O_p\left(T^{-(1/2)}\lambda_T^{-(q-q^*)}\delta_T^{-5}\right) + O_p(\lambda_T^r\delta_T^{-5}).$$

Remarks on Lemmas N1–N3

(a) Lemma N1 is needed to show that $\mathcal{J}_{T,S}$ in equation (20) converges in probability to $\mathcal{J}$, where $\mathcal{J}$ has been defined in Theorem 1 (see, also, equation (A13) below). Lemmas N2 and N3 are needed to show that the terms $\mathcal{I}^i_{1T,S}$ and $\mathcal{I}^i_{2T,S}$ in the first order conditions (A11)–(A12a)–(A12b) below converge in distribution to the Gaussian terms provided in equations (A14a)–(A14b) below.

(b) The bandwidth conditions in Assumption 9(b) guarantee that the suprema in Lemmas N1–N3 go to 0 in probability, as shown in the next remarks.

Remarks on Assumption 9(b). It is easily seen that all the suprema in Lemmas N1–N3 go to 0 in probability under the conditions that $\lambda_T \to 0$ and $T^{\frac{1}{2}}\lambda_T^{q+1}\delta_T^4 \to \infty$ in Assumption 9(b). The only nontrivial conditions that must be shown to hold are that in Lemma N3, (i) $\lambda_T^r\delta_T^{-5} \to 0$ and (ii) $T^{1/2}\lambda_T^{q-q^*}\delta_T^5 \to \infty$. But by the second part of Assumption 9(b), $(T^{\frac{1}{2}}\lambda_T^{q-q^*}\delta_T^5)\lambda_T^{q^*+1}\delta_T^{-1} \to \infty$. Hence, under the second part of Assumption 9(b), we have that $T^{\frac{1}{2}}\lambda_T^{q-q^*}\delta_T^5 \to \infty$ holds if $\lambda_T^{q^*+1}\delta_T^{-1} \to 0$. So we must simultaneously have $\lambda_T^r\delta_T^{-5} \to 0$ and $\lambda_T^{q^*+1}\delta_T^{-1} \to 0$, that is $\delta_T^{-1}\lambda_T^{\min\{q^*+1,\,\frac{1}{5}r\}} \to 0$.

Proof of asymptotic normality. By Assumption 7(a), the order of derivation and integration in $\nabla_\theta L^{\mathrm{CD}}_{T,S}(\theta)$ can be interchanged (see Newey and McFadden, 1994, lemma 3.6, pp. 2152–2153). Therefore, the CD-SNE satisfies the following first order conditions:

$$\mathbf{0}_n = \frac{1}{S}\sum_{i=1}^{S}\iint\left[\frac{\pi^i_{2T}(z,v;\theta_{T,S})}{\pi^i_{1T}(v;\theta_{T,S})} - \frac{\pi_{2T}(z,v)}{\pi_{1T}(v)}\right]\nabla_\theta\pi_{T,S}(z \mid v;\theta_{T,S})w_T(z,v)\mathbb{T}^2_{T,S}(v;\theta_{T,S})dzdv$$

$$+\iint\left[\pi_{T,S}(z \mid v;\theta_{T,S}) - \pi_T(z \mid v)\right]^2 w_T(z,v)\mathbb{T}_{T,S}(v;\theta_{T,S})\nabla_\theta\mathbb{T}_{T,S}(v;\theta_{T,S})dzdv. \tag{A10}$$

In Al-M08 (Section C.2), we demonstrate that the effects of the gradient $\nabla_\theta\mathbb{T}_{T,S}(v;\theta_{T,S})$ are asymptotically negligible. Precisely, an expansion of the first order conditions in (A10) around θ_0 leaves,

$$\mathbf{0}_n = \frac{1}{S}\sum_{i=1}^{S}\sqrt{T}\iint\left[\frac{\pi^i_{2T}(z,v;\theta_0)}{\pi^i_{1T}(v;\theta_0)} - \frac{\pi_{2T}(z,v)}{\pi_{1T}(v)}\right]\nabla_\theta\pi_{T,S}(z \mid v;\theta_0)w_T(z,v)\mathbb{T}^2_{T,S}(v;\theta_0)dzdv + o_p(1)$$

$$+\left[\iint\left|\nabla_\theta\pi_{T,S}(z \mid v;\theta_0)\mathbb{T}_{T,S}(v;\theta_0)\right|_2 w_T(z,v)dzdv + o_p(1)\right]\sqrt{T}(\theta_{T,S}-\theta_0).$$

Lengthy computations in Al-M08 (Section C.2) then lead to:

$$\mathbf{0}_n = \frac{1}{S}\sum_{i=1}^{S}(\mathcal{I}^i_{1T,S}+\mathcal{I}^i_{2T,S}) - (\mathcal{I}^0_{1T,S}+\mathcal{I}^0_{2T,S}) + [\mathcal{J}_{T,S}+o_p(1)]\sqrt{T}(\theta_{T,S}-\theta_0), \tag{A11}$$

where, for $i = 0, 1, \ldots, S$,

$$\mathcal{I}^i_{1T,S} \equiv \iint\frac{\nabla_\theta\pi_{T,S}(z \mid v;\theta_0)w_T(z,v)}{\pi^i_{1T}(v;\theta_0)}\mathbb{T}^2_{T,S}(v;\theta_0)dA^i_T(z,v) \tag{A12a}$$

$$\mathcal{I}^i_{2T,S} \equiv \iint\frac{\nabla_\theta\pi_{T,S}(z \mid v;\theta_0)E(\pi_{2T}(z,v))w_T(z,v)}{\pi^i_{1T}(v;\theta_0)\pi_{1T}(v)}\mathbb{T}^2_{T,S}(v;\theta_0)dzdA^i_T(v) \tag{A12b}$$

and where $A^i_T(z,v)$ and $A^i_T(v)$ are as in the definitions (21) in the main text.

By Lemma N1, and extensive computations in Al-M08 (Section C.2),

$$\mathcal{J}_{T,S} \equiv \iint\left|\nabla_\theta\pi_{T,S}(z \mid v;\theta_0)\mathbb{T}_{T,S}(v;\theta_0)\right|_2 w_T(z,v)dzdv \xrightarrow{p} \mathcal{J} \equiv \iint|\nabla_\theta\pi(z \mid v;\theta_0)|_2\, w(z,v)dzdv. \tag{A13}$$

Next, let $F_2(x) = \int_{-\infty}^{x} \pi_2(u;\theta_0)du$ and, for all $i = 0, 1, \ldots, S$, $F_{2T}^i(x) = \int_{-\infty}^{x} \pi_{2T}^i(u;\theta_0)du$ ($x = [z \;\; v] \in \mathbb{R}^q$), where $\pi_{2T}^0(x;\theta_0) \equiv \pi_{2T}(x)$. Let $\mathbb{G}$ be a measurable V-C subgraph class of uniformly bounded functions (see, for example, Arcones and Yu (1994, Definition 2.2 p. 51)). By Arcones and Yu (1994, corollary 2.1 pp. 59–60), for each $G \in \mathbb{G}$, and $x_t = [z_t \;\; v_t]$, $T^{-1/2}\sum_{t=1+l}^{T}[G(x_t) - EG]$ converges in law to a Gaussian process under Assumptions 2 and 3. Now $\lambda_T^{-q} K_q((x_t - x)/\lambda_T) \in \mathbb{G}(x \in \mathbb{R}^q)$. Therefore, under Assumptions 2 and 3, the terms $A_T^i(x) \equiv A_T^i(z, v) = \int_{-\infty}^{z}\int_{-\infty}^{v} dA_T^i(s', s)$ in the first definition of (21) in the main text,

$$A_T^i(x) \equiv \sqrt{T}[F_{2T}^i(x) - E(F_{2T}^i(x))] \Rightarrow \omega_i^0(F_2(x)), \quad i = 0, 1, \ldots, S,$$

where $\omega_i^0(F_2(x))$ are independent Gaussian processes with covariance kernel,

$$\mathbb{C}_q(x, x') \equiv \min\{F_2(x), F_2(x')\}[1 - F_2(x')] + \sum_{k=1}^{\infty}[F_2^k(x, x') + F_2^k(x', x) - 2F_2(x)F_2(x')],$$

and $F_2^k(x, x') \equiv P(x_0 \le x, x_k \le x')$, for $(x, x') \in \mathbb{R}^q \times \mathbb{R}^q$.

Similarly, let $F_1(v) = \int_{-\infty}^{v} \pi_1(u;\theta_0)du$ and, for all $i = 0, 1, \ldots, S$, $F_{1T}^i(v) = \int_{-\infty}^{v} \pi_{1T}^i(u;\theta_0)du$ ($v \in \mathbb{R}^{q-q^*}$), where $\pi_{1T}^0(v;\theta_0) \equiv \pi_{1T}(v)$. Under Assumptions 2 and 3, the terms $A_T^i(v) = \int_{-\infty}^{v} dA_T^i(s)$ in the second definition of (21) in the main text,

$$A_T^i(v) \equiv \sqrt{T}[F_{1T}^i(v) - E(F_{1T}^i(v))] \Rightarrow \hat{\omega}_i^0(F_1(v)), \quad i = 0, 1, \ldots, S,$$

where $\hat{\omega}_i^0(F_1(v))$ are independent Gaussian processes with covariance kernel $\mathbb{C}_{q-q^*}(v, v')$, $(v, v') \in \mathbb{R}^{q-q^*} \times \mathbb{R}^{q-q^*}$. Hence, by Lemmas N2 and N3 and computations in Al-M08 (Section C.2), the terms $\mathcal{I}_{1T,S}^i$ and $\mathcal{I}_{2T,S}^i$ in (A12a)–(A12b) satisfy

$$\mathcal{I}_{1T,S}^i \xrightarrow{d} \mathcal{I}_1^i \equiv \iint \eta(z, v)d\omega_i^0(F_2(z, v)) \tag{A14a}$$

$$\mathcal{I}_{2T,S}^i \xrightarrow{d} \mathcal{I}_2^i \equiv \int \gamma(v)d\hat{\omega}_i^0(F_1(v)) \tag{A14b}$$

where $\eta(z, v)$ and $\gamma(v)$ are as in the main text (see equations (22)), and are reported for convenience below:

$$\eta(z, v) = \frac{\nabla_\theta \pi(z \mid v;\theta_0)w(z, v)}{\pi_1(v;\theta_0)}; \quad \gamma(v) = \int \frac{\nabla_\theta \pi(z \mid v;\theta_0)\pi_2(z, v;\theta_0)w(z, v)}{\pi_1(v;\theta_0)^2}dz. \tag{A15}$$

The terms $\mathcal{I}_1^i$, $i = 0, 1, \ldots, S$, are all independent and asymptotically centered Gaussian. Therefore, by equation (A11), equation (A13), and by the Slutzky's theorem, $\sqrt{T}(\theta_{T,S} - \theta_0)$ is asymptotically centred normally distributed with variance,

$$V_S \equiv \mathcal{J}^{-1}\mathrm{var}\left(\frac{1}{S}\sum_{i=1}^{S}(\mathcal{I}_1^i + \mathcal{I}_2^i) - (\mathcal{I}_1^0 + \mathcal{I}_2^0)\right)\mathcal{J}^{\top -1} = \mathcal{J}^{-1}\left(1 + \frac{1}{S}\right)\mathrm{var}(\mathcal{I}_1^0 + \mathcal{I}_2^0)\mathcal{J}^{\top -1},$$

where the variance terms $\mathrm{var}(\mathcal{I}_1^0 + \mathcal{I}_2^0)$ reported in Theorem 1 are finite by the mixing condition in Assumption 2 and the assumption that $E[\|\Upsilon(z_t, v_t)\|^{\vartheta}]^{1/\vartheta} < \infty$, for some $\vartheta > 2$ (by, for example, Politis and Romano, 1994, thm. 2.3, p. 466), and follow by the same computations in Aït-Sahalia (1994) (proof of thm. 1, pp. 21–22) and Aït-Sahalia (1996) (proof of equation (12), pp. 420–421).

APPENDIX B. PROOF OF THEOREM 2

B.1. *Consistency*

Similarly as in Appendix A.1, we produce the arguments that apply to the case in which the bandwidth sequence satisfies Assumption 10(a). Accordingly, we will make a repeated use of Lemmas C1(a), C2(a), and C3(a). The case of a bandwidth sequence that satisfies Assumption 10(b) (which is used to prove asymptotic normality in Appendix B.2 below) is dealt with similarly, by replacing Lemma Cℓ(a) with Lemma Cℓ(b), $\ell = 1, 2, 3$.

We begin with two preliminary results.

Lemma C2. *Let Assumptions* 1(a), 2, *and* 3 *hold, and set* $\mathcal{A}_T \equiv \{(z,v) \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} : \pi_{2T}(z,v) > \alpha_T\}$, *where* $\alpha_T \to 0$, $T^{\frac{1}{2}}\lambda_T^q \alpha_T^3 \to \infty$ *and* $\lambda_T^{q^*}\alpha_T \to 0$. *We have*

(a) *Let* $\lambda_T \to \bar{\lambda}$, *where* $0 \leq \bar{\lambda} < \infty$; *then,*

$$\sup_{(z,v)\in\mathcal{A}_T}\left[\frac{1}{m_2(z,v;\theta_0)}\left|\frac{\pi_{1T}(v)}{\pi_T(z\mid v)} - \frac{m_1(v;\theta_0)}{m(z\mid v;\theta_0)}\right|\right] \xrightarrow{p} 0.$$

(b) *Let* $\lambda_T \to 0$ *and* $\alpha_T^3 \lambda_T^{-r} \to \infty$; *then,*

$$\sup_{(z,v)\in\mathcal{A}_T}\left[\frac{1}{\pi_2(z,v;\theta_0)}\left|\frac{\pi_{1T}(v)}{\pi_T(z\mid v)} - \frac{\pi_2(v;\theta_0)}{\pi(z\mid v;\theta_0)}\right|\right] \xrightarrow{p} 0.$$

In the next lemma, $\mathcal{B}_T$ is the same set introduced in Lemma C1, and δ_T is the same trimming sequence introduced in Assumptions 8 and 10. Moreover, $\mathcal{A}_T$ and α_T in the lemma below are as in Lemma C2.

Lemma C3. *Let Assumptions* 1(a), 2, *and* 3 *hold, and let* $\alpha_T \to 0$, $\delta_T \to 0$, $T^{\frac{1}{2}}\lambda_T^q \alpha_T^2 \delta_T \to \infty$ *and* $T^{\frac{1}{2}}\lambda_T^{q-q^*}\alpha_T^2\delta_T^2 \to \infty$. *We have*

(a) *Let* $\lambda_T \to \bar{\lambda}$, *where* $0 \leq \bar{\lambda} < \infty$; *for each* $i = 1, \ldots, S$, *and* $\theta \in \Theta$,

$$\sup_{(z,v)\in\mathcal{A}_T\cap\mathcal{B}_T}\left[\frac{1}{m_2(z,v;\theta_0)m(z\mid v;\theta_0)}\left|\frac{\pi_{2T}^i(z,v;\theta)}{\pi_{1T}^i(v;\theta)} - m(z\mid v;\theta)\right|\right] \xrightarrow{p} 0.$$

(b) *Let* $\lambda_T \to 0$, $\alpha_T^2\delta_T\lambda_T^{-r} \to \infty$ *and* $\alpha_T^2\delta_T^2\lambda_T^{-r} \to \infty$; *then, for each* $i = 1, \ldots, S$, *and* $\theta \in \Theta$,

$$\sup_{(z,v)\in\mathcal{A}_T\cap\mathcal{B}_T}\left[\frac{1}{m_2(z,v;\theta_0)m(z\mid v;\theta_0)}\left|\frac{\pi_{2T}^i(z,v;\theta)}{\pi_{1T}^i(v;\theta)} - \pi(z\mid v;\theta)\right|\right] \xrightarrow{p} 0.$$

Remarks on Assumption 10(a). It is easily seen that the bandwidth conditions in Lemmas C2(a) and C3(a) hold if $\alpha_T \to 0$, $\delta_T \to 0$, $T^{\frac{1}{2}}\lambda_T^q\alpha_T^3 \to \infty$ and $T^{\frac{1}{2}}\lambda_T^q\alpha_T^2\delta_T^2 \to \infty$. In turn, these conditions are satisfied if, for some constant κ, $\delta_T/\alpha_T \to \kappa$, and $T^{\frac{1}{2}}\lambda_T^q\alpha_T^4 \to \infty$, as required by Assumption 10(a).

We are ready to produce the consistency proof. By the remarks on the proof of Proposition 1 in Appendix A.1, and equation (A2), we only have to show that for all $\theta \in \Theta$,

$$\iint a_{iT,S}(z,v;\theta)dzdv \xrightarrow{p} 0, \quad i = 1,2, \tag{B1}$$

where the terms $a_{iT,S}$ are defined as in Appendix A.1 (equations (A3)), but with weighting function $w_T(z,v) = [\pi_{1T}(v)/\pi_T(z\mid v)]\mathbb{T}_{2T}(z,v)$ and $w(z,v) = m_1(v;\theta_0)/m(z\mid v;\theta_0)$. We proceed as in Appendix A.1, and study these two integrals separately.

– For all $(z,v,\theta) \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} \times \Theta$,

$$\begin{aligned}
&a_{1T,S}(z,v;\theta)\\
&\leq |\pi_{T,S}(z\mid v;\theta) - \pi_T(z\mid v)||m(z\mid v;\theta) - m(z\mid v;\theta_0)|\left|\frac{\pi_{1T}(v)}{\pi_T(z\mid v)} - \frac{m_1(v;\theta_0)}{m(z\mid v;\theta_0)}\right|\mathbb{T}_{T,S}(v;\theta)\mathbb{T}_{2T}(z,v)\\
&\quad + |\pi_{T,S}(z|v;\theta) - \pi_T(z\mid v)|\mathbb{T}_{T,S}(v;\theta)|m(z\mid v;\theta) - m(z\mid v;\theta_0)|\frac{m_1(v;\theta_0)}{m(z\mid v;\theta_0)}[1 - \mathbb{T}_{2T}(z,v)]\\
&\leq \ell_{1T,S}(z,v;\theta)\ell_{2T}(z,v;\theta)m_2(z,v;\theta_0)|m(z\mid v;\theta) - m(z\mid v;\theta_0)|\\
&\quad + \ell_{2T}(z,v;\theta)[m(z\mid v;\theta) - m(z\mid v;\theta_0)]^2 m_2(z,v;\theta_0)\mathbb{T}_{T,S}(v;\theta)\\
&\quad + \ell_{3T,S}(z,v;\theta)|m(z\mid v;\theta) - m(z\mid v;\theta_0)|m_2(z,v;\theta_0)m_1(v;\theta_0)[1 - \mathbb{T}_{2T}(z,v)]\\
&\quad + \frac{m_1(v;\theta_0)}{m(z\mid v;\theta_0)}[m(z\mid v;\theta) - m(z\mid v;\theta_0)]^2\mathbb{T}_{T,S}(v;\theta)[1 - \mathbb{T}_{2T}(z,v)]\\
&\equiv a_{11T,S}(z,v;\theta) + a_{12T,S}(z,v;\theta) + a_{13T,S}(z,v;\theta) + a_{14T,S}(z,v;\theta),
\end{aligned}$$

where

$$\ell_{1T,S}(z,v;\theta) \equiv \left[\frac{1}{S}\sum_{i=1}^{S}|\pi_T^i(z \mid v;\theta) - m(z \mid v;\theta)| + |\pi_T(z \mid v) - m(z \mid v;\theta_0)|\right]\mathbb{T}_{T,S}(v;\theta)$$

$$\ell_{2T}(z,v;\theta) \equiv \frac{1}{m_2(z,v;\theta_0)}\left|\frac{\pi_{1T}(v)}{\pi_T(z \mid v)} - \frac{m_1(v;\theta_0)}{m(z \mid v;\theta_0)}\right|\mathbb{T}_{2T}(z,v)$$

$$\ell_{3T,S}(z,v;\theta) \equiv \frac{\ell_{1T,S}(z,v;\theta)}{m_2(z,v;\theta_0)m(z \mid v;\theta_0)}.$$

We have that for all $\theta \in \Theta$: (i) $\iint a_{11T,S} \xrightarrow{p} 0$, by Lemma C1(a) in Appendix A.1, by Lemma C2, and by boundedness and integrability of the function $m_2(z,v;\theta)|m(z \mid v;\theta) - m(z \mid v;\theta_0)|$; (ii) $\iint a_{12T,S} \xrightarrow{p} 0$, by Lemma C2(a), and by boundedness and integrability of the function $[m(z \mid v;\theta) - m(z \mid v;\theta_0)]^2 m_2(z,v;\theta)$; (iii) $\iint a_{13T,S} \xrightarrow{p} 0$, by Lemmas C1(a), C2(a) and C3(a). As regards the $\iint a_{14T,S}$ term, note that the function,

$$m(z \mid v;\theta_0)^{-1}[m(z \mid v;\theta) - m(z \mid v;\theta_0)]^2 m_1(v;\theta_0)$$

is the integrand of the asymptotic criterion $L^{\mathrm{CD}}(\theta)$ in (15), with $w(z,v) = m_1(v;\theta_0)^2/m_2(z,v;\theta_0)$, which is bounded and integrable by the assumption that $L^{\mathrm{CD}}(\theta)$ is continuous and bounded on Θ (Assumption 5). Moreover, for all $(z,v,\theta) \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} \times \Theta$, we have that $|\mathbb{T}_{T,S}(v;\theta)[1 - \mathbb{T}_{2T}(z,v)]| \leq 1$, and $[1 - \mathbb{T}_{2T}(z,v)] \xrightarrow{p} 0$ pointwise. Hence $\iint a_{14T,S} \xrightarrow{p} 0$ for all $\theta \in \Theta$ and, hence, for all $\theta \in \Theta$,

$$\iint a_{1T,S}(z,v;\theta)dzdv \xrightarrow{p} 0. \tag{B2}$$

– For all $(z,v,\theta) \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} \times \Theta$,

$$\begin{aligned}
&a_{2T,S}(z,v;\theta)\\
&\leq [\phi_{T,S}(z,v;\theta) + \phi(z,v;\theta)][|\pi_{T,S}(z \mid v;\theta) - m(z \mid v;\theta)| + |\pi_T(z \mid v) - m(z \mid v;\theta_0)|]\mathbb{T}_{T,S}(v;\theta)\\
&\quad + [\phi_{T,S}(z,v;\theta) + \phi(z,v;\theta)][m(z \mid v;\theta) - m(z \mid v;\theta_0)][1 - \mathbb{T}_{T,S}(v;\theta)]\\
&\equiv a_{21T,S}(z,v;\theta) + a_{22T,S}(z,v;\theta).
\end{aligned}$$

For all $(z,v,\theta) \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} \times \Theta$,

$$\begin{aligned}
&a_{21T,S}(z,v;\theta)\\
&\leq |\pi_{T,S}(z \mid v;\theta) - m(z \mid v;\theta)|\mathbb{T}_{T,S}(v;\theta)\frac{1}{m_2(z,v;\theta_0)}\left|\frac{\pi_{1T}(v)}{\pi_T(z \mid v)} - \frac{m_1(v;\theta_0)}{m(z \mid v;\theta_0)}\right|\mathbb{T}_{2T}(z,v)\\
&\quad \times m_2(z,v;\theta_0)\left[|\pi_{T,S}(z \mid v;\theta) - m(z \mid v;\theta)| + |\pi_T(z \mid v) - m(z \mid v;\theta_0)|\right]\mathbb{T}_{T,S}(v;\theta)\\
&\quad + |\pi_T(z \mid v) - m(z \mid v;\theta_0)|\mathbb{T}_{T,S}(v;\theta)\frac{1}{m_2(z,v;\theta_0)}\left|\frac{\pi_{1T}(v)}{\pi_T(z \mid v)} - \frac{m_1(v;\theta_0)}{m(z \mid v;\theta_0)}\right|\mathbb{T}_{2T}(z,v)\\
&\quad \times m_2(z,v;\theta_0)[|\pi_{T,S}(z \mid v;\theta) - m(z \mid v;\theta)| + |\pi_T(z \mid v) - m(z \mid v;\theta_0)|]\mathbb{T}_{T,S}(v;\theta)\\
&\quad + |m(z \mid v;\theta) - m(z \mid v;\theta_0)|\mathbb{T}_{T,S}(v;\theta)\frac{1}{m_2(z,v;\theta_0)}\left|\frac{\pi_{1T}(v)}{\pi_T(z \mid v)} - \frac{m_1(v;\theta_0)}{m(z \mid v;\theta_0)}\right|\mathbb{T}_{2T}(z,v)\\
&\quad \times m_2(z,v;\theta_0)[|\pi_{T,S}(z \mid v;\theta) - m(z \mid v;\theta)| + |\pi_T(z \mid v) - m(z \mid v;\theta_0)|]\mathbb{T}_{T,S}(v;\theta)\\
&\quad + |m(z \mid v;\theta) - m(z \mid v;\theta_0)|m_1(v;\theta_0)m_2(z,v;\theta_0)\\
&\quad \times \frac{1}{m(z \mid v;\theta_0)m_2(z,v;\theta_0)}[|\pi_{T,S}(z \mid v;\theta) - m(z \mid v;\theta)| + |\pi_T(z \mid v) - m(z \mid v;\theta_0)|]\mathbb{T}_{T,S}(v;\theta)\\
&\quad + |\pi_{T,S}(z \mid v;\theta) - \pi_T(z \mid v)|\mathbb{T}_{T,S}(v;\theta)\frac{m_1(v;\theta_0)}{m(z \mid v;\theta_0)}\mathbb{T}_{2T}(z,v)
\end{aligned}$$

$$\times [|\pi_{T,S}(z \mid \upsilon;\theta) - m(z \mid \upsilon;\theta)| + |\pi_T(z \mid \upsilon) - m(z \mid \upsilon;\theta_0)|]\mathbb{T}_{T,S}(\upsilon;\theta)$$
$$\equiv a_{211T,S}(z,\upsilon;\theta) + a_{212T,S}(z,\upsilon;\theta) + a_{213T,S}(z,\upsilon;\theta) + a_{214T,S}(z,\upsilon;\theta) + a_{215T,S}(z,\upsilon;\theta).$$

We have that for all $\theta \in \Theta$: (i) $\iint \sum_{j=1,2,3} a_{21jT,S} \xrightarrow{p} 0$ by Lemma C1(a)in Appendix A.1, by Lemma C2(a), and by boundedness and integrability of $m(z \mid \upsilon;\theta)$ and $m_2(z,\upsilon;\theta)$; (ii) $\iint a_{214T,S} \xrightarrow{p} 0$ by Lemma C3(a) and boundedness and integrability of $m(z \mid \upsilon;\theta)$ and $m_2(z,\upsilon;\theta)$; (iii) $\iint a_{215T,S} \xrightarrow{p} 0$ by Lemmas C1(a) and C3(a). Therefore, for all $\theta \in \Theta$, $\iint a_{21T,S} \xrightarrow{p} 0$.
Next, for all $(z,\upsilon,\theta) \in \mathbb{R}^{q^*} \times \mathbb{R}^{q-q^*} \times \Theta$,

$$a_{22T,S}(z,\upsilon;\theta) \leq \left\{ a^*_{22T,S}(z,\upsilon;\theta) m_2(z,\upsilon;\theta_0) + [m(z \mid \upsilon;\theta) - m(z \mid \upsilon;\theta_0)]^2 \frac{m_1(\upsilon;\theta_0)}{m(z \mid \upsilon;\theta_0)} \right\} [1 - \mathbb{T}_{T,S}(\upsilon;\theta)],$$

where

$$a^*_{22T,S}(z,\upsilon;\theta)$$
$$\equiv [|\pi_{T,S}(z \mid \upsilon;\theta) - m(z \mid \upsilon;\theta)| + |\pi_T(z \mid \upsilon) - m(z \mid \upsilon;\theta_0)| + |m(z \mid \upsilon;\theta) - m(z \mid \upsilon;\theta_0)|]\mathbb{T}_{T,S}(\upsilon;\theta)$$
$$\times \frac{1}{m_2(z,\upsilon;\theta_0)} \left| \frac{\pi_{1T}(\upsilon;\theta)}{\pi_T(z \mid \upsilon)} - \frac{m_1(\upsilon;\theta_0)}{m(z \mid \upsilon;\theta_0)} \right| \mathbb{T}_{2T}(z,\upsilon) |m(z \mid \upsilon;\theta) - m(z \mid \upsilon;\theta_0)|$$
$$+ \frac{1}{m_2(z,\upsilon;\theta_0) m(z \mid \upsilon;\theta_0)} |\pi_{T,S}(z \mid \upsilon;\theta) - \pi_T(z \mid \upsilon)| \mathbb{T}_{T,S}(\upsilon;\theta) m_1(\upsilon;\theta_0)$$
$$\times |m(z \mid \upsilon;\theta) - m(z \mid \upsilon;\theta_0)| \mathbb{T}_{2T}(z,\upsilon).$$

As in Appendix A.1, we have that for all $\theta \in \Theta$, $1 - \mathbb{T}_{T,S}(\upsilon;\theta) \xrightarrow{p} 0$ pointwise. Since for all $(\upsilon,\theta) \in \mathbb{R}^{q-q^*} \times \Theta$, $1 - \mathbb{T}_{T,S}(\upsilon;\theta) \leq 1$, and the functions $m_2(z,\upsilon;\theta_0)$ and $m(z \mid \upsilon;\theta_0)^{-1}[m(z \mid \upsilon;\theta) - m(z \mid \upsilon;\theta_0)]^2 m_1(\upsilon;\theta_0)$ are bounded and integrable (by the assumption that the asymptotic criterion $L^{\mathrm{CD}}(\theta)$ is continuous and bounded on Θ), then, $\iint a_{22T,S} \xrightarrow{p} 0$ for all $\theta \in \Theta$. Hence, for all $\theta \in \Theta$,

$$\iint a_{2T,S}(z,\upsilon;\theta) dz d\upsilon \xrightarrow{p} 0. \tag{B3}$$

Hence, equation (B1) holds by equations (B2) and (B3).

B.2. *Asymptotic normality*

In the following two lemmas, $\mathcal{B}_T$ is the same set introduced in Lemma C1, and δ_T is the same trimming sequence introduced in Assumptions 8 and 10. Moreover, $\mathcal{A}_T$ and δ_T in the lemma below are as in Lemma C2.

Lemma N4. *Let the assumptions in Lemma* N1 *hold. Let* $\upsilon \mapsto \xi_{1T}(\upsilon)$ $(\upsilon \in \mathbb{R}^{q-q^*})$ *be a sequence of real, bounded functions satisfying* $\sup_{\upsilon \in \mathbb{R}^{q-q^*}} |\xi_{1T}(\upsilon) - \xi_1(\upsilon)| = O_p(T^{-\frac{1}{2}} \lambda_T^{-(q-q^*)}) + O_p(\lambda_T^r)$, *for some bounded function* ξ_1. *Then, for all* $\theta \in \Theta$ *and* $j = 1, \ldots, n$,

$$\sup_{(z,\upsilon) \in \mathcal{A}_T \cap \mathcal{B}_T} \left| \frac{\nabla_{\theta_j} \pi_{T,S}(z \mid \upsilon;\theta_0) \pi_{1T}(\upsilon) \xi_{1T}(\upsilon)}{\pi_{2T}(z,\upsilon)} - \frac{\nabla_{\theta_j} \pi(z \mid \upsilon;\theta_0) \xi_1(\upsilon)}{\pi(z \mid \upsilon;\theta_0)} \right|$$
$$= O_p\left(T^{-\frac{1}{2}} \lambda_T^{-q-1} \alpha_T^{-1} \delta_T^{-2}\right) + O_p\left(T^{-\frac{1}{2}} \lambda_T^{-(q-q^*)-1} \alpha_T^{-1} \delta_T^{-2}\right) + O_p\left(T^{-\frac{1}{2}} \lambda_T^{-(q-q^*)} \alpha_T^{-1} \delta_T^{-3}\right)$$
$$+ O_p\left(\lambda_T^r \alpha_T^{-1} \delta_T^{-3}\right) + O_p\left(T^{-\frac{1}{2}} \lambda_T^{-(q-q^*)} \alpha_T^{-1}\right) + O_p\left(T^{-\frac{1}{2}} \lambda_T^{-q} \alpha_T^{-2}\right) + O_p\left(\lambda_T^r \alpha_T^{-2}\right).$$

Lemma N5. *Let the assumptions in Lemma* N1 *hold, and let* $\xi_{1T}(\upsilon)$ *be the sequence of functions in Lemma* N4.

Then, for all $i = 1, \ldots, S$, *and* $j = 1, \ldots, n$,

$$\sup_{(z,v)\in\mathcal{A}_T\cap\mathcal{B}_T}\left|\frac{\nabla_{\theta_j}\pi_{T,S}(z\mid v;\theta_0)E\left(\pi_{2T}(z,v)\right)\xi_{1T}(v)\pi_{1T}(v)}{\pi^i_{1T}(v;\theta_0)\pi_{2T}(z,v)}-\nabla_{\theta_j}\pi(z\mid v;\theta_0)\xi_1(v)\right|$$

$$= O_p\left(T^{-\frac{1}{2}}\lambda_T^{-q-1}\alpha_T^{-1}\delta_T^{-3}\right)+O_p\left(T^{-\frac{1}{2}}\lambda_T^{-q}\alpha_T^{-2}\delta_T^{-1}\right)+O_p\left(T^{-\frac{1}{2}}\lambda_T^{-(q-q^*)}\alpha_T^{-1}\delta_T^{-4}\right)$$

$$+O_p\left(\lambda_T^r\alpha_T^{-1}\delta_T^{-4}\right)+O_p\left(\lambda_T^r\alpha_T^{-2}\delta_T^{-1}\right)+O_p\left(\lambda_T^r\delta_T^{-1}\alpha_T^{-1}\right).$$

Remarks on Lemmas N4–N5.

(a) Lemma N4 is needed to show that the term $\mathcal{J}_{T,S}$ defined in equation (B5a) below converges in probability to the term $\mathcal{J}$ defined in equation (B6a) below. Lemma N5 is needed to show that the terms $\mathcal{I}^i_{1T,S}$ and $\mathcal{I}^i_{2T,S}$ defined in equations (B5b) and (B5c) below converge in distribution to the Gaussian terms provided in equations (B6b) and (B6c) below.
(b) The bandwidth conditions in Assumption 10(b) ensure that the suprema in Lemmas N4 and N5 go to 0 in probability, as shown in the next remarks.

Remarks on Assumption 10(b). Below, we show that under Assumption 10(b), the conditions $\alpha_T^3\lambda_T^{-r}\to\infty$ (in Lemma C2(b)) and $\alpha_T^2\delta_T\lambda_T^{-r}\to\infty$ and $\alpha_T^2\delta_T^2\lambda_T^{-r}\to\infty$ (in Lemma C3(b)) hold true. It is also easily seen that all the suprema in Lemmas N4 and N5 go to 0 in probability under the conditions that $\lambda_T\to 0$, $\alpha_T\to 0$, $\delta_T\to 0$, and,

(i) $T^{\frac{1}{2}}\lambda_T^q\alpha_T^3\to\infty$
(ii) $T^{\frac{1}{2}}\lambda_T^q\alpha_T^2\delta_T\to\infty$
(iii) $T^{\frac{1}{2}}\lambda_T^{q+1}\alpha_T\delta_T^3\to\infty$
(iv) $T^{\frac{1}{2}}\lambda_T^{q-q^*}\alpha_T\delta_T^4\to\infty$
(v) $\lambda_T^r\alpha_T^{-3}\to 0$
(vi) $\lambda_T^r\alpha_T^{-2}\delta_T^{-2}\to 0$
(vii) $\lambda_T^r\alpha_T^{-1}\delta_T^{-4}\to 0$

If $\alpha_T\to 0$, $\delta_T\to 0$, and $\delta_T/\alpha_T\to\kappa$, as required by Assumption 10, the previous conditions can be simplified to, (i) $T^{\frac{1}{2}}\lambda_T^q\alpha_T^4\to\infty$ (also required by Assumption 10(b)); (ii) $T^{\frac{1}{2}}\lambda_T^{q-q^*}\alpha_T^5\to\infty$, and (iii) $\lambda_T^r\alpha_T^{-5}\to 0$. By the same arguments produced in the Remarks on Assumption 10(a) in Appendix B.1, one has that (ii) and (iii) are satisfied if $\alpha_T^{-1}\lambda_T^{\min\{q^*+1,\frac{1}{5}r\}}\to 0$, as required by Assumption 10(b). Clearly, these conditions also imply that $\alpha_T^3\lambda_T^{-3}\to\infty$ (in Lemma C2(b)) and $\alpha_T^2\delta_T\lambda_T^{-r}\to\infty$ (in Lemma C3(b)) (with $\delta_T/\alpha_T\to\kappa$), as we initially claimed.

Proof of asymptotic normality. The proof is only sketched as it relies on arguments similar to those we produced in Appendix A.3. The extensive details of the proof are in Al-M08 (section D.2). Let $\xi(z,v)\equiv\pi_2(z,v;\theta_0)w(z,v)/\pi_1(v;\theta_0)^2$, and consider the definition of $\gamma(v)$ in Appendix A.3 (see equation (A15)). By plugging $\xi(z,v)$ into equation (A15) yields

$$\gamma(v)=\int\nabla_\theta\pi(z\mid v;\theta_0)\xi(z,v)dz. \tag{B4}$$

Next, let

$$\mathbb{W}_T\equiv\left\{w_T(z,v):w_T(z,v)=\xi_{1T}(v)\frac{\pi_{1T}(v)^2}{\pi_{2T}(z,v)}\mathbb{T}_{2T}(z,v)\right\},$$

where the function ξ_{1T} satisfies the conditions in Lemma N4. We study the asymptotic behaviour of the CD-SNE for the weighting functions $w_T\in\mathbb{W}_T$. For every $w_T\in\mathbb{W}_T$, the terms $\mathcal{I}^i_{1T,S}$, $\mathcal{I}^i_{2T,S}$ and $\mathcal{J}_{T,S}$ in equations (A13), (A14a), and

(A14b) of Appendix A.3 become

$$\mathcal{J}_{T,S} = \iint \left| \frac{\nabla_\theta \pi_{T,S}(z \mid v;\theta_0)\pi_{1T}(v)}{\pi_{2T}(z,v)} \mathbb{T}_{T,S}(v;\theta_0)\sqrt{\mathbb{T}_{2T}(z,v)} \right|_2 \xi_{1T}(v)\pi_{2T}(z,v)dzdv \tag{B5a}$$

$$\mathcal{I}^i_{1T,S} = \iint \frac{\nabla_\theta \pi_{T,S}(z \mid v;\theta_0)\pi_{1T}(v)^2}{\pi^i_{1T}(v;\theta_0)\pi_{2T}(z,v)} \xi_{1T}(v)\mathbb{T}_{2T}(z,v)\mathbb{T}^2_{T,S}(v;\theta_0)dA^i_T(z,v) \tag{B5b}$$

$$\mathcal{I}^i_{2T,S} = \iint \frac{\nabla_\theta \pi_{T,S}(z \mid v;\theta_0)E(\pi_{2T}(z,v))\pi_{1T}(v)}{\pi^i_{1T}(v;\theta_0)\pi_{2T}(z,v)} \xi_{1T}(v)\mathbb{T}_{2T}(z,v)\mathbb{T}^2_{T,S}(v;\theta_0)dzdA^i_T(v). \tag{B5c}$$

By Lemmas N4 and N5, the bandwidth conditions in Assumption 10(b), and lengthy computations in Al-M08 (section D.2),

$$\mathcal{J}_{T,S} \xrightarrow{p} \mathcal{J} \equiv \iint \left| \frac{\nabla_\theta \pi(z \mid v;\theta_0)}{\pi(z \mid v;\theta_0)} \right|_2 \xi_1(v)\pi_2(z,v;\theta_0)dzdv \tag{B6a}$$

$$\mathcal{I}^i_{1T,S} \xrightarrow{d} \mathcal{I}^i_1 \equiv \iint \frac{\nabla_\theta \pi(z \mid v;\theta_0)}{\pi(z \mid v;\theta_0)} \xi_1(v)d\omega^0_i(F(z,v)), \qquad i = 0,1,\ldots,S \tag{B6b}$$

$$\mathcal{I}^i_{2T,S} \xrightarrow{d} \mathcal{I}^i_2 \equiv \int \gamma(v)d\hat{\omega}^0_i(F_1(v)), \quad i = 0,1,\ldots,S \tag{B6c}$$

Moreover, for any $w_T \in \mathbb{W}_T$, the limiting function in (B4) $\xi(z,v) = \xi_1(v)$ and, hence, $\gamma(v) = \xi_1(v)\int \nabla_\theta \pi(z \mid v;\theta_0)dz = 0$, for all $v \in \mathbb{R}^{q-q^*}$, or $\mathcal{I}^i_2 \equiv 0$, for $i = 0,1,\ldots,S$. Therefore, the next result follows by the same arguments in Appendix A.3, the assumption that $E[\|\nabla_\theta \log \pi(z_t|v_t;\theta_0)\|^\vartheta]^{1/\vartheta} < \infty$, for some $\vartheta > 2$ (and boundedness of $\xi_1(v)$), and the mixing condition in Assumption 2:

Proposition 2. *Under the Assumptions of Theorem* 2*, the CD-SNE with weighting functions $w_T \in \mathbb{W}_T$ is consistent and asymptotically normal with variance/covariance matrix*

$$\left(1+\frac{1}{S}\right)\left(\mathrm{var}(\Phi_t) + \sum_{k=1}^{\infty}\left[\mathrm{cov}(\Phi_t, \Phi_{t+k}) + \mathrm{cov}(\Phi_{t+k}, \Phi_t)\right]\right),$$

where $\Phi_t \equiv \Phi(z_t, v_t)$ and,

$$\Phi(z,v) \equiv \left[\iint \left| \frac{\nabla_\theta \pi(s'|s;\theta_0)}{\pi(s'|s;\theta_0)} \right|_2 \xi_1(s)\pi_2(s',s;\theta_0)ds'ds\right]^{-1} \frac{\nabla_\theta \pi(z \mid v;\theta_0)}{\pi(z \mid v;\theta_0)}\xi_1(v).$$

Theorem 2 is a special case of Proposition 2 with $\xi_1(\cdot) = \xi_{1T}(\cdot) \equiv 1$ and $(z,v) = (y_2, y_1)$. The efficiency claim follows by the standard score martingale difference argument (see, for example, Wooldridge, 1994, lemma 5.2, p. 2677).

APPENDIX C. ASYMPTOTICS FOR THE ESTIMATOR IN SECTION 2.3

In Al-M08 (section C.3), we show that the estimator $\hat{\theta}_{T,S}$ in (13) is weakly consistent and asymptotically normal with variance/covariance matrix equal to,

$$\left(1+\frac{1}{S}\right)\hat{\mathcal{J}}^{-1}\hat{V}\hat{\mathcal{J}}^{\top -1},$$

where

$$\hat{\mathcal{J}} = \sum_{k=1}^{l}\int_{\mathbb{R}^{2q^*}} |\nabla_\theta \pi(y^{\mathrm{o}}|y^{\mathrm{o}}_{-k};\theta_0)|_2 w(y^{\mathrm{o}}, y^{\mathrm{o}}_{-k})dy^{\mathrm{o}}dy^{\mathrm{o}}_{-k},$$

$$\hat{V} = \mathrm{var}(\hat{\Upsilon}(y^{\mathrm{o}}_t, \cdots, y^{\mathrm{o}}_{t-l}))$$

$$+\sum_{k=1}^{\infty}[\mathrm{cov}(\hat{\Upsilon}(y^{\mathrm{o}}_t, \cdots, y^{\mathrm{o}}_{t-l}), \hat{\Upsilon}(y^{\mathrm{o}}_{t+k}, \cdots, y^{\mathrm{o}}_{t+k-l})) + \mathrm{cov}(\hat{\Upsilon}(y^{\mathrm{o}}_{t+k}, \cdots, y^{\mathrm{o}}_{t+k-l}), \hat{\Upsilon}(y^{\mathrm{o}}_t, \cdots, y^{\mathrm{o}}_{t-l}))],$$

and

$$\hat{\Upsilon}(y_t^{\mathrm{o}}, \cdots, y_{t-l}^{\mathrm{o}}) = \sum_{k=1}^{l} [\eta(y_t^{\mathrm{o}}, y_{t-k}^{\mathrm{o}}) + \gamma(y_{t-k}^{\mathrm{o}})],$$

where $\eta(\cdot)$ and $\gamma(\cdot)$ are as in (22), with y_t^{o} and y_{t-k}^{o} replacing z_t and v_t.

APPENDIX D. PROOF OF THEOREM 3

The proof of Theorem 3 proceeds along similar lines as those in the proof of Theorems 1 and 2, and is provided in Al-M08 (section B).

Acknowledgements. We wish to thank the editor, Bernard Salanié, and two anonymous referees for very helpful comments and suggestions; and Yacine Aït-Sahalia, Torben Andersen, Alessandro Beber, Marine Carrasco, Mikhail Chernov, Frank Diebold, Walter Distaso, Cristian Huse, Dennis Kristensen, Oliver Linton, Nour Meddahi, Angelo Melino, Alex Michaelides, Eric Renault, Michael Rockinger, Christopher Sims, Sam Thompson, seminar participants at CORE, the LSE, the Norwegian School of Economics and Business, Princeton University, the University of Pennsylvania and, especially, Valentina Corradi for her valuable comments and Fabio Fornari who worked with us in the initial stages of the paper. The usual disclaimer applies.

REFERENCES

AI, C. (1997), "A Semiparametric Maximum Likelihood Estimator", *Econometrica*, **65**, 933–963.
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).
AïT-SAHALIA, Y., FAN, J. and PENG, H. (2005), "Nonparametric Transition-Based Tests for Diffusions" (Working Paper, Princeton University).
ALTISSIMO, F. and MELE, A. (2008), "Simulated Nonparametric Estimation of Dynamic Models: Unpublished Appendix" (Working Paper, London School of Economics).
ANDREWS, D. W. K. (1995), "Nonparametric Kernel Estimation for Semiparametric Models", *Econometric Theory*, **11**, 560–596.
ANTOINE, B., BONNAL, H. and RENAULT, E. (2007), "On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood", *Journal of Econometrics*, **138**, 461–487.
ARCONES, M. A. and YU, B. (1994), "Central Limit Theorems for Empirical and U-Processes of Stationary Mixing Sequences", *Journal of Theoretical Probability*, **7**, 47–71.
BERAN, R. (1977), "Minimum Hellinger Distance Estimates for Parametric Models", *Annals of Statistics*, **5**, 445–463.
BICKEL, P. J. and ROSENBLATT, M. (1973), "On Some Global Measures of the Deviations of Density Function Estimates", *Annals of Statistics*, **1**, 1071–1095.
BIERENS, H. J. (1983), "Uniform Consistency of Kernel Estimators of a Regression Function Under Generalized Conditions", *Journal of the American Statistical Association*, **78**, 699–707.
BILLIO, M. and MONFORT, A. (2003), "Kernel-Based Indirect Inference", *Journal of Financial Econometrics*, **1**, 297–326.
CARRASCO, M. and FLORENS, J.-P. (2004), "On the Asymptotic Efficiency of GMM" (Working Paper, University of Rochester).
CARRASCO, M., CHERNOV, M., FLORENS, J.-P. and GHYSELS, E. (2006), "Efficient Estimation of Jump-Diffusions and General Dynamic Models with a Continuum of Moment Conditions", *Journal of Econometrics* (forthcoming).
CHEN, X., LINTON, O. and ROBINSON, P. M. (2001), "The Estimation of Conditional Densities", *Journal of Statistical Planning and Inference*, Special Issue in Honor of George Roussas, 71–84.
CRESSIE, N. and READ, T. R. C. (1984), "Multinomial Goodness-of-Fit Tests", *Journal of the Royal Statistics Society, Series B*, **46**, 440–464.
DUFFIE, D. and SINGLETON, K. J. (1993), "Simulated Moments Estimation of Markov Models of Asset Prices", *Econometrica*, **61**, 929–952.
ELERIAN, O., CHIB, S. and SHEPHARD, N. (2001), "Likelihood Inference for Discretely Observed Nonlinear Diffusions", *Econometrica*, **69**, 959–993.
FAN, Y. (1994), "Testing the Goodness-of-Fit of a Parametric Density Function by Kernel Method", *Econometric Theory*, **10**, 316–356.
FERMANIAN, J.-D. and SALANIÉ, B. (2004), "A Nonparametric Simulated Maximum Likelihood Estimation Method", *Econometric Theory*, **20**, 701–734.
GALLANT, A. R. (2001), "Effective Calibration" (Working Paper, University of North Carolina).

GALLANT, A. R. and TAUCHEN, G. (1996), "Which Moments to Match?", *Econometric Theory*, **12**, 657–681.
GOURIÉROUX, C., MONFORT, A. and RENAULT, E. (1993), "Indirect Inference", *Journal of Applied Econometrics*, **8**, S85–S118.
HAJIVASSILIOU, V. and McFADDEN, D. (1998), "The Method of Simulated Scores for the Estimation of Limited-Dependent Variable Models", *Econometrica*, **66**, 863–896.
HANSEN, L. and SCHEINKMAN, J. A. (1995), "Back to the Future: Generating Moment Implications for Continuous-Time Markov Processes", *Econometrica*, **63**, 767–804.
HÄRDLE, W. and MAMMEN, E. (1993), "Comparing Nonparametric versus Parametric Regression Fits", *Annals of Statistics*, **21**, 1926–1947.
HONG, Y. and WHITE, H. (2005), "Asymptotic Distribution Theory for Nonparametric Entropy Measures of Serial Dependence", *Econometrica*, **73**, 837–901.
IMBENS, G. W., SPADY, R. H. and JOHNSON, P. (1998), "Information Theoretic Approaches to Inference in Moment Conditions Models", *Econometrica*, **66**, 333–357.
KITAMURA, Y., TRIPATHI, G. and AHN, H. (2004), "Empirical Likelihood-Based Inference in Conditional Moment Restriction Models", *Econometrica*, **72**, 1667–1714.
KLOEDEN, P. E. and PLATEN, E. (1999), *Numerical Solutions of Stochastic Differential Equations* (Berlin: Springer Verlag).
KRISTENSEN, D. and SHIN, Y. (2006), "Estimation of Dynamic Models with Nonparametric Simulated Maximum Likelihood" (Working Paper, Columbia University).
LAROQUE, G. and SALANIÉ, B. (1989), "Estimation of Multimarket Fix-Price Models: An Application of Pseudo-Maximum Likelihood Methods", *Econometrica*, **57**, 831–860.
LAROQUE, G. and SALANIÉ, B. (1993), "Simulation-Based Estimation of Models with Lagged Latent Variables", *Journal of Applied Econometrics*, **8**, S119–S133.
LAROQUE, G. and SALANIÉ, B. (1994), "Estimating the Canonical Disequilibrium Model: Asymptotic Theory and Finite Sample Properties", *Journal of Econometrics*, **62**, 165–210.
LEE, L. F. (1995), "Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models", *Econometric Theory*, **11**, 437–483.
LEE, B.-S. and INGRAM, B. F. (1991), "Simulation Estimation of Time-Series Models", *Journal of Econometrics*, **47**, 197–207.
LINDSAY, B. G. (1994), "Efficiency versus Robustness: The Case for Minimum Hellinger Distance and Related Methods", *Annals of Statistics*, **22**, 1081–1114.
LINTON, O. and XIAO, Z. (2000), "Second Order Approximation for Adaptive Regression Estimators" (Working Paper, London School of Economics).
McFADDEN, D. (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration", *Econometrica*, **57**, 995–1026.
NEWEY, W. K. (1991), "Uniform Convergence in Probability and Stochastic Equicontinuity", *Econometrica*, **59**, 1161–1167.
NEWEY, W. K. and McFADDEN, D. L. (1994), "Large Sample Estimation and Hypothesis Testing", in R. F. Engle and D. L. McFadden (eds.) *Handbook of Econometrics*, Vol. 4, ch. 36 (Amsterdam: Elsevier) 2111–2245.
PAGAN, A. and ULLAH, A. (1999), *Nonparametric Econometrics* (Cambridge: Cambridge University Press).
PAKES, A. and POLLARD, D. (1989), "Simulation and the Asymptotics of Optimization Estimators", *Econometrica*, **57**, 1027–1057.
PASTORELLO, S., PATILEA, V. and RENAULT, E. (2003), "Iterative and Recursive Estimation in Structural Non Adaptive Models", *Journal of Business and Economic Statistics*, **21**, 449–509.
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.
POLITIS, D. N. and ROMANO, J. P. (1994), "Limit Theorems for Weakly Dependent Hilbert Space Valued Random Variables with Applications to the Stationary Bootstrap", *Statistica Sinica*, **4**, 461–476.
SANTA-CLARA, P. (1995), "Simulated Likelihood Estimation of Diffusions With an Application to the Short Term Interest Rate" (Ph.D. dissertation, INSEAD).
SILVERMAN, B. W. (1986), *Density Estimation for Statistics and Data Analysis* (London: Chapman and Hall).
SINGLETON, K. J. (2001), "Estimation of Affine Asset Pricing Models Using the Empirical Characteristic Function", *Journal of Econometrics*, **102**, 111–141.
SINGLETON, K. J. (2006), *Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment* (Princeton: Princeton University Press).
SMITH, A. (1993), "Estimating Nonlinear Time Series Models Using Simulated Vector Autoregressions", *Journal of Applied Econometrics*, **8**, S63–S84.
TAUCHEN, G. (1997), "New Minimum Chi Square Methods in Empirical Finance", in D. Kreps and K. Wallis (eds.) *Advances in Econometrics 7th World Congress* (Cambridge: Cambridge University Press) 279–317.
WHITE, H. (1994), *Estimation, Inference and Specification Analysis* (Cambridge: Cambridge University Press).
WOOLDRIDGE, J. M. (1994), "Estimation and Inference for Dependent Processes", in R. F. Engle and D. L. McFadden (eds.) *Handbook of Econometrics*, Vol. 4, ch. 45 (Amsterdam: Elsevier) 2639–2738.