

# Estimation of Dynamic Models with Nonparametric Simulated Maximum Likelihood

Dennis Kristensen\*      Yongseok Shin†

September 2008

## Abstract

We propose a simulated maximum likelihood estimator for dynamic models based on nonparametric kernel methods. Our method is designed for models without latent dynamics from which one can simulate observations but cannot obtain a closed-form representation of the likelihood function. Using the simulated observations, we nonparametrically estimate the density—which is unknown in closed form—by kernel methods, and then construct a likelihood function that can be maximized. We prove for dynamic models that this nonparametric simulated maximum likelihood (NPSML) estimator is consistent and asymptotically efficient. NPSML is applicable to general classes of models and is easy to implement in practice.

---

\*Department of Economics, Columbia University (e-mail: [dk2313@columbia.edu](mailto:dk2313@columbia.edu)).

†Department of Economics, Washington University in St. Louis and University of Wisconsin-Madison (e-mail: [yshin@wustl.edu](mailto:yshin@wustl.edu)).

# 1 Introduction

We propose a simulated maximum likelihood estimator for dynamic models based on nonparametric kernel methods. Our method is designed for models without latent dynamics from which one can simulate observations but cannot obtain a closed-form representation of the likelihood function. For any given parameter value, conditioning on available information, we draw  $N$  i.i.d. simulated observations from the model. We then use these simulated observations to nonparametrically estimate the conditional density—unknown in closed form—by kernel methods. The kernel estimate converges to the true conditional density as  $N$  goes to infinity, enabling us to approximate the true density arbitrarily well with a sufficiently large  $N$ . We then construct the likelihood and search over the parameter space to obtain a maximum likelihood estimator—nonparametric simulated maximum likelihood (NPSML) estimator.

NPSML was introduced by Fermanian and Salanié (2004), who obtained theoretical results only for static models. In this paper, we develop and generalize their method to dynamic models, including nonstationary and time-inhomogeneous ones. We give general conditions for the NPSML estimator to be consistent and have the same asymptotic distribution as the infeasible maximum likelihood estimator (MLE). For the stationary case, we also analyze the impact of simulations on the bias and variance of the NPSML estimator.

NPSML can be used for estimating general classes of models, such as structural Markov decision processes and discretely-sampled diffusions. As for Markov decision processes, the transition density of endogenous state variables embodies an optimal policy function of a dynamic programming problem, and hence does not typically have a closed-form representation (Doraszelski and Pakes, 2007; Rust, 1994). However, we can closely approximate the optimal policy function numerically, and simulate observations from the model for NPSML. Similarly, as for the estimation of continuous-time stochastic models with discretely-sampled data, the transition densities are well-defined, but only in few special cases can we derive closed-form expressions for them. Again, diffusion processes can be approximated with various discretization schemes to a given level of precision, and hence we can simulate observations from the model which are then used for NPSML.

For the classes of models that NPSML addresses, there are two categories of existing approaches. The first is based on moment matching, and includes simulated methods of moments (Duffie and Singleton, 1993; Lee and Ingram, 1991; McFadden, 1989; Pakes and Pollard, 1989), indirect inference (Gouriéroux et al., 1993; Smith, 1993), and efficient methods of moments (Gallant and Tauchen, 1996). These are all general-purpose methods, but cannot attain asymptotic efficiency—even for models that are Markov in observables—unless the true score is encompassed by the target moments (Tauchen, 1997). More recently, Altissimo and Mele (2008) and Carrasco et al. (2007) developed general-purpose estimators based on matching a continuum of moments that are asymptotically as efficient as maximum likelihood estimators for fully observed systems. One attractive feature

of NPSML—which it shares with Altissimo and Mele (2008) and Carrasco et al. (2007)—is that asymptotic efficiency is attained without having to judiciously choose an auxiliary model. For NPSML, the researcher has to choose a kernel and a bandwidth for the nonparametric estimation of transition densities. However, there exist many data-driven methods that guide the researcher in this regard such that our method can be made fully automated while yielding full efficiency. Another advantage is that, unlike most of the above methods (Altissimo and Mele, 2008; Carrasco et al., 2007; Gallant and Tauchen, 1996; Gouriéroux et al., 1993; Smith, 1993), NPSML can handle nonstationary and time-inhomogeneous dynamics.

The approaches in the second category approximate the likelihood function itself, and hence is more closely related to NPSML. Examples of this approach include the simulated likelihood method (Lee, 1995), and the method of simulated scores (Hajivassiliou and McFadden, 1998), both of which are designed for limited dependent variable models. Another set of examples are various maximum likelihood methods for discretely sampled diffusions (Aït-Sahalia, 2002, 2004; Brandt and Santa-Clara, 2002; Elerian et al., 2001; Pedersen, 1995a,b; Sandmann and Koopman, 1998).<sup>1</sup> While all these methods result in asymptotically efficient estimators, they are designed only for specific classes of models—i.e. limited dependent variable models or diffusions, and cannot be adapted easily to other classes of models. NPSML is for general purposes in both theoretical and practical senses. Theoretically, we establish its asymptotic properties under fairly weak regularity conditions allowing for a wide range of different models. At the practical level, when the model specification changes, only the part of the computer code that generates simulated observations needs to be modified, leaving other parts (e.g. kernel estimation of conditional density or numerical maximization of likelihood) unchanged.

Throughout this paper, we assume that it is possible to simulate the current variables of the model conditioning on finitely-many past observations. This excludes cases with latent dynamics since these cannot be simulated one step at a time. Extensions to methods with built-in nonlinear filters that explicitly account for latent dynamics are worked out in a companion paper (Kristensen and Shin, 2007) building on the main results obtained here.

The rest of the paper is organized as follows. In the next section, we set up our framework to present the simulated conditional density and the associated NPSML estimator. In Section 3, we derive the asymptotic properties of the NPSML estimator under regularity conditions. Section 4 provides a detailed description on implementing NPSML with a numerical example, and Section 5 concludes.

---

<sup>1</sup>Obviously, we are citing only a small subset of methods for diffusion estimation—namely, those that maximize approximated likelihood and that are hence most closely related to NPSML. It should be noted that, unlike the others, Aït-Sahalia (2002, 2004) use analytic expansions of the transition density and forgo simulations. Markov chain Monte Carlo methods are widely used for Bayesian estimation of diffusions. Elerian, Chib, and Shephard (2001) is a representative example, and Johannes and Polson (2005) provide a broad overview of such methods.

## 2 Nonparametric Simulated Maximum Likelihood

### 2.1 Construction of NPSML Estimator

Suppose that we have  $T$  observations,  $\{(y_t, x_t)\}_{t=1}^T$ ,  $y_t \in \mathbb{R}^k$  and  $x_t \in \mathcal{X}_t$ . The space  $\mathcal{X}_t$  can be time-varying. We assume that the data has been generated by a fully parametric model:

$$y_t = g_t(x_t, \varepsilon_t; \theta), \quad t = 1, \dots, T, \quad (1)$$

where  $\theta \in \Theta \subseteq \mathbb{R}^d$  is an unknown parameter vector, and  $\varepsilon_t | x_t \sim F_\varepsilon$ . Assume that  $F_\varepsilon$  is known and does not depend on  $\theta$ .<sup>2</sup> One general class of such models is the one with  $x_t \equiv y_{t-1}$ , such that  $\{y_t\}$  is a (possibly time-inhomogeneous) Markov process. In this case (1) is a fully specified model. However, we allow  $x_t$  to contain other (exogenous) variables than lagged values of  $y_t$ , in which case (1) is only a partially specified model. Also, we allow the process  $z_t$  to be nonstationary, either due to unit-root-type behaviour or due to time-dependence of  $g_t$ .

The model is assumed to have an associated conditional density  $p_t(y|x; \theta)$ . That is,

$$P(y_t \in A | x_t = x) = \int_A p_t(y|x; \theta) dy, \quad t = 1, \dots, T,$$

for any Borel set  $A \subseteq \mathbb{R}^k$ . A natural estimator of  $\theta$  is then the maximizer of the conditional log-likelihood:

$$\tilde{\theta} = \arg \max_{\theta \in \Theta} L_T(\theta), \quad L_T(\theta) = \sum_{t=1}^T \log p_t(y_t | x_t; \theta).$$

If the model (1) is fully specified, i.e.  $x_t = y_{t-1}$ , then this is the full likelihood of the model conditional on the starting value. If on the other hand,  $x_t$  contains other variables than lagged values of  $y_t$ ,  $L_T(\theta)$  is a partial likelihood.

Suppose now that  $p_t(y|x; \theta)$  does not have a closed-form representation, and thus the maximum likelihood estimation of  $\theta$  is not feasible. In terms of the model (1), this normally occurs when either the inverse of  $g_t(x_t, \varepsilon_t; \theta)$  w.r.t.  $\varepsilon_t$  does not exist, or that the inverse does not have a closed-form expression.<sup>3</sup> Such a situation may arise, for example, when the function  $g$  involves a solution to a dynamic programming problem, or when we are dealing with discretely-sampled diffusions. In such cases, although  $p_t(y|x; \theta)$  is not available in closed form, we are still able to generate simulated observations from the model: A solution to a dynamic programming problem can be represented numerically, and a diffusion can be approximated by various discretization schemes up to a given level of precision.

---

<sup>2</sup>We can actually allow the distribution  $F_\varepsilon$  to be time-varying and dependent on  $x_t$  as well—i.e.  $F_\varepsilon(\cdot) \equiv F_\varepsilon(\cdot, t, x_t)$ . For simplicity, we do not consider such cases here.

<sup>3</sup>If the inverse has a closed-form expression, we have  $p_t(y|x; \theta) = p_\varepsilon(g_t^{-1}(y, x; \theta)) \left| \frac{\partial g_t^{-1}(y, x; \theta)}{\partial y} \right|$ , and the likelihood is easily evaluated.

We here propose a general method to obtain a simulated conditional density, which in turn will be used to obtain a simulated version of the maximum likelihood estimator. For any given  $t \geq 1$ ,  $y \in \mathbb{R}^k$ ,  $x \in \mathcal{X}_t$ , and  $\theta \in \Theta$ , we wish to compute a simulated version of  $p_t(y|x;\theta)$ . To this end, we first generate  $N$  i.i.d. draws from  $F_\varepsilon$ ,  $\{\varepsilon_i\}_{i=1}^N$ , through a random number generator, and use these to obtain:

$$Y_{t,i}^{x,\theta} = g_t(x, \varepsilon_i; \theta), \quad i = 1, \dots, N.$$

By construction, the  $N$  simulated i.i.d. random variables,  $\{Y_{t,i}^{x,\theta}\}_{i=1}^N$ , follow the target distribution:  $Y_{t,i}^{x,\theta} \sim p_t(\cdot|x;\theta)$ ,  $i = 1, \dots, N$ . They can therefore be used to estimate  $p_t(y|x;\theta)$  with kernel methods. Define:

$$\hat{p}_t(y|x;\theta) = \frac{1}{N} \sum_{i=1}^N K_h(Y_{t,i}^{x,\theta} - y), \quad (2)$$

where  $K_h(v) = K(v/h)/h^k$ ,  $K : \mathbb{R}^k \mapsto \mathbb{R}$  is a kernel, and  $h > 0$  a bandwidth.<sup>4</sup> Under regularity conditions on  $p_t$  and  $K$ , we obtain:

$$\hat{p}_t(y|x;\theta) = p_t(y|x;\theta) + O_P(1/\sqrt{Nh^k}) + O_P(h^2), \quad N \rightarrow \infty,$$

where the remainder terms are  $o_P(1)$  if  $h \rightarrow 0$  and  $Nh^k \rightarrow \infty$ .

Once (2) has been used to obtain the simulated conditional density, we can now use it to construct the following simulated MLE of  $\theta_0$ :

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \hat{L}_T(\theta), \quad \hat{L}_T(\theta) = \sum_{t=1}^T \log \hat{p}_t(y_t|x_t;\theta).$$

When searching for  $\hat{\theta}$  through numerical optimization, we use the same draws for all values of  $\theta$ . We may also use the same batch of draws from  $F_\varepsilon(\cdot)$ ,  $\{\varepsilon_i\}_{i=1}^N$ , across different values of  $t$  and  $x$ .

Since  $\hat{p}_t(y|x;\theta) \xrightarrow{P} p_t(y|x;\theta)$ , the simulated likelihood  $\hat{L}_T(\theta) \xrightarrow{P} L_T(\theta)$  as  $N \rightarrow \infty$  for a given  $T \geq 1$  under regularity conditions. The main theoretical results of this paper demonstrate that  $\hat{\theta}$  inherits the properties of the infeasible MLE,  $\tilde{\theta}$ , as  $T, N \rightarrow \infty$ , under suitable conditions.

Let us note the following two points. Firstly, the usual curse of dimensionality for nonparametric estimators depends only on  $k \equiv \dim(y_t)$  here, and the dimension of  $x_t$  is irrelevant in itself. Secondly, because we use i.i.d. draws, the density estimator is not affected by the dependence structure in the observed data. In particular, our estimator works equally well whether the observed data are i.i.d. or nonstationary.

Numerical optimization is facilitated if  $\hat{L}_T(\theta)$  is continuous and differentiable in  $\theta$ . With (2), if  $K$  and  $\theta \mapsto g_t(x, \varepsilon; \theta)$  are  $r \geq 0$  times continuously differentiable, then  $\hat{L}_T(\theta)$  has the same

---

<sup>4</sup>Here and in the following, we will use  $K$  to denote a generic kernel.

property. This follows from the chain rule and the fact that we use the same random draws  $\{\varepsilon_i\}_{i=1}^N$  for all values of  $\theta$ .

A disadvantage of our estimator is that, for a finite  $N$  and a fixed  $h > 0$ , the simulated log-likelihood function is a biased estimate of the actual one. To obtain consistency, we will have to let  $N \rightarrow \infty$  which is a feature that is shared by most other simulated likelihood methods.<sup>5</sup> This is in contrast to, for example, simulated methods of moment, where unbiased estimators of moments can be constructed, and consistency therefore be obtained for a fixed  $N$ . In addition to this, we also have to require  $h \rightarrow 0$  to obtain consistency. However, if one is willing to make a stronger assumption about the identification of the model, this issue can be partially avoided. For example, in the stationary case, the standard identification assumption is

$$\mathbb{E} [\log p(y_t|x_t; \theta)] < \mathbb{E} [\log p(y_t|x_t; \theta_0)], \quad \theta \neq \theta_0.$$

A stronger identification condition implying the former is

$$\mathbb{E} \left[ \log \left( \int K(v) p(y_t + hv|x_t; \theta) dv \right) \right] < \mathbb{E} \left[ \log \left( \int K(v) p(y_t + hv|x_t; \theta_0) dv \right) \right], \quad \theta \neq \theta_0,$$

for all  $0 \leq h \leq \bar{h}$  for some  $\bar{h} > 0$ .<sup>6</sup> Under this identification condition, one can show consistency of our estimator for any fixed  $0 < h \leq \bar{h}$  as  $N \rightarrow \infty$ . A similar identification condition can be found in Altissimo and Mele (2008). Still, for a fixed  $h > 0$  the resulting estimator will no longer have full efficiency. To obtain this, one has to let  $h \rightarrow 0$ .

While we here focus on the kernel estimator, one can use other nonparametric density estimators as well. Examples are the semi-nonparametric estimators of Fenton and Gallant (1996), Gallant and Nychka (1987), Phillips (1983) and Wahba (1981); the log-spline estimator of Stone (1990); and the wavelet estimator of Donoho et al. (1996). What is needed is that the nonparametric estimator converges towards the true density sufficiently fast.

**Example: Discretely-Observed Jump Diffusion.** Consider an  $\mathbb{R}^k$ -dimensional continuous-time stochastic process  $\{y_t : t \geq 0\}$  that solves:

$$dy_t = \mu(t, y_t; \theta) dt + \Sigma(t, y_t; \theta) dW_t + J_t dQ_t. \quad (3)$$

Here, the model contains both continuous and jump components.  $W_t \in \mathbb{R}^l$  is a standard Brownian motion, while  $Q_t$  is an independent pure jump process with stochastic intensity  $\lambda(t, y_t; \theta)$  and jump

---

<sup>5</sup>See Lee and Song (2006) for an exception.

<sup>6</sup>This follows from the following inequality:

$$\begin{aligned} \mathbb{E} [\log p(y_t|x_t; \theta)] &= \lim_{h \rightarrow 0} \mathbb{E} \left[ \log \left( \int K(v) p(y_t + hv|x_t; \theta) dv \right) \right] < \lim_{h \rightarrow 0} \mathbb{E} \left[ \log \left( \int K(v) p(y_t + hv|x_t; \theta_0) dv \right) \right] \\ &= \mathbb{E} [\log p(y_t|x_t; \theta_0)]. \end{aligned}$$

size 1. The functions  $\mu : [0, \infty) \times \mathbb{R}^k \times \Theta \mapsto \mathbb{R}^k$  and  $\Sigma : [0, \infty) \times \mathbb{R}^k \times \Theta \mapsto \mathbb{R}^{k \times k}$  is the drift and the diffusion term respectively, while  $J_t$  measures the jump sizes and has density  $v(t, y_t; \theta)$ .

Such jump diffusions are widely used in finance to model the dynamics of stock prices, interest rates, exchange rates and so on (Sundaresan, 2000). Suppose we have a sample  $y_1, \dots, y_T$ —without loss of generality, we normalize the time interval between observations to 1—and wish to estimate  $\theta$  by maximum likelihood. Although under regularity conditions (Lo, 1988) the transition density  $P(y_{t+1} \in A | y_t = x) = \int_A p_t(y|x; \theta) dy$  is well-defined, it cannot be written in closed form.<sup>7</sup> However, discretization schemes (Bruti-Liberati and Platen, 2007; Kloeden and Platen, 1992) can be used to simulate observations from the model for any given level of accuracy, hence enabling NPSML. We re-visit this example in Section 4 where we provide a detailed description of implementing NPSML in practice.

## 2.2 Extensions and Alternative Schemes

**Discrete Random Variables.** Discrete random variables can be accommodated within our framework. Suppose  $y_t$  contains both continuous and discrete random variables. For example,  $y_t = (y_{1t}, y_{2t}) \in \mathbb{R}^{k+l}$  where  $y_{1t} \in \mathbb{R}^k$  is a continuous random variable while  $y_{2t} \in \mathcal{Y}_2 \subset \mathbb{R}^l$  is a random variable with (potentially infinite number of) discrete outcomes,  $\mathcal{Y}_2 = \{y_{2,1}, y_{2,2}, \dots\}$ . We could then use a mixed kernel to estimate  $p_t(y|x)$ . For given simulated observations  $Y_{t,i}^{x,\theta} = (Y_{1t,i}^{x,\theta}, Y_{2t,i}^{x,\theta})$ ,  $i = 1, \dots, N$ :

$$\hat{p}_t(y_1, y_2 | x; \theta) = \frac{1}{N} \sum_{i=1}^N K_h(Y_{1t,i}^{x,\theta} - y_1) \mathbb{I}\{Y_{2t,i}^{x,\theta} = y_2\}, \quad (y_1, y_2) \in \mathbb{R}^{k+l}, \quad (4)$$

where  $\mathbb{I}\{\cdot\}$  is the indicator function and  $K : \mathbb{R}^k \mapsto \mathbb{R}$  is the kernel from before. However, the resulting simulated log-likelihood will be discontinuous and optimization may be difficult. One could replace the indicator function used for the discrete component with a smoother. Examples of smoothers can be found in Cai et al. (2001) and Li and Racine (2007, Chapter 2). These will increase bias but reduce variance of the estimator, and at the same time lead to a continuous function. However, in general,  $Y_{2t,i}^{x,\theta}$  itself will not be continuous so either way, with a discrete component,  $\hat{L}_T(\theta)$  based on (4) is no longer continuous w.r.t.  $\theta$ .

Instead, we will here assume that there exists a function  $Y_{2t,i}^{x,\theta}(y_2) = g_2(y_2, x, \varepsilon; \theta)$  that is smooth in  $\theta$  such that

$$\mathbb{E} \left[ Y_{2t,i}^{x,\theta}(y_2) | Y_{1t,i}^{x,\theta} = y_1 \right] = p_t(y_2 | y_1, x; \theta). \quad (5)$$

---

<sup>7</sup>Schaumburg (2001) and Yu (2007) use analytic expansions to approximate the transition density for univariate and multivariate jump diffusions, respectively. Their asymptotic result requires that the sampling interval shrink to zero. The theoretical results of the simulated MLE of Brandt and Santa-Clara (2002) or Pedersen (1995a,b) need to be substantially modified before they can be applied to generalized Lévy processes.

Thus,  $Y_{2t,i}^{x,\theta}$  now denotes a simulated value of the associated density, and not the outcome of the dependent variable. We then propose to estimate the joint density by

$$\hat{p}_t(y_1, y_2|x; \theta) = \frac{1}{N} \sum_{i=1}^N K_h(Y_{1t,i}^{x,\theta} - y_1) Y_{2t,i}^{x,\theta}(y_2). \quad (6)$$

To motivate the above assumption and the resulting estimator, we first note that a discrete random variable can always be represented as  $y_{2,t} = D(z_t)$  for some continuous variables  $z_t \in \mathbb{R}^m$  and some function  $D: \mathbb{R}^m \mapsto \mathcal{Y}_2$  which we, for the sake of the argument, assume does not depend on  $(t, x, \theta)$ . For example, most limited dependent variables can be written on this form, c.f. Manrique and Shephard (1998) and the references therein. We assume that  $z_t$  satisfies  $z_t = g_Z(x, \varepsilon; \theta)$  for some function  $g_Z$  that can be written on closed form, and has associated conditional density  $p_{z_t|x_t}(z|x)$ . Clearly,  $p_t(y_2|x) = P(y_{2,t} = y_2|x_t = x)$  satisfies

$$p_t(y_2|x) = P(z_t \in D^{-1}(y_2) | x_t = x) = \int_{D^{-1}(y_2)} p_{z_t|x_t}(z|x) dz.$$

The last integral is equal to  $\int_{\mathbb{R}^m} \frac{p_t(z|x)}{p_D(z|y_2)} p_D(z|y_2) dz$  for any density  $p_D(z|y_2)$  with support  $D^{-1}(y_2)$ . If  $p_{z_t|x_t}(z|x)$  is known on closed form, this integral can then be simulated by

$$\hat{p}_t^{(1)}(y_2|x) = \frac{1}{N} \sum_{i=1}^N \mathcal{K}^{(1)}(\tilde{Z}_i^{\theta, y_2}, y_2|x), \quad \mathcal{K}^{(1)}(\tilde{Z}_{t,i}^{x,\theta}, y_2) = \frac{p_{z_t|x_t}(\tilde{Z}_i^{\theta, y_2}|x)}{p_D(\tilde{Z}_i^{\theta, y_2}|y_2)}, \quad (7)$$

where  $\tilde{Z}_i^{\theta, y_2} \stackrel{iid}{\sim} p_D(z|y_2)$ , as is standard in the estimation of limited dependent variable models.

If  $p_{z|x}(z|x)$  cannot be written on closed form, we propose to instead use

$$\hat{p}_{z_t|x_t}(z|x) = \frac{1}{N} \sum_{i=1}^N K_b(Z_{t,i}^{x,\theta} - z),$$

where  $Z_{t,i}^{x,\theta} \stackrel{iid}{\sim} p_{z_t|x_t}(z|x)$  and  $b > 0$  is another bandwidth. If  $\int_{D^{-1}(y_2)} K_b(Z_{t,i}^{x,\theta} - z) dz$  can be written on closed form, we follow Fermanian and Salanié (2004, pp. 709–710 and 724–725) and use:

$$\hat{p}_t^{(2)}(y_2|x) = \frac{1}{N} \sum_{i=1}^N \mathcal{K}_b^{(2)}(Z_{t,i}^{x,\theta}, y_2), \quad \mathcal{K}_b^{(2)}(Z_{t,i}^{x,\theta}, y_2) = \int_{D^{-1}(y_2)} K_b(Z_{t,i}^{x,\theta} - z) dz. \quad (8)$$

If this not the case, we can use

$$\hat{p}_t^{(3)}(y_2|x) = \sum_{i=1}^N \hat{\mathcal{K}}_b^{(2)}(Z_{t,i}^{x,\theta}, y_2), \quad \hat{\mathcal{K}}_b^{(2)}(Z_{t,i}^{x,\theta}, y_2) = \frac{1}{N} \sum_{j=1}^N \frac{K_b(Z_{t,i}^{x,\theta} - \tilde{Z}_j^{\theta, y_2})}{p_D(\tilde{Z}_j^{\theta, y_2}|y_2)}. \quad (9)$$

In all three case, we can write the resulting simulated joint density on the form (6) by choosing  $Y_{2t,i}^{x,\theta}(y_2) = \mathcal{K}^{(1)}(\tilde{Z}_i^{\theta, y_2}, y_2|x)$ ,  $Y_{2t,i}^{x,\theta}(y_2) = \mathcal{K}_b^{(2)}(Z_{t,i}^{x,\theta}, y_2)$  and  $Y_{2t,i}^{x,\theta}(y_2) = \hat{\mathcal{K}}_b^{(2)}(Z_{t,i}^{x,\theta}, y_2)$  respectively. Here,  $\theta \mapsto Y_{2t,i}^{x, y_2, \theta}$  is smooth with a bias that disappears as  $b \rightarrow 0$  and variance that is bounded in  $b$ . Thus, the order of the variance of  $\hat{L}_T(\theta)$  is not affected by any added discrete variables, and the curse of dimensionality remains of order  $k = \dim(y_{1t})$ .

**Time-Homogeneous Processes.** If the data-generating process is time-homogeneous such that  $p_t(y|x;\theta) \equiv p(y|x;\theta)$ , and we can simulate a trajectory  $\{Y_t^\theta, X_t^\theta : i = t, \dots, \tilde{N}\}$  from the model—as is the case with most simulation-based methods used for dynamic models, then the following alternative is available:

$$\check{p}(y|x;\theta) = \frac{\sum_{t=1}^{\tilde{N}} K_h(Y_t^\theta - y)K_h(X_t^\theta - x)}{\sum_{t=1}^{\tilde{N}} K_h(X_t^\theta - x)}. \quad (10)$$

This estimator is used in for example Altissimo and Mele (2008) and Hurn et al. (2003). It potentially saves time since one can use the same simulated data points to approximate the conditional density at all data points. So we only generate  $\tilde{N}$  simulated observations here to obtain the simulated likelihood at a given parameter value, while in the time-inhomogeneous case we need to simulate  $T \times N$  values. On the other hand, the convergence of  $\check{p}$  will be slower due to (i) the dimension of  $(Y_t^\theta, X_t^\theta)$  being greater than that of  $Y_t^\theta$ , and (ii) the dependence between  $(Y_s^\theta, X_s^\theta)$  and  $(Y_t^\theta, X_t^\theta)$ ,  $s \neq t$ . So one will have to choose a larger  $\tilde{N}$  for the simulated conditional density in (10) relative to the one in (2).

Typically, one will have to assume a stationary solution to the dynamic system under consideration for  $\check{p} \xrightarrow{P} p$ , and either have to start the simulation from the stationary distribution, or assume that the simulated process converges towards the stationary distribution at a suitable rate. For the latter to hold, one will need to impose some form of mixing condition on the process, as in Altissimo and Mele (2008) and Duffie and Singleton (1993). Then a large value of  $N$  is needed to ensure that the simulated process is sufficiently close to its stationary distribution—that is, one has to allow for a burn-in.

The estimator in (10) may work under nonstationarity as well. Recently, a number of papers have considered kernel estimation of nonstationary Markov processes. The kernel estimator proves to be consistent and asymptotically mixed-normally distributed when the Markov process is recurrent (Bandi and Phillips, 2003; Karlsen and Tjøstheim, 2001). However, the convergence rate will be path-dependent and relatively slow. So, for strongly dependent and nonstationary processes, it will be preferable to use the estimator in (2).

In the remainder of this paper we focus on (2). The properties of (10) can be obtained by following the same strategy of proof as the one we employ for (2). The only difference is that, to obtain  $\check{p} \xrightarrow{P} p$ , one has to take into account the dependence of the simulated values. A sufficient set of conditions for  $\check{p}(y|x;\theta) \xrightarrow{P} p(y|x;\theta)$  uniformly in  $y, x$  and  $\theta$  when the dynamics of the parametric model is near-epoch dependent can be found in Andrews (1995, Corollary 2).

**Quasi Maximum Likelihood Estimation.** The use of our approximation method is not limited to actual MLEs. In many situations, one can define quasi- or pseudo-likelihood which, even though it is not the true likelihood, identifies the parameters of the true model. One obvious example of

this is the standard regression model, where the MLE based on Gaussian errors (i.e. the least-squares estimator) proves to be robust to deviations from the normality assumption. Another example is estimation of (G)ARCH models using quasi-maximum likelihood—e.g. Lee and Hansen (1994). These are cases where the quasi-likelihood can be written explicitly. If one cannot find explicit expressions of the quasi-likelihood, one can instead employ our estimator, simulating from the quasi-model: Suppose for example that data has been generated by the model (1), but the distribution of the errors  $F_\varepsilon$  is unknown. We could then choose a suitable distribution  $G_\varepsilon$ , draw  $\{\varepsilon_i\}_{i=1}^N$  from  $G_\varepsilon$  and then proceed as in Section 2.1. The resulting estimator would no longer be a simulated MLE but rather a simulated QMLE. In this setting, the asymptotic distribution has to be adjusted to accommodate for the fact that we are no longer using the true likelihood function to estimate the parameters. This obviously extends to the case of misspecified models as in White (1984).

The above procedure is one example of how our simulation method can be applied to non- and semiparametric estimation problems where an infinite-dimensional component of the model is unknown. Another example is the situation where data has been generated by the model (1) with known distribution  $F_\varepsilon$ , but now  $\theta = (\alpha, \gamma)$  where  $\alpha$  and  $\gamma$  are finite- and infinite-dimensional parameters respectively. An application of our method in this setting can be found in Kristensen (2008a) where  $\gamma$  is a density. Again, our asymptotic results have to be adjusted to allow for  $\theta$  to contain infinite-dimensional parameters.

### 3 Asymptotic Properties of the NPSMLE

Given the convergence of the simulated conditional density towards the true one, we expect that the NPSMLE  $\hat{\theta}$  based on the simulated density in equation (6) will have the same asymptotic properties as the infeasible MLE  $\tilde{\theta}$  for a suitably chosen sequence  $N = N(T)$  and  $h = h(N)$  (and  $b = b(N)$  when an additional kernel is used). We give two sets of results. The first establishes that  $\hat{\theta}$  is first-order asymptotic equivalent to  $\tilde{\theta}$  under general conditions, allowing for nonstationarity. Under additional assumptions, including stationarity, we derive approximate expressions of the bias and variance components of  $\hat{\theta}$  relative to the actual MLE due to the simulations, and give results for the higher-order asymptotic properties of  $\hat{\theta}$ .

We allow for a mixed discrete and continuous distribution of the response variable, and write  $y_t = (y_{1t}, y_{2t}) \in \mathcal{Y}_1 \times \mathcal{Y}_2$ , where  $\mathcal{Y}_1 = \mathbb{R}^k$  and  $\mathcal{Y}_2 = \{y_{2,1}, y_{2,2}, \dots\}$ . Here,  $y_{1t}$  has a continuous distribution, while  $y_{2t}$  is discrete with  $y_{2,i} \in \mathbb{R}^l$ . The joint distribution can be written as  $p_t(y_1, y_2|x; \theta) = p_t(y_2|y_1, x; \theta)p_t(y_1|x; \theta)$  where  $p_t(y_{2,i}|y_1, x; \theta)$  are conditional probabilities satisfying  $\sum_{i=1}^m p_t(y_{2,i}|y_1, x; \theta) = 1$ , while  $p_t(y_1|x; \theta)$  is a conditional density w.r.t. the Lebesgue measure. Also, let  $p_t(y_{2,i}|x; \theta)$  denote the conditional probabilities of  $y_{2t}|x_t = x$ .

The asymptotics are derived under the following simulation scheme,

$$Y_{1t,i}^{x_t,\theta} = g_{1,t}(x_t, \varepsilon_i; \theta), \quad (11)$$

$$Y_{2t,i}^{x_t,\theta}(y_{2t}) = g_{2,t}(y_{2t}, x_t, \varepsilon_i; \theta), \quad (12)$$

for  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , where  $\{\varepsilon_i\}_{i=1}^N$  are i.i.d. draws from  $F_\varepsilon$ , such that equation (5) holds. The condition in equation (5) is met when  $Y_{2t,i}^{x_t,\theta}(y_2) = \mathcal{K}^{(1)}(\tilde{Z}_i^{\theta,y_2}, y_2|x)$  with  $\mathcal{K}^{(1)}$  given in equation (7), while it only holds approximately for  $\mathcal{K}^{(2)}$  and  $\hat{\mathcal{K}}^{(2)}$  defined in equations (8)–(9) due to biases induced by the use of kernel smoothing in those two cases. We handle these two cases in Theorem 4 where results for approximate simulations are given.

Note that we here use the same errors to generate the simulations over time. An alternative simulation scheme would be to draw a new batch of errors for each observation  $x_t$ ,  $Y_{t,i}^{x_t,\theta} = g_t(x_t, \varepsilon_{t,i}; \theta)$ ,  $i = 1, \dots, \tilde{N}$ , such that the total number of simulations would be  $\tilde{N} \times T$ ,  $\{\varepsilon_{i,t}\}_{i=1}^{\tilde{N}}$ ,  $t = 1, \dots, T$ . Under regularity conditions, the NPSMLE based on this simulation scheme would have similar asymptotic properties as the one based on the simulations in equations (11)–(12). However, as demonstrated in Lee (1992), choosing  $N = \tilde{N}T$ , the variance of the NPSMLE based on equations (11)–(12) will be smaller.<sup>8</sup>

In order for  $\hat{\theta}$  to be asymptotically equivalent to  $\tilde{\theta}$ , we need  $\hat{p} \xrightarrow{P} p$  sufficiently fast in some suitable function norm. To establish this, we verify the general conditions for uniform rates of kernel estimators found in Kristensen (2008b). These general conditions are verified by the following set of regularity conditions regarding the model and its associated conditional density.

**A.1** The functions  $(x, t, \theta) \mapsto g_{1,t}(x, \varepsilon; \theta)$  and  $(x, t, \theta) \mapsto g_{2,t}(y_2, x, \varepsilon; \theta)$  are continuously differentiable for all  $y_2$  and  $\varepsilon$  such that for some function  $\Lambda(\cdot)$  and constants  $\lambda_1, \lambda_2 \geq 0$ ,

$$\|g_{1,t}(x, \varepsilon; \theta)\| \leq \Lambda(\varepsilon) \left[ 1 + \|x\|^{\lambda_{1,1}} + t^{\lambda_{1,2}} \right], \quad \|g_{2,t}(y_2, x, \varepsilon; \theta)\| \leq \Lambda(\varepsilon) \left[ 1 + \|x\|^{\lambda_{2,1}} + t^{\lambda_{2,2}} \right],$$

and  $\mathbb{E}[\Lambda(\varepsilon)^s] < \infty$  for some  $s > 2$ . The derivatives of  $g_1$  and  $g_2$  w.r.t.  $(x, t, \theta)$  satisfy the same bounds.

**A.2** The conditional density  $p_t(y_1, y_2|x; \theta)$  is continuous w.r.t.  $\theta \in \Theta$ , and  $r \geq 2$  times continuously differentiable w.r.t.  $y_1$  with bounded derivatives such that with  $\bar{B}(x, t) = \bar{B} \left( 1 + \|x\|^{\lambda_{0,1}} + t^{\lambda_{0,2}} \right)$ , for some constants  $\bar{B} > 0$  and  $\lambda_1, \lambda_2 \geq 0$ , the following bounds hold uniformly over  $(t, y_1, y_2, x, \theta)$ :

$$\sum_{|\alpha|=r} \left| \frac{\partial^r p_t(y_1, y_2|x; \theta)}{\partial y_1^\alpha} \right| \leq \bar{B}(x, t), \quad \|y_1\|^k p_t(y_1, y_2|x; \theta) \leq \bar{B}(x, t). \quad (13)$$

---

<sup>8</sup>The results of Lee (1992) are for discrete choice models, but we conjecture that his results can be extended to general simulated MLE.

**A.3**  $\theta \mapsto g_{1,t}(x, \varepsilon; \theta)$  and  $\theta \mapsto g_{2,t}(x, y_2, \varepsilon; \theta)$  are twice continuous differentiable for all  $t, x, \varepsilon$  with their derivatives satisfying the same moment conditions as  $g_1$  and  $g_2$  in (A.1).

**A.4**  $\partial p_t(y|x; \theta)/\partial \theta$  and  $\partial^2 p_t(y|x; \theta)/(\partial \theta \partial \theta')$  are  $r \geq 2$  times continuously differentiable w.r.t.  $y_1$  with bounded derivatives such that they satisfy the same bounds in equation (13) as  $p$ .

Assumptions (A.1)–(A.2) are used to establish uniform convergence of  $\hat{p}$  by verifying the general conditions in Kristensen (2008b), c.f. Lemma 11. Assumption (A.1) imposes restrictions on the two data-generating functions  $g_1$  and  $g_2$ . The smoothness conditions are rather weak, and satisfied by most models, while the polynomial bounds imposed on the two functions can be exchanged for other bounds, but will complicate some of the conditions imposed below. Note that the moment conditions in (A.1) do not concern the observed process  $\{(y_t, x_t)\}$ , only the errors  $\varepsilon$  that we draw when simulating. If for example,  $\Lambda(\varepsilon) \leq \|\varepsilon\|^q$ , then the moment condition is satisfied if  $\mathbb{E}[\|\varepsilon\|^{qs}] < \infty$ . Thus, in this case, the moment condition only rule out models driven by fat-tailed errors. If the model is time-homogenous,  $\lambda_{k,2} = 0$ ,  $k = 1, 2$ .

Assumption (A.2) restricts the conditional density that we are trying to estimate. The smoothness assumptions imposed on  $p$  in (A.2) in conjunction with the use of higher-order kernels reduces the bias of  $\hat{p}$ . The bounds are imposed to obtain a uniform bound of the variance of  $\hat{p}$ . Again, the assumptions are quite weak and are satisfied by many models. If the model is time-homogenous,  $\lambda_{0,2} = 0$ .

Assumptions (A.3) and (A.4) will only be used when examining the effect of the simulations on the asymptotic variance of the estimator. These two conditions yield uniform convergence of  $\partial \hat{p}_t(y|x; \theta)/\partial \theta$  and  $\partial^2 \hat{p}_t(y|x; \theta)/(\partial \theta \partial \theta')$ , which in turn allows us to analyze the first and second derivatives of the simulated log-likelihood (Lemma 12).

Our conditions are slightly stronger than the ones found in Fermanian and Salanié (2004, Conditions M.1–2 and L.1–3). There, weaker bounds and smoothness conditions are imposed on the function  $g$ , while their restrictions on the conditional density are very similar to ours.

The kernel  $K$  is assumed to belong to the following class of so-called higher-order or bias-reducing kernels.

**K.1** The kernel  $K$  satisfies:

1.  $|K(u)| \leq \bar{K} < \infty$  and  $\int |K(u)| du \leq \mu < \infty$ . There exist  $\Lambda, L < \infty$  such that either (i)  $K(u) = 0$  for  $\|u\| > L$  and  $|K(u) - K(u')| \leq \Lambda \|u - u'\|$ , or (ii)  $K(u)$  is differentiable with  $|\partial K(u)/\partial u| \leq \Lambda$ . For some  $a > 1$ ,  $|\partial^i K(u)/\partial u^i| \leq \Lambda \|u\|^{-a}$  for  $\|u\| \geq L$  and  $i = 0, 1$ .
2. For some  $r \geq 1$ :  $\int K(u) u^i du = 0$ ,  $i = 1, \dots, r - 1$ , and  $\int K(u) |u|^r du < \infty$ .

**K.2** The first and second derivative of  $K$  exist and also satisfy (K.1.1).

This is a broad class of kernels allowing for unbounded support. For example, the Gaussian kernel satisfies (K.1) with  $r = 2$ . When  $r > 2$ ,  $K$  is a so-called higher-order kernel that reduces the bias of  $\hat{p}$  and its derivatives, and thereby obtains a faster rate of convergence. The smoothness of  $p$  as measured by its number of derivatives,  $r$ , determines the degree of bias reduction. The additional assumption (K.2) is used in conjunction with (A.3)–(A.4) to show that the first and the second derivatives of  $\hat{p}$  w.r.t.  $\theta$  also converge uniformly.

Next, we impose regularity conditions on the model to ensure that the actual MLE is asymptotically well-behaved. We first introduce the relevant terms driving the asymptotics of the MLE. We first normalize the log-likelihood by some factor  $\nu_T \rightarrow \infty$ :

$$L_T(\theta) = \frac{1}{\nu_T} \sum_{t=1}^T \log p_t(y_t|x_t; \theta).$$

This normalizing factor  $\nu_T$  is introduced to ensure that  $L_T(\theta)$  is well-behaved asymptotically and that certain functions of data are suitably bounded, c.f. (C.1)–(C.5) below. It is only important for the theoretical derivations, and not relevant for the actual implementation of our estimator since  $\nu_T$  does not depend on  $\theta$ . The choice of  $\nu_T$  depends on the dynamics of the model. The standard choice is  $\nu_T = T$  as is, for example, the case when the model is stationary. In order to allow for non-standard behaviour of the likelihood, e.g. unit root-type asymptotics, we don't impose this restriction though.

Assuming that  $L_T(\theta)$  is three times differentiable, c.f. (C.4) below, we can define:

$$\begin{aligned} S_T(\theta) &= \frac{\partial L_T(\theta)}{\partial \theta} = \frac{1}{\nu_T} \sum_{t=1}^T \frac{\partial \log p_t(y_t|x_t; \theta)}{\partial \theta} \in \mathbb{R}^d, \\ H_T(\theta) &= \frac{\partial^2 L_T(\theta)}{\partial \theta \partial \theta'} = \frac{1}{\nu_T} \sum_{t=1}^T \frac{\partial^2 \log p_t(y_t|x_t; \theta)}{\partial \theta \partial \theta'} \in \mathbb{R}^{d \times d}, \\ G_{T,i}(\theta) &= \frac{\partial^3 L_T(\theta)}{\partial \theta \partial \theta' \partial \theta_i} = \frac{1}{\nu_T} \sum_{t=1}^T \frac{\partial^3 \log p_t(y_t|x_t; \theta)}{\partial \theta \partial \theta' \partial \theta_i} \in \mathbb{R}^{d \times d}. \end{aligned}$$

The information is then defined as:

$$i_T(\theta) = \frac{1}{\nu_T} \sum_{t=1}^T \mathbb{E} \left[ \frac{\partial \log p_t(y_t|x_t; \theta)}{\partial \theta} \frac{\partial \log p_t(y_t|x_t; \theta)}{\partial \theta'} \right] = \mathbb{E}[H_T(\theta)] \in \mathbb{R}^{d \times d}.$$

We also define the diagonal matrix  $\mathcal{I}_T(\theta) = \text{diag}\{i_T(\theta)\} \in \mathbb{R}^{d \times d}$ , and:

$$U_T(\theta) = \mathcal{I}_T^{-1/2}(\theta) S_T(\theta), \quad V_T(\theta) = \mathcal{I}_T^{-1/2}(\theta) H_T(\theta) \mathcal{I}_T^{-1/2}(\theta), \quad W_{T,i}(\theta) = \mathcal{I}_T^{-1/2}(\theta) G_{T,i}(\theta) \mathcal{I}_T^{-1/2}(\theta).$$

We then impose the following conditions on the actual log-likelihood function and the associated MLE where  $\mathcal{I}_T \equiv \mathcal{I}_T(\theta_0)$ :

**C.1** The parameter space is given by a sequence of local neighbourhoods,

$$\Theta_T = \left\{ \theta : \|\mathcal{I}_T^{1/2}(\theta - \theta_0)\| < \epsilon \right\} \subseteq \mathbb{R}^d,$$

for some  $\epsilon > 0$  with  $\mathcal{I}_T^{-1} = O_P(1)$ .

**C.2**  $\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) = o_P(1)$ .

**C.3**  $\xi \mapsto L_T(\theta_0 + \mathcal{I}_T^{-1/2}\xi)$  is stochastically equicontinuous for  $\|\xi\| < \epsilon$ .

**C.4**  $L_T(\theta)$  is three times continuously differentiable with its derivatives satisfying:

1.  $(U_T(\theta_0), V_T(\theta_0)) \xrightarrow{d} (0, H_\infty)$ , and  $(\sqrt{\nu_T}U_T(\theta_0), V_T(\theta_0)) \xrightarrow{d} (S_\infty, H_\infty)$ , with  $H_\infty > 0$  a.s.;
2.  $\max_{j=1, \dots, d} \sup_{\theta \in \Theta_T} \|W_{j,T}(\theta)\| = O_P(1)$ .

**C.5** The following bounds hold for some  $\delta, q > 0$ :

1.  $\sup_{\theta \in \Theta_T} \nu_T^{-q} \sum_{t=1}^T |\log p_t(y_t|x_t; \theta)|^{1+\delta} = O_P(1)$ ;
2.  $\nu_T^{-q} \sum_{t=1}^T \|x_t\|^{1+\delta} = O_P(1)$  and  $\nu_T^{-q} \sum_{t=1}^T \Lambda_{1t}^2(\varepsilon_t) = O_P(1)$ .

The above conditions (C.1)–(C.4) are standard conditions found in the literature on non-ergodic models—e.g. Basawa and Scott (1983); Jeganathan (1995); Saikkonen (1995). For general non-ergodic models, simple conditions for (C.2)–(C.5) are not available and they have to be verified on a case-by-case basis. For the stationary case, (C.2)–(C.5) are implied by primitive conditions as found below.

The specification of the parameter space in (C.1) to be a sequence of non-increasing sets is introduced to allow for non-ergodic models. Currently, to the best of our knowledge, there exists no general results on the properties of MLEs for general dynamic models over a fixed parameter space. Park and Phillips (2001) give results with fixed parameter space for the case of nonlinear regression with integrated time series. There, it is required that the individual components of the estimator all converge with the same rate. In contrast, we here allow for different convergence rates given by  $\mathcal{I}_T^{1/2}$ . This is usually the case for non-ergodic models as for example in the error-correction model discussed below.

Assumption (C.2) gives us consistency of the actual MLE, while (C.3) is used in the proof of  $\hat{\theta} \xrightarrow{P} \tilde{\theta}$ . See Saikkonen (1995) for further details. Assumption (C.4) is a strengthening of (C.2)–(C.3), c.f. Lemma 7. It implies consistency and that the asymptotic distribution of the MLE is given by:

$$\sqrt{\nu_T} \mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) \xrightarrow{d} H_\infty^{-1} S_\infty,$$

c.f. Lemmas 5–6.<sup>9</sup>

Assumption (C.5) imposes bounds on a number of sample averages. They are used when showing that the simulated and the actual log-likelihood are asymptotically identical. Note that the factor  $\nu_T$  in (C.5) is the same as the one we normalized the log-likelihood with. The exponent  $q > 0$  can be chosen to ensure that both the log-likelihood and the sample averages in (C.5) are well-behaved.

In the ergodic case, we can appeal to standard results for stochastic equicontinuity—e.g. Newey (1991)—to obtain that (C.3) and (C.5) hold with  $\nu_T = T$  and  $q = 1$ , given that  $\mathbb{E}[\|x_t\|^{1+\delta}] < \infty$  and  $\mathbb{E}[\sup_{\theta \in \Theta} |\log p(y_t|x_t; \theta)|^{1+\delta}] < \infty$ . See Corollary 2 below and its proof for further details. Furthermore,  $i_T(\theta_0) = i(\theta_0) + o_P(1)$  with  $i(\theta_0) = \mathbb{E}[\partial^2 \log p(y_t|x_t; \theta_0)/(\partial\theta\partial\theta')]$ , such that  $\mathcal{I}_T$  can be chosen as the constant  $\text{diag}\{i(\theta)\}$ . This in turn implies that  $\Theta_T$  is a fixed compact parameter set, and we get standard  $\sqrt{\nu_T} = \sqrt{T}$ -convergence towards a normal distribution. Thus, in the case of stationarity, (C.1)–(C.5) are more or less identical to the ones imposed in Fermanian and Salanié (2004, Conditions L.1–3).

In the general case, one should choose  $\nu_T$  as the square of the slowest rate of convergence of the vector of MLEs. There is a tension between (C.1) and (C.5) in terms of the choice of  $\nu_T$ . We cannot choose  $\nu_T \rightarrow \infty$  too fast, since then  $\|\mathcal{I}_T\| \rightarrow 0$  (in which case no information regarding  $\theta_0$  is available) and this is ruled out by (C.1). On the other hand, we have to choose  $\nu_T^q \rightarrow \infty$  sufficiently fast to ensure that the bounds in (C.5) hold. By choosing  $q > 0$  sufficiently large, (C.1) and (C.5) will both be satisfied. However, a large value of  $q$  implies that we have to use a larger number of simulations for the NPSMLE to be asymptotically equivalent to the MLE, c.f. (B.1)–(B.2) below.

As an example of non-standard asymptotics of the MLE, consider a linear error-correction model,

$$\Delta y_t = \alpha\beta' y_{t-1} + \Omega^{1/2} \varepsilon_t, \quad \varepsilon_t \sim N(0, I_k).$$

We can split the parameter vector into short-run,  $\theta_1 = (\alpha, \text{vech}(\Omega))$ , and long-run parameters,  $\theta_2 = \beta$ . The MLE  $\tilde{\theta}_1$  converges with  $\sqrt{T}$ -speed towards a normal distribution, while  $\tilde{\theta}_2$  is super-consistent with  $T(\tilde{\theta}_2 - \theta_2)$  converging towards a Dickey-Fuller type distribution. In that situation, we choose  $\sqrt{\nu_T} = \sqrt{T}$ , but now  $i_T(\theta_0)$  and therefore  $\mathcal{I}_T$ , is not asymptotically constant. As demonstrated in Saikkonen (1995), this model satisfies (C.2)–(C.4). Furthermore,  $x_t = y_{t-1}$  satisfies  $T^{-2} \sum_{t=1}^T \|x_t\|^{1+\delta} = O_P(1)$  so we can choose  $q = 2$ . We also refer to Kristensen and Rahbek (2008) and Park and Phillips (2001) where (C.2)–(C.5) are verified for some non-linear, non-stationary models.

Finally, we need to introduce trimming of the approximate log-likelihood to obtain uniform convergence of  $\log \hat{p}_t$  as is standard in the literature on semiparametric estimators. We redefine our

---

<sup>9</sup>Basawa and Scott (1983) and Jeganathan (1995) show what  $S_\infty$  and  $H_\infty$  look like in various cases.

simulated log-likelihood as

$$\hat{L}_T(\theta) = \frac{1}{\nu_T} \sum_{t=1}^T \tau_a(\hat{p}_t(y_t|x_t; \theta)) \log \hat{p}_t(y_t|x_t; \theta),$$

where  $\tau_a(\cdot)$  is continuously differentiable trimming function satisfying  $\tau_a(z) = 1$  if  $|z| > a$ , and 0 if  $|z| < a/2$ , with a trimming sequence  $a = a(N) \rightarrow 0$ . One could here simply use the indicator function for the trimming, but then  $\hat{L}_T(\theta)$  would no longer be differentiable, and differentiability is useful when using numerical optimization algorithms to solve for  $\hat{\theta}$ .

We impose the following restrictions on how the bandwidth  $h$  and trimming sequence  $a$  can converge to zero in conjunction with  $N, T \rightarrow \infty$ :

**B.** With  $q, \delta > 0$  given in (C.5),  $\bar{\lambda}_k = \lambda_{0,k} + \lambda_{1,k} + \lambda_{2,k}$ ,  $k = 1, 2$ , where  $\lambda_{i,1}, \lambda_{i,2} \geq 0$ ,  $i = 0, 1, 2$ , are given in (A.1)-(A.2) and for some  $\gamma > 0$ :

1.  $|\log a| \nu_T^{q-1} N^{-\gamma(1+\delta)} \rightarrow 0$ ;  $|\log(4a)|^{-\delta} \nu_T^{q-1} \rightarrow 0$ ;  $T \nu_T^{-1} a^{-1} [N^{\gamma \bar{\lambda}_1} + T^{\bar{\lambda}_2}] \log(N) / \sqrt{N h^k} \rightarrow 0$ ; and  $T \nu_T^{-1} a^{-1} [N^{\gamma \lambda_{0,1}} + T^{\lambda_{0,2}}] \rightarrow 0$ .
2.  $|\log a| \nu_T^q N^{-\gamma(1+\delta)} \rightarrow 0$ ;  $|\log(4a)|^{-\delta} \nu_T^q \rightarrow 0$ ;  $T \nu_T^{-1/2} a^{-1} [N^{\gamma \bar{\lambda}_1} + T^{\bar{\lambda}_2}] \log(N) / \sqrt{N h^k} \rightarrow 0$ ; and  $T \nu_T^{-1/2} a^{-1} [N^{\gamma \lambda_{0,1}} + T^{\lambda_{0,2}}] \rightarrow 0$ .

Condition (B.1) is imposed when showing consistency of the NPSMLE, while (B.2) will imply that the NPSMLE has the same asymptotic distribution as the MLE. The parameter  $\gamma > 0$  can be chosen freely, however it has to be chosen small enough such that, for example,  $T a^{-1} N^{\gamma \bar{\lambda}_1 - 1} \log(N)^2 / h^k \rightarrow 0$  as required in (B.2). We observe that large values of  $q$  and/or  $\bar{\lambda}_1, \bar{\lambda}_2$  implies that  $N$  has to diverge at a faster rate relative to  $T$ . In practice, this means that a larger number of simulations have to be used for a given  $T$  to obtain a precise estimate. The joint requirements imposed on  $a$ ,  $h$  and  $N$  are fairly complex, and it is not obvious how to choose these nuisance parameters for a given sample size  $T$ . This is a problem shared by, for example, semiparametric estimators that rely on a preliminary kernel estimator. We refer to Ichimura and Todd (2007) for an in-depth discussion of these matters.

Our strategy of proof is based on some apparently new results for approximate estimators, c.f. Appendix A. In particular, Theorems 8–9 establish that the NPSMLE and the MLE will be asymptotically first-order equivalent if  $\hat{L}_T(\theta)$  converges uniformly towards  $L_T(\theta)$  at the right rate. This makes our proofs considerably less burdensome than those found in other studies of simulation-based estimators—e.g. Altissimo and Mele (2008); Fermanian and Salanié (2004)—since we do not need to show that the simulated score and Hessian also converge.

**Theorem 1** *Assume that (A.1)–(A.2), (K.1) and (C.5) hold. Then the NPSMLE based on (6) satisfies:*

- (i) Under (C.1)–(C.3):  $\mathcal{L}_T^{1/2}(\hat{\theta} - \theta_0) = o_P(1)$  for any sequences  $N \rightarrow \infty$ ,  $h, a \rightarrow 0$  satisfying (B.1).  
(ii) Under (C.4):  $\sqrt{\nu_T} \mathcal{L}_T^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} H_\infty^{-1} S_\infty$  for any sequences  $N$ ,  $h$  and  $a$  satisfying (B.2).

When the data generating process is stationary and ergodic, the following more primitive conditions can be shown to imply (C.1)–(C.5) and (B.1)–(B.2).

**Corollary 2** *Assume that  $\{(y_t, x_t)\}$  is stationary and ergodic, that (A.1)–(A.2) and (K.1) hold, and:*

- (i)  $\mathbb{E}[\|x_t\|^{1+\delta}] < \infty$ ,  $|\log p(y|x; \theta)| \leq b_1(y|x)$ ,  $\forall \theta \in \Theta$ , with  $\mathbb{E}[b_1(y_t|x_t)^{1+\delta}] < \infty$  and  $\Theta$  compact;  
(ii)  $\mathbb{E}[\log p(y_t|x_t; \theta)] < \mathbb{E}[\log p(y_t|x_t; \theta_0)]$ ,  $\forall \theta \neq \theta_0$ .

Then  $\hat{\theta} \xrightarrow{P} \theta_0$  as  $\sqrt{\log(N)/N} h^{-k-1} a^{-1} \rightarrow 0$ ,  $h^r a^{-1} \rightarrow 0$ , and  $N^{-2\gamma} h^{-k} \log a \rightarrow 0$  for some  $\gamma > 0$ .

If furthermore:

- (iv)  $i(\theta_0) = \mathbb{E} \left[ \frac{\partial \log p(y_t|x_t; \theta_0)}{\partial \theta} \frac{\partial \log p(y_t|x_t; \theta_0)}{\partial \theta'} \right]$  exists and is nonsingular;  
(v)  $\left\| \frac{\partial^2 \log p(y|x; \theta)}{\partial \theta \partial \theta'} \right\| \leq b_2(y|x)$  uniformly in a neighborhood of  $\theta_0$  with  $\mathbb{E}[b_2(y_t|x_t)] < \infty$ ;

then  $\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, i(\theta_0)^{-1})$  as  $T\sqrt{\log(N)/N} h^{-k-1} a^{-1} \rightarrow 0$ ,  $Th^r a^{-1} \rightarrow 0$ ,  $TN^{-2\gamma} h^{-k} \log a \rightarrow 0$ ,  $T(\log a)^{-1} \rightarrow 0$ ,  $TN^{-\gamma} \rightarrow 0$ .

When  $N$ ,  $h$  and  $a$  satisfy (B.1)–(B.2), the simulated and actual MLE are asymptotically first order equivalent. However, in finite sample, the NPSMLE will most likely suffer from additional biases and variance. To highlight these potential effects, we further examine the properties of the NPSMLE when (B.1)–(B.2) are not satisfied. To this end, we have to invoke the additional smoothness conditions on  $g$  and  $p$  stated in (A.3)–(A.4) since we need to be able to analyze the first and second derivative of  $\hat{L}_T(\theta)$ .

Under the additional smoothness conditions, the first two derivatives of  $g_{1t}(x, \varepsilon; \theta)$  and  $g_{2t}(y_2, x, \varepsilon; \theta)$  w.r.t.  $\theta$  exist, and we can compute the first two derivatives of our density estimator:

$$\frac{\partial \hat{p}_t(y|x; \theta)}{\partial \theta} = \frac{1}{Nh} \sum_{i=1}^N \left\{ K_h(Y_{1t,i}^{x_t, \theta} - y_1) \dot{Y}_{2t,i}^{x_t, \theta}(y_2) + K_h'(Y_{1t,i}^{x_t, \theta} - y_1) \dot{Y}_{1t,i}^{x_t, \theta} Y_{2t,i}^{x_t, \theta}(y_2) \right\}, \quad (14)$$

$$\begin{aligned} \frac{\partial^2 \hat{p}_t(y|x; \theta)}{\partial \theta \partial \theta'} &= \frac{1}{Nh} \sum_{i=1}^N K_h'(Y_{1t,i}^{x_t, \theta} - y_1) \left\{ 2\dot{Y}_{1t,i}^{x_t, \theta} \dot{Y}_{2t,i}^{x_t, \theta}(y_2)' + \ddot{Y}_{1t,i}^{x_t, \theta} Y_{2t,i}^{x_t, \theta}(y_2) \right\} \\ &\quad + \frac{1}{Nh} \sum_{i=1}^N \left\{ K_h(Y_{1t,i}^{x_t, \theta} - y_1) \dot{Y}_{2t,i}^{x_t, \theta}(y_2) + K_h''(Y_{1t,i}^{x_t, \theta} - y_1) \dot{Y}_{1t,i}^{x_t, \theta} (\dot{Y}_{1t,i}^{x_t, \theta})' Y_{2t,i}^{x_t, \theta}(y_2) \right\}, \end{aligned} \quad (15)$$

where

$$\dot{Y}_{1i,t}^{x,\theta} = \frac{\partial g_t(x, \varepsilon_i; \theta)}{\partial \theta}, \quad \ddot{Y}_{1i,t}^{x,\theta} = \frac{\partial^2 g_t(x, \varepsilon_i; \theta)}{\partial \theta \partial \theta'},$$

and similarly for  $\dot{Y}_{2i,t}^{x,\theta}$  and  $\ddot{Y}_{2i,t}^{x,\theta}$ . Lemma 12 shows that these are uniformly consistent estimates of the actual derivatives of the conditional density  $p_t$ . We can in turn use these to obtain a simulated version of the score,

$$\hat{S}_T(\theta) = \frac{1}{\nu_T} \sum_{t=1}^T \frac{\partial \hat{p}_t(y_t|x_t; \theta)}{\partial \theta} \left\{ \frac{\tau_a(\hat{p}_t(y_t|x_t; \theta))}{\hat{p}_t(y_t|x_t; \theta)} + \tau'_a(\hat{p}_t(y_t|x_t; \theta)) \log \hat{p}_t(y_t|x_t; \theta) \right\}, \quad (16)$$

and the Hessian (see the proof of Theorem 3 for the expression). We then follow Lee (1999) and consider a second order functional Taylor expansion of  $\hat{S}_T(\theta)$  w.r.t.  $\hat{p}$ . This takes the form:

$$\hat{S}_T(\theta_0) = S_T(\theta_0) + \nabla S_{T,N}[\hat{p} - p] + \nabla^2 S_{T,N}[\hat{p} - p, \hat{p} - p] + R_{T,N}, \quad (17)$$

where  $\nabla S_{T,N}[\hat{p} - p]$  and  $\nabla^2 S_{T,N}[\hat{p} - p, \hat{p} - p]$  are the first and the second order functional differentials w.r.t.  $p$ , while  $R_{T,N}$  is the remainder term. The expressions of these can be found in the proof of Theorem 3.

We then wish to analyze the asymptotic behaviour of each of the four terms on the right hand side. To conduct our analysis, which involves  $U$ -statistics, we restrict our attention to the stationary and  $\beta$ -mixing case. See e.g. Ango Nze and Doukhan (2004) for an introduction to this concept. We also assume a bounded support of  $x_t$ , and that  $p(y|x; \theta)$  is uniformly bounded away from zero thereby obviating trimming. Under these and other regularity conditions, we obtain that asymptotically the two first terms in the expansion in equation (17) satisfy (c.f. the proof of Theorem 3):

$$\sqrt{T} S_T(\theta_0) + \sqrt{T} \nabla S_{T,N}[\hat{p} - p] \propto \sqrt{T} h^r \mu_1 + Z_1 + \sqrt{\frac{T}{N}} Z_2.$$

Here, the first term is a bias component incurred by kernel estimation, while the two remaining ones are variance components:  $Z_1$  and  $Z_2$  are two independent variables where  $Z_1 \sim \mathcal{N}(0, i(\theta_0)^{-1})$  is the variance component of the observed data, while  $Z_2 \sim \mathcal{N}(0, \text{Var}(\bar{\psi}_2(\varepsilon_i)))$  is the variance component of the simulations. Here,

$$\bar{\psi}_2(\varepsilon_i) = \mathbb{E} \left[ \frac{\dot{Y}_{2,i}^{x_t}(y_{2t})}{p(y_{2t}|x_t)} \middle| \varepsilon_i \right] - \mathbb{E} \left[ \frac{s(Y_{1i}^{x_t}, y_{2t}|x_t) Y_{2,i}^x(y_{2t})}{p(y_{2t}|x_t)} \middle| \varepsilon_i \right],$$

where  $s(y_1, y_2|x)$  denotes the score at  $\theta = \theta_0$ , and  $p(y_{2,t}|x_t)$  the conditional distribution of  $y_{2,t}|x_t = x$ . The second order term also contains a bias component,

$$\sqrt{T} \nabla^2 S_{T,N}[\hat{p} - p, \hat{p} - p] \propto \frac{\sqrt{T}}{N h^{k+1}} \mu_2 + O_P(\sqrt{T} h^{2r}),$$

which all non-linear, simulation-based estimators will suffer from, while the remainder term is of a lower order:

$$\sqrt{T}R_{T,N} \propto O_P \left( \sqrt{T} / (Nh^{k+2})^{3/2} \right) + O_P \left( \sqrt{T}h^{3r} \right).$$

The two leading bias terms in the above expressions,  $\mu_1$  and  $\mu_2$ , are given by:

$$\mu_1 = \int \left[ \int \partial_y^r \partial_\theta p(y|x; \theta) dy \right] p(x) dx - \int \left[ \int s(y|x; \theta) \partial_y^r p(y|x; \theta) dy \right] p(x) dx, \quad (18)$$

$$\mu_2 = \mathbb{E} \left[ \frac{\dot{Y}_{1,i}^{x_t} Y_{2,i}^x (y_{2t})^2}{p(Y_{1,i}^{x_t}, y_{2t}|x_t) p(y_{2t}|x_t)} \right] \int K'(v) K(v) dv. \quad (19)$$

From these results, we conclude that if  $\sqrt{T}h^r \rightarrow 0$  and  $\sqrt{T}/(Nh^{k+1}) \rightarrow 0$ , all bias terms vanish and  $\sqrt{T}(\hat{\theta} - \theta_0)$  follows a normal distribution centered around zero. On the other hand, if either  $\sqrt{T}h^r$  or  $\sqrt{T}/(Nh^{k+1})$  does not vanish, a bias term will be present and the asymptotic distribution will not be centered around zero. Also, there will be an increase in variance due to the presence of  $Z_2$ .

**Theorem 3** *Assume that:*

- (i)  $\{(y_t, x_t)\}$  is stationary and  $\beta$ -mixing with geometrically decreasing mixing coefficients;
- (ii) (A.1)–(A.5) and (K.1) hold, and  $\theta \mapsto g(x, e; \theta)$  is twice differentiable with both derivatives satisfying (A.5);
- (iii) (i)–(v) of Corollary 2 hold;
- (iv)  $x_t$  is bounded and  $\inf_{y_1, y_2, x, \theta} p(y_1, y_2|x; \theta) > 0$ .

Then, if  $\sqrt{T}h^r \rightarrow c_1 \geq 0$  and  $\sqrt{T}/(Nh^{k+1}) \rightarrow c_2 \geq 0$ ,

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N} \left( \bar{c}, i(\theta_0)^{-1} \left[ i(\theta_0) + \frac{T}{N} \text{Var}(\bar{\psi}_2(\varepsilon_i)) \right] i(\theta_0)^{-1} \right),$$

where  $\bar{c} = c_1\mu_1 + c_2\mu_2$ , with  $\mu_1$  and  $\mu_2$  as in equations (18)–(19).

If  $\sqrt{T}/(Nh^{k+1}) \rightarrow c_2$ , then  $T/N \rightarrow 0$  such that the limiting distribution in Theorem 3 is equivalent to  $\mathcal{N}(\bar{c}, i(\theta_0)^{-1})$ . We have here kept the factor  $1 + T/N$  in the asymptotic variance to give a better description of the finite-sample performance of the NPSMLE.

For the case where an unbiased estimator of the density is available and a new batch of simulations is used for each observation, Lee (1999) derives results similar to Theorem 3.

**Estimation of Variance.** To do any finite-sample inference, an estimator of the asymptotic distribution (which depends on the unknown parameter  $\theta$ ) is needed. A general method is simply to simulate the score (and potentially also the observed information) for a sufficiently large  $T$  and evaluate at  $\theta = \hat{\theta}$ . These can then be used to approximate  $H_\infty^{-1}S_\infty$ . The computation of the score and Hessian can be done in several ways. If the model satisfies (A.3), the estimators of the score and Hessian given in (16) and the proof of Theorem 3 are available. In the general case, a simple approach is to use numerical derivatives. Define:

$$\frac{\partial \hat{p}_t(y|x; \theta)}{\partial \theta_k} = \frac{\hat{p}_t(y|x; \theta + \delta e_k) - \hat{p}_t(y|x; \theta - \delta e_k)}{2\delta},$$

where  $e_k$  is the  $k$ th column of the identity matrix. We have:

$$\begin{aligned} & \frac{\partial \hat{p}_t(y|x; \theta)}{\partial \theta_k} - \frac{\partial p_t(y|x; \theta)}{\partial \theta_k} \\ &= \frac{\hat{p}_t(y|x; \theta + \delta e_k) - p_t(y|x; \theta + \delta e_k)}{2\delta} - \frac{\hat{p}_t(y|x; \theta - \delta e_k) - p_t(y|x; \theta - \delta e_k)}{2\delta} \\ &+ \left\{ \frac{p_t(y|x; \theta + \delta e_k) - p_t(y|x; \theta - \delta e_k)}{2\delta} - \frac{\partial p_t(y|x; \theta)}{\partial \theta_k} \right\}. \end{aligned}$$

Now letting  $\delta = \delta(N) \rightarrow 0$  as  $N \rightarrow \infty$  at a suitable rate, all three terms are  $o_P(1)$ . A consistent estimator of the second derivative can be obtained in a similar fashion. These can in turn be used to construct estimators of the information and score.

**Approximate Simulations.** In many cases, the model in (1) is itself intractable, such that one cannot directly simulate from the model, and one only has an approximation of the model at hand. For example, solutions to dynamic programming problems are typically approximated numerically, and sample paths of diffusions must be approximated by some discretization scheme. We here derive the asymptotics of the approximate NPSMLE based on simulations from a sequence of approximate models. Assuming that the approximation error from using the approximate model relative to the true one can be made arbitrarily small, we demonstrate that the approximate NPSMLE has the same asymptotic properties as the actual MLE.

Suppose we only have the following approximations of  $g_{1t}$  and  $g_{2t}$ ,  $g_{M,1t}(x, \varepsilon; \theta)$  and  $g_{M,2t}(y_2, x, \varepsilon; \theta)$  at our disposal, where  $g_{M,kt} \rightarrow g_{kt}$ ,  $k = 1, 2$ , as  $M \rightarrow \infty$  in a suitable function norm specified below in condition (M.1). We then redefine the simulated conditional density as:

$$\hat{p}_{M,t}(y|x; \theta) = \frac{1}{N} \sum_{i=1}^N K_h(\hat{Y}_{1t,i}^{x,\theta} - y_1) \hat{Y}_{2t,i}^{x,\theta}(y_2),$$

where  $\hat{Y}_{t,i}^{x,\theta}$  is generated by the approximate model,

$$\hat{Y}_{1t,i}^{x,\theta} = g_{M,1t}(x, \varepsilon_i; \theta), \quad \hat{Y}_{2t,i}^{x,\theta}(y_2) = g_{M,2t}(y_2, x, \varepsilon_i; \theta), \quad i = 1, \dots, N.$$

Let  $\hat{\theta}_M$  be the associated approximate NPSMLE,

$$\hat{\theta}_M = \arg \max \hat{L}_{M,T}(\theta) \quad \hat{L}_{M,T}(\theta) = \sum_{t=1}^T \tau_a(\hat{p}_{M,t}(y_t|x_t; \theta)) \log \hat{p}_{M,t}(y_t|x_t; \theta).$$

We give regularity conditions under which  $\hat{\theta}_M$  has the same asymptotic properties as  $\hat{\theta}$  which is based on simulations from the true model. We impose the following condition on the sequence of approximate models, and on the rates of  $N$ ,  $h$ ,  $a$  relative to the approximation error.

**M.1** The sequence of approximate models  $\{g_M\}$  satisfies for some constants  $B_k, \lambda_{3,k}, \lambda_{4,k} \geq 0$ ,  $k = 1, 2$ :

$$\begin{aligned} \mathbb{E} \left[ \sup_{\theta \in \Theta} \|g_{M,1t}(x, \varepsilon; \theta) - g_{1t}(x, \varepsilon; \theta)\| \right] &\leq B_1 \left( 1 + \|x\|^{\lambda_{3,1}} + t^{\lambda_{3,2}} \right) \delta_{M,1}, \\ \mathbb{E} \left[ \sup_{\theta \in \Theta} \|g_{M,2t}(y_2, x, \varepsilon; \theta) - g_{2t}(y_2, x, \varepsilon; \theta)\| \right] &\leq B_2 \left( 1 + \|x\|^{\lambda_{4,1}} + t^{\lambda_{4,2}} \right) \delta_{M,2}, \end{aligned}$$

where  $\delta_{M,k} \rightarrow 0$  as  $M \rightarrow \infty$ .

**B.1'**  $T\nu_T^{-1}a^{-1}h^{-1} [N^{\gamma\lambda_{3,1}} + T^{\lambda_{3,2}}] \delta_{M,1} \rightarrow 0$  and  $T\nu_T^{-1}a^{-1} [N^{\gamma\lambda_{4,1}} + T^{\lambda_{4,2}}] \delta_{M,2} \rightarrow 0$ .

**B.2'**  $T\nu_T^{-1/2}a^{-1}h^{-1} [N^{\gamma\lambda_{3,1}} + T^{\lambda_{3,2}}] \delta_{M,1} \rightarrow 0$  and  $T\nu_T^{-1/2}a^{-1} [N^{\gamma\lambda_{4,1}} + T^{\lambda_{4,2}}] \delta_{M,2} \rightarrow 0$ .

**Theorem 4** *Assume that the conditions of Theorem 1 hold together with (M.1). Then the approximate NPSMLE satisfies:*

1.  $\mathcal{I}_T^{1/2}(\hat{\theta}_M - \theta_0) = o_P(1)$  for any sequences  $N$ ,  $h$ , and  $a$  satisfying (B.1) and (B.1');
2.  $\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\hat{\theta}_M - \theta_0) \xrightarrow{d} H_\infty^{-1}S_\infty$  for any sequences  $N$ ,  $h$ , and  $a$  satisfying (B.2) and (B.2').

One can exchange (M.1) for

$$\mathbb{E} \left[ \sup_{\theta \in \Theta} |\log p_{M,t}(y_{1t}, y_{2t}|x_t; \theta) - \log p_t(y_{1t}, y_{2t}|x_t; \theta)| \right] \leq \delta_M,$$

in which case 1 and 2 of Theorem 4 still go through under (B.1''),  $\delta_M \rightarrow 0$ , and (B.2''),  $\sqrt{\nu_T}\delta_M \rightarrow 0$  respectively. However, since  $p_{M,t}(y_t|x_t; \theta)$  and  $p_t(y_t|x_t; \theta)$  in most cases are difficult to analyze, condition (M.1) is in general easier to verify.

## 4 Implementing NPSML

One of the merits of NPSML is its general applicability. The applications include Markov decision processes (Pakes, 1994; Rust, 1994), and discretely-sampled diffusions, where  $p_t(y_t|x_t; \theta)$  typically does not have a closed-form representation but observations can still be simulated for NPSML.

In Section 4.1 we re-visit the jump-diffusion example of Section 2, and provide a detailed description on the implementation of NPSML in practice. We focus on this particular example for two reasons. Firstly, diffusion models can be described more concisely than a typical Markov decision model, which requires a detailed description of the economic environment. Secondly, the literature on estimating general jump diffusions has largely sidestepped maximum likelihood—See the discussion in Footnote 7. In this context, this estimation exercise showcases the usefulness of NPSML.

NPSML being for general purposes, other applications can be implemented in a similar way. At the implementation stage, only the part of the computer code that generates simulated observations needs to be modified. In Section 4.2, we briefly discuss how NPSML can be used for estimating generic Markov decision processes.

### 4.1 Discretely-Observed Diffusions

We consider a bivariate version of the model in (3).

$$dy_{1,t} = \left( \mu - \frac{\exp(y_{2,t})}{2} \right) dt + \exp\left(\frac{y_{2,t}}{2}\right) dW_{1,t} + \log(1 + J_t) dQ_t, \quad (20)$$

$$dy_{2,t} = (\alpha_0 - \alpha_1 y_{2,t}) dt + \alpha_2 dW_{2,t}. \quad (21)$$

This specification is used by Andersen, Benzoni, and Lund (2002) to model daily stock (S&P 500) returns. In their paper,  $y_{2,t}$  is an unobservable stochastic volatility process, and they use EMM for estimation. Here we assume that both  $y_{1,t}$  and  $y_{2,t}$  are observable. One interpretation is that we infer the volatility from derivative prices as in Ait-Sahalia and Kimmel (2007). Note that it is not our intention to replicate either paper.

The factors  $W_{1,t}$  and  $W_{2,t}$  are standard one-dimensional Brownian motions with correlation  $\rho$  between them.  $Q_t$  is a pure jump process with jump size 1, independent of  $W_{1,t}$  and  $W_{2,t}$ , and its jump intensity is given by  $\lambda_0$ . The jump size  $J_t$  is assumed to be log-normally distributed:

$$\log(1 + J_t) \sim \mathcal{N}(-0.5\gamma^2, \gamma^2). \quad (22)$$

Note that the parameter vector  $\theta \in \mathbb{R}^7$  is  $(\mu, \alpha_0, \alpha_1, \alpha_2, \gamma, \rho, \lambda_0)$ .

Ideally, we would like to give precise conditions under which the general jump diffusion (3) satisfies (A.1)–(A.4) and (C.1)–(C.5). However, this proves very difficult without imposing strong conditions ruling out standard models considered in empirical finance, including the current example

(20)–(21). Sufficient conditions for the existence of a twice-differentiable transition density for the general jump diffusion can be found in Bichteler et al. (1987) and Lo (1988), but these are rather restrictive and require, among other things, that the drift and diffusion term be linearly bounded and infinitely differentiable. The asymptotic properties of the MLE of general jump diffusions are not very well-understood yet due to the problems of not having the transition density on closed form. In a few special cases, its properties can be derived, e.g. Aït-Sahalia (2002).

In what follows, we first generate a continuous sample path  $\{(y_{1,t}, y_{2,t}) \in \mathbb{R}^2 : 0 \leq t \leq T\}$  from the true parameter values given in Table 1 (second column). We then assume that we observe this process only discretely, for  $t = 0, 1, \dots, T$ . Note that the discrete observations are temporally equidistant, with the interval length normalized to 1. We use these discrete observations  $\{(y_{1,t}, y_{2,t}) : t = 0, 1, \dots, T\}$  as our data. To generate this data series, we use the Euler scheme with the observation interval divided into 100 subintervals to approximate the diffusion process.

Then we use NPSML while forgoing our knowledge of the parameter values used for data generation. The first step of NPSML involves generating simulated observations from the model for any given  $\theta$ , and we use the Euler scheme to approximate the data generating process.<sup>10</sup> Given  $(y_{1,s}, y_{2,s})$  for some period  $s$ , we divide the interval between  $s + 1$  and  $s$  into  $M$  subintervals. In our benchmark estimation, we use  $M = 10$ . We recursively compute for  $m = 1, \dots, M$ :

$$\begin{aligned} u_{1,m}^i &= u_{1,m-1}^i + \left( \mu - \frac{\exp(u_{2,m-1}^i)}{2} \right) \frac{1}{M} + \exp\left( \frac{u_{2,m-1}^i}{2} \right) \frac{\widetilde{W}_{1,m}^i}{\sqrt{M}} + \log(1 + J_m^i) U_m^i, \\ u_{2,m}^i &= u_{2,m-1}^i + (\alpha_0 - \alpha_1 u_{2,m-1}^i) \frac{1}{M} + \alpha_2 \frac{W_{2,m}^i}{\sqrt{M}}, \end{aligned}$$

with  $u_{1,0}^i = y_{1,s}$  and  $u_{2,0}^i = y_{2,s}$  for all  $i = 1, \dots, N$ ;  $J_m^i$  is an i.i.d. random variable with its distribution given in (22);  $U_m^i$  is an i.i.d. binomial random variable, with  $Prob(U_m^i = 1) = \frac{\lambda_0}{M}$ ,<sup>11</sup>  $\widetilde{W}_{m,i}^1 = \sqrt{1 - \rho^2} W_{1,m}^i + \rho W_{2,m}^i$ , where  $W_{1,m}^i$  and  $W_{2,m}^i$  are i.i.d. standard normal random variables. The subscript  $i$  indexes simulations. In our benchmark estimation, we use  $N = 1,000$ .

With the simulated observations  $Y_{s+1,i}^\theta \equiv (u_{1,M}^i, u_{2,M}^i)$  for  $i = 1, \dots, N$ , we use (2) to obtain:

$$\hat{p}_{s+1}(y_{1,s+1}, y_{2,s+1} | y_{1,s}, y_{2,s}; \theta) = \frac{1}{N} \sum_{i=1}^N K_h \left( Y_{s+1,i}^\theta - (y_{1,s+1}, y_{2,s+1}) \right). \quad (23)$$

We use multiplicative Gaussian kernel  $K_h(\cdot) = K_{h_1}(\cdot)K_{h_2}(\cdot)$ , with the bandwidths  $h_1$  and  $h_2$  (for  $y_1$  and  $y_2$  respectively) given by the rule of thumb of Scott (1992, p.152). In particular,  $h_k = \hat{\sigma}_k N^{-1/6}$

<sup>10</sup>We are approximating a continuous-time process using a discretization scheme, and hence need to appeal to Theorem 4. Bruti-Liberati and Platen (2007) and Kloeden and Platen (1992) give conditions under which the discrete-time approximation satisfies condition (M.1). See also Detemple et al. (2006).

<sup>11</sup>To draw the binomial random variable  $U$ , we first generate a uniform  $[0,1]$  random number and determine whether it is less than  $Prob(U_m^i = 1)$ .

for  $k = 1, 2$ , where  $\hat{\sigma}_k$  is the sample standard deviation of  $y_{k,s}$  in the data. Note that we do not take advantage of the cross-correlation between  $y_{1,s}$  and  $y_{2,s}$  in the data, and instead use a simpler kernel and bandwidth.

With the estimated  $\hat{p}_t$  for  $t = 1, 2, \dots, T$ , we can evaluate the conditional likelihood, which is then maximized over the parameter space. As is typical for simulation-based estimations, when we maximize the likelihood function, we use the same set of random numbers for any  $\theta$ .<sup>12</sup>

Parameter	True Value	(1)	(2)	(3)	(4)
$\mu$	0.0304	0.0305 (0.0022,0.0524)	0.0305 (0.0162,0.0533)	0.0307 (0.0012,0.0661)	0.0306 (0.0084,0.0448)
$\alpha_0$	-0.0120	-0.0148 (-0.0181,-0.0100)	-0.0152 (-0.0186,-0.0111)	-0.0144 (-0.0188,-0.0088)	-0.0149 (-0.0185,-0.0100)
$\alpha_1$	0.0145	0.0161 (0.0116,0.2061)	0.0164 (0.0121,0.0207)	0.0160 (0.0114,0.0214)	0.0161 (0.0120,0.0214)
$\alpha_2$	0.1153	0.1147 (0.1107,0.1168)	0.1139 (0.1092,0.1185)	0.1167 (0.1118,0.1198)	0.1127 (0.1089,0.1156)
$\gamma$	0.0150	0.0199 (0.0060,0.0542)	0.0310 (0.0000,0.0368)	0.0100 (0.0017,0.0158)	0.0121 (0.0085,0.0126)
$\rho$	-0.6125	-0.7291 (-0.7595,-0.6863)	-0.7526 (-0.7984,-0.7012)	-0.6933 (-0.7189,-0.6592)	-0.7740 (-0.8064,-0.7344)
$\lambda_0$	0.0200	0.0169 (0.0101,0.0213)	0.0133 (0.0086,0.0175)	0.0196 (0.0122,0.0197)	0.0166 (0.0104,0.0222)

**Table 1: Estimation Results.** In each cell, the mean of the 100 point estimates in the simulation study is reported in the top half. In the bottom half, the 90% confidence interval constructed from the point estimates is reported. Column (1) is our benchmark with  $N = 1,000$  and the rule-of-thumb bandwidths. Column (2) reports the results with  $N = 750$ . Column (3) is for  $N = 1,000$  and bandwidths that are 20 percent narrower than those in the benchmark. Column (4) is for  $N = 1,000$  and bandwidths that are 20 percent wider than those in the benchmark.

In our simulation study, we draw 100 sample paths of length  $T = 1,000$  each, and estimate each sample path with NPSML. In column (1) of Table 1, we report the mean of the 100 point estimates for each parameter, and the 90% confidence interval constructed from these point estimates, with  $N = 1,000$  and the rule-of-thumb bandwidths. The NPSML performs reasonably well, although the correlation coefficient  $\rho$  is systemically underestimated. One remarkable outcome is that the jump parameters— $\gamma$  and  $\lambda_0$ —are rather precisely estimated, even though there are only 20 or so

<sup>12</sup>In the case of the binomial random variable  $U$ , we fix the realization of the underlying uniform random variable. For different  $\theta$ — $\lambda_0$ , to be exact,  $U$  itself may have different realizations.

jump realizations in each sample path.<sup>13</sup>

To assess how sensitive the estimation results are to the choice of  $N$  (number of artificial observations used for density estimation) and the kernel bandwidths, we try different  $N$  and bandwidths. In column (2), we reduce the number of artificial observations to  $N = 750$ . In column (3), we use  $N = 1,000$ , but reduce both bandwidths by 20 percent. Finally, in column (4), we use  $N = 1,000$  and bandwidths that are 20 percent greater than those in the benchmark.

When  $N$  is reduced to 750—column (2), the mean of the estimates move further away from the true parameter value. However, there is no clear increase or decrease in the dispersion of the parameter estimates. The results in column (3) are of particular interest to us. Our theoretical results suggest that bandwidths should be chosen to go to zero at a faster rate than in the standard cases. With a little under-smoothing as in column (3), the mean of the estimates are closer to the true parameter value than in the benchmark. Note the estimates of  $\rho$  in particular. On the other hand, the results with over-smoothing as in column (4) do not compare favorably with the benchmark results. We, in accordance with our theory, recommend a bandwidth narrower than what is given by the rule of thumb in actual implementations.

## 4.2 Markov Decision Processes and Dynamic Games

Another class of economic models that NPSML can readily be applied to is Markov decision processes: See Rust (1994) for an overview. In these models, the transition density is given by

$$p(y_t|x_t; \theta) = \int p(y_t|x_t, u_t)q(u_t)du_t,$$

where  $p(y_t|x_t, u_t)$  is typically governed by an optimal decision rule of a dynamic programming problem. The integral on the right-hand side does not have a closed-form representation, except in few special cases. However, conditioning on  $x_t$ , one can simulate  $u_t$  and hence  $y_t$ , and use kernel methods to estimate  $p(y_t|x_t)$ . Therefore, NPSML is feasible.

NPSML can also be used to estimate a related class of economic models: Markov-perfect equilibria of dynamic games. Ericson and Pakes (1995) provide a canonical framework for this literature: a dynamic model of oligopolistic industry with entry and exit. The equilibrium transition probability of this model is given by

$$p_t(\omega_{t+1}|\omega_t; \theta), \quad \omega \in \mathcal{Z}^n,$$

where  $\mathcal{Z}$  is a finite set of integers. The transition probability depends on individual firm-specific shocks, industry-wide shocks, and Markov-perfect strategies of firms regarding entry, exit and

---

<sup>13</sup>We ran the same exercise with trimming of the approximate log-likelihood. The results, with  $N$  being as large as 1,000, were virtually the same as in column (1).

investment.<sup>14</sup> Firms' strategies represent an optimal decision rule of a dynamic programming problem. Clearly, the transition probability does not have a closed-form representation, but it is still possible to simulate observations from the model conditioning on  $\omega_t$ . Thus, NPSML is feasible.<sup>15</sup> The computational burden of such models grow quickly with  $n$ . Doraszelski and Judd (2008) show how one can avoid this problem by casting the problems in continuous time. NPSML is readily applicable to such continuous-time dynamic stochastic games as well.

## 5 Concluding Remarks

We have generalized the nonparametric simulated maximum likelihood estimator of Fermanian and Salanié (2004) to deal with dynamic models, including nonstationary and time-inhomogeneous ones. Theoretical conditions in terms of the number of simulations and the bandwidth are given ensuring that the NPSML estimator inherits the asymptotic properties of the infeasible MLE.

This method is applicable to general classes of models, and can be implemented with ease. Our finite-sample simulation study demonstrates that the method works well in practice.

One limitation of the paper is that we only consider the cases where it is possible to simulate the dependent variable conditional on finitely-many past observations. This excludes the cases with latent dynamics. Extensions to methods with built-in nonlinear filters that explicitly account for latent variable dynamics are worked out in a companion paper (Kristensen and Shin, 2007) based on the main results given here.

---

<sup>14</sup>In this class of models, conditioning on  $\omega_t$ ,  $\omega_{t+1}$  depends not only on individual actions but also on idiosyncratic and aggregate shocks. To obtain the transition probability, all the shocks need to be integrated out.

<sup>15</sup>In solving individual firms' dynamic programming problem, one needs to know their continuation value, and hence the transition probability. Therefore, for a given  $\theta$ , one needs to compute a fixed point in  $Pr(\omega_{t+1}|\omega_t)$ . See Doraszelski and Pakes (2007) for a more detailed discussion.

## A Some General Results for Approximate Estimators

We first establish some general results for approximate MLEs. These results will then be applied to show the desired results for the proposed NPSML considered here.

In the following, let  $\hat{\theta} = \arg \max_{\theta \in \Theta} \hat{L}_T(\theta)$  and  $\tilde{\theta} = \arg \max_{\theta \in \Theta} L_T(\theta)$ , where  $L_T(\theta) = \nu_T^{-1} \sum_{t=1}^T \log p_t(y_t|x_t; \theta)$  is the true but infeasible log-likelihood, and  $\hat{L}_T(\theta) = \hat{L}_{T,N}(\theta)$  is a sequence of approximations to  $L_T(\theta)$  (not necessarily the simulated version proposed here). We then first establish the asymptotic properties of the true MLE under (C.1)–(C.4). Next, we give a general set of conditions for the approximate estimator,  $\hat{\theta}$ , to be asymptotically equivalent to  $\tilde{\theta}$ .

### A.1 Asymptotics of True MLE

**Lemma 5** *Assume that (C.1) and (C.4) hold. Then with probability tending to one, there exists a unique minimum point  $\tilde{\theta}$  of  $L_T(\theta)$  in  $\Theta_T$  which solves  $S_T(\tilde{\theta}) = 0$ . It satisfies  $\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) = o_P(1)$ .*

**Proof** Use a second order Taylor expansion to obtain for any bounded sequence  $\xi_T \in \mathbb{R}^d$  such that  $\theta_0 + \mathcal{I}_T^{-1/2}\xi_T \in \Theta_T$ ,

$$L_T(\theta_0 + \mathcal{I}_T^{-1/2}\xi_T) - L_T(\theta_0) = U_T(\theta_0)\xi_T + \frac{1}{2}\xi_T'V_T(\bar{\theta})\xi_T,$$

for some  $\bar{\theta} \in [\theta_0, \theta_0 + \mathcal{I}_T^{-1/2}h_T] \in \Theta_T$ . By another application of Taylor's Theorem,

$$\begin{aligned} |\xi_T'V_T(\bar{\theta})\xi_T - \xi_T'V_T(\theta_0)\xi_T| &= \left| \xi_T'\mathcal{I}_T^{-1/2} [H_T(\bar{\theta}) - H_T(\theta_0)] \mathcal{I}_T^{-1/2}\xi_T \right| \\ &\leq \max_{i=1, \dots, d} \sup_{\theta \in \Theta_T} |\xi_T'W_T(\theta)\xi_T| \times \|\mathcal{I}_T^{-1/2}\xi_T\| \\ &= O_P(\|\mathcal{I}_T^{-1/2}\xi_T\|) = o_P(1), \end{aligned}$$

where we have used (C.4.2) and the fact that  $\|\mathcal{I}_T^{-1/2}\xi_T\| = o_P(1)$ . Thus,

$$L_T(\theta_0 + \mathcal{I}_T^{-1/2}\xi_T) - L_T(\theta_0) = U_T(\theta_0)\xi_T + \frac{1}{2}\xi_T'V_T(\theta_0)\xi_T + o_P(1) = \frac{1}{2}\xi_T'H_\infty\xi_T + o_P(1),$$

where the second equality follows by (C.4.1). Since  $\xi_T'H_\infty\xi_T > 0$  a.s.,  $L_T(\theta)$  is convex with probability tending to one in the neighbourhood  $\Theta_T$ , and so a unique minimizer  $\tilde{\theta} \in \Theta_T$  exists which solves the first-order condition,  $S_T(\tilde{\theta}) = 0$ . We can choose  $\epsilon > 0$  in the definition of  $\Theta_T$  arbitrarily small, and conclude that the solution to  $S_T(\tilde{\theta}) = 0$  will still lie in  $\Theta_T$ . Thus,  $\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) = o_P(1)$ .  $\blacksquare$

**Lemma 6** *Assume that (C.1) and (C.4) hold. Then the MLE  $\tilde{\theta}$  satisfies*

$$\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) \xrightarrow{d} H_\infty^{-1}S_\infty.$$

**Proof** By Lemma 5, we know that  $\tilde{\theta}$  is consistent and solves the first order condition. A first order Taylor expansion of the score and using (C.4.2) together with the same arguments as in the proof of Lemma 5 yield

$$\sqrt{\nu_T}U_T(\theta_0) = V_T(\tilde{\theta})\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) = V_T(\theta_0)\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) + o_P(1)$$

such that, by (C.4.1),  $\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) = H_\infty^{-1}S_\infty + o_P(1)$ . ■

The following lemma states that (C.3) holds under (C.4).

**Lemma 7** *Assume that (C.4) holds. Then  $\xi \mapsto L_T(\theta_0 + \mathcal{I}_T^{-1/2}\xi)$  is stochastically equicontinuous.*

**Proof** Under the assumptions made,

$$L_T(\theta_0 + \mathcal{I}_T^{-1/2}\xi_1) - L_T(\theta_0 + \mathcal{I}_T^{-1/2}\xi_2) = \frac{1}{2}(\xi_1 - \xi_2)'H_\infty(\xi_1 - \xi_2) + o_P(1),$$

such that for any deterministic sequence  $\delta_n \rightarrow 0^+$ ,

$$\sup_{\|\xi_1 - \xi_2\| < \delta_n} \|L_T(\theta_0 + \mathcal{I}_T^{-1/2}\xi_1) - L_T(\theta_0 + \mathcal{I}_T^{-1/2}\xi_2)\| \leq \frac{1}{2}\|H_\infty\|\delta_n^2 + o_P(1) \rightarrow^P 0. \quad \blacksquare$$

## A.2 Asymptotics of Approximate MLE

**Theorem 8** *Assume that (C.1)–(C.3) hold and (\*)  $\sup_{\theta \in \Theta_T} |\hat{L}_T(\theta) - L_T(\theta)| = o_P(1)$  as  $T \rightarrow \infty$  for a sequence  $N = N(T) \rightarrow \infty$ . Then  $\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) = o_P(1)$ .*

**Proof** Define  $\hat{\xi}_T = \mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0)$  and  $\tilde{\xi}_T = \mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0)$ . We then wish to show that for any  $\eta > 0$ ,

$$P(\|\mathcal{I}_T^{1/2}(\hat{\theta} - \tilde{\theta})\| > \eta) = P(\|\tilde{\xi}_T - \hat{\xi}_T\| > \eta) \rightarrow 0, \quad T \rightarrow \infty.$$

Let  $\eta > 0$  be given. Then by (C.3) there exists a  $\delta > 0$  such that,  $\|\tilde{\xi}_T - \hat{\xi}_T\| > \eta$  implies  $|L_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) - L_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T)| \geq \delta$  with probability tending to 1. Thus, as  $T \rightarrow \infty$ ,

$$P(\|\tilde{\xi}_T - \hat{\xi}_T\| > \varepsilon) \leq P(|L_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) - L_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T)| \geq \delta).$$

We then have to show that the right-hand side converges to zero. Since  $\tilde{\theta}$  is the maximizer of  $L_T(\theta)$ , we know that  $L_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) \leq L_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T)$ . Thus,

$$\begin{aligned} |L_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) - L_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T)| &= L_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T) - L_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) \\ &= \left\{ \hat{L}_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) - L_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) \right\} \\ &\quad + \left\{ L_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T) - \hat{L}_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T) \right\}, \end{aligned}$$

where, by (\*),

$$\hat{L}_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) - L_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) \leq \sup_{\theta \in \Theta} |L_T(\theta) - \hat{L}_T(\theta)| = o_P(1)$$

while, by the definition of  $\hat{\theta}$  and again using (\*),

$$\begin{aligned} L_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T) - \hat{L}_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) &\leq L_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T) - \hat{L}_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{h}_T) \\ &\leq \sup_{\theta \in \Theta} |L_T(\theta) - \hat{L}_T(\theta)| = o_P(1). \quad \blacksquare \end{aligned}$$

Next, we state two results for the approximate estimator to have the same asymptotic distribution as the actual MLE. Theorem 9 establishes this result only requiring that the approximate likelihood function  $\sup_{\theta \in \Theta_T} |\hat{L}_T(\theta) - L_T(\theta)| = o_P(1/\sqrt{\nu_T})$ . Theorem 10 imposes stronger smoothness conditions, requiring that  $\hat{L}_T(\theta)$  be once differentiable; on the other hand we only require  $\sup_{\theta \in \Theta_T} \|\partial \hat{L}_T(\theta)/\partial \theta - \partial L_T(\theta)/\partial \theta\| = o_P(1/\|\sqrt{\nu_T}\mathcal{I}_T^{1/2}\|)$  which is a weaker convergence restriction than  $o_P(1/\sqrt{\nu_T})$ , since  $\|\mathcal{I}_T^{-1/2}\| = O(1)$ . So there is a trade-off between smoothness and rate of convergence.

**Theorem 9** *Assume that (C.1) and (C.4) hold. Then:*

- (i) *If  $\sup_{\theta \in \Theta_T} |\hat{L}_T(\theta) - L_T(\theta)| = o_P(1)$ , then  $\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) = o_P(1)$ .*
- (ii) *If  $\sup_{\theta \in \Theta_T} |\hat{L}_T(\theta) - L_T(\theta)| = o_P(1/\sqrt{\nu_T})$ , then  $\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) \xrightarrow{d} H_\infty^{-1}S_\infty$  and  $\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\hat{\theta} - \tilde{\theta}) = o_P(1)$ . In particular,  $\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} H_\infty^{-1}S_\infty$ .*

**Proof** Under (C.1) and (C.4), it holds that  $\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) = o_P(1)$  and that (C.3) is satisfied by Lemmas 5 and 7 respectively. Thus, the conditions of Theorem 8 holds which then yields (i).

To prove (ii), we use the same arguments as in the proof of Lemma 5 to obtain:

$$\begin{aligned} L_T(\tilde{\theta}) - L_T(\hat{\theta}) &= L_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T) - L_T(\theta_0 + \mathcal{I}_T^{-1/2}\hat{\xi}_T) \\ &= \frac{1}{2}(\tilde{\xi}_T - \hat{\xi}_T)' \mathcal{I}_T^{-1/2} H_T(\bar{\theta}) \mathcal{I}_T^{-1/2} (\tilde{\xi}_T - \hat{\xi}_T) \\ &= \frac{1}{2}(\tilde{\xi}_T - \hat{\xi}_T)' H_\infty (\tilde{\xi}_T - \hat{\xi}_T) + o_P(1), \end{aligned}$$

for some  $\bar{\theta} \in [\tilde{\theta}, \hat{\theta}]$ , since  $S_T(\tilde{\theta}) = S_T(\theta_0 + \mathcal{I}_T^{-1/2}\tilde{\xi}_T) = 0$  by the definition of  $\tilde{\theta}$ . We now use the exact same argument as in the proof of Theorem 8 for the left-hand side to obtain that

$$\sqrt{\nu_T}\{L_T(\tilde{\theta}) - L_T(\hat{\theta})\} = \sqrt{\nu_T}\{L_T(\tilde{\theta}) - \hat{L}_T(\hat{\theta})\} + \sqrt{\nu_T}\{\hat{L}_T(\hat{\theta}) - L_T(\hat{\theta})\} = o_P(1).$$

Thus,

$$\|\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\hat{\theta} - \tilde{\theta})\|^2 \leq \|H_\infty^{-1}\| \|\sqrt{\nu_T}(\tilde{\xi}_T - \hat{\xi}_T)'\| \|H_\infty(\tilde{\xi}_T - \hat{\xi}_T)\| = o_P(1). \quad \blacksquare$$

**Theorem 10** *Assume that (C.1) and (C4) hold together with:*

- (i)  $\theta \mapsto \hat{L}_T(\theta)$  is once differentiable in  $\Theta_T$ .

(ii) There exists a sequence  $N = N(T) \rightarrow \infty$  such that  $\sup_{\theta \in \Theta_T} \|\hat{S}_T(\theta) - S_T(\theta)\| = o_P(1/\|\sqrt{\nu_T} \mathcal{I}_T^{1/2}\|)$ .

Then  $\sqrt{\nu_T} \mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} H_\infty^{-1} S_\infty$  for this sequence  $N$ .

**Proof** By a standard Taylor expansion,

$$\sqrt{\nu_T} \mathcal{I}_T^{-1/2} S_T(\hat{\theta}) = \sqrt{\nu_T} \mathcal{I}_T^{-1/2} S_T(\theta_0) + \mathcal{I}_T^{-1/2} H_T(\theta_0) \mathcal{I}_T^{-1/2} \sqrt{\nu_T} \mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) + o_P(1),$$

where

$$\|\sqrt{\nu_T} \mathcal{I}_T^{1/2} S_T(\hat{\theta})\| = \|\sqrt{\nu_T} \mathcal{I}_T^{1/2} \{S_T(\hat{\theta}) - \hat{S}_T(\hat{\theta})\}\| \leq \sup_{\theta \in \Theta_T} \|\sqrt{\nu_T} \mathcal{I}_T^{1/2} \{S_T(\theta) - \hat{S}_T(\theta)\}\| = o_P(1).$$

Thus,

$$\sqrt{\nu_T} \mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) = \mathcal{I}_T^{-1/2} H_T(\theta_0) \mathcal{I}_T^{-1/2} \left\{ \sqrt{\nu_T} \mathcal{I}_T^{1/2} S_T(\theta_0) + o_P(1) \right\} \xrightarrow{d} H_\infty^{-1} S_\infty. \quad \blacksquare$$

## B Properties of Simulated Conditional Density

We here establish uniform convergence of  $\hat{p}_t$  given in equation (6) and its derivatives w.r.t.  $\theta$ .

**Lemma 11** Assume that (A.1)–(A.2) and (K.1) hold. Then  $\hat{p}_t$  given in (6) satisfies for all  $y_2 \in \mathcal{Y}_2$  and any compact set  $\Theta$ :

$$\begin{aligned} & \sup_{1 \leq t \leq T} \sup_{y_1 \in \mathbb{R}^k} \sup_{\|x\| \leq d_n} \sup_{\theta \in \Theta} |\hat{p}_t(y_1, y_2|x; \theta) - p_t(y_1, y_2|x; \theta)| \\ &= O_P \left( \left[ d_n^{\bar{\lambda}_1} + T^{\bar{\lambda}_2} \right] \log(N) / \sqrt{N h^k} \right) + O_P \left( \left[ d_n^{\lambda_{0,1}} + T^{\lambda_{0,2}} \right] h^r \right), \end{aligned}$$

where  $\bar{\lambda}_k = \lambda_{0,k} + \lambda_{1,k} + \lambda_{2,k}$ ,  $k = 1, 2$ .

**Proof** Define  $\gamma = (x, \theta, t) \in \Gamma = \mathcal{X}_t \times \Theta \times \{1, 2, 3, \dots\}$ . Write  $\hat{p}(y_1, y_2; \gamma) = \hat{p}(y_1, y_2|x; \theta)$  and  $p(y_1, y_2; \gamma) = p(y_1, y_2|x; \theta)$ . We split up into a bias and a variance component:

$$\begin{aligned} & \sup_{\|(x, \theta)\| \leq d_n} \sup_{t \leq T} |\hat{p}(y_1, y_2; \gamma) - p(y_1, y_2; \gamma)| \\ & \leq \sup_{\|(x, \theta)\| \leq d_n} \sup_{t \leq T} |\mathbb{E}[\hat{p}(y_1, y_2; \gamma)] - p(y_1, y_2; \gamma)| + \sup_{\|(x, \theta)\| \leq d_n} \sup_{t \leq T} |\hat{p}(y_1, y_2; \gamma) - \mathbb{E}[\hat{p}(y_1, y_2; \gamma)]| \\ & = : \text{Bias}(y_1, y_2; \gamma) + \text{Var}(y_1, y_2; \gamma). \end{aligned}$$

Using standard arguments for kernel estimators, the bias term can be shown to satisfy

$$|\text{Bias}(y_1, y_2; \gamma)| \leq h^r \int |K(v)| |v|^r dv \times \sup_{\|(x, \theta)\| \leq d_n} \sup_{t \leq T} \left| \frac{\partial^r p(y_1; \gamma)}{\partial y_1^r} \right| + o(h^r).$$

Thus, using the bound imposed on the  $r$ th derivative,  $|\text{Bias}(y_1, y_2; \gamma)| = O\left(\left[d_n^{\lambda_{0,1}} + T^{\lambda_{0,2}}\right] h^r\right)$  uniformly over  $(y_1, \gamma)$ . To establish the uniform rate of the variance term, we apply the result of Kristensen (2008b, Theorem 1) for averages of the form

$$\hat{\Psi}(x; \gamma) = \frac{1}{nh^d} \sum_{i=1}^n Y_i(\gamma) G\left(\frac{X_i(\gamma) - x}{h}\right), \quad (24)$$

for some kernel-type function  $G$ . With  $Y_i(\gamma) = Y_{2t,i}^{x,\theta}(y_2)$ ,  $X_i(\gamma) = Y_{1t,i}^{x,\theta}$  and  $G = K$ , our simulated density can be written on this form. We then verify that his conditions (A.1)–(A.5) are satisfied under our Assumptions (A.1)–(A.2) and his conditions (A.6) on  $G$  is implied by our (K.1). It's clear that our (K.1) implies his (A.6). Also, Kristensen (2008b, Assumptions A.1) is trivially satisfied since we have i.i.d. draws. Kristensen (2008b, Assumptions A.2) follows from our Assumption (A.1). Finally, the bounds in Kristensen (2008b, Assumptions A.4–A.5) in our case becomes, using Kristensen (2008b, Remark 2.2):

$$\begin{aligned} \tilde{B}_0 &= p(y_1; \gamma), \\ \tilde{B}_1 &= \|y_1\|^k \mathbb{E} \left[ \left| Y_{2t,i}^{x,\theta}(y_2) \right| \middle| Y_{1t,i}^{x,\theta} = y_1 \right] p(y_1; \gamma), \\ \tilde{B}_2 &= \|y_1\|^k \mathbb{E} \left[ \left| \dot{Y}_{2t,i}^{x,\theta}(y_2) \right| \middle| Y_{1t,i}^{x,\theta} = y_1 \right] p(y_1; \gamma), \\ \tilde{B}_3 &= \|y_1\|^k \mathbb{E} \left[ \left| Y_{2t,i}^{x,\theta}(y_2) \dot{Y}_{1t,i}^{x,\theta} \right| \middle| Y_{1t,i}^{x,\theta} = y_1 \right] p(y_1; \gamma), \end{aligned}$$

where  $\dot{Y}_{1t,i}^{x,\theta}$  and  $\dot{Y}_{2t,i}^{x,\theta}(y_2)$  are the derivatives w.r.t.  $(x, \theta, t)$ . By Assumption (A.2),  $\tilde{B}_0 = O\left(1 + \|x\|^{\lambda_{0,1}} + t^{\lambda_{0,2}}\right)$  while, using (A.1),

$$\begin{aligned} \mathbb{E} \left[ \left| Y_{2t,i}^{x,\theta}(y_2) \right| \middle| Y_{1t,i}^{x,\theta} = y_1 \right] &= \int_{\{e: g_1(e; \gamma) = y_1\}} |g_2(y_2, e; \gamma)| dF_\varepsilon(e) \\ &\leq \int |g_2(y_2, e; \gamma)| dF_\varepsilon(e) \\ &= \mathbb{E} [|g_{2,t}(y_2, \varepsilon; \gamma)|] \\ &\leq \mathbb{E} [\Lambda(\varepsilon)] \left[ 1 + \|x\|^{\lambda_{2,1}} + t^{\lambda_{2,2}} \right], \end{aligned}$$

and similarly for the two other conditional expectations in  $\tilde{B}_2$  and  $\tilde{B}_3$ . Thus,

$$\tilde{B}_k \leq \mathbb{E} [\Lambda(\varepsilon)] \left( 1 + \|x\|^{\lambda_1} + t^{\lambda_2} \right) \|y_1\|^q p(y_1; \gamma) = O\left( 1 + \|x\|^{\lambda_{0,1} + \lambda_{2,1}} + t^{\lambda_{0,2} + \lambda_{2,2}} \right),$$

for  $k = 1, 2$ , where the second equality follows from (A.2), while  $\tilde{B}_3 = O\left( 1 + \|x\|^{\bar{\lambda}_1} + t^{\bar{\lambda}_2} \right)$ .  $\blacksquare$

**Lemma 12** *Assume that (A.1)–(A.4) and (K.1) hold. Then  $\partial^i \hat{p}_t / \partial \theta^i$ ,  $i = 1, 2$ , given in (14) satisfy for all  $y_2 \in \mathcal{Y}_2$  and any compact set  $\Theta$ :*

$$\begin{aligned} &\sup_{1 \leq t \leq T} \sup_{y_1 \in \mathbb{R}^k} \sup_{\|x\| \leq d_n} \sup_{\theta \in \Theta} \left| \frac{\partial^i \hat{p}_t(y_1, y_2 | x; \theta)}{\partial \theta^i} - \frac{\partial^i p_t(y_1, y_2 | x; \theta)}{\partial \theta^i} \right| \\ &= O_P\left(\left[d_n^{\bar{\lambda}_1} + T^{\bar{\lambda}_2}\right] \log(N) / \sqrt{N h^{k+i}}\right) + O_P\left(\left[d_n^{\lambda_{0,1}} + T^{\lambda_{0,2}}\right] h^r\right). \end{aligned}$$

**Proof** We only give a proof for the first derivative. The proof for the second one follows along the same lines. We proceed as in the proof of Lemma 11. From the expression in equation (14),

$$\begin{aligned} \mathbb{E} \left[ \frac{\partial \hat{p}_t(y_1, y_2 | x; \theta)}{\partial \theta} \right] &= \frac{1}{h^k} \int K \left( \frac{Y_{1t,i}^{x,\theta} - y_1}{h} \right) \dot{Y}_{2t,i}^{x,\theta}(y_2) dF_\varepsilon(\varepsilon) \\ &\quad + \frac{1}{h^{k+1}} \int K' \left( \frac{Y_{1t,i}^{x,\theta} - y_1}{h} \right) \dot{Y}_{1t,i}^{x,\theta} Y_{2t,i}^{x,\theta}(y_2) dF_\varepsilon(\varepsilon), \end{aligned}$$

where, uniformly over  $(t, x, \theta)$ ,

$$\begin{aligned} \frac{1}{h^k} \int K \left( \frac{Y_{1t,i}^{x,\theta} - y_1}{h} \right) \dot{Y}_{2t,i}^{x,\theta}(y_2) dF_\varepsilon(\varepsilon) &= \int K(v) p(y_1 + vh | x; \theta) \frac{\partial p_t(y_2 | y_1 + vh, x; \theta)}{\partial \theta} dv \\ &= \frac{\partial p_t(y_2 | y_1, x; \theta)}{\partial \theta} p_t(y_1 | x; \theta) + O\left(\left[d_n^{\lambda_1} + T^{\lambda_2}\right] h^r\right), \\ \frac{1}{h^{k+1}} \int K' \left( \frac{Y_{1t,i}^{x,\theta} - y_1}{h} \right) \dot{Y}_{1t,i}^{x,\theta} Y_{2t,i}^{x,\theta}(y_2) dF_\varepsilon(\varepsilon) &= p_t(y_2 | y_1, x; \theta) \frac{\partial p_t(y_1 | x; \theta)}{\partial \theta} + O\left(\left[d_n^{\lambda_1} + T^{\lambda_2}\right] h^r\right). \end{aligned}$$

For the variance component, we again apply the results of Kristensen (2008b). With  $\gamma = (x, \theta, t)$  and  $X_{n,i}(\gamma) = Y_{t,i}^{x,\theta}$ ,  $\partial \hat{p}_t / \partial \theta$  can be written as the sum of two kernel averages, each of the form (24); the first with  $Y_{n,i}(\gamma) = \dot{Y}_{1t,i}^{x,\theta} Y_{2t,i}^{x,\theta}$  and  $G = K'$ , and the second with  $Y_{n,i}(\gamma) = \dot{Y}_{2t,i}^{x,\theta}$  and  $G = K$ . Under the conditions imposed on the model, Assumptions (A.1)–(A.5) of Kristensen (2008b) hold.

## C Proofs

**Proof of Theorem 1** The first part of the result will follow if we can verify the conditions in Theorem 8. In order to do this, we introduce an additional trimming function,  $\tilde{\tau}_{a,t} = \tau_a(\hat{p}_t(y_t | x_t; \theta)) \mathbb{I}\{\|x_t\| \leq N^\gamma\}$ , where  $\mathbb{I}\{\cdot\}$  is the indicator function and  $\gamma > 0$  is chosen as in (B.1), and two trimming sets,

$$A_{1,t}(\varepsilon) = \{\hat{p}_t(y_t | x_t; \theta) \geq \varepsilon a, \|x_t\| \leq N^\gamma\}, \quad A_{2,t}(\varepsilon) = \{p_t(y_t | x_t; \theta) \geq \varepsilon a, \|x_t\| \leq N^\gamma\},$$

for any  $\varepsilon > 0$ . Defining  $A_t(\varepsilon) = A_{1,t}(\varepsilon) \cap A_{2,t}(\varepsilon)$ , it follows by the same arguments as in Andrews (1995, p. 588),  $A_{2,t}(2\varepsilon) \subseteq A_{1,t}(\varepsilon) \subseteq A_t(\varepsilon/2)$  w.p.a.1 as  $N \rightarrow \infty$  under (B.1). Thus,  $\mathbb{I}_{A_{2,t}(4)} \leq \mathbb{I}_{A_{1,t}(2)} \leq \tilde{\tau}_{a,t} \leq \mathbb{I}_{A_{1,t}(1/2)} \leq \mathbb{I}_{A_t(1/4)}$ .

We then split up  $\hat{L}_T(\theta) - L_T(\theta)$  into three terms,

$$\begin{aligned} \hat{L}_T(\theta) - L_T(\theta) &= \frac{1}{\nu_T} \sum_{t=1}^T [\tau_a(\hat{p}_t(y_t | x_t; \theta)) - \tilde{\tau}_{a,t}] \log \hat{p}_t(y_t | x_t; \theta) \\ &\quad + \frac{1}{\nu_T} \sum_{t=1}^T \tilde{\tau}_{a,t} [\log \hat{p}_t(y_t | x_t; \theta) - \log p_t(y_t | x_t; \theta)] + \frac{1}{\nu_T} \sum_{t=1}^T [\tilde{\tau}_{a,t} - 1] \log p_t(y_t | x_t; \theta) \\ &=: B_1(\theta) + B_2(\theta) + B_3(\theta), \end{aligned}$$

and then show that  $\sup_{\theta \in \Theta} |B_i(\theta)| = o_P(1)$ ,  $i = 1, 2, 3$ . By (C.5),

$$|B_1(\theta)| \leq \frac{|\log a|}{\nu_T} \sum_{t=1}^T \mathbb{I}\{\|x_t\| > N^\gamma\} \leq \frac{|\log a| \nu_T^{q-1}}{N^{\gamma(1+\delta)}} \frac{1}{\nu_T^q} \sum_{t=1}^T \|x_t\|^{1+\delta} \leq \frac{|\log a| \nu_T^{q-1}}{N^{\gamma(1+\delta)}} \times O_P(1),$$

while,

$$\begin{aligned} |B_2(\theta)| &\leq \frac{1}{\nu_T} \sum_{t=1}^T \mathbb{I}_{A_t(1/4)} |\log \hat{p}_t(y_t|x_t; \theta) - \log p_t(y_t|x_t; \theta)| \\ &\leq \frac{T}{a\nu_T} \times \sup_{1 \leq t \leq T} \sup_{\theta \in \Theta} \sup_{y \in \mathbb{R}^k} \sup_{\|x\| \leq N^\gamma} |\hat{p}_t(y_t|x_t; \theta) - p_t(y_t|x_t; \theta)|. \end{aligned}$$

The final term is bounded by

$$\begin{aligned} |B_3(\theta)| &\leq \frac{1}{\nu_T} \sum_{t=1}^T |\tilde{\tau}_{a,t} - 1| |\log p_t(y_t|x_t; \theta)| \\ &\leq \frac{1}{\nu_T} \sum_{t=1}^T \mathbb{I}\{p_t(y_t|x_t; \theta) < 4a\} |\log p_t(y_t|x_t; \theta)| + \frac{1}{\nu_T} \sum_{t=1}^T \mathbb{I}\{\|x_t\| > N^\gamma\} |\log p_t(y_t|x_t; \theta)| \\ &=: B_{3,1}(\theta) + B_{3,2}(\theta). \end{aligned}$$

First, as  $a \rightarrow 0$ ,

$$\begin{aligned} |B_{3,1}(\theta)| &\leq \frac{1}{\nu_T} \sum_{t=1}^T \mathbb{I}\{p_t(y_t|x_t; \theta) < 4a\} |\log p_t(y_t|x_t; \theta)| \\ &= \frac{1}{\nu_T} \sum_{t=1}^T \mathbb{I}\{|\log p_t(y_t|x_t; \theta)| > |\log(4a)|\} |\log p_t(y_t|x_t; \theta)| \\ &\leq |\log(4a)|^{-\delta} \nu_T^{q-1} \frac{1}{\nu_T^q} \sum_{t=1}^T |\log p_t(y_t|x_t; \theta)|^{1+\delta} \\ &= |\log(4a)|^{-\delta} \nu_T^{q-1} \times O_P(1). \end{aligned}$$

where we have used (C.5). Similarly, by (C.5),

$$\begin{aligned} |B_{3,2}(\theta)| &\leq \frac{1}{\nu_T} \sum_{t=1}^T \mathbb{I}\{\|x_t\| > N^\gamma\} |\log p_t(y_t|x_t; \theta)| \\ &\leq \left\{ \frac{1}{\nu_T} \sum_{t=1}^T \mathbb{I}\{\|x_t\| > N^\gamma\} \right\}^{\delta/(1+\delta)} \left\{ \frac{1}{\nu_T} \sum_{t=1}^T |\log p_t(y_t|x_t; \theta)|^{1+\delta} \right\}^{1/(1+\delta)} \\ &\leq \frac{\nu_T^{q-1}}{N^{\gamma(1+\delta)}} \left\{ \frac{1}{\nu_T^q} \sum_{t=1}^T \|x_t\|^{1+\delta} \right\}^{\delta/(1+\delta)} \left\{ \frac{1}{\nu_T^q} \sum_{t=1}^T |\log p_t(y_t|x_t; \theta)|^{1+\delta} \right\}^{1/(1+\delta)} \\ &= \frac{\nu_T^{q-1}}{N^{\gamma(1+\delta)}} \times O_P(1). \end{aligned}$$

The consistency result now follows from Theorem 8 together with Lemma 11 and (B.1).

To show the second result, we merely have to strengthen the convergence of  $\hat{L}_T(\theta)$  to take place with rate  $\nu_T$ , c.f. Theorem 9(ii). One can still apply the above bounds which now have to go to zero with rate  $\nu_T$ . This is ensured by (B.2). ■

**Proof of Corollary 2** We verify that (C.1)–(C.5) hold under the conditions imposed in the corollary. First, with  $\nu_T = T$  and  $q = 1$ , we obtain by LLN for mixing sequences that  $i_T(\theta_0) = i(\theta_0) + o_P(1)$  with  $i(\theta_0) = \mathbb{E}[\partial^2 \log p(y_t|x_t; \theta_0)/(\partial\theta\partial\theta')]$ , such that  $\mathcal{I}_T$  can be chosen as the constant  $\mathcal{I} = \text{diag}\{i(\theta)\}$ . Thus, there is a one-to-one deterministic correspondence between the mapping  $\xi \mapsto L_T(\theta_0 + \mathcal{I}_T^{-1/2}\xi)$  and  $L_T(\theta)$  and we can restrict our attention to the latter. From e.g. Tauchen (1985) that  $\sup_{\theta \in \Theta} |L_T(\theta) - L(\theta)| = o_P(1)$  with  $L(\theta) = \mathbb{E}[\log p(y_t|x_t; \theta)]$  continuous under Condition (i). Thus, by Newey (1991),  $L_T(\theta)$  is stochastically equicontinuous and we have verified (C.3). Similarly, (C.5) follows by the (uniform) LLN,

$$\sup_{\theta \in \Theta_T} \left| T^{-1} \sum_{t=1}^T |\log p_t(y_t|x_t; \theta)|^{1+\delta} - \mathbb{E} \left[ |\log p(y_t|x_t; \theta)|^{1+\delta} \right] \right| \rightarrow^P 0,$$

$$T^{-1} \sum_{t=1}^T \|x_t\|^{1+\delta} - \mathbb{E} \left[ \|x_t\|^{1+\delta} \right] \rightarrow^P 0, \quad \nu_T^{-q} \sum_{t=1}^T \Lambda_1^2(\varepsilon_t) - \mathbb{E} \left[ \Lambda_1^2(\varepsilon_t) \right] \rightarrow^P 0.$$

To verify (C.4), appeal to the CLT for mixing sequences to obtain:

$$\sqrt{\nu_T} U_T(\theta_0) = \mathcal{I}^{-1/2} \times T^{-1/2} \sum_{t=1}^T \frac{\partial \log p_t(y_t|x_t; \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \rightarrow^d N \left( 0, \mathcal{I}^{-1/2} i(\theta_0) \mathcal{I}^{-1/2} \right),$$

while by the LLN,

$$V_T(\theta_0) = \mathcal{I}^{-1/2} \times T^{-1} \sum_{t=1}^T \frac{\partial^2 \log p_t(y_t|x_t; \theta)}{\partial \theta \partial \theta'} \Big|_{\theta=\theta_0} \times \mathcal{I}^{-1/2} \rightarrow^P -\mathcal{I}^{-1/2} i(\theta_0) \mathcal{I}^{-1/2}.$$

Finally,

$$\max_{j=1, \dots, d} \sup_{\theta \in \Theta} \|W_{j,T}(\theta)\| \leq C \sup_{\theta \in \Theta} \left\| T^{-1} \sum_{t=1}^T \frac{\partial^2 \log p_t(y_t|x_t; \theta)}{\partial \theta \partial \theta'} \Big|_{\theta=\theta_0} - i(\theta_0) \right\| + C i(\theta_0) = O_P(1). \quad \blacksquare$$

**Proof of Theorem 3** Since  $p(y|x; \theta)$  is bounded away from zero, we may here re-define the simulated likelihood as  $\hat{L}_T(\theta) = T^{-1} \sum_{t=1}^T \log \hat{p}(y_t|x_t; \theta)$  such that the associated score takes the form

$$\hat{S}_T(\theta) = T^{-1} \sum_{t=1}^T \frac{1}{\hat{p}(y_t|x_t; \theta)} \frac{\partial \hat{p}(y_t|x_t; \theta)}{\partial \theta}.$$

By the mean value theorem, for some  $\bar{\theta}$  on the line segment between  $\hat{\theta}$  and  $\theta_0$ .

$$0 = \hat{S}_T(\theta_0) + \hat{H}_T(\bar{\theta})(\hat{\theta} - \theta_0).$$

We then analyze the two terms,  $\hat{S}_T(\theta_0)$  and  $\hat{H}_T(\bar{\theta})$ , in turn. Define  $p_t(\theta) = p(y_t|x_t; \theta)$ ,  $\hat{p}_t(\theta) = \hat{p}(y_t|x_t; \theta)$  and  $s_t(\theta) = \partial_\theta \log p_t(\theta)$ .

We first consider  $\hat{S}_T(\theta_0)$  and suppress the dependence on  $\theta_0$  since this is fixed. The expansion in equation (17) now follows by Lee (1999, Proposition 1) with

$$\begin{aligned}\nabla S_{T,N}[\hat{p} - p] &= T^{-1} \sum_{t=1}^T \frac{1}{p_t} \{\partial_\theta \hat{p}_t - \partial_\theta p_t\} - T^{-1} \sum_{t=1}^T \frac{s_t}{p_t} \{\hat{p}_t - p_t\}, \\ \nabla^2 S_{T,N}[\hat{p} - p, \hat{p} - p] &= -T^{-1} \sum_{t=1}^T \frac{1}{p_t^2} \{\partial_\theta \hat{p}_t - \partial_\theta p_t\} \{\hat{p}_t - p_t\} + T^{-1} \sum_{t=1}^T \frac{s_t}{p_t^2} \{\hat{p}_t - p_t\}^2, \\ R_{T,N} &= -T^{-1} \sum_{t=1}^T \frac{1}{\hat{p}_t p_t^2} \{\partial_\theta \hat{p}_t - \partial_\theta p_t\} \{\hat{p}_t - p_t\}^2 + T^{-1} \sum_{t=1}^T \frac{s_t}{\hat{p}_t p_t^2} \{\hat{p}_t - p_t\}^3.\end{aligned}$$

We split up the first differential into a bias and a variance component,

$$\nabla S_{T,N}[\hat{p} - p] = \nabla S_{T,N}[\mathbb{E}[\hat{p}] - p] + \nabla S_{T,N}[\hat{p} - \mathbb{E}[\hat{p}]],$$

where  $\mathbb{E}[\hat{p}]$  denotes the expectation of  $\hat{p}$  conditional on data. Using standard bias expansions for kernel estimators,

$$\begin{aligned}\hat{p}(y_1, y_2|x) - \mathbb{E}[\hat{p}(y_1, y_2|x)] &= h^r \partial_{y_1}^r p(y_1, y_2|x) + o(h^r), \\ \partial_\theta \hat{p}(y_1, y_2|x) - \mathbb{E}[\partial_\theta \hat{p}(y_1, y_2|x)] &= h^r \partial_{y_1}^r \partial_\theta p(y_1, y_2|x) + o(h^r),\end{aligned}$$

implying that the bias component satisfies

$$\begin{aligned}\nabla S_{T,N}[\mathbb{E}[\hat{p}] - p] &= h^r T^{-1} \sum_{t=1}^T \frac{\partial_y^r \partial_\theta p_t}{p_t} - h^r T^{-1} \sum_{t=1}^T \frac{s_t}{p_t} \partial_y^r p_t + o_P(h^r) \\ &= \mu_1 h^r + o_P(h^r),\end{aligned}$$

with  $\mu_1$  given in equation (18). Next, define

$$\psi(z_t, \varepsilon_i) = \frac{1}{p_t} \delta_1(z_t, \varepsilon_i) - \frac{s_t}{p_t} \delta_2(z_t, \varepsilon_i),$$

where  $z_t = (y_t, x_t)$  and

$$\begin{aligned}\delta_1(z_t, \varepsilon_i) &= \frac{1}{h^{k+1}} K' \left( \frac{Y_{1,i}^{x_t} - y_{1t}}{h} \right) \dot{Y}_{1,i}^{x_t} Y_{2,i}^{x_t}(y_{2t}) + \frac{1}{h^k} K \left( \frac{Y_{1,i}^{x_t} - y_{1t}}{h} \right) \dot{Y}_{2,i}^{x_t}(y_{2t}), \\ \delta_2(z_t, \varepsilon_i) &= \frac{1}{h^k} K \left( \frac{Y_{1,i}^{x_t} - y_{1t}}{h} \right) Y_{2,i}^{x_t}(y_{2t}).\end{aligned}$$

Here,  $Y_{1,i}^x = g_1(x, \varepsilon_i; \theta_0)$  and  $\dot{Y}_{1,i}^x = \partial_\theta g_1(x, \varepsilon_i; \theta_0)$  and similarly for  $Y_{2,i}^x(y_2)$  and  $\dot{Y}_{2,i}^x(y_2)$ . With  $\psi_1(z_t) = \mathbb{E}[\psi(z_t, \varepsilon_i)|z_t]$ ,  $\psi_2(\varepsilon_i) = \mathbb{E}[\psi(z_t, \varepsilon_i)|\varepsilon_i]$  and  $\bar{\psi} = \mathbb{E}[\psi(z_t, \varepsilon_i)]$ , we can then write  $\nabla S_{T,N}[\hat{p} - \mathbb{E}[\hat{p}]]$  as

$$\nabla S_{T,N}[\hat{p} - \mathbb{E}[\hat{p}]] = \frac{1}{TN} \sum_{t=1}^T \sum_{i=1}^N \{\psi(z_t, \varepsilon_i) - \psi_1(z_t)\} = \frac{1}{N} \sum_{i=1}^N \{\psi_2(\varepsilon_i) - \bar{\psi}\} + A_{T,N},$$

where  $A_{T,N} = \frac{1}{TN} \sum_{t=1}^T \sum_{i=1}^N \{\psi(z_t, \varepsilon_i) - \psi_1(z_t) - \psi_2(\varepsilon_i) + \bar{\psi}\}$ . By standard results for  $U$ -statistics of  $\beta$ -mixing sequences, c.f. Denker and Keller (1983), we obtain that  $A_{T,N} = O_P(1/\sqrt{NT})$ , while  $\psi_2(\varepsilon_i) - \bar{\psi} = \bar{\psi}_2(\varepsilon_i) - \mathbb{E}[\bar{\psi}_2(\varepsilon_i)] + o_P(1/\sqrt{N})$ . This follows from

$$\begin{aligned} \mathbb{E} \left[ \frac{\delta_1(z_t, \varepsilon_i)}{p(y_t|x_t)} \middle| \varepsilon_i \right] &= \sum_{y_2 \in \mathcal{Y}_2} \int \dot{Y}_{1,i}^{x,\theta} Y_{2,i}^x(y_2) \left\{ \frac{1}{h^{k+1}} \int K' \left( \frac{Y_{1,i}^x - y_1}{h} \right) dy_1 \right\} dF_x(x) \\ &\quad + \sum_{y_2 \in \mathcal{Y}_2} \int \dot{Y}_{2,i}^x(y_2) \left\{ \frac{1}{h^k} \int K \left( \frac{Y_{1,i}^x - y_1}{h} \right) dy_1 \right\} dF_x(x) \\ &= \sum_{y_2 \in \mathcal{Y}_2} \int \dot{Y}_{2,i}^x(y_2) dF_x(x) \\ &= \mathbb{E} \left[ \frac{\dot{Y}_{2,i}^x(y_2)}{p(y_2|x)} \middle| \varepsilon_i \right] \end{aligned}$$

where we have used that  $h^{-(k+1)} \int K' \left( \frac{Y_i^x - y}{h} \right) dy = h^{-k} \int K'(v) dv = 0$ , and

$$\begin{aligned} \mathbb{E} \left[ \frac{s(y_t|x_t)}{p(y_t|x_t)} \delta_2(z_t, \varepsilon_i) \middle| \varepsilon_i \right] &= \sum_{y_2 \in \mathcal{Y}_2} \int \left\{ \frac{1}{h^k} \int s(y_1, y_2|x) K \left( \frac{Y_{1,i}^x - y_1}{h} \right) dy_1 \right\} Y_{2,i}^x(y_2) dF_x(x) \\ &= \sum_{y_2 \in \mathcal{Y}_2} \int s(Y_{1,i}^x, y_2|x) Y_{2,i}^x(y_2) dF_x(x) + O(h^r) \\ &= \mathbb{E} \left[ \frac{s(Y_{1,i}^x, y_2|x) Y_{2,i}^x(y_2)}{p(y_2|x)} \middle| \varepsilon_i \right] + O(h^r). \end{aligned}$$

The second order differential can be written as:

$$\begin{aligned} \nabla^2 S_{T,N} [\hat{p} - p, \hat{p} - p] &= \nabla^2 S_{T,N} [\mathbb{E}[\hat{p}] - p, \mathbb{E}[\hat{p}] - p] + \nabla^2 S_{T,N} [\hat{p} - \mathbb{E}[\hat{p}], \hat{p} - \mathbb{E}[\hat{p}]] \\ &\quad + 2\nabla^2 S_{T,N} [\mathbb{E}[\hat{p}] - p, \hat{p} - \mathbb{E}[\hat{p}]]. \end{aligned}$$

Since the cross-term is of a smaller order than the first two ones, we can ignore this and set it to zero in the following. Again using the bias expansion for kernel estimators and appealing to the LLN for stationary and ergodic sequences, the bias component satisfies

$$\begin{aligned} &\nabla^2 S_{T,N} [\mathbb{E}[\hat{p}] - p, \mathbb{E}[\hat{p}] - p] \\ &= -\frac{1}{T} \sum_{t=1}^T \frac{1}{p_t^2} \{ \mathbb{E}[\partial_\theta \hat{p}_t] - \partial_\theta p_t \} \{ \mathbb{E}[\hat{p}_t] - p_t \} + \frac{1}{T} \sum_{t=1}^T \frac{s_t}{p_t^2} \{ \mathbb{E}[\hat{p}_t] - p_t \}^2 \\ &= O_P(h^{2r}). \end{aligned}$$

The variance component can be written as

$$\nabla^2 S_{T,N} [\hat{p} - \mathbb{E}[\hat{p}], \hat{p} - \mathbb{E}[\hat{p}]] = \frac{1}{TN^2} \sum_{t=1}^T \sum_{i=1}^N \phi(z_t, \varepsilon_i, \varepsilon_i) + \frac{1}{TN^2} \sum_{t=1}^T \sum_{i \neq j} \phi(z_t, \varepsilon_i, \varepsilon_j) \quad (25)$$

$$=: \frac{1}{N} U_{T,N} + B_{T,N}, \quad (26)$$

with

$$\phi(z_t, \varepsilon_i, \varepsilon_j) = -\frac{1}{p_t^2} \delta_1(z_t, \varepsilon_i) \delta_2(z_t, \varepsilon_j) + \frac{s_t}{p_t^2} \delta_2(z_t, \varepsilon_i) \delta_2(z_t, \varepsilon_j).$$

The first term,  $U_{T,N}$ , is again a second order two-sample  $U$ -statistic while  $B_{T,N}$  is a third order one. We first analyze  $U_{T,N}$ . We have  $\mathbb{E}[\phi(z_t, \varepsilon_i, \varepsilon_i) | z_t] = 0$  and, using the same arguments as before,

$$\begin{aligned} \mathbb{E} \left[ \frac{1}{p_t^2} \delta_1(z_t, \varepsilon_i) \delta_2(z_t, \varepsilon_i) | \varepsilon_i \right] &= \frac{1}{h^{k+1}} \mathbb{E} \left[ \frac{\dot{Y}_{1,i}^{x_t} Y_{2,i}^x(y_{2t})^2}{p(Y_{1,i}^{x_t}, y_{2t} | x_t) p(y_{2t} | x_t)} \middle| \varepsilon_i \right] \int K^*(v) dv \\ &\quad + \frac{1}{h^k} \mathbb{E} \left[ \frac{Y_{2,i}^{x_t}(y_{2t}) \dot{Y}_{2,i}^x(y_{2t})}{p(Y_{1,i}^{x_t}, y_{2t} | x) p(y_{2t} | x_t)} \middle| \varepsilon_i \right] \int K^2(v) dv, \end{aligned}$$

where  $K^*(v) = K(v) K'(v)$ , and

$$\mathbb{E} \left[ \frac{s_t}{p_t^2} \delta_2^2(z_t, \varepsilon_i) | \varepsilon_i \right] = \frac{1}{h^k} \mathbb{E} \left[ \frac{Y_{2,i}^{x_t}(y_{2t}) \dot{Y}_{2,i}^x(y_{2t})}{p(Y_{1,i}^{x_t}, y_{2t} | x) p(y_{2t} | x_t)} \middle| \varepsilon_i \right] \int K^2(v) dv,$$

such that

$$\mathbb{E}[\phi(z_t, \varepsilon_i, \varepsilon_i)] = h^{-(k+1)} \mu_2 + h^{-k} \tilde{\mu}_2 + o(h^{-(k+1)}),$$

where  $\mu_2$  is given in equation (19) and  $\tilde{\mu}_2$  is the sum of the two other expectations above. Thus,  $U_{T,N}/N = 1/(Nh^{k+1}) \mu_2 + o_P(1/(Nh^{k+1}))$ . Next, we observe that:

$$\mathbb{E}[\phi(z_t, \varepsilon_i, \varepsilon_j) | z_t, \varepsilon_i] = \mathbb{E}[\phi(z_t, \varepsilon_i, \varepsilon_j) | z_t, \varepsilon_j] = 0,$$

while  $\mathbb{E}[\phi(z_t, \varepsilon_i, \varepsilon_j) | \varepsilon_i, \varepsilon_j] = o(1)$  as  $h \rightarrow 0$  such that the corresponding  $U$ -statistic  $B_{T,N}$  is second-order degenerate, implying  $B_{T,N} = O_P(1/(N^{3/2}h^{k+1}))$ . Finally, appealing to Lemma 12 and the LLN,

$$\begin{aligned} \sqrt{T} \|R_{T,N}\| &\leq T^{-1/2} \sum_{t=1}^T \frac{1}{|\hat{p}_t| p_t^2} \|\partial_\theta \hat{p}_t - \partial_\theta p_t\| \{\hat{p}_t - p_t\}^2 + T^{-1/2} \sum_{t=1}^T \frac{s_t}{|\hat{p}_t| p_t^2} |\hat{p}_t - p_t|^3 \\ &= O_P(\sqrt{T} h^{3r}) + O_P(\sqrt{T}/(Nh^{k+2})^{3/2}). \end{aligned}$$

Next, we consider the Hessian:

$$\hat{H}_T(\theta) = \frac{1}{T} \sum_{t=1}^T \frac{\partial_{\theta\theta}^2 \hat{p}_t}{\hat{p}_t} + \frac{1}{T} \sum_{t=1}^T \frac{\partial_\theta \hat{p}_t \partial_\theta \hat{p}_t'}{\hat{p}_t^2}.$$

We first write

$$\|\hat{H}_T(\bar{\theta}) - H_0\| \leq \sup_{\theta \in \Theta} \|\hat{H}_T(\theta) - H_T(\theta)\| + \sup_{\theta \in \Theta} \|H_T(\theta) - H(\theta)\| + \|H(\bar{\theta}) - H(\theta_0)\|.$$

The second and the third terms are both  $o_P(1)$  by standard ULLN results for stationary and ergodic sequences under the conditions we have imposed on  $\partial^2 \log p(y|x; \theta) / \partial \theta \partial \theta'$  (Newey, 1991). Regarding the first term, write

$$\begin{aligned} \hat{H}_T(\theta) - H_T(\theta) &= \frac{1}{T} \sum_{t=1}^T \left\{ \frac{\partial_{\theta\theta}^2 \hat{p}_t(\theta)}{\hat{p}_t(\theta)} - \frac{\partial_{\theta\theta}^2 p_t(\theta)}{p_t(\theta)} \right\} + \frac{1}{T} \sum_{t=1}^T \left\{ \frac{\partial_{\theta} \hat{p}_t(\theta) \partial_{\theta} \hat{p}_t(\theta)'}{\hat{p}_t^2(\theta)} - \frac{\partial_{\theta} p_t(\theta) \partial_{\theta} p_t(\theta)'}{p_t^2(\theta)} \right\} \\ &=: A_1 + A_2. \end{aligned}$$

The term  $A_1$  can be written as

$$\begin{aligned} A_1 &= \frac{1}{T} \sum_{t=1}^T \frac{1}{p_t(\theta)} \left[ 1 - \frac{1}{\hat{p}_t(\theta)} (\hat{p}_t(\theta) - p_t(\theta)) \right] \times \left[ (\partial_{\theta\theta}^2 \hat{p}_t(\theta) - \partial_{\theta\theta}^2 p_t(\theta)) - \frac{\partial_{\theta\theta}^2 p_t(\theta)}{p_t(\theta)} (\hat{p}_t(\theta) - p_t(\theta)) \right] \\ &= \frac{1}{T} \sum_{t=1}^T \frac{1 + o_P(1)}{p_t(\theta)} (\partial_{\theta\theta}^2 \hat{p}_t(\theta) - \partial_{\theta\theta}^2 p_t(\theta)) + \frac{1}{T} \sum_{t=1}^T \frac{\partial_{\theta\theta}^2 p_t(\theta) (1 + o_P(1))}{p_t^2(\theta)} (\hat{p}_t(\theta) - p_t(\theta)) \\ &= (1 + o_P(1)) \left\{ \frac{1}{T} \sum_{t=1}^T \frac{1}{p_t(\theta)} \right\} \sup_t \|\partial_{\theta\theta}^2 \hat{p}_t(\theta) - \partial_{\theta\theta}^2 p_t(\theta)\| \\ &\quad + (1 + o_P(1)) \left\{ \frac{1}{T} \sum_{t=1}^T \frac{\partial_{\theta\theta}^2 p_t(\theta)}{p_t^2(\theta)} \right\} \sup_t |\hat{p}_t(\theta) - p_t(\theta)| \\ &= o_P(1), \end{aligned}$$

where we have used the LLN together with  $\sup_t |\hat{p}_t(\theta) - p_t(\theta)| = o_P(1)$ ,  $\sup_t \|\partial_{\theta\theta}^2 \hat{p}_t(\theta) - \partial_{\theta\theta}^2 p_t(\theta)\| = o_P(1)$  and  $\inf_{t,\theta} p_t(\theta) > 0$ . We can show that  $A_2 = o_P(1)$  by the same arguments.

Combining the results for the score and the Hessian, with  $Z_T = \sqrt{T} \{S_T(\theta) + \nabla S_{T,N}[\hat{p} - \mathbb{E}[\hat{p}]]\}$ ,

$$\sqrt{T}(\hat{\theta} - \theta_0) = (H_0 + o_P(1))^{-1} \left\{ Z_T + \sqrt{T} h^r \mu_1 + \sqrt{T} / (Nh^k) \mu_2 + o_P(1) \right\} \xrightarrow{d} \mathcal{N} \left( \bar{c}_T, i(\theta_0)^{-1} + \frac{T}{N} \Omega \right),$$

where  $\Omega = i(\theta_0)^{-1} \text{Var}(\bar{\psi}_2(\varepsilon_i)) i(\theta_0)^{-1}$ .  $\blacksquare$

**Proof of Theorem 4** We follow the exact same arguments as in the proof of Theorem 3, except that we now have

$$\begin{aligned} \sup_{y \in \mathbb{R}^k} \sup_{\|x\| \leq N^\gamma} \sup_{\theta \in \Theta} |\hat{p}_{M,t}(y|x; \theta) - p_t(y|x; \theta)| &\leq \sup_{y \in \mathbb{R}^k} \sup_{\|x\| \leq N^\gamma} \sup_{\theta \in \Theta} |\hat{p}_{M,t}(y|x; \theta) - \hat{p}_t(y|x; \theta)| \\ &\quad + \sup_{y \in \mathbb{R}^k} \sup_{\|x\| \leq N^\gamma} \sup_{\theta \in \Theta} |\hat{p}_t(y|x; \theta) - p_t(y|x; \theta)|, \end{aligned}$$

where  $\hat{p}_t(y|x; \theta)$  is the (infeasible) kernel estimator based on simulations from the true model. The rate of convergence of the second term is given by Lemma 11, while the first term satisfies by (M.1),

$$\begin{aligned} |\hat{p}_{M,t}(y|x; \theta) - \hat{p}_t(y|x; \theta)| &= \frac{1}{N} \sum_{i=1}^N \left| K_h(\hat{Y}_{1t,i}^{x,\theta} - y_1) \hat{Y}_{2t,i}^{x,\theta}(y_2) - K_h(Y_{1t,i}^{x,\theta} - y_1) Y_{2t,i}^{x,\theta}(y_2) \right| \\ &\leq \frac{1}{Nh} \sum_{i=1}^N \left| K_h'(\bar{Y}_{t,i}^{x,\theta} - y) \right| \left| \hat{Y}_{2t,i}^{x,\theta}(y_2) \right| \left| \hat{Y}_{1t,i}^{x,\theta} - Y_{1t,i}^{x,\theta} \right| \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{N} \sum_{i=1}^N \left| K_h(Y_{1t,i}^{x,\theta} - y_1) \right| \left| \hat{Y}_{2t,i}^{x,\theta}(y_2) - Y_{2t,i}^{x,\theta}(y_2) \right| \\
& \leq \frac{\delta_{M,1}}{h} B_1 \left( 1 + \|x\|^{\lambda_{3,1}} + t^{\lambda_{3,2}} \right) \\
& \quad + \delta_{M,2} B_2 \left( 1 + \|x\|^{\lambda_{4,1}} + t^{\lambda_{4,2}} \right) + O_P(1),
\end{aligned}$$

where  $\bar{Y}_{t,i}^{x,\theta} = \lambda_{t,i}^{x,\theta} \hat{Y}_{t,i}^{x,\theta} + (1 - \lambda_{t,i}^{x,\theta}) Y_{t,i}^{x,\theta}$ ,  $\lambda_{t,i}^{x,\theta} \in [0, 1]$ . Thus,

$$\begin{aligned}
& \frac{T}{a\nu_T} \times \sup_{1 \leq t \leq T} \sup_{\theta \in \Theta} \sup_{y \in \mathbb{R}^k} \sup_{\|x\| \leq N^\gamma} |\hat{p}_{M,t}(y|x; \theta) - \hat{p}_t(y|x; \theta)| \\
& = O_P \left( \frac{T}{a\nu_T} \frac{\delta_{M,1}}{h} [N^{\gamma\lambda_{3,1}} + T^{\lambda_{3,2}}] \right) + O_P \left( \frac{T}{a\nu_T} \delta_{M,2} [N^{\gamma\lambda_{4,1}} + T^{\lambda_{4,2}}] \right),
\end{aligned}$$

where the right-hand side has to go to zero to obtain consistency. This holds by (B.1'). For first-order equivalence, we require that the right-hand side vanish with rate  $\sqrt{\nu_T}$ . This holds by (B.2').

■

## References

- AÏT-SAHALIA, Y. (2002): “Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approximation Approach,” *Econometrica*, 70, 223–262.
- (2004): “Closed-Form Likelihood Expansions for Multivariate Diffusions,” Manuscript, Princeton University.
- AÏT-SAHALIA, Y. AND R. KIMMEL (2007): “Maximum Likelihood Estimation of Stochastic Volatility Models,” *Journal of Financial Economics*, 83, 413–452.
- ALTISSIMO, F. AND A. MELE (2008): “Simulated Nonparametric Estimation of Dynamic Models,” *Review of Economic Studies*, forthcoming.
- ANDERSEN, T. G., L. BENZONI, AND J. LUND (2002): “An Empirical Investigation of Continuous-Time Equity Return Models,” *Journal of Finance*, 57, 1239–1284.
- ANDREWS, D. W. K. (1995): “Nonparametric Kernel Estimation for Semiparametric Models,” *Econometric Theory*, 11, 560–596.
- ANGO NZE, P. AND P. DOUKHAN (2004): “Weak Dependence: Models and Applications to Econometrics,” *Econometric Theory*, 20, 995–1045.
- BANDI, F. M. AND P. C. B. PHILLIPS (2003): “Fully Nonparametric Estimation of Scalar Diffusion Models,” *Econometrica*, 71, 241–283.
- BASAWA, I. V. AND D. J. SCOTT (1983): *Asymptotic Optimal Inference for Non-Ergodic Models*, New York: Springer-Verlag.
- BICHTELER, K., J.-B. GRAVERAUX, AND J. JACOD (1987): *Mallivin Calculus for Processes with Jumps*, Gordon and Breach Science Publishers.
- BRANDT, M. W. AND P. SANTA-CLARA (2002): “Simulated Likelihood Estimation of Diffusions with an Application to Exchange Rates Dynamics in Incomplete Markets,” *Journal of Financial Economics*, 63, 161–210.
- BRUTI-LIBERATI, N. AND E. PLATEN (2007): “Approximation of Jump Diffusions in Finance and Economics,” *Computational Economics*, 29, 283–312.
- CAI, Z., Q. YAO, AND W. ZHANG (2001): “Smoothing for Discrete-Valued Time Series,” *Journal of the Royal Statistical Society (B)*, 63, 357–375.
- CARRASCO, M., M. CHERNOV, J.-P. FLORENS, AND E. GHYSELS (2007): “Efficient estimation of general dynamic models with a continuum of moment conditions,” *Journal of Econometrics*, 140, 529–573.

- DENKER, M. AND G. KELLER (1983): “On U-Statistics and v. Mises’ Statistics for Weakly Dependent Processes,” *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 64, 505–522.
- DETEMPLE, J., R. GARCIA, AND M. RINDISBACHER (2006): “Asymptotic Properties of Monte Carlo Estimators of Diffusion Processes,” *Journal of Econometrics*, 134, 1–68.
- DONOHO, D. L., I. M. JOHNSTONE, G. KERKYACHARIAN, AND D. PICARD (1996): “Density Estimation by Wavelet Tresholding,” *Annals of Statistics*, 24, 508–539.
- DORASZELSKI, U. AND K. JUDD (2008): “Avoiding the Curse of Dimensionality in Dynamic Stochastic Games,” Manuscript, Harvard University.
- DORASZELSKI, U. AND A. PAKES (2007): “A Framework for Applied Dynamic Analysis in IO,” in *Handbook of Industrial Organization*, ed. by M. Armstrong and R. Porter, Amsterdam: Elsevier B.V., vol. 3, 1887–1966.
- DUFFIE, D. AND K. J. SINGLETON (1993): “Simulated Moments Estimation of Markov Models of Asset Prices,” *Econometrica*, 61, 929–952.
- ELERIAN, O., S. CHIB, AND N. SHEPHARD (2001): “Likelihood Inference for Discretely Observed Non-linear Diffusions,” *Econometrica*, 69, 959–993.
- ERICSON, R. AND A. PAKES (1995): “Markov-Perfect Industry Dynamics: A Framework for Empirical Work,” *Review of Economic Studies*, 62, 53–82.
- FENTON, V. M. AND A. R. GALLANT (1996): “Notes and Comments: Convergence Rates of SNP Density Estimators,” *Econometrica*, 64, 719–727.
- FERMANIAN, J.-D. AND B. SALANIÉ (2004): “A Nonparametric Simulated Maximum Likelihood Estimation Method,” *Econometric Theory*, 20, 701–734.
- GALLANT, A. R. AND D. W. NYCHKA (1987): “Semi-Nonparametric Maximum Likelihood Estimation,” *Econometrica*, 55, 363–390.
- GALLANT, A. R. AND G. TAUCHEN (1996): “Which Moments to Match?” *Econometric Theory*, 12, 657–681.
- GOURIÉROUX, C., A. MONFORT, AND É. RENAULT (1993): “Indirect Inference,” *Journal of Applied Econometrics*, 8, S85–S118.
- HAJIVASSILIOU, V. A. AND D. L. MCFADDEN (1998): “The Method of Simulated Scores for the Estimation of LDV Models,” *Econometrica*, 66, 863–896.
- HURN, A. S., K. A. LINDSAY, AND V. L. MARTIN (2003): “On the Efficacy of Simulated Maximum Likelihood for Estimating the Parameters of Stochastic Differential Equations,” *Journal of Time Series Analysis*, 24, 45–63.

- ICHIMURA, H. AND P. E. TODD (2007): “Implementing Nonparametric and Semiparametric Estimators,” in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Amsterdam: North Holland, vol. 6, 5369–5468.
- JEGANATHAN, P. (1995): “Some Aspects of Asymptotic Theory with Applications to Time Series Models,” *Econometric Theory*, 11, 818–887.
- JOHANNES, M. AND N. POLSON (2005): “MCMC Methods for Continuous-Time Financial Econometrics,” in *Handbook of Financial Econometrics*, ed. by Y. Aït-Sahalia and L. P. Hansen, Amsterdam: North Holland.
- KARLSEN, H. A. AND D. TJØSTHEIM (2001): “Nonparametric Estimation in Null Recurrent Time Series,” *Annals of Statistics*, 29, 372–416.
- KLOEDEN, P. E. AND E. PLATEN (1992): *Numerical Solution of Stochastic Differential Equations*, New York: Springer-Verlag.
- KRISTENSEN, D. (2008a): “Pseudo-Maximum-Likelihood Estimation in Two Classes of Semiparametric Diffusion Models,” Manuscript, Department of Economics, Columbia University.
- (2008b): “Uniform Convergence Rates of Kernel Estimators with Heterogenous, Dependent Data,” Manuscript, Department of Economics, Columbia University.
- KRISTENSEN, D. AND A. RAHBEK (2008): “Likelihood-Based Inference for Cointegration with Nonlinear Error-Correction,” *Journal of Econometrics*, forthcoming.
- KRISTENSEN, D. AND Y. SHIN (2007): “Estimation of Hidden Markov Models with Nonparametric Simulated Maximum Likelihood,” Manuscript, Department of Economics, Columbia University.
- LEE, B.-S. AND B. F. INGRAM (1991): “Simulation Estimation of Time-Series Models,” *Journal of Econometrics*, 47, 197–205.
- LEE, D. AND K. SONG (2006): “A Consistent Simulated MLE for Discrete Choices When the Number of Simulations is Finite,” Manuscript, Department of Economics, University of Pennsylvania.
- LEE, L.-F. (1992): “On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models,” *Econometric Theory*, 8, 518–552.
- (1995): “Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models,” *Econometric Theory*, 11, 437–483.
- (1999): “Statistical Inference with Simulated Likelihood Functions,” *Econometric Theory*, 15, 337–360.
- LEE, S.-W. AND B. E. HANSEN (1994): “Asymptotic Theory for the GARCH(1,1) Quasi-Maximum Likelihood Estimator,” *Econometric Theory*, 10, 29–52.

- LI, Q. AND J. RACINE (2007): *Nonparametric Econometrics: Theory and Practice*, New Jersey: Princeton University Press.
- LO, A. W. (1988): “Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data,” *Econometric Theory*, 4, 231–247.
- MANRIQUE, A. AND N. SHEPHARD (1998): “Simulation-Based Likelihood Inference for Limited Dependent Processes,” *Econometrics Journal*, 1, 174–202.
- MCFADDEN, D. (1989): “A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration,” *Econometrica*, 57, 995–1026.
- NEWHEY, W. K. (1991): “Uniform Convergence in Probability and Stochastic Equicontinuity,” *Econometrica*, 59, 1161–1167.
- PAKES, A. (1994): “The Estimation of Dynamic Structural Models: Problems and Prospects, Part II: Mixed Continuous-Discrete Models and Market Interactions,” in *Advances in Econometrics: Proceedings of the Sixth World Congress of the Econometric Society*, ed. by J.-J. Laffont and C. A. Sims, Cambridge University Press, 171–259.
- PAKES, A. AND D. POLLARD (1989): “Simulation and the Asymptotics of Optimization Estimators,” *Econometrica*, 57, 1027–1057.
- PARK, J. Y. AND P. C. B. PHILLIPS (2001): “Nonlinear Regressions with Integrated Time Series,” *Econometrica*, 69, 117–161.
- PEDERSEN, A. R. (1995a): “Consistency and Asymptotic Normality of an Approximate Maximum Likelihood Estimator for Discretely Observed Diffusion Processes,” *Bernoulli*, 1, 257–279.
- (1995b): “A New Approach to Maximum-Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations,” *Scandinavian Journal of Statistics*, 22, 55–71.
- PHILLIPS, P. C. B. (1983): “ERA’s: A New Approach to Small Sample Theory,” *Econometrica*, 51, 1505–1525.
- RUST, J. (1994): “Structural Estimation of Markov Decision Processes,” in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden, Amsterdam: North Holland, vol. 4, 3081–3143.
- SAIKKONEN, P. (1995): “Problems with the Asymptotic Theory of Maximum Likelihood Estimation in Integrated and Cointegrated Systems,” *Econometric Theory*, 11, 888–911.
- SANDMANN, G. AND S. J. KOOPMAN (1998): “Estimation of Stochastic Volatility Models via Monte Carlo Maximum Likelihood,” *Journal of Econometrics*, 87, 271–301.
- SCHAUMBURG, E. (2001): “Maximum Likelihood Estimation of Lévy Type SDEs,” Ph.D. thesis, Princeton University.

- SCOTT, D. W. (1992): *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: Wiley.
- SMITH, JR., A. (1993): “Estimating Nonlinear Time-Series Models Using Simulated Vector Autoregressions,” *Journal of Applied Econometrics*, 8, S63–S84.
- STONE, C. J. (1990): “Large-Sample Inference for Log-Spline Models,” *Annals of Statistics*, 18, 717–741.
- SUNDARESAN, S. M. (2000): “Continuous-Time Methods in Finance: A Review and an Assessment,” *Journal of Finance*, 55, 1569–1622.
- TAUCHEN, G. (1985): “Diagnostic Testing and Evaluation of Maximum Likelihood Models,” *Journal of Econometrics*, 30, 415–443.
- (1997): “New Minimum Chi-Square Methods in Empirical Finance,” in *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress*, ed. by D. M. Kreps and K. F. Wallis, Cambridge: Cambridge University Press, 279–317.
- WAHBA, G. (1981): “Data-Based Optimal Smoothing of Orthogonal Series Density Estimates,” *Annals of Statistics*, 9, 146–156.
- WHITE, JR., H. L. (1984): “Maximum Likelihood Estimation of Misspecified Dynamic Models,” in *Misspecification Analysis*, ed. by T. K. Dijkstra, New York: Springer-Verlag, 1–19.
- YU, J. (2007): “Closed-Form Likelihood Approximation and Estimation of Jump-Diffusions with an Application to the Realignment Risk of the Chinese Yuan,” *Journal of Econometrics*, Forthcoming.