Journal of Econometrics 167 (2012) 76–94

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Estimation of dynamic models with nonparametric simulated maximum likelihood☆

Dennis Kristensen[a,b,c], Yongseok Shin[d,e,*]

[a] Department of Economics, University College London, United Kingdom
[b] Department of Economics, Columbia University, United States
[c] CREATES, Aarhus University, Denmark
[d] Department of Economics, Washington University in St. Louis, United States
[e] Federal Reserve Bank of St. Louis, United States

ARTICLE INFO

Article history:
Received 19 January 2006
Received in revised form
9 June 2011
Accepted 21 September 2011
Available online 10 November 2011

ABSTRACT

We propose an easy-to-implement simulated maximum likelihood estimator for dynamic models where no closed-form representation of the likelihood function is available. Our method can handle any simulable model without latent dynamics. Using simulated observations, we nonparametrically estimate the unknown density by kernel methods, and then construct a likelihood function that can be maximized. We prove that this nonparametric simulated maximum likelihood (NPSML) estimator is consistent and asymptotically efficient. The higher-order impact of simulations and kernel smoothing on the resulting estimator is also analyzed; in particular, it is shown that the NPSML does not suffer from the usual curse of dimensionality associated with kernel estimators. A simulation study shows good performance of the method when employed in the estimation of jump–diffusion models.

1. Introduction

We propose a simulated maximum likelihood estimator for dynamic models based on nonparametric kernel methods. Our method is designed for models where no closed-form representation of the likelihood function is available. Our method can handle any simulable model without latent dynamics. For any given parameter value, conditioning on available past 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 estimator (NPSMLE). NPSML was introduced by Fermanian and Salanié (2004), who obtained theoretical results only for static models. In this paper, we generalize their method to dynamic models, including nonstationary and time-inhomogeneous processes. We give general conditions for the NPSMLE 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 NPSMLE. In particular, we show that the estimator does not suffer from the curse of dimensionality despite the use of kernel smoothers. Finally, we show that the theoretical results remain valid even if only simulations from an approximate model are available.

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 (Rust, 1994; Doraszelski and Pakes, 2007). 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, a large class of continuous-time processes, including jump–diffusions, can be approximated with various discretization schemes to a given level of precision, and we can simulate observations from the model which are then used for NPSML.

☆ We thank the seminar participants at Berkeley, BU, Brown, Columbia, LSE, NYU, Rice, and Stanford for many useful comments. We also thank the referees who offered exceptionally thorough and helpful comments. Kyu-Chul Jung provided excellent research assistance. Kristensen gratefully acknowledges the financial support of the National Science Foundation (SES-0961596) and the Danish Research Foundation (through a grant to CREATES).

* Corresponding author at: Department of Economics, Washington University in St. Louis, United States Tel.: +1 314 935 7138.
E-mail addresses: d.kristensen@ucl.ac.uk (D. Kristensen), yshin@wustl.edu (Y. Shin).

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.09.042

Indeed, we investigate the performance of NPSML when applied to jump–diffusion models with particular attention to the impact of the number of simulations and bandwidth. We find that NPSML performs well even for a moderate number of simulations and that it is quite robust to the choice of bandwidth.

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 (Lee and Ingram, 1991; Duffie and Singleton, 1993; Creel and Kristensen, 2009), indirect inference (Gouriéroux et al., 1993; Smith, 1993; Creel, 2011), 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, Carrasco et al. (2007) and Altissimo and Mele (2009) developed general-purpose estimators based on matching a continuum of moments that are asymptotically as efficient as MLEs for fully-observed systems. One attractive feature of NPSML – which it shares with Carrasco et al. (2007) and Altissimo and Mele (2009) – 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 fully automated to yield full efficiency (Jones et al., 1996). Another advantage is that, unlike most of the above methods, 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 is various maximum likelihood methods for discretely sampled diffusions (Pedersen, 1995a,b; Sandmann and Koopman, 1998; Elerian et al., 2001; Aït-Sahalia, 2002, 2008; Brandt and Santa-Clara, 2002). 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 simulates observations needs to be modified, leaving other parts (e.g., kernel estimation of conditional density or numerical maximization of likelihood) unchanged.

The basic implementation of our method requires that it is possible to simulate the current variables of the model conditioning on finitely-many past observations. This excludes models with latent dynamics since these cannot be simulated one step at a time. Nevertheless, our method can be modified to handle latent dynamics, but this modified version will not obtain full efficiency (Section 2.2). Extensions of our method that obtain full efficiency in the presence of latent dynamics are worked out in a companion paper (Brownlees et al., 2011), 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 NPSMLE. In Section 3, we derive the asymptotic properties of the NPSMLE under regularity conditions. Section 4 provides a detailed description on implementing NPSML with numerical examples, and Section 5 concludes the paper.

2. Nonparametric simulated maximum likelihood

2.1. Construction of NPSMLE

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 is generated by a fully parametric model:

$$y_t = g_t(x_t, \varepsilon_t; \theta), \quad t = 1, \ldots, T, \tag{1}$$

where $\theta \in \Theta \subseteq \mathbb{R}^d$ is an unknown parameter vector, and ε_t is an i.i.d. sequence with known distribution F_ε. Without loss of generality, we assume that F_ε does not depend on t or θ. Our setting accommodates Markov models where $x_t \equiv y_{t-1}$, 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 y_t, in which case (1) is only a partially-specified model. Also, we allow the processes (y_t, x_t) to be nonstationary, for example due to unit-root-type behavior or deterministic time trends.

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

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

for any Borel set $A \subseteq \mathbb{R}^k$. A natural estimator of θ 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 model (1) is fully specified, i.e. x_t only contains lagged y_t, 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 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 θ is not feasible. In terms of model (1), this occurs when either the inverse of $g_t(x_t, \varepsilon_t; \theta)$ w.r.t. ε_t does not exist, or when the inverse does not have a closed-form expression.[1] 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.

Here we propose a general method to obtain a simulated conditional density, which in turn will be used to obtain a simulated version of the MLE. For any given $1 \leq t \leq T$, $y_t \in \mathbb{R}^k$, $x_t \in \mathcal{X}_t$, and $\theta \in \Theta$, we wish to compute a simulated version of $p_t(y_t|x_t; \theta)$. To this end, we first generate N i.i.d. draws from F_ε, $\{\varepsilon_i\}_{i=1}^N$, and use these to compute

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

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

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

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

where $K_h(\upsilon) = K(\upsilon/h)/h^k$, $K : \mathbb{R}^k \mapsto \mathbb{R}$ is a kernel, and $h > 0$ a bandwidth.[2] Under regularity conditions on p_t and K, we obtain

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

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

Once (2) has been used to obtain the simulated conditional density, we can now construct the following simulated MLE of θ_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 θ. We may also use the same batch of draws from $F_\varepsilon(\cdot)$, $\{\varepsilon_i\}_{i=1}^N$, across t. Numerical optimization is facilitated if $\hat{L}_T(\theta)$ is continuous and differentiable in θ. 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 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 θ.

Since $\hat{p}_t(y_t|x_t;\theta) \xrightarrow{P} p_t(y_t|x_t;\theta)$, $\hat{L}_T(\theta) \xrightarrow{P} L_T(\theta)$ as $N \to \infty$ 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 $T, N \to \infty$, under suitable conditions.

The precision of $\hat{\theta}$ relative to $\tilde{\theta}$ clearly depends on the quality of the approximation of $p_t(y|x;\theta)$ by $\hat{p}_t(y|x;\theta)$. Let us note the following important points concerning the impact of the simulated density. First, 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 whether the observed data are i.i.d. or nonstationary. Second, the simulated density, $\hat{p}_t(y|x;\theta)$, suffers from the usual curse of dimensionality for kernel density estimators with its variance being of order $1/(Nh^k)$. The curse of dimensionality only depends on $k \equiv \dim(y_t)$ here since we do not smooth over x_t, and so the dimension of x_t is irrelevant in itself. Still one could be concerned that for high-dimensional models (i.e., a large k), a much larger number of simulations must be used to control the variance component of the resulting estimator relative to, say, the standard simulated method of moments (where the variance is of order $1/N$). However, as we demonstrate in Section 3, this is fortunately not the case: the summation of the log-densities in the computation of $\hat{L}_T(\theta)$ functions as an additional smoothing device such that the additional variance of $\hat{L}_T(\theta)$ due to simulations recovers the standard parametric rate $1/N$. This a well-known phenomenon in the literature on two-step semiparametric estimators: they can obtain parametric rates despite the fact that they depend on first-step nonparametric estimators; e.g. Kristensen (2010).

On the other hand, a disadvantage of our estimator is that for a fixed N and $h > 0$, the simulated log-likelihood function is a biased estimate of the actual one. First, to obtain consistency, we will have to let $N \to \infty$ which is a feature shared by most non-linear simulation-based likelihood methods.[3] In addition, we have to let $h \to 0$ which is an additional complication relative to other simulation-based estimators where no such nuisance parameter appears. In particular, one has to choose h for a given sample and simulation size. However, if one is willing to make a stronger assumption about the identification of the model, the choice of the bandwidth may be less important. For example, in the stationary case, the standard identification assumption states that $\mathbb{E}[\log p(y_t|x_t;\theta)] < \mathbb{E}[\log p(y_t|x_t;\theta_0)]$ for $\theta \neq \theta_0$. A stronger identification condition implying the former is

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

for all $0 \leq h \leq \bar{h}$ for some $\bar{h} > 0$.[4] Under the latter identification condition, one can show consistency of our estimator for any fixed $0 < h \leq \bar{h}$ as $N \to \infty$. A similar identification condition can be found in Altissimo and Mele (2009). Still, for a fixed $h > 0$, the resulting estimator will no longer enjoy full efficiency. To obtain this, one has to let $h \to 0$. Moreover, the above argument assumes knowledge of the threshold $\bar{h} > 0$ for a given model, and so, in practice, the bandwidth selection problem still remains. However, it suggests that one can still identify parameters in large samples when a given $h > 0$ is chosen and that the NPSMLE will be fairly robust to the choice of h. This is supported by our simulation study, which shows that the NPSML performs well within a fairly broad range of bandwidth choices. Still a careful choice of the bandwidth will in general lead to better performance of the estimator.

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

Example: discretely-observed jump–diffusion. Consider an $\mathbb{R}^k$-dimensional continuous-time stochastic process $\{y_t : t \geq 0\}$ that solves the following stochastic differential equation:

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

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

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 $y_1, \ldots, y_T$ – without loss of generality, we normalize the time interval between observations to 1 – and wish to estimate θ by maximum likelihood. Although under regularity conditions (Lo, 1988) the transition density $p_t(y|x;\theta)$ satisfying $P(y_{t+1} \in A|y_t = x) = \int_A p_t(y|x;\theta)dy$ is well-defined, it cannot, in general, be written in closed form, which in turn complicates estimation.[5] However, discretization schemes (Kloeden and Platen, 1992; Bruti-Liberati and Platen, 2007) can be used to simulate observations from the model for any given level of accuracy, enabling NPSML. We re-visit this example in Section 4, where we provide a detailed description of implementing NPSML in practice.

[2] Here and in the following, we will use K to denote a generic kernel.

[3] See Lee and Song (2009) for an exception.

[4] This follows from the following inequality:

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

[5] Schaumburg (2001) and White (2007), building on the approach of Aït-Sahalia (2002), 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 simulated MLE of Pedersen (1995a,b) or Brandt and Santa-Clara (2002) needs to be substantially modified before they can be applied to Lévy processes.

2.2. Extensions and alternative schemes

Latent dynamics. Our method can be modified to handle latent dynamics. Suppose y_t is generated from

$$[y_t, w_t] = g(y_{t-1}, w_{t-1}, \varepsilon_t; \theta),$$

where w_t is unobserved/latent and $\varepsilon_t \overset{i.i.d.}{\sim} F_\varepsilon$. The full likelihood function will require computation of conditional densities on the form $p(y_t|y_{t-1}, y_{t-2}, \ldots, y_0; \theta)$, which in general is complicated due to the expanding information set. We can however construct a simulated version of the following "limited-information" likelihood (LIL) given by $L_T(\theta) = \sum_{t=1}^T \log p(y_t|x_t; \theta)$ where x_t is a set of conditioning variables chosen by the econometrician, say, $x_t = (y_{t-1}, \ldots, y_{t-m})$ for some $m \geq 1$. There will be an efficiency loss from estimating θ using this LIL relative to the full likelihood, but the LIL is easier to implement. First, simulate a (long) trajectory $\{Y_t^\theta\}_{t=1}^{\widetilde{N}}$ by

$$[Y_t^\theta, W_t^\theta] = g(Y_{t-1}^\theta, W_{t-1}^\theta, \varepsilon_t; \theta), \quad t = 1, \ldots, \widetilde{N},$$

where $\{\varepsilon_t\}_{t=1}^{\widetilde{N}}$ are i.i.d. draws from F_ε. We can then use these simulations to construct a simulated version of $p(y_t|x_t; \theta)$ by the following kernel estimator of the conditional density,

$$\check{p}(y|x; \theta) = \frac{\sum_{t=1}^{\widetilde{N}} K_h(Y_t^\theta - y)K_h(X_t^\theta - x)}{\sum_{t=1}^{\widetilde{N}} K_h(X_t^\theta - x)}. \tag{4}$$

where $X_t^\theta = (Y_{t-1}^\theta, \ldots, Y_{t-m}^\theta)$. Similar ideas were utilized in Altissimo and Mele (2009) and Creel and Kristensen (2009).

A disadvantage of the above method is that the convergence of $\check{p}$ relative to $\hat{p}$ will be slower due to (i) the dimension of (Y_t^θ, X_t^θ) can potentially be quite large and (ii) the simulated variables are now dependent. One will have to choose a larger $\widetilde{N}$ for the simulated conditional density in (4) relative to the one in (2). To handle (ii), one will typically have to assume a stationary solution to the dynamic system under consideration, and either 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 (2009) and Creel and Kristensen (2009). Then a large value of $\widetilde{N}$ can 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 (4) 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 (Karlsen and Tjøstheim, 2001; Bandi and Phillips, 2003). However, the convergence rate will be path-dependent and relatively slow.

In the remainder of this paper, we focus on (2). The properties of (4) can be obtained by following the same proof strategy as the one we employ for (2). The only difference is that, to obtain $\check{p} \xrightarrow{P} p$ in the sup norm, one has to take into account the dependence of the simulated values. This can be done along the lines of Creel and Kristensen (2009) where kernel regressions and simulations are combined to compute GMM estimators for dynamic latent variable models.

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, $\mathcal{Y}_2 = \{y_{2,1}, y_{2,2}, \ldots\}$. We could then use a mixed kernel to estimate $p_t(y|x)$. For given simulated observations $Y_{t,i}^\theta = (Y_{1t,i}^\theta, Y_{2t,i}^\theta)$, $i = 1, \ldots, N$:

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

$$(y_{1t}, y_{2t}) \in \mathbb{R}^{k+l}, \tag{5}$$

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, ch. 2). These will increase bias but reduce variance of the estimator, and at the same time lead to a continuous function. However, in general, $\theta \mapsto Y_{2t,i}^\theta$ itself will not be continuous, and, with a discrete component, $\hat{L}_T(\theta)$ based on (5) is no longer continuous w.r.t. θ.

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

$$\mathbb{E}[Y_{2t,i}^\theta|Y_{1t,i}^\theta = y_{1t}] = p_t(y_{2t}|y_{1t}, x_t; \theta). \tag{6}$$

Thus, $Y_{2t,i}^\theta$ 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

$$\hat{p}_t(y_{1t}, y_{2t}|x_t; \theta) = \frac{1}{N}\sum_{i=1}^N K_h(Y_{1t,i}^\theta - y_{1t})Y_{2t,i}^\theta. \tag{7}$$

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, θ). For example, most limited dependent variables can be written in this form; c.f. Manrique and Shephard (1998) and the references therein. We assume that z_t satisfies $z_t = g_Z(x_t, \varepsilon_t; \theta)$ for some function g_Z that can be written in closed form, and has associated conditional density $p_{z_t|x_t}(z|x)$. Clearly, $p_t(y_2|x) = P(y_{2t} = y_2|x_t = x)$ 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 in closed form, this integral can then be simulated by

$$\hat{p}_t(y_{2t}|x_t) = \frac{1}{N}\sum_{i=1}^N \mathcal{K}^{(1)}(\tilde{Z}_{t,i}^\theta|y_{2t}, x_t),$$

$$\mathcal{K}^{(1)}(\tilde{Z}_{t,i}^\theta|y_{2t}, x_t) = \frac{p_{z_t|x_t}(\tilde{Z}_{t,i}^\theta|x_t)}{p_D(\tilde{Z}_{t,i}^\theta|y_{2t})}, \tag{8}$$

where $\tilde{Z}_{t,i}^\theta \overset{iid}{\sim} p_D(z|y_{2t})$, as is standard in the estimation of limited dependent variable models.

If $p_{z|x}(z|x)$ cannot be written in closed form, we propose to use $\hat{p}_{z_t|x_t}(z_t|x_t) = \frac{1}{N}\sum_{i=1}^N K_b(Z_{t,i}^\theta - z)$, where $Z_{t,i}^\theta = g_Z(x_t, \varepsilon_i; \theta)$ and $b > 0$ is another bandwidth. If $\int_{D^{-1}(y_2)} K_b(Z_{t,i}^\theta - z)dz$ can be written in closed form, we follow Fermanian and Salanié (2004, pp. 709–710 and 724–725) and use

$$\hat{p}_t(y_{2t}|x_t) = \frac{1}{N}\sum_{i=1}^N \mathcal{K}_b^{(2)}(Z_{t,i}^\theta, y_{2t}),$$

$$\mathcal{K}_b^{(2)}(Z_{t,i}^\theta, y_{2t}) = \int_{D^{-1}(y_{2t})} K_b(Z_{t,i}^\theta - z)dz. \tag{9}$$

If this is not the case, we can use

$$\hat{p}_t(y_{2t}|x_t) = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathcal{K}}_b^{(2)}(Z_{t,i}^\theta, y_{2t}),$$
$$\hat{\mathcal{K}}_b^{(2)}(Z_{t,i}^\theta, y_{2t}) = \frac{1}{N}\sum_{j=1}^{N} \frac{K_b(Z_{t,i}^\theta - \tilde{Z}_{t,j}^\theta)}{p_D(\tilde{Z}_{t,j}^\theta|y_{2t})}. \tag{10}$$

In all three cases, we can write the resulting simulated joint density in Eq. (7) by choosing $Y_{2t,i}^\theta = \mathcal{K}^{(1)}(\tilde{Z}_{t,i}^\theta, y_{2t}|x_t)$, $Y_{2t,i}^\theta = \mathcal{K}_b^{(2)}(Z_{t,i}^\theta, y_{2t})$ and $Y_{2t,i}^\theta = \hat{\mathcal{K}}_b^{(2)}(Z_{t,i}^\theta, y_{2t})$, respectively. Here, $\theta \mapsto Y_{2t,i}^\theta$ is smooth with a bias that disappears as $b \to 0$ 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 variable, and the curse of dimensionality remains of order $k = \dim(y_{1t})$.

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 model (1), but the data-generating distribution of the errors is unknown. We could then choose a suitable distribution F_ε, draw $\{\varepsilon_i\}_{i=1}^N$ from F_ε 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 should be adjusted to accommodate the fact that we are not using the true likelihood 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 model (1) with known distribution F_ε, but now $\theta = (\alpha, \gamma)$, where α and γ are finite- and infinite-dimensional parameters, respectively. An application of our method in this setting can be found in Kristensen (2010) where γ is a density. Again, our asymptotic results have to be adjusted to allow for θ to contain infinite-dimensional parameters.

3. Asymptotic properties of NPSMLE

Given the convergence of the simulated conditional density towards the true one, we expect that the NPSMLE $\hat{\theta}$ based on the simulated kernel density estimator will have the same asymptotic properties as the infeasible MLE $\tilde{\theta}$ for a suitably chosen sequence $N = N(T)$ and $h = h(N)$. 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 expressions of the leading bias and variance components of $\hat{\theta}$ relative to the actual MLE due to simulations and kernel smoothing, 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 $\mathcal{Y}_1 \subseteq \mathbb{R}^k$ and $\mathcal{Y}_2 = \{y_{2,1}, y_{2,2}, \ldots\} \subset \mathbb{R}^l$. Here, y_{1t} has a continuous distribution, while y_{2t} is discrete. 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 $p_t(y_{2,j}|y_1, x;\theta)$ are conditional probabilities satisfying $\sum_{j=1}^l p_t(y_{2,j}|y_1, x;\theta) = 1$, while $p_t(y_1|x;\theta)$ is a conditional density w.r.t. the Lebesgue measure. Also, let $p_t(y_{2,j}|x;\theta)$ denote the conditional probabilities of $y_{2t}|x_t = x$.

The asymptotics are derived for the kernel estimator given in Eq. (7) where

$$Y_{1t,i}^\theta := g_{1,t}(x_t, \varepsilon_i;\theta), \tag{11}$$
$$Y_{2t,i}^\theta := g_{2,t}(y_{2t}, x_t, \varepsilon_i;\theta), \tag{12}$$

for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where $\{\varepsilon_i\}_{i=1}^N$ are i.i.d. draws from F_ε, such that Eq. (6) holds. Recall that $Y_{2t,i}^\theta$ denotes a simulated value of the associated density, and not the outcome of the dependent variable. The condition in Eq. (6) is met when $Y_{2t,i}^\theta = \mathcal{K}^{(1)}(\tilde{Z}_{t,i}^\theta, y_{2t}|x_t)$ with $\mathcal{K}^{(1)}$ given in Eq. (8), while it only holds approximately for $\mathcal{K}^{(2)}$ and $\hat{\mathcal{K}}^{(2)}$ defined in Eqs. (9) and (10) due to biases induced by the use of kernel smoothing. We handle these two cases in Theorem 3.4, 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, $Y_{t,i}^\theta = g_t(x_t, \varepsilon_{t,i};\theta)$, $i = 1, \ldots, \widetilde{N}$, such that the total number of simulations would be $\widetilde{N} \times T$, $\{\varepsilon_{i,t}\}_{i=1}^{\widetilde{N}}$, $t = 1, \ldots, T$. Under regularity conditions, the NPSMLE based on this simulation scheme would have similar asymptotic properties as the one based on the simulations in Eqs. (11) and (12). However, as demonstrated in Lee (1992), choosing $N = \widetilde{N}T$, the variance of the NPSMLE based on Eqs. (11) and (12) will be smaller.[6]

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 (2009). These general conditions are satisfied under 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 ε such that for some function $\Lambda(\cdot)$ and constants $\lambda_{i,j} \geq 0$, $i, j = 1, 2$,

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

and $\mathbb{E}[\Lambda(\varepsilon)^s] < \infty$ for some $s > 2$. The derivatives of g_1 and g_2 w.r.t. (x, t, θ) 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 the r-th derivative being uniformly continuous. There exist constants $\bar{B}_0 > 0$ and $\lambda_{0,1}, \lambda_{0,2} \geq 0$, such that the following bounds hold uniformly over (t, y_1, y_2, x, θ) with $\bar{B}(x, t) = \bar{B}_0(1 + \|x\|^{\lambda_{0,1}} + t^{\lambda_{0,2}})$:

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

[6] 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 continuously differentiable for all t, x, ε with their derivatives satisfying the same moment conditions as g_1 and g_2 in Assumption 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 Eq. (13) as p.

Assumptions A.1 and A.2 are used to establish uniform convergence of $\hat{p}$ (Lemma B.1). 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 condition in A.1 does not concern the observed process $\{(y_t, x_t)\}$, only the errors ε that we draw when simulating. If for example, $\Lambda(\varepsilon) \propto \|\varepsilon\|^q$, then the moment condition is satisfied if $\mathbb{E}[\|\varepsilon\|^{qs}] < \infty$. Thus, in this case, the moment condition only rules out models driven by fat-tailed errors. If the model is time-homogeneous, $\lambda_{i,2} = 0, i = 1, 2$.

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 (which are introduced below) controls 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-homogeneous, $\lambda_{0,2} = 0$.

Assumptions A.3 and A.4 will only be used when examining the higher-order impact of simulations and kernel smoothing on our 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 B.2).

Our conditions are slightly stronger than the ones found in Fermanian and Salanié (2004, 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:
- (a) $\sup_{u\in\mathbb{R}^k}|K(u)| < \infty$ and $\int_{\mathbb{R}^k}|K(u)|du < \infty$. There exist $C, L < \infty$ such that either (i) $K(u) = 0$ for $\|u\| > L$ and $|K(u) - K(u')| \leq C\|u - u'\|$, or (ii) $K(u)$ is differentiable with $\sup_{u\in\mathbb{R}^k}|\partial K(u)/\partial u| \leq C$. For some $a > 1$, $|\partial^{|\alpha|}K(u)/\partial u^\alpha| \leq C\|u\|^{-a}$ for $\|u\| \geq L$ and all $1 \leq |\alpha| \leq 2$.
- (b) $\int_{\mathbb{R}^k} K(u)du = 1$ and for some $r \geq 1$: $\int_{\mathbb{R}^k} K(u)u^\alpha du = 0$, $1 \leq |\alpha| \leq r - 1$, and $\int_{\mathbb{R}^k} K(u)\|u\|^r du < \infty$.

K.2 The first and the second derivatives of K 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 Assumptions A.3 and A.4 to show that the first and the second derivatives of $\hat{p}$ w.r.t. θ 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 $\nu_T \to \infty$:

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

This normalizing factor ν_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.4 below. It is only important for the theoretical derivations, and not relevant for the actual implementation of our estimator since ν_T does not depend on θ. The choice of ν_T depends on the dynamics of the model. The standard choice is $\nu_T = T$, as is the case when the model is stationary. In order to allow for non-standard behavior of the likelihood due to, for example, stochastic and deterministic trends, we do not impose this restriction though.

We also redefine the simulated version of the likelihood: in order to obtain uniform convergence of $\log\hat{p}_t(y|x;\theta)$, we need to introduce trimming of the approximate log-likelihood as is standard in the literature on semiparametric estimators. The trimmed and normalized version of the simulated log-likelihood is given 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 a continuously differentiable trimming function satisfying $\tau_a(z) = 1$ if $|z| > a$, and 0 if $|z| < a/2$, with a trimming sequence $a = a(N) \to 0$. Here one could 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 $\hat{\theta}$.

Assuming that $L_T(\theta)$ is three times differentiable (c.f. Assumption C.3 below), we can define

$$S_T(\theta) = \frac{\partial L_T(\theta)}{\partial\theta} = \frac{1}{\nu_T}\sum_{t=1}^{T}\frac{\partial\log p_t(y_t|x_t;\theta)}{\partial\theta} \in \mathbb{R}^d,$$

$$H_T(\theta) = \frac{\partial^2 L_T(\theta)}{\partial\theta\partial\theta'} = \frac{1}{\nu_T}\sum_{t=1}^{T}\frac{\partial^2\log p_t(y_t|x_t;\theta)}{\partial\theta\partial\theta'} \in \mathbb{R}^{d\times d},$$

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

The information is then defined as

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

We also define the diagonal matrix $\mathcal{I}_T(\theta) = \mathrm{diag}\{i_T(\theta)\} \in \mathbb{R}^{d\times d}$, where $\mathrm{diag}\{i_T(\theta)\}$ denotes the diagonal elements of the matrix $i_T(\theta)$, and

$$\begin{aligned} &U_T(\theta) = \mathcal{I}_T^{-\frac{1}{2}}(\theta)S_T(\theta), \qquad V_T(\theta) = \mathcal{I}_T^{-\frac{1}{2}}(\theta)H_T(\theta)\mathcal{I}_T^{-\frac{1}{2}}(\theta), \\ &W_{T,i}(\theta) = \mathcal{I}_T^{-\frac{1}{2}}(\theta)G_{T,i}(\theta)\mathcal{I}_T^{-\frac{1}{2}}(\theta). \end{aligned} \tag{14}$$

With $\mathcal{I}_T \equiv \mathcal{I}_T(\theta_0)$, we then impose the following conditions on the actual log-likelihood function and the associated MLE which ensure consistency and a well-defined asymptotic distribution of the actual MLE, $\tilde{\theta}$.

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

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

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

C.2 For any $\eta > 0$, there exists a $\delta > 0$ such that

$$\lim_{T\to\infty} P\left(\sup_{\|\mathcal{I}_T^{1/2}(\theta_0-\theta)\|>\eta}\{L_T(\theta_0) - L_T(\theta)\} \geq \delta\right) = 1.$$

C.3 $L_T(\theta)$ is three times continuously differentiable with its derivatives satisfying:
  (a) $(\sqrt{\nu_T}U_T(\theta_0), V_T(\theta_0)) \xrightarrow{d} (S_\infty, H_\infty)$ with $H_\infty < 0$ a.s.;
  (b) $\max_{j=1,\ldots,d} \sup_{\theta\in\Theta_T} \|W_{j,T}(\theta)\| = O_P(1)$.

C.4 The following bounds hold for some $\delta, q > 0$:
  (a) $\sup_{\theta\in\Theta_T} \nu_T^{-q} \sum_{t=1}^T |\log p_t(y_t|x_t;\theta)|^{1+\delta} = O_P(1)$;
  (b) $\nu_T^{-q} \sum_{t=1}^T \|x_t\|^{1+\delta} = O_P(1)$ and $\nu_T^{-q} \sum_{t=1}^T \Lambda^2(\varepsilon_t) = O_P(1)$.

The above conditions C.1–C.4 are generalized versions of the conditions normally required for consistency and asymptotic normality of MLEs in stationary and ergodic models. For general non-ergodic models, simple conditions for C.2–C.4 are not available and they have to be verified on a case-by-case basis. For the stationary case, they are implied by primitive conditions as found below in Corollary 3.2.

The specification of the parameter space in C.1 to be a sequence of non-increasing compact sets is introduced to allow for non-ergodic models. Given C.1, we re-define $\tilde\theta$ and $\hat\theta$ as the maximizers of the exact and approximate likelihood over Θ_T. This, in particular, means that we will only conduct our analysis of the NPSMLE in a (possibly shrinking) neighborhood of the true parameter value. A global analysis would have been preferable, but currently, to the best of our knowledge, there exists no general result on the properties of MLEs for non-ergodic models over a fixed parameter space. For specific models, global results exist such as those found in Park and Phillips (2000, 2001), but these results appear difficult to extend to a more general setup, and usually the analysis of non-ergodic models is done locally; see e.g. Kristensen and Rahbek (2010).

Assumption C.2 gives us consistency of the actual MLE; c.f. Lemma A.1. It is a combined uniform convergence and identification condition. It is for example implied by uniform convergence of the log-likelihood towards some population function which in turn identifies the true value of the parameter; see for example Amemiya (1983). Note that the condition may potentially hold even if the log-likelihood is not continuous, since its asymptotic limit will in great generality be so.

Assumption C.3 is a further strengthening of C.2 stating that the score and Hessian converge in distribution after suitable normalization. This condition, in conjunction with C.1, implies both consistency and that the asymptotic distribution of the MLE is given as $\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\tilde\theta - \theta_0) \xrightarrow{d} -H_\infty^{-1}S_\infty$, c.f. Lemma A.3.[7]

Assumption C.4 imposes bounds on a number of sample averages. They are used to show that the trimming of the simulated log-likelihood is asymptotically negligible for suitable choices of the trimming parameter a. Note that the factor ν_T in C.4 is the same as the one we normalized the log-likelihood with. The exponent $q > 0$ should be chosen to ensure that both the log-likelihood and the sample averages in C.4 are well-behaved.

In the ergodic case, we can appeal to standard results for stochastic equicontinuity (e.g. Proposition 1 in Kristensen and Rahbek (2005)) to obtain that C.4 holds 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 3.2 below and its proof for further details. Furthermore, $i_T(\theta_0) = i(\theta_0) + o_P(1)$ with $i(\theta) = \mathbb{E}[\partial^2 \log p(y_t|x_t;\theta)/(\partial\theta\partial\theta')]$, such that $\mathcal{I}_T$ can be chosen as the constant $\mathrm{diag}\{i(\theta_0)\}$. This in turn implies that Θ_T is a fixed compact parameter set, and we get the standard $\sqrt{T}$-convergence towards a normal distribution. Thus, in the case of stationarity, C.1–C.4 are more or less identical to the ones imposed in Fermanian and Salanié (2004, L.1–3).

In the general case, one should choose ν_T as the square of the slowest rate of convergence of the vector of MLEs. There is a tension between C.1 and C.4 in terms of the choice of ν_T. We cannot choose $\nu_T \to \infty$ too fast, since then $\|\mathcal{I}_T\| \to 0$ (in which case no information regarding θ_0 is available) and this is ruled out by C.1. On the other hand, we have to choose $\nu_T^q \to \infty$ sufficiently fast to ensure that the bounds in C.4 hold. By choosing $q > 0$ sufficiently large, C.1 and C.4 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 and B.2 below.

As an example of non-standard asymptotics of the MLE, consider a linear error-correction model, $\Delta y_t = \alpha\beta' y_{t-1} + \Omega^{1/2}\varepsilon_t$, where $\varepsilon_t \sim N(0, I_k)$. We can split the parameter vector into short-run, $\theta_1 = (\alpha, vech(\Omega))$, 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 superconsistent with $T(\tilde\theta_2 - \theta_2)$ converging towards a Dickey–Fuller type distribution. In this case, we choose $\sqrt{\nu_T} = \sqrt{T}$, and so $i_T(\theta_0)$ 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 $T^{-2}\sum_{t=1}^T \|x_t\|^{1+\delta} = O_P(1)$, and we can choose $q = 2$. We also refer to Park and Phillips (2001) and Kristensen and Rahbek (2010) where C.2–C.4 are verified for some non-linear, non-stationary models.

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

B. With $q, \delta > 0$ given in Condition C.4, $\bar\lambda_k = \lambda_{0,k} + \lambda_{1,k} + \lambda_{2,k}$, $k = 1, 2$, where $\lambda_{i,1}, \lambda_{i,2} \geq 0$, $i = 0, 1, 2$, are given in Assumptions A.1 and A.2 and for some $\gamma > 0$:
  (a) $|\log a|\nu_T^{q-1}N^{-\gamma(1+\delta)} \to 0$; $|\log(4a)|^{-\delta}\nu_T^{q-1} \to 0$; $T\nu_T^{-1}a^{-1}[N^{\gamma\bar\lambda_1} + T^{\bar\lambda_2}]\log(N)/\sqrt{Nh^k} \to 0$; and $T\nu_T^{-1}a^{-1}[N^{\gamma\lambda_{0,1}} + T^{\lambda_{0,2}}]h^r \to 0$.
  (b) $|\log a|\nu_T^q N^{-\gamma(1+\delta)} \to 0$; $|\log(4a)|^{-\delta}\nu_T^q \to 0$; $T\nu_T^{-1/2}a^{-1}[N^{\gamma\bar\lambda_1} + T^{\bar\lambda_2}]\log(N)/\sqrt{Nh^k} \to 0$; and $T\nu_T^{-1/2}a^{-1}[N^{\gamma\lambda_{0,1}} + T^{\lambda_{0,2}}]h^r \to 0$.

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. We observe that large values of q and/or $\bar\lambda_1, \bar\lambda_2$ imply 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. Fortunately, our simulation results indicate that standard bandwidth selection rules together with a bit of undersmoothing in general deliver satisfactory results.

Our strategy of proof is based on some apparently new results for approximate estimators (Appendix A). In particular, Theorems A.4 and A.5 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 a sufficiently fast rate. This makes our proofs considerably less burdensome than those found in other studies of simulation-based estimators (e.g. Fermanian and Salanié (2004) and Altissimo and Mele (2009)) since we do not need to analyze the simulated score and Hessian.

Theorem 3.1. *Assume that* A.1, A.2, *and* K.1 *hold. Then the NPSMLE* $\hat\theta$ *based on* (7) *satisfies the following.*

[7] Basawa and Scott (1983) and Jeganathan (1995) show what S_∞ and H_∞ look like in various cases.

(i) *Under Conditions* C.1, C.2, *and* C.4: $\mathcal{I}_T^{1/2}(\hat{\theta}-\theta_0)=o_P(1)$ *for any sequences* $N\to\infty$, *and* $h, a\to 0$ *satisfying* B.1.

(ii) *Under Conditions* C.1, C.3, *and* C.4: $\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\hat{\theta}-\theta_0)\xrightarrow{d}-H_\infty^{-1}S_\infty$ *for any sequences* $N\to\infty$, *and* $h, a\to 0$ *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.4.

Corollary 3.2. *Assume that* (y_t, x_t) *is stationary and ergodic, and that* A.1, A.2, K.1, *and* B.1 *hold with* $q=1$, $\nu_T=T$, $\bar{\lambda}_2=\lambda_{0,2}=0$ *and*

(i) $\mathbb{E}[\|x_t\|^{1+\delta}]<\infty$, $|\log p(y|x;\theta)|\le b_1(y|x)$, $\forall\theta\in\Theta$, *with* $\mathbb{E}[b_1(y_t|x_t)^{1+\delta}]<\infty$ *and* Θ *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 $\hat{\theta}\xrightarrow{P}\theta_0$.

If furthermore B.2 *holds with* $q=1$, $\nu_T=T$ *and* $\bar{\lambda}_2=\lambda_{0,2}=0$ *together with*

(iii) $i(\theta_0)=\mathbb{E}[\frac{\partial\log p(y_t|x_t;\theta_0)}{\partial\theta}\frac{\partial\log p(y_t|x_t;\theta_0)}{\partial\theta'}]$ *exists and is nonsingular;*

(iv) $\|\frac{\partial^2\log p(y|x;\theta)}{\partial\theta\partial\theta'}\|\le b_2(y|x)$ *uniformly in a neighborhood of* θ_0 *with* $\mathbb{E}[b_2(y_t|x_t)]<\infty$;

then $\sqrt{T}(\hat{\theta}-\theta_0)\xrightarrow{d}\mathcal{N}(0, i(\theta_0)^{-1})$.

If for simplicity we set $\gamma=0$ in B.2 and disregard the conditions on the trimming parameter a, then roughly speaking, the NPSMLE will be first-order equivalent to the exact MLE in the stationary case if $Th^{2r}\to 0$ and $T/(Nh^k)\to 0$ reflecting the bias and variance due to kernel smoothing and simulations. The variance requirement seems to indicate that the usual curse of dimensionality inherent in kernel density estimation is present. However, this is caused by the initial error bounds used to establish Corollary 3.2 being overly conservative. In the following, we will obtain more precise error bounds which show that the curse of dimensionality is significantly less severe. Moreover, these refined error bounds allow us to better gauge which additional biases and variances the NPSMLE suffers from due to simulations and kernel smoothing. These can potentially be used to adjust confidence bands based on the NPSMLE to take into account the additional simulation errors.

Since the higher-order analysis involves the first and the second derivatives of $\hat{L}_T(\theta)$, we have to invoke the additional smoothness conditions on g and p stated in A.3 and A.4. Under the additional smoothness conditions, the first two derivatives of $g_{1,t}(x,\varepsilon;\theta)$ and $g_{2,t}(y_2,x,\varepsilon;\theta)$ w.r.t. θ exist, and so the first two derivatives of our density estimator are well-defined:

$$\begin{aligned}\frac{\partial\hat{p}_t(y_t|x_t;\theta)}{\partial\theta}&=\frac{1}{N}\sum_{i=1}^N\{\dot{Y}_{2t,i}^{\theta}K_h(Y_{1t,i}^{\theta}-y_1)\\&\quad+\dot{Y}_{1t,i}^{\theta}K_h^{(1)}(Y_{1t,i}^{\theta}-y_1)Y_{2t,i}^{\theta}\},\end{aligned}\tag{15}$$

$$\begin{aligned}\frac{\partial^2\hat{p}_t(y_t|x_t;\theta)}{\partial\theta\partial\theta_j}&=\frac{1}{N}\sum_{i=1}^N\{\ddot{Y}_{2t,i}^{\theta_j}K_h(Y_{1t,i}^{\theta}-y_1)\\&\quad+\dot{Y}_{2t,i}^{\theta}K_h^{(1)}(Y_{1t,i}^{\theta}-y_1)'\dot{Y}_{1t,i}^{\theta_j}\}\\&\quad+\frac{1}{N}\sum_{i=1}^N\{\ddot{Y}_{1t,i}^{\theta_j}K_h^{(1)}(Y_{1t,i}^{\theta}-y_1)Y_{2t,i}^{\theta}\\&\quad+\dot{Y}_{1t,i}^{\theta}K_h^{(2)}(Y_{1t,i}^{\theta}-y_1)\dot{Y}_{1t,i}^{\theta_j}Y_{2t,i}^{\theta}\}\\&\quad+\frac{1}{N}\sum_{i=1}^N\dot{Y}_{1t,i}^{\theta}K_h^{(1)}(Y_{1t,i}^{\theta}-y_1)\dot{Y}_{2t,i}^{\theta_j},\end{aligned}\tag{16}$$

for $j=1,\dots,d$, where $K_h^{(i)}(y_1)=K^{(i)}(y_1/h)/h^{k+i}$, $i=1,2$, with $K^{(1)}(y_1)=\partial K(y_1)/(\partial y_1)\in\mathbb{R}^k$ and $K^{(2)}(y_1)=\partial^2K(y_1)/(\partial y_1\partial y_1')\in\mathbb{R}^{k\times k}$, while $\dot{Y}_{1t,i}^{\theta}=(\dot{Y}_{1t,i}^{\theta_1},\dots,\dot{Y}_{1t,i}^{\theta_d})'\in\mathbb{R}^{d\times k}$ with

$$\dot{Y}_{1t,i}^{\theta_j}=\frac{\partial g_{1,t}(x_t,\varepsilon_i;\theta)}{\partial\theta_j}\in\mathbb{R}^k,\qquad\ddot{Y}_{1t,i}^{\theta_j}=\frac{\partial^2 g_{1,t}(x_t,\varepsilon_i;\theta)'}{\partial\theta\partial\theta_j}\in\mathbb{R}^{d\times k},$$

and similarly for $\dot{Y}_{2t,i}^{\theta}\in\mathbb{R}^d$ and $\ddot{Y}_{2t,i}^{\theta_j}\in\mathbb{R}^d$. Lemma B.2 shows that these are uniformly consistent estimates of the actual derivatives of the conditional density p_t. The corresponding simulated version of the score is given by

$$\begin{aligned}\hat{S}_T(\theta)&=\frac{1}{\nu_T}\sum_{t=1}^T\frac{\partial\hat{p}_t(y_t|x_t;\theta)}{\partial\theta}\left\{\frac{\tau_a(\hat{p}_t(y_t|x_t;\theta))}{\hat{p}_t(y_t|x_t;\theta)}\right.\\&\quad\left.+\tau_a'(\hat{p}_t(y_t|x_t;\theta))\log\hat{p}_t(y_t|x_t;\theta)\right\}.\end{aligned}\tag{17}$$

An expression for the simulated version of the Hessian can be found in the proof of Theorem 3.3. We then follow Kristensen and Salanie (2010) and consider a second order functional Taylor expansion of $\hat{S}_T(\theta)$ w.r.t. $\hat{p}$. This takes the form

$$\begin{aligned}\hat{S}_T(\theta_0)&=S_T(\theta_0)+\nabla S_{T,N}[\hat{p}-p]\\&\quad+\nabla^2S_{T,N}[\hat{p}-p,\hat{p}-p]+R_{T,N},\end{aligned}\tag{18}$$

where $\nabla S_{T,N}[\hat{p}-p]$ and $\nabla^2S_{T,N}[\hat{p}-p,\hat{p}-p]$ are the first- and 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 3.3, where the properties of the first- and second-order terms are analyzed.

To facilitate our analysis, which involves U-statistics, we restrict our attention to the stationary and β-mixing case. See e.g. Ango Nze and Doukhan (2004) for an introduction to this concept. We also assume that $p(y|x;\theta)$ is uniformly bounded away from zero thereby obviating trimming. Under these and other regularity conditions, we show that the two first terms in the expansion in Eq. (18) satisfy (c.f. the proof of Theorem 3.3)

$$\sqrt{T}S_T(\theta_0)+\sqrt{T}\nabla S_{T,N}[\hat{p}-p]\simeq\sqrt{T}h^r\mu_1+Z_1+\sqrt{\frac{T}{N}}Z_2,\tag{19}$$

where higher-order terms have been left out. Here, the first term is a bias component incurred by kernel smoothing, while the two remaining ones are variance components: Z_1 and Z_2 are two independent variables where $Z_1\sim\mathcal{N}(0,i(\theta_0))$ is the variance component of the observed data, while $Z_2\sim\mathcal{N}(0,\mathrm{Var}(\psi_2(\varepsilon_i)))$ is the variance component of the simulations. The variance of Z_2 is given by

$$\psi_2(\varepsilon_i)=\mathbb{E}\left[\frac{\dot{Y}_{2t,i}^{\theta_0}}{p(y_{2t}|x_t)}\middle|\varepsilon_i\right]-\mathbb{E}\left[\frac{s(Y_{1t,i}^{\theta_0},y_{2t}|x_t)Y_{2t,i}^{\theta_0}}{p(y_{2t}|x_t)}\middle|\varepsilon_i\right],\tag{20}$$

where $s(y_1,y_2|x)$ denotes the score at $\theta=\theta_0$.

The second order term also contains a bias component which all non-linear, simulation-based estimators suffer from,

$$\sqrt{T}\nabla^2S_{T,N}[\hat{p}-p,\hat{p}-p]\simeq\frac{\sqrt{T}}{Nh^{k+1}}\mu_2+O_P(\sqrt{T}h^{2r}),\tag{21}$$

while the remainder term is of a lower order,

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

The two leading bias terms in the above expressions, μ_1 and μ_2, are given by

$$\mu_1 = \sum_{|\alpha|=r} \int\int \left\{ \frac{\partial^{|\alpha|+1} p(y_t|x_t;\theta)}{\partial\theta \partial y_t^\alpha} - s(y_t|x_t;\theta) \frac{\partial^{|\alpha|} p(y_t|x_t;\theta)}{\partial y_t^\alpha} \right\} p(x_t) dx_t dy_t \in \mathbb{R}^d, \quad (22)$$

$$\mu_2 = \mathbb{E}\left[\frac{\dot{Y}_{1t,i}^{\theta_0} (Y_{2t,i}^{\theta_0})^2}{p(Y_{1t,i}^{\theta_0}, y_{2t}|x_t) p(y_{2t}|x_t)} \right] \int K(v) K^{(1)}(v) dv \in \mathbb{R}^d, \quad (23)$$

where $p(x_t)$ denotes the marginal density of x_t.

This shows that the overall bias of the estimator due to simulations and kernel smoothing is $i^{-1}(\theta_0)\{h^r \mu_1 + 1/(Nh^{k+1})\mu_2\}$, while an additional variance term relative to the exact MLE shows up and is given by $T/N \times i^{-1}(\theta_0)\mathrm{Var}(\bar{\psi}_2(\varepsilon_i))i^{-1}(\theta_0)$. Thus, if $\sqrt{T}h^r \to 0$ and $\sqrt{T}/(Nh^{k+1}) \to 0$, all bias terms vanish and $\sqrt{T}(\hat{\theta} - \theta_0)$ follows a normal distribution centered around zero. If furthermore $T/N \to 0$, no additional variance will be present and the NPSMLE is first-order equivalent to the true MLE. 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, if $T/N \not\to 0$ there will be an increase in variance due to the presence of Z_2. One could potentially reduce (or even remove) some of these bias and variance components by employing the techniques of Kristensen and Salanie (2010) who develop higher-order improvements of simulation-based estimators.

We collect the results in the following theorem.

Theorem 3.3. *Assume that*

(i) $\{(y_t, x_t)\}$ *is stationary and* β*-mixing with geometrically decreasing mixing coefficients;*
(ii) A.1–A.4 *and* K.1–K.2 *hold;*
(iii) (i)–(iv) *of Corollary* 3.2 *hold;*
(iv) x_t *is bounded and* $\inf_{y_1,y_2,x,\theta} p(y_1, y_2|x;\theta) > 0$.

Then, if $\sqrt{T}h^r \to c_1 \geq 0$, $\sqrt{T}/(Nh^{k+1}) \to c_2 \geq 0$ *and* $T/N \to c_3 \geq 0$,

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(\bar{c}, i^{-1}(\theta_0)[i(\theta_0) + c_3 \mathrm{Var}(\psi_2(\varepsilon_i))]i^{-1}(\theta_0)),$$

where $\bar{c} = i^{-1}(\theta_0)\{c_1\mu_1 + c_2\mu_2\}$, *with* μ_1 *and* μ_2 *as in Eqs.* (22) *and* (23).

The requirement (iv) is only imposed to simplify the proofs which otherwise would get overly long and complicated. We expect that the above result will still hold with the requirement (iv) replaced by additional restrictions on the trimming parameter a.

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 3.3.

Estimation of asymptotic distribution. To do any finite-sample inference, an estimator of the asymptotic distribution is needed. A general Monte Carlo method would be to simulate a large number of independent, long trajectories from the model and for each trajectory compute the corresponding score and Hessian at $\theta = \hat{\theta}$. This would yield an approximation of the limiting distribution, $-H_\infty^{-1} S_\infty$. 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 Eq. (17) and the proof of Theorem 3.3 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 kth column of the identity matrix. We have

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

Now letting $\delta = \delta(N) \to 0$ as $N \to \infty$ 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 exact model. Suppose that one, on the other hand, has an approximation of the model at one's disposal. For example, solutions to dynamic programming problems are typically approximated numerically, and sample paths of diffusions are approximated by discretization schemes.

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 $g_{M,2t}(y_2, x, \varepsilon;\theta)$ available, where $g_{M,kt} \to g_{kt}$, $k = 1, 2$, as $M \to \infty$ in L_1-norm. We then redefine the simulated conditional density as

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

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

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

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

$$\hat{\theta}_M = \arg\max_{\theta\in\Theta_T} \hat{L}_{M,T}(\theta),$$

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

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 $B_k, \lambda_{3,k}, \lambda_{4,k} \geq 0$, $k = 1, 2$:

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

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

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

B.1′ $T\nu_T^{-1}a^{-1}h^{-1}[N^{\gamma\lambda_{3,1}} + T^{\lambda_{3,2}}]\delta_{M,1} \to 0$ and $T\nu_T^{-1}a^{-1}[N^{\gamma\lambda_{4,1}} + T^{\lambda_{4,2}}]\delta_{M,2} \to 0$.

B.2′ $Tv_T^{-1/2}a^{-1}h^{-1}[N^{\gamma\lambda_{3,1}} + T^{\lambda_{3,2}}]\delta_{M,1} \to 0$ and $Tv_T^{-1/2}a^{-1}[N^{\gamma\lambda_{4,1}} + T^{\lambda_{4,2}}]\delta_{M,2} \to 0$.

Assumption M.1 simply states that the approximation error is bounded as a polynomial in x and t. Assumptions B.1′ and B.2′ require that coefficients of this polynomial error bound, $\delta_{M,1}$ and $\delta_{M,2}$, go to zero sufficiently fast as $M \to \infty$ to control for the impact of the approximation error. Kloeden and Platen (1992) and Bruti-Liberati and Platen (2007) give primitive conditions under which the discrete-time approximation satisfies M.1; see also Detemple et al. (2006).

Under Assumptions M.1 and B.1′ (B.2′), we show that the approximate NPSMLE is asymptotically equivalent to the NPSMLE up to order $\mathcal{I}_T^{1/2}$ ($\sqrt{\nu_T}\mathcal{I}_T^{1/2}$). Combining this result with Theorem 3.1, we obtain the following.

Theorem 3.4. *Assume that the conditions of Theorem 3.1 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* $M, N \to \infty$, *and* $h, a \to 0$ *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* $M, N \to \infty$, *and* $h, a \to 0$ *satisfying* B.2 *and* B.2′.

One can exchange M.1 for

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

in which case Theorem 3.4 holds with B.1′ and B.2′ being replaced by the simpler conditions B.1″ : $\delta_M \to 0$ and B.2″ : $\sqrt{\nu_T}\delta_M \to 0$, respectively. However, since $p_{M,t}(y_t|x_t;\theta)$ and $p_t(y_t|x_t;\theta)$, in general, are difficult to analyze, condition M.1 is easier to verify.

4. Implementing NPSML

One of the merits of NPSML is its general applicability. The applications include Markov decision processes 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.

The first example (Section 4.1.1) is the short-term interest rate model of Cox et al. (1985). This univariate diffusion has a known transition density, and therefore has been a popular benchmark of numerous diffusion estimation strategies (Durham and Gallant, 2002). We provide a detailed enumerate on the implementation of NPSML in practice, and then test the validity of our approach by comparing it to the true MLE. In Section 4.1.2, we re-visit the jump–diffusion example of Section 2. The literature on estimating general jump–diffusions has largely sidestepped maximum likelihood. In this context, this estimation exercise showcases the usefulness of NPSML. In Section 4.2, we briefly discuss how NPSML can be used for estimating generic Markov decision processes. We discuss diffusion models in detail here because they can be described more concisely than a typical Markov decision model, which requires a detailed enumerate of the economic environment. We refer to Kristensen and Schjerning (2011) for implementation and analysis of smoothed maximum-likelihood estimators in a discrete Markov decision model.

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.

4.1. Discretely-observed jump–diffusions

4.1.1. Cox–Ingersoll–Ross model

Cox et al. model short-term interest rates as a square-root diffusion process:

$$dy_t = \beta(\alpha - y_t)dt + \sigma\sqrt{y_t}dW_t.$$

We collect the unknown parameters in $\theta = (\alpha, \beta, \sigma)'$, which can be estimated with maximum likelihood. Conveniently, the transition density for a discretely-sampled path is known. When $t > s$,

$$p(y_t, t|y_s, s;\theta) = ce^{-w-v}\left(\frac{v}{w}\right)^{q/2} I_q(2\sqrt{wv}), \tag{24}$$

where $c = 2\beta/(\sigma^2(1 - e^{-\beta(t-s)}))$, $w = cy_s e^{-\beta(t-s)}$, $v = cy_t$, $q = 2\alpha\beta/\sigma^2 - 1$, and $I_q(\cdot)$ is the modified Bessel function of the first kind of order q. Note that non-negativity of y_t requires $2\alpha\beta \ge \sigma^2$.

Following the benchmark in Durham and Gallant (2002), we generate 1000 artificial sample paths of y_t according to the true transition density with $(\alpha, \beta, \sigma) = (0.06, 0.5, 0.15)$. Each path is then discretely observed 300 times, with time distance $t - s = 1/12$ between observations. These can be thought of monthly observations over 25 years. We estimate the diffusion parameters with true maximum likelihood and then with our NPSML.

In the implementation of NPSML, we forgo our knowledge of (24). Given a set of parameter values and y_s, we simulate paths using the Euler scheme; c.f. Kloeden and Platen (1992).[8] We divide the interval $t - s$ into M subintervals, and recursively compute for $m = 0, \ldots, M - 1$:

$$u_{m+1}^i = u_m^i + \beta(\alpha - u_m^i)\delta + \sigma\sqrt{u_m^i}\delta^{1/2}W_{m+1}^i,$$

where $u_0^i = y_s$, $\delta = \frac{t-s}{M}$, and W_{m+1}^i's are i.i.d. standard normal random variables. Then, we set $\hat{Y}_{t,i}^\theta = u_M^i$, the i-th simulated observation of y_t conditional on y_s. Once we have generated $\hat{Y}_{t,i}^\theta$ for $i = 1, \ldots, N$, then we can estimate the transition density by $\hat{p}_M(y_t, t|y_s, s) = \sum_{i=1}^N K_h(\hat{Y}_{t,i}^\theta - y_t)/N$, where K is chosen as a Gaussian kernel. It is now straightforward to construct $\hat{L}_T(\theta)$ and maximize it over Θ. We do not use trimming in our construction of the likelihood.

To study the finite-sample properties of our estimates as we increase the number of simulated observations per data point, we run NPSML for four different N's ($N = 100, 250, 500, 625$). Another choice to be made in implementing NPSML is bandwidth h. Silverman's rule of thumb gives us $h = 0.0035$, using the estimated standard deviation of $y_{t+1} - y_t$ and $N = 100$. In each estimation with a given h and N, we hold fixed h while we maximize over the parameter space. To assess the effect of bandwidth choice in NPSML, we consider three different h's ($h = 0.0030, 0.0035, 0.0040$). In the simulation stage of NPSML, we use Euler scheme with $M = 10$ to approximate the continuous-time process, and also adopt antithetic methods to reduce the simulation variance.

The results of the simulation study are shown in Table 1. The true MLE result is shown in the top panel. In the middle panel, we fix the number of simulated observations at $N = 500$, and vary the bandwidth for NPSML. In the bottom panel, we fix the bandwidth at $h = 0.0035$, and vary the number of simulations for NPSML. The true parameter values are $(\alpha, \beta, \sigma) = (0.06, 0.5, 0.15)$. We report the bias, standard deviation, and the root mean-squared error (RMSE) from the 1000 estimations.

For all parameters, NPSML has larger biases and standard deviations (and hence RMSEs) than does the true MLE, with the exception being the smaller bias in the NPSML estimates of β. In the bottom panel, we see that the biases and standard deviations decrease as we increase the number of simulated observations N. The exception is the bias of the σ estimates, which increases and then decreases in absolute value with N.

[8] We are approximating a continuous-time process using a discretization scheme, and hence need to appeal to Theorem 3.4.

Table 1
Estimation results for the Cox–Ingersoll–Ross model.

			α	β	σ
True MLE		Bias	−0.0008	0.0805	0.0002
		Std. dev.	0.0069	0.1214	0.0034
		RMSE	0.0069	0.1456	0.0034
NPSMLE $N = 500$	$h = 0.0030$	Bias	0.0036	0.0749	−0.0060
		Std. dev.	0.0143	0.1888	0.0066
		RMSE	0.0147	0.2031	0.0089
	$h = 0.0035$	Bias	0.0028	0.0778	−0.0093
		Std. dev.	0.0132	0.1857	0.0064
		RMSE	0.0135	0.2013	0.0113
	$h = 0.0040$	Bias	0.0021	0.0793	−0.0129
		Std. dev.	0.0122	0.1819	0.0064
		RMSE	0.0124	0.1984	0.0144
NPSMLE $h = 0.0035$	$N = 100$	Bias	0.0071	0.0740	−0.0043
		Std. dev.	0.0180	0.2115	0.0076
		RMSE	0.0193	0.2241	0.0087
	$N = 250$	Bias	0.0031	0.0817	−0.0078
		Std. dev.	0.0127	0.1908	0.0068
		RMSE	0.0131	0.2076	0.0103
	$N = 625$	Bias	0.0006	0.0650	−0.0090
		Std. dev.	0.0076	0.1269	0.0043
		RMSE	0.0076	0.1426	0.0100

Changing bandwidths produces a more complex pattern. For α, both biases and standard deviations decrease as we increase h. For β, biases increase but standard deviations fall as h goes up, although RMSE goes down like standard deviations. For σ, the magnitude of biases increases with h, while standard deviations barely move leading to an overall increase in RMSE as h increases. However, the changes in bias and standard deviations are not very large, and we conclude that the NPSMLE appears to be quite robust towards the choice of bandwidth.

Comparing the magnitudes of biases and standard deviations from NPSML with those of the true MLE, one can get a sense of how the additional biases and variances as stated in Theorem 3.3 are affected by our choice of bandwidth h and the number of simulated observations N.

4.1.2. Jump–diffusion

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

$$dy_{1,t} = \left(\mu - \frac{\exp(y_{2,t})}{2}\right)dt + \exp\left(\frac{y_{2,t}}{2}\right)dW_{1,t} + \log(1+J_t)dQ_t, \tag{25}$$

$$dy_{2,t} = (\alpha_0 - \alpha_1 y_{2,t})dt + \alpha_2 dW_{2,t}. \tag{26}$$

This specification is used by Andersen et al. (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 Aït-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 ρ 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 λ_0. The jump size J_t is assumed to be log-normally distributed:

$$\log(1+J_t) \sim \mathcal{N}(-0.5\gamma^2, \gamma^2). \tag{27}$$

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

Ideally, we would like to give precise conditions under which the general jump–diffusion (3) satisfies Assumptions A.1–A.4 and C.1–C.4. However, this proves very difficult without imposing strong conditions ruling out standard models considered in empirical finance, including the current example (25) and (26). 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 terms 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 in closed form. Only in a few special cases, its properties can be derived; see e.g. Aït-Sahalia (2002).

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

Then we use NPSML without using 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 θ, and we use the Euler scheme to approximate the data generating process. 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 $m = 1, \ldots, M$:

$$u^i_{1,m} = u^i_{1,m-1} + \left(\mu - \frac{\exp(u^i_{2,m-1})}{2}\right)\frac{1}{M} + \exp\left(\frac{u^i_{2,m-1}}{2}\right)\frac{\widetilde{W}^i_{1,m}}{\sqrt{M}} + \log(1+J^i_m)U^i_m,$$

$$u^i_{2,m} = u^i_{2,m-1} + (\alpha_0 - \alpha_1 u^i_{2,m-1})\frac{1}{M} + \alpha_2\frac{W^i_{2,m}}{\sqrt{M}},$$

with $u^i_{1,0} = y_{1,s}$ and $u^i_{2,0} = y_{2,s}$ for all $i = 1, \ldots, N$; J^i_m is an i.i.d. random variable with its distribution given in (27); U^i_m is an i.i.d. binomial random variable, with $Prob(U^i_m = 1) = \frac{\lambda_0}{M}$; $\widetilde{W}^1_{m,i} = \sqrt{1-\rho^2}W^i_{1,m} + \rho W^i_{2,m}$, where $W^i_{1,m}$ and $W^i_{2,m}$ are i.i.d. standard normal random variables. The subscript i indexes simulations. In our benchmark estimation, we use $N = 1000$.

With the (approximate) simulated observations $\hat{Y}^\theta_{s+1,i} \equiv (u^i_{1,M}, u^i_{2,M})$ for $i = 1, \ldots, N$ with $N = 1000$, we use (2) to obtain

$$\hat{p}_M(y_{1,s+1}, y_{2,s+1}|y_{1,s}, y_{2,s}; \theta) = \frac{1}{N}\sum_{i=1}^N K_h(\hat{Y}^\theta_{s+1,i} - (y_{1,s+1}, y_{2,s+1}))$$

where K is a multiplicative Gaussian kernel, $K_h(\cdot) = K_{h_1}(\cdot)K_{h_2}(\cdot)$. The bandwidths h_1 and h_2 (for y_1 and y_2 respectively) are chosen by the rule of thumb of Scott (1992, p. 152). Again, we hold fixed h_1 and h_2 while we maximize over the parameter space.

With the estimated $\hat{p}_t$ for $t = 1, 2, \ldots, T$, 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 θ.[9]

In our simulation study, we draw 100 sample paths of length $T = 1000$ each, and estimate each sample path with NPSML.

[9] In the case of the binomial random variable U, we fix the realization of the underlying uniform random variable. For different θ-λ_0, to be exact, U itself may have different realizations.

Table 2
Estimation results for jump–diffusion. 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 $N = 1000$ and the rule-of-thumb bandwidths. Column (2) reports the results with $N = 750$. Column (3) is for $N = 1000$ and bandwidths that are 20% narrower than those in the benchmark. Column (4) is for $N = 1000$ and bandwidths that are 20% wider than those in the benchmark.

Parameter	True value	(1)	(2)	(3)	(4)
μ	0.0304	0.0305	0.0305	0.0307	0.0306
		(0.0022, 0.0524)	(0.0162, 0.0533)	(0.0012, 0.0661)	(0.0084, 0.0448)
α_0	−0.0120	−0.0148	−0.0152	−0.0144	−0.0149
		(−0.0181, −0.0100)	(−0.0186, −0.0111)	(−0.0188, −0.0088)	(−0.0185, −0.0100)
α_1	0.0145	0.0161	0.0164	0.0160	0.0161
		(0.0116, 0.2061)	(0.0121, 0.0207)	(0.0114, 0.0214)	(0.0120, 0.0214)
α_2	0.1153	0.1147	0.1139	0.1167	0.1127
		(0.1107, 0.1168)	(0.1092, 0.1185)	(0.1118, 0.1198)	(0.1089, 0.1156)
γ	0.0150	0.0199	0.0310	0.0100	0.0121
		(0.0060, 0.0542)	(0.0000, 0.0368)	(0.0017, 0.0158)	(0.0085, 0.0126)
ρ	−0.6125	−0.7291	−0.7526	−0.6933	−0.7740
		(−0.7595, −0.6863)	(−0.7984, −0.7012)	(−0.7189, −0.6592)	(−0.8064, −0.7344)
λ_0	0.0200	0.0169	0.0133	0.0196	0.0166
		(0.0101, 0.0213)	(0.0086, 0.0175)	(0.0122, 0.0197)	(0.0104, 0.0222)

In column (1) of Table 2, we report the mean of the 100 point estimates for each parameter, and the 90% confidence interval constructed from the point estimates, with $N = 1000$ and the rule-of-thumb bandwidths. The NPSML performs reasonably well, although the correlation coefficient ρ is systemically underestimated. One remarkable outcome is that the jump parameters (γ and λ_0) are rather precisely estimated, even though there are only 20 or so jump realizations in each sample path.[10]

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 $N = 1000$, but reduce both bandwidths by 20%. Finally, in column (4), we use $N = 1000$ and bandwidths that are 20% greater than those in the benchmark.

When N is reduced to 750—column (2), the mean estimates move further away from the true parameter values. However, there is no clear increase or decrease in the dispersion of the 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 estimates are closer to the true parameter values than in the benchmark. Note the estimates of ρ, in particular. On the other hand, the results with over-smoothing as in column (4) do not compare favorably with the benchmark results with slightly more bias. We, in accordance with our theory, recommend a bandwidth narrower that what is given by the rule of thumb in actual implementations. At the same time, we would like to emphasize that the estimator in general is quite robust to the choice of bandwidth.

4.2. Markov decision processes and dynamic games

Another class of economic models that NPSML can 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 (Kristensen and Schjerning, 2011).

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 states, and n is the number of state variables. The transition probability depends on individual firm-specific shocks, industry-wide shocks, and Markov-perfect strategies of firms regarding entry, exit and investment.[11] 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 ω_t.[12] Thus, NPSML is feasible. 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 applicable to such continuous-time dynamic stochastic games as well.

5. Concluding remarks

We have generalized the NPSML 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 NPSMLE 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 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 (Brownlees et al., 2011), based on the main results given here.

[10] 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).

[11] In this class of models, conditioning on ω_t, ω_{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.

[12] In solving individual firms' dynamic programming problem, one needs to know their continuation value, and hence the transition probability. Therefore, for a given θ, one needs to compute a fixed point in $p_t(\omega_{t+1}|\omega_t)$.

Appendix 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 our proposed NPSMLE. In the rest of this part of the appendix, we consider the fully general situation where $\hat{L}_T(\theta) = \hat{L}_{T,N}(\theta)$ is a sequence of approximations to $L_T(\theta)$ (not necessarily the nonparametric simulated one proposed in the main text) of some infeasible (quasi-)log-likelihood function $L_T(\theta)$. We will analyze the impact of the approximation on the corresponding approximate estimator $\hat{\theta}$.

We first establish the asymptotic properties of the true MLE $\tilde{\theta}$ under Assumptions C.1–C.3. We then 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 A.1. *Assume that* C.1 *and* C.2 *hold. Then* $\jmath_T^{1/2}(\tilde{\theta} - \theta_0) = o_P(1)$.

Proof. We introduce a normalized version of the parameter $\theta \in \Theta_T$, $\xi := \jmath_T^{1/2}(\theta - \theta_0)$, and define the corresponding likelihood in terms of this new parameterization, $Q_T(\xi) := L_T(\theta_0 + \jmath_T^{-1/2}\xi)$. The claim will now follow from $\tilde{\xi} := \jmath_T^{1/2}(\tilde{\theta} - \theta_0) \xrightarrow{P} 0$. However, under C.1 and C.2, it is easily seen that $Q_T(\xi)$ satisfies the conditions of Theorem 3.4 of White (1994) with $\bar{Q}_T(\xi) = Q_T(\xi)$ from which the desired result follows. □

Lemma A.2. *Assume that* C.1 *holds. Then* C.3 *implies* C.2 .

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

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

for some $\bar{\theta} \in [\theta_0, \theta_0 + \jmath_T^{-1/2}\xi_T] \in \Theta_T$, and with $U_T(\theta)$ and $V_T(\theta)$ defined in Eq. (14). By another application of Taylor's theorem,

$$\begin{aligned}
&|\xi_T' V_T(\bar{\theta})\xi_T - \xi_T' V_T(\theta_0)\xi_T| \\
&= |\xi_T' \jmath_T^{-1/2}[H_T(\bar{\theta}) - H_T(\theta_0)]\jmath_T^{-1/2}\xi_T| \\
&\le \left\{ \max_{i=1,\ldots,d} \sup_{\theta\in\Theta_T} \xi_T' W_{T,i}(\theta)\xi_T \right\} \times \|\jmath_T^{-1/2}\xi_T\| = o_P(1),
\end{aligned}$$

where we have used C.3.2 and the fact that $\|\jmath_T^{-1/2}\xi_T\| = o_P(1)$. Thus,

$$\begin{aligned}
L_T(\theta_0 + \jmath_T^{-1/2}\xi_T) - L_T(\theta_0) &= U_T(\theta_0)\xi_T + \frac{1}{2}\xi_T' V_T(\theta_0)\xi_T + o_P(1) \\
&= \frac{1}{2}\xi_T' H_\infty \xi_T' + o_P(1),
\end{aligned}$$

where the second equality follows from C.3.1. Since $\xi_T' H_\infty \xi_T' < 0$ a.s., $L_T(\theta)$ is concave with a unique maximum at $\theta = \theta_0$ in Θ_T (with probability tending to one). In particular, C.2 holds. □

Lemma A.3. *Assume that* C.1 *and* C.3 *hold. Then the MLE* $\tilde{\theta}$ *satisfies*

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

Proof. By Lemmas A.1 and A.2, we know that $\tilde{\theta}$ is consistent. A first-order Taylor expansion of the score and C.3.2 together with the same arguments as in the proof of Lemma A.2 yield

$$\begin{aligned}
\sqrt{\nu_T}U_T(\theta_0) &= -V_T(\bar{\theta})\sqrt{\nu_T}\jmath_T^{1/2}(\tilde{\theta} - \theta_0) \\
&= -V_T(\theta_0)\sqrt{\nu_T}\jmath_T^{1/2}(\tilde{\theta} - \theta_0) + o_P(1),
\end{aligned}$$

and the result now follows from C.3.1. □

A.2. Asymptotics of approximate MLE

Theorem A.4. *Assume that* C.1 *and* C.2 *hold and* $\sup_{\theta\in\Theta_T} |\hat{L}_T(\theta) - L_T(\theta)| = o_P(1)$ *as* $T \to \infty$ *for a sequence* $N = N(T) \to \infty$. *Then* $\jmath_T^{1/2}(\hat{\theta} - \theta_0) = o_P(1)$.

Proof. We wish to show that for any $\eta > 0$,

$$P(\|\jmath_T^{1/2}(\hat{\theta} - \theta_0)\| > \eta) \to 0, \quad T \to \infty.$$

Let $\eta > 0$ be given. Then by C.2 there exists a $\delta > 0$ such that, $L_T(\theta_0) - L_T(\hat{\theta}) \ge \delta$ with probability tending to 1. Thus, as $T \to \infty$,

$$P(\|\jmath_T^{1/2}(\hat{\theta} - \theta_0)\| > \eta) \le P(L_T(\theta_0) - L_T(\hat{\theta}) \ge \delta).$$

We then have to show that the right-hand side converges to zero. To this end, write

$$L_T(\theta_0) - L_T(\hat{\theta}) = \{L_T(\theta_0) - \hat{L}_T(\theta_0)\} + \{\hat{L}_T(\theta_0) - L_T(\hat{\theta})\},$$

where,

$$L_T(\theta_0) - \hat{L}_T(\theta_0) \le \sup_{\theta\in\Theta_T} |L_T(\theta) - \hat{L}_T(\theta)| = o_P(1),$$

while, by the definition of $\hat{\theta}$,

$$\begin{aligned}
\hat{L}_T(\theta_0) - L_T(\hat{\theta}) &\le \hat{L}_T(\hat{\theta}) - L_T(\hat{\theta}) \\
&\le \sup_{\theta\in\Theta_T} |L_T(\theta) - \hat{L}_T(\theta)| = o_P(1). \quad \square
\end{aligned}$$

Next, we state two results for the approximate estimator to have the same asymptotic distribution as the actual MLE. Theorem A.5 establishes this result only requiring that the approximate likelihood function satisfies $\sup_{\theta\in\Theta_T} |\hat{L}_T(\theta) - L_T(\theta)| = o_P(1/\sqrt{\nu_T})$. Theorem A.6 imposes stronger smoothness conditions, requiring that $\hat{L}_T(\theta)$ be twice differentiable with derivatives $\hat{S}_T(\theta)$ and $\hat{H}_T(\theta)$; on the other hand, we only require $\|\hat{S}_T(\theta_0) - S_T(\theta_0)\| = o_P(1/\|\sqrt{\nu_T}\jmath_T^{1/2}\|)$ and $\sup_{\theta\in\Theta_T} \|\hat{H}_T(\theta) - H_T(\theta)\| = o_P(\|\jmath_T^{-1}\|)$ which are weaker convergence restrictions than $o_P(1/\sqrt{\nu_T})$, since $\|\jmath_T^{-1/2}\| = O(1)$. So there is a trade-off between smoothness and the convergence rate.

Theorem A.5. *Assume that* C.1 *and* C.3 *hold, and* $\sup_{\theta\in\Theta_T} |\hat{L}_T(\theta) - L_T(\theta)| = o_P(1/\sqrt{\nu_T})$ *for some sequence* $N = N(T) \to \infty$. *Then,* $\sqrt{\nu_T}\jmath_T^{1/2}(\hat{\theta} - \tilde{\theta}) = o_P(1)$. *In particular,* $\sqrt{\nu_T}\jmath_T^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} -H_\infty^{-1}S_\infty$.

Proof. Define $\tilde{\xi}_T = \jmath_T^{1/2}(\tilde{\theta} - \theta_0)$ and $\hat{\xi}_T = \jmath_T^{1/2}(\hat{\theta} - \theta_0)$. By a second-order Taylor expansion of $L_T(\hat{\theta})$ around $\tilde{\theta}$ together with the fact that $S_T(\tilde{\theta}) = 0$ yields, for some $\bar{\theta} \in [\tilde{\theta}, \hat{\theta}]$,

$$\begin{aligned}
L_T(\tilde{\theta}) - L_T(\hat{\theta}) &= L_T(\theta_0 + \jmath_T^{-1/2}\tilde{\xi}_T) - L_T(\theta_0 + \jmath_T^{-1/2}\hat{\xi}_T) \\
&= \frac{1}{2}(\tilde{\xi}_T - \hat{\xi}_T)' \jmath_T^{-1/2} H_T(\bar{\theta}) \jmath_T^{-1/2}(\tilde{\xi}_T - \hat{\xi}_T).
\end{aligned}$$

By the same arguments as in the proof of Lemma A.2, we obtain

$$L_T(\tilde{\theta}) - L_T(\hat{\theta}) = \frac{1}{2}(\tilde{\xi}_T - \hat{\xi}_T)' H_\infty (\tilde{\xi}_T - \hat{\xi}_T) + o_P(1/\sqrt{\nu_T}).$$

Write the left-hand side as

$$\begin{aligned}
\sqrt{\nu_T}\{L_T(\tilde{\theta}) - L_T(\hat{\theta})\} &= \sqrt{\nu_T}\{L_T(\tilde{\theta}) - \hat{L}_T(\tilde{\theta})\} \\
&\quad + \sqrt{\nu_T}\{\hat{L}_T(\tilde{\theta}) - L_T(\hat{\theta})\}
\end{aligned}$$

where

$$\sqrt{\nu_T}\{L_T(\tilde{\theta}) - \hat{L}_T(\tilde{\theta})\} \le \sqrt{\nu_T} \sup_{\theta\in\Theta_T} |\hat{L}_T(\theta) - L_T(\theta)| = o_P(1),$$

and, using that $\hat{\theta}$ is the maximizer of $\hat{L}_T(\theta)$,

$$\begin{aligned}\sqrt{\nu_T}\{\hat{L}_T(\tilde{\theta}) - L_T(\hat{\theta})\} &\leq \sqrt{\nu_T}\{\hat{L}_T(\hat{\theta}) - L_T(\hat{\theta})\} \\ &\leq \sqrt{\nu_T}\sup_{\theta\in\Theta_T}|\hat{L}_T(\theta) - L_T(\theta)| = o_P(1).\end{aligned}$$

Thus,

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

Theorem A.6. *Assume that* C.1 *and* C.3 *hold together with the following.*

(i) $\theta \mapsto \hat{L}_T(\theta)$ *is twice differentiable in* Θ_T.

(ii) *There exists a sequence* $N = N(T) \to \infty$ *such that* $\|\mathcal{I}_T^{1/2}\{\hat{S}_T(\theta_0) - S_T(\theta_0)\}\| = o_P(1/\sqrt{\nu_T})$, *and* $\|\mathcal{I}_T^{-1}\{H_T(\theta_0) - \hat{H}_T(\theta_0)\}\| = o_P(1)$.

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 standard Taylor expansions,

$$\begin{aligned}0 &= \sqrt{\nu_T}\mathcal{I}_T^{-1/2}S_T(\tilde{\theta}) \\ &= \sqrt{\nu_T}\mathcal{I}_T^{1/2}S_T(\theta_0) + \mathcal{I}_T^{-1/2}H_T(\theta_0)\mathcal{I}_T^{-1/2}\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) \\ &\quad + o_P(1),\end{aligned}$$

and

$$\begin{aligned}0 &= \sqrt{\nu_T}\mathcal{I}_T^{-1/2}\hat{S}_T(\hat{\theta}) \\ &= \sqrt{\nu_T}\mathcal{I}_T^{1/2}\hat{S}_T(\theta_0) + \mathcal{I}_T^{-1/2}\hat{H}_T(\theta_0)\mathcal{I}_T^{-1/2}\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) \\ &\quad + o_P(1).\end{aligned}$$

Subtracting the two equations, and using that $\mathcal{I}_T^{-1/2}H_T(\theta_0)\mathcal{I}_T^{-1/2} \xrightarrow{d} H_\infty > 0$, we obtain

$$\begin{aligned}0 &= \sqrt{\nu_T}\mathcal{I}_T^{1/2}\{S_T(\theta_0) - \hat{S}_T(\theta_0)\} + H_\infty\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\tilde{\theta} - \hat{\theta}) \\ &\quad + \mathcal{I}_T^{-1/2}\{H_T(\theta_0) - \hat{H}_T(\theta_0)\}\mathcal{I}_T^{-1/2}\sqrt{\nu_T}\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0).\end{aligned}$$

The result now follows from condition (ii) and Lemma A.3. $\square$

Appendix B. Properties of simulated conditional density

We here establish uniform convergence of $\hat{p}_t$ given in Eq. (7) and its derivatives w.r.t. θ.

Lemma B.1. *Assume that* A.1, A.2 *and* K.1 *hold. Then* $\hat{p}_t$ *in* (7) *satisfies for all* $y_2 \in \mathcal{Y}_2$ *and any compact set* Θ:

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

where $\bar{\lambda}_i = \lambda_{0,i} + \lambda_{1,i} + \lambda_{2,i}$, $i = 1, 2$.

Proof. Define $\gamma = (x, \theta, t) \in \Gamma = \mathcal{X}_t \times \Theta \times \{1, 2, 3, \ldots\}$. 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:

$$\begin{aligned}\hat{p}(y_1, y_2; \gamma) - p(y_1, y_2; \gamma) &= \{\mathbb{E}[\hat{p}(y_1, y_2; \gamma)] - p(y_1, y_2; \gamma)\} \\ &\quad + \{\hat{p}(y_1, y_2; \gamma) - \mathbb{E}[\hat{p}(y_1, y_2; \gamma)]\} \\ &=: \text{Bias}(y_1, y_2; \gamma) + \text{Var}(y_1, y_2; \gamma).\end{aligned}$$

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

$$\begin{aligned}|\text{Bias}(y_1, y_2; \gamma)| &\leq h^r\int|K(v)|\,\|v\|^r dv \times \left|\frac{\partial^r p_t(y_1, y_2; \gamma)}{\partial y_1^r}\right| \\ &\quad + o(h^r).\end{aligned}$$

Thus, using the bound imposed on the r-th derivative, $|\text{Bias}(y_1, y_2; \gamma)| = O([d_n^{\lambda_{0,1}} + T^{\lambda_{0,2}}]h^r)$ uniformly over (y_1, γ). To establish the uniform rate of the variance term, we apply the result of Kristensen (2009, Theorem 1) for averages of the form

$$\hat{\Psi}(x;\gamma) = \frac{1}{nh^d}\sum_{i=1}^n Y_i(\gamma)G\left(\frac{X_i(\gamma) - x}{h}\right), \tag{28}$$

for some kernel-type function G. With $Y_i(\gamma) = g_{2,t}(y_2, x, \varepsilon_i;\theta)$, $X_i(\gamma) = g_{1,t}(x, \varepsilon_i;\theta)$ and $G = K$, our simulated density can be written in this form. We then verify that Kristensen's conditions A.1–A.6 are satisfied under our assumptions. His A.1 is trivially satisfied since we have i.i.d. draws, while his A.6 imposed on G is implied by our K.1. The bounds in his A.4 and A.5 become in our case, using his Remark 2.2:

$$\begin{aligned}&\widetilde{B}_0 = p(y_1;\gamma), \qquad \widetilde{B}_1 = \|y_1\|^k\mathbb{E}[|Y_i(\gamma)|\,|X_i(\gamma) = y_1]p(y_1;\gamma), \\ &\widetilde{B}_2 = \|y_1\|^k\mathbb{E}[\|\partial_\gamma Y_i(\gamma)\|\,|X_i(\gamma) = y_1]p(y_1;\gamma), \\ &\widetilde{B}_3 = \|y_1\|^k\mathbb{E}[|Y_i(\gamma)|\,\|\partial_\gamma X_i(\gamma)\|\,|X_i(\gamma) = y_1]p(y_1;\gamma),\end{aligned}$$

where $\partial_\gamma Y_i(\gamma)$ and $\partial_\gamma X_i(\gamma)$ denote their derivatives w.r.t. $\gamma = (x, \theta, t)$. By A.2, $\widetilde{B}_0 = O(1 + \|x\|^{\lambda_{0,1}} + t^{\lambda_{0,2}})$ while, using A.1,

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

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

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

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

Lemma B.2. *Assume that* A.1–4 *and* K.1 *hold. Then* $\partial^i\hat{p}_t/\partial\theta^i$, $i = 1, 2$, *given in* (15) *and* (16) *satisfy for all* $y_2 \in \mathcal{Y}_2$ *and any compact set* Θ:

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

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 B.1. With $Y_{1t,i}^\theta = g_{1,t}(x_t, \varepsilon_i;\theta)$ and $Y_{2t,i}^\theta = g_{2,t}(y_{2t}, x_t, \varepsilon_i;\theta)$, it follows from the expression in Eq. (15) that conditional on (y_{1t}, y_{2t}, x_t):

$$\begin{aligned}\mathbb{E}\left[\frac{\partial\hat{p}_t(y_{1t}, y_{2t}|x_t;\theta)}{\partial\theta}\right] &= \frac{1}{h^k}\int\dot{Y}_{2t,i}^\theta K\left(\frac{Y_{1t,i}^\theta - y_{1t}}{h}\right) \\ &\quad \times dF_\varepsilon(\varepsilon) + \frac{1}{h^{k+1}}\int\dot{Y}_{1t,i}^\theta K^{(1)} \\ &\quad \times\left(\frac{Y_{1t,i}^\theta - y_{1t}}{h}\right)Y_{2t,i}^\theta dF_\varepsilon(\varepsilon),\end{aligned}$$

where, uniformly over (t, x_t, θ),

$$\frac{1}{h^k}\int\dot{Y}_{2t,i}^\theta K\left(\frac{Y_{1t,i}^\theta - y_{1t}}{h}\right)dF_\varepsilon(\varepsilon)$$

$$= \int K(v)p(y_{1t}+vh|x_t;\theta)\frac{\partial p_t(y_{2t}|y_{1t}+vh,x_t;\theta)}{\partial\theta}dv$$
$$= \frac{\partial p_t(y_{2t}|y_{1t},x;\theta)}{\partial\theta}p_t(y_{1t}|x_t;\theta)+O([d_n^{\lambda_1}+T^{\lambda_2}]h^r),$$
$$\frac{1}{h^{k+1}}\int \dot{Y}^\theta_{1t,i}K^{(1)}\left(\frac{Y^\theta_{1t,i}-y_{1t}}{h}\right)Y^\theta_{2t,i}dF_\varepsilon(\varepsilon)$$
$$= p_t(y_{2t}|y_{1t},x_t;\theta)\frac{\partial p_t(y_{1t}|x_t;\theta)}{\partial\theta}+O([d_n^{\lambda_1}+T^{\lambda_2}]h^r).$$

For the variance component, we again apply the results of Kristensen (2009). With $\gamma=(x,\theta,t)$ and $X_{n,i}(\gamma)=g_{1,t}(x,\varepsilon_i;\theta)$, $\partial\hat{p}_t/\partial\theta$ can be written as the sum of two kernel averages, each of form (28); the first with $Y_{n,i}(\gamma)=\dot{g}_{1,t}(x,\varepsilon_i;\theta)g_{2,t}(y_2,x,\varepsilon_i;\theta)$ and $G=K^{(1)}$, and the second with $Y_{n,i}(\gamma)=\dot{g}_{2,t}(y_2,x,\varepsilon_i;\theta)$ and $G=K$. Under the conditions imposed on our model, his A.1–A.5 hold. □

Appendix C. Proofs

Proof of Theorem 3.1. The first part of the result will follow if we can verify the conditions in Theorem A.4. In order to do this, we introduce an additional trimming function, $\tilde{\tau}_{a,t}=\tau_a(\hat{p}_t(y_t|x_t;\theta))\mathbb{I}\{\|x_t\|\le N^\gamma\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function and $\gamma>0$ is chosen as in B.1, and two trimming sets,

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

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 $N\to\infty$ under B.1. Thus, $\mathbb{I}_{A_{2,t}(4)}\le\mathbb{I}_{A_{1,t}(2)}\le\tilde{\tau}_{a,t}\le\mathbb{I}_{A_{1,t}(1/2)}\le\mathbb{I}_{A_t(1/4)}$.

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

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

and then show that $\sup_{\theta\in\Theta_T}|B_i(\theta)|=o_P(1)$, $i=1,2,3$. By C.4,

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

while,

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

The final term is bounded by

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

First, as $a\to0$,

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

where we have used C.5. Similarly, again by C.5,

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

The consistency result follows from Theorem A.4 together with Lemma B.1 and condition B.1.

To show the second result, we merely have to strengthen the convergence of $\hat{L}_T(\theta)$ to take place with rate ν_T; c.f. Theorem A.5. One can still apply the above bounds which now have to go to zero with rate ν_T. This is ensured by B.2. □

Proof of Corollary 3.2. We verify that C.1–C.4 hold with $\nu_T=T$ and $q=1$ under the conditions imposed in the corollary. First, we obtain by LLN for mixing sequences that $\imath_T(\theta_0)=i(\theta_0)+o_P(1)$ with $i(\theta_0)=\mathbb{E}[\partial^2\log p(y_t|x_t;\theta_0)/(\partial\theta\partial\theta')]$, such that $\imath_T$ can be chosen as the constant $\imath=\mathrm{diag}\{i(\theta)\}$. There is a one-to-one deterministic correspondence between the mapping $\xi\mapsto L_T(\theta_0+\imath_T^{-1/2}\xi)$ and $L_T(\theta)$ and we can restrict our attention to the latter.

To check C.2, first note that by the uniform LLN $\sup_{\theta\in\Theta}|L_T(\theta)-L(\theta)|=o_P(1)$ with $L(\theta)=\mathbb{E}[\log p(y_t|x_t;\theta)]$ being continuous under Condition (i); see, for example, Kristensen and Rahbek (2005, Proposition 1). Thus,

$$\begin{aligned}L_T(\tilde{\theta})-L_T(\theta) &= \{L(\tilde{\theta})-L(\theta)\}+\{L_T(\tilde{\theta})-L(\tilde{\theta})\}-\{L_T(\theta)-L(\theta)\}\\ &= \{L(\tilde{\theta})-L(\theta)\}+o_P(1)\\ &= \{L(\theta_0)-L(\theta)\}+o_P(1).\end{aligned}$$

Since Θ is compact and θ_0 is the unique maximum of $L(\theta)$, for any given $\delta>0$, there exists $\eta>0$ such that $\sup_{\|\theta_0-\theta\|>\eta}\{L(\theta_0)-L(\theta)\}\ge\delta$. This proves that C.2 holds.

To verify C.3, appeal to the Martingale CLT for mixing sequences to obtain

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

while by the LLN,

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

Finally,

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

Condition C.4 holds by another application of the (uniform) LLN:

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

$$T^{-1}\sum_{t=1}^{T}\|x_t\|^{1+\delta} \xrightarrow{P} \mathbb{E}[\|x_t\|^{1+\delta}],$$

$$T^{-1}\sum_{t=1}^{T}\Lambda_1^2\varepsilon_t \xrightarrow{P} \mathbb{E}[\Lambda_1^2(\varepsilon_t)]. \quad \square$$

Proof of Theorem 3.3. Since $p(y|x;\theta)$ is bounded away from zero, we can leave out the trimming and re-define the simulated likelihood as $\hat{L}_T(\theta) = T^{-1}\sum_{t=1}^{T}\log\hat{p}(y_t|x_t;\theta)$. The associated simulated 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 θ_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 $s_t(\theta) = \partial\log p_t(\theta)/(\partial\theta)$. Also, we will use the notation $\partial_\theta p_t(\theta) = \partial p_t(\theta)/(\partial\theta) \in \mathbb{R}^d$, $\partial^{\alpha,1}_{y_1,\theta}p_t(\theta) = \partial p_t(\theta)/(\partial y_1^\alpha\partial\theta) \in \mathbb{R}^d$, and similarly for other functions. When a function is evaluated at θ_0, we will frequently suppress the dependence on θ.

We first analyze $\hat{S}_T(\theta_0)$: by the same arguments as in Lee (1999, Proposition 1), the expansion given in Eq. (18) holds with

$$\nabla S_{T,N}[\hat{p}-p] = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{p_t}\{\partial_\theta\hat{p}_t - \partial_\theta p_t\} - \frac{s_t}{p_t}\{\hat{p}_t - p_t\}\right],$$

$$\nabla^2 S_{T,N}[\hat{p}-p,\hat{p}-p] = \frac{1}{T}\sum_{t=1}^{T}\left[-\frac{1}{p_t^2}\{\partial_\theta\hat{p}_t - \partial_\theta p_t\} \times \{\hat{p}_t - p_t\} + \frac{s_t}{p_t^2}\{\hat{p}_t - p_t\}^2\right],$$

$$R_{T,N} = \frac{1}{T}\sum_{t=1}^{T}\left[-\frac{1}{\hat{p}_t p_t^2}\{\partial_\theta\hat{p}_t - \partial_\theta p_t\}\{\hat{p}_t - p_t\}^2 + \frac{s_t}{\hat{p}_t p_t^2}\{\hat{p}_t - p_t\}^3\right].$$

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

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

where $\mathbb{E}[\cdot|\mathcal{Z}_T]$ denotes expectations conditional on data, $\mathcal{Z}_T = \{z_1,\ldots,z_T\}$ where $z_t = (y_t, x_t)$. Using standard bias expansions for kernel estimators,

$$p(y_{1,t},y_{2,t}|x_t) - \mathbb{E}[\hat{p}(y_{1,t},y_{2,t}|x_t)|z_t] = h^r\sum_{|\alpha|=r}\partial^\alpha_{y_1}p(y_{1,t},y_{2,t}|x_t) + o(h^r),$$

$$\partial_\theta p(y_{1,t},y_{2,t}|x_t) - \mathbb{E}[\partial_\theta\hat{p}(y_{1,t},y_{2,t}|x_t)|z_t] = h^r\sum_{|\alpha|=r}\partial^{\alpha,1}_{y_1,\theta}p(y_{1,t},y_{2,t}|x_t) + o(h^r),$$

implying that the bias component satisfies

$$\nabla S_{T,N}[\mathbb{E}[\hat{p}|\mathcal{Z}_T]-p] = h^r T^{-1}\sum_{t=1}^{T}\sum_{|\alpha|=r}\left[\frac{\partial^{\alpha,1}_{y,\theta}p_t}{p_t} - \frac{s_t}{p_t}\partial^\alpha_y p_t\right] + o_P(h^r) = \mu_1 h^r + o_P(h^r),$$

with μ_1 given in Eq. (22). Next, define

$$\psi_N(z_t,\varepsilon_i) = \frac{1}{p_t}\delta_{N,1}(z_t,\varepsilon_i) - \frac{s_t}{p_t}\delta_{N,2}(z_t,\varepsilon_i),$$

$$\delta_{N,1}(z_t,\varepsilon_i) = \frac{1}{h^{k+1}}\dot{Y}_{1t,i}K^{(1)}\left(\frac{Y_{1t,i}-y_{1t}}{h}\right)Y_{2t,i} + \frac{1}{h^k}\dot{Y}_{2t,i}K\left(\frac{Y_{1t,i}-y_{1t}}{h}\right),$$

$$\delta_{N,2}(z_t,\varepsilon_i) = \frac{1}{h^k}K\left(\frac{Y_{1t,i}-y_{1t}}{h}\right)Y_{2t,i}.$$

With $\psi_{N,1}(z_t) = \mathbb{E}[\psi_N(z_t,\varepsilon_i)|z_t]$, we can then write the variance component as

$$\nabla S_{T,N}[\hat{p}-\mathbb{E}[\hat{p}|\mathcal{Z}_T]] = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\{\psi_N(z_t,\varepsilon_i) - \psi_{N,1}(z_t)\},$$

which, since $\mathcal{Z}_T$ and $\mathcal{E}_N = \{\varepsilon_1,\ldots,\varepsilon_N\}$ are mutually independent, we recognize as a so-called two-sample U-statistic; see, for example, Lehmann (1951). With $\psi_{N,2}(\varepsilon_i) := \mathbb{E}[\psi_N(z_t,\varepsilon_i)|\varepsilon_i]$, and

$$w_N(z_t) := \frac{1}{N}\sum_{i=1}^{N}\{\psi_N(z_t,\varepsilon_i) - \psi_{N,1}(z_t) - \psi_{N,2}(\varepsilon_i) + \mathbb{E}[\psi(z_t,\varepsilon_i)]\},$$

we can decompose the statistic as

$$\nabla S_{T,N}[\hat{p}-\mathbb{E}[\hat{p}]] = \frac{1}{N}\sum_{i=1}^{N}\{\psi_{N,2}(\varepsilon_i) - \mathbb{E}[\psi(z_t,\varepsilon_i)]\} + \frac{1}{T}\sum_{t=1}^{T}w_N(z_t) =: U_{1,T,N} + W_{1,T,N}.$$

The second term will now be shown to be negligible. Conditioned on $\mathcal{E}_N$, $\{w_N(z_t): t=1,\ldots,T\}$ is a stationary and mixing sequence for any given N. Moreover,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}w_N(z_t)|\mathcal{E}_N\right] = \mathbb{E}[w_N(z_t)|\mathcal{E}_N] = \frac{1}{N}\sum_{i=1}^{N}\{\mathbb{E}[\psi_N(z_t,\varepsilon_i)|\varepsilon_i] - \mathbb{E}[\psi_{N,1}(z_t)]$$

$$- \psi_{N,2}(\varepsilon_i) + \mathbb{E}[\psi(z_t, \varepsilon_i)]\}$$
$$= \frac{1}{N}\sum_{i=1}^{N}\{\psi_{N,2}(\varepsilon_i) - \psi_{N,2}(\varepsilon_i)\} = 0.$$

This implies that $\mathbb{E}[W_{1,T,N}] = 0$, and, by the law of total variance,

$$\begin{aligned}\mathrm{Var}(W_{1,T,N}) &= \mathbb{E}[\mathrm{Var}(W_{1,T,N}|\mathcal{E}_N)] + \mathrm{Var}(\mathbb{E}[W_{1,T,N}|\mathcal{E}_N]) \\ &= \mathbb{E}[\mathrm{Var}(W_{1,T,N}|\mathcal{E}_N)].\end{aligned}$$

Here, by standard results for the variance of mixing sequences (see, for example, Kristensen and Salanie (2010, Lemma 5)),

$$\mathrm{Var}(W_{1,T,N}|\mathcal{E}_N) \le \frac{C}{T}\mathbb{E}[\|w_N(z_t)\|^{2+\delta}|\mathcal{E}_N]^{2/(2+\delta)},$$

for some constant $C > 0$ which only depends on the mixing coefficients of $\{z_t\}$. By standard arguments for i.i.d. sample averages of kernel smoothers,

$$\begin{aligned}\mathbb{E}[\|w_N(z_t)\|^{2+\delta}] &\le \frac{C}{N^{1+\delta/2}}\mathbb{E}[\|\psi_N(z_t, \varepsilon_i)\|^{2(1+\delta)}] \\ &= O\left(\frac{1}{N^{1+\delta/2}h^{k+2\delta}}\right).\end{aligned}$$

In total, $W_{1,T,N} = O_P(1/\sqrt{TNh^{2(k+2\delta)/(2+\delta)}})$. To analyze $U_{1,T,N}$, first note that

$$\begin{aligned}\mathbb{E}\left[\frac{\delta_{N,1}(z_t, \varepsilon_i)}{p(y_t|x_t)}|\varepsilon_i\right] &= \sum_{y_{2t}\in\mathcal{Y}_2}\int \dot{Y}_{1t,i}Y_{2t,i} \\ &\quad\times\left\{\frac{1}{h^{k+1}}\int K^{(1)}\left(\frac{Y_{1t,i}-y_{1t}}{h}\right)dy_{1t}\right\} \\ &\quad\times dF_x(x_t) + \sum_{y_{2t}\in\mathcal{Y}_2}\int \dot{Y}_{2t,i} \\ &\quad\times\left\{\frac{1}{h^{k}}\int K\left(\frac{Y_{1t,i}-y_{1t}}{h}\right)dy_{1t}\right\}dF_x(x_t) \\ &= \mathbb{E}\left[\frac{\dot{Y}_{2t,i}}{p(y_{2t}|x_t)}|\varepsilon_i\right], \qquad (29)\end{aligned}$$

where we have used $h^{-(k+1)}\int K^{(1)}(\frac{Y_{1t,i}-y_{1t}}{h})dy_{1t} = h^{-k}\int K^{(1)}(v)dv = 0$. Second,

$$\begin{aligned}&\mathbb{E}\left[\frac{s(y_t|x_t)}{p(y_t|x_t)}\delta_{N,2}(z_t, \varepsilon_i)\middle|\varepsilon_i\right] \\ &= \sum_{y_{2t}\in\mathcal{Y}_2}\int\left\{\frac{1}{h^k}\int s(y_{1t}, y_{2t}|x_t)\left(\frac{Y_{1t,i}-y_{1t}}{h}\right)dy_{1t}\right\} \\ &\quad\times Y_{2t,i}dF_x(x_t) \\ &= \sum_{y_{2t}\in\mathcal{Y}_2}\int s(Y_{1t,i}, y_2|x)Y_{2t,i}dF_x(x_t) + O(h^r) \\ &= \mathbb{E}\left[\frac{s(Y_{1t,i}, y_{2t}|x_t)Y_{2t,i}}{p(y_{2t}|x_t)}|\varepsilon_i\right] + O(h^r). \qquad (30)\end{aligned}$$

Thus, $U_{1,T,N} = \sum_{i=1}^{N}\{\psi_2(\varepsilon_i) - \mathbb{E}[\psi_2(\varepsilon_i)]\}/N + O_P(h^r)$, with $\psi_2(\varepsilon_i)$ defined in Eq. (20). By the CLT for mixing sequences, $\sqrt{N}U_{1,T,N} \xrightarrow{d} N(0, \mathrm{Var}(\psi_2(\varepsilon_i)))$. This establishes Eq. (19).

Next, we analyze the second order differential. Write

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

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

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

The variance component can be written as

$$\begin{aligned}&\nabla^2 S_{T,N}[\hat{p}-\mathbb{E}[\hat{p}|\mathcal{Z}_T], \hat{p}-\mathbb{E}[\hat{p}|\mathcal{Z}_T]] \\ &= \frac{1}{TN^2}\sum_{t=1}^{T}\sum_{i=1}^{N}\phi_N(z_t, \varepsilon_i, \varepsilon_i) + \frac{1}{TN^2}\sum_{t=1}^{T}\sum_{i\neq j}\phi_N(z_t, \varepsilon_i, \varepsilon_j) \\ &=: \frac{1}{N}U_{2,T,N} + W_{2,T,N},\end{aligned}$$

with

$$\begin{aligned}\phi_N(z_t, \varepsilon_i, \varepsilon_j) &= -\frac{1}{p_t^2}\{\delta_{N,1}(z_t, \varepsilon_i) - \mathbb{E}[\delta_{N,1}(z_t, \varepsilon_i)|z_t]\}\{\delta_{N,2}(z_t, \varepsilon_j) \\ &\quad-\mathbb{E}[\delta_{N,2}(z_t, \varepsilon_j)|z_t]\} + \frac{s_t}{p_t^2}\{\delta_{N,2}(z_t, \varepsilon_i) \\ &\quad-\mathbb{E}[\delta_{N,2}(z_t, \varepsilon_i)|z_t]\}\{\delta_{N,2}(z_t, \varepsilon_j) \\ &\quad-\mathbb{E}[\delta_{N,2}(z_t, \varepsilon_j)|z_t]\}.\end{aligned}$$

The first term, $U_{2,T,N}$, is again a second order two-sample U-statistic while $W_{2,T,N}$ is a third order one. To analyze $U_{2,T,N}$, we proceed in the same manner as with $\nabla S_{T,N}[\hat{p}-\mathbb{E}[\hat{p}]]$. By the Hoeffding decomposition,

$$\begin{aligned}U_{2,T,N} &= \mathbb{E}[\phi_N(z_t, \varepsilon_i, \varepsilon_i)] + \frac{1}{N}\sum_{i=1}^{N}\{\phi_{N,1}(\varepsilon_i) - \mathbb{E}[\phi_{N,1}(\varepsilon_i)]\} \\ &\quad+\frac{1}{T}\sum_{t=1}^{T}\{\phi_{N,2}(z_t) - \mathbb{E}[\phi_{N,2}(z_t)]\} + \tilde{R}_{T,N},\end{aligned}$$

where $\phi_{N,1}(\varepsilon_i) = \mathbb{E}[\phi_N(z_t, \varepsilon_i, \varepsilon_i)|\varepsilon_i]$, $\phi_2(z_t) = \mathbb{E}[\phi_N(z_t, \varepsilon_i, \varepsilon_i)|z_t]$ and $\tilde{R}_{T,N}$ is the remainder term. By the same arguments as used for $W_{1,T,N}$, it follows that $\tilde{R}_{T,N} = O_P\left(1/\sqrt{TNh^{k+\delta}}\right)$, while by the CLT for stationary and mixing sequences the two sample averages in the above expression are of order $O_P\left(1/\sqrt{T}\right)$ and $O_P\left(1/\sqrt{N}\right)$ respectively. Using the same arguments as before,

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

where $K^*(v) = K(v)K^{(1)}(v)$, and

$$\begin{aligned}&\mathbb{E}\left[\frac{s_t}{p_t^2}\delta_{N,2}^2(z_t, \varepsilon_i)|\varepsilon_i\right] \\ &\simeq \frac{1}{h^k}\mathbb{E}\left[\frac{Y_{2t,i}\dot{Y}_{2t,i}}{p(Y_{1t,i}, y_2|x)p(y_{2t}|x_t)}\middle|\varepsilon_i\right]\int K^2(v)dv,\end{aligned}$$

where we have left out higher order terms. Thus,

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

where μ_2 is given in Eq. (23) and $\tilde{\mu}_2$ is the sum of the two other expectations above. Next,

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

while $\mathbb{E}[\phi_N(z_t, \varepsilon_i, \varepsilon_j)|\varepsilon_i, \varepsilon_j] = o(1)$ as $h \to 0$ such that the corresponding U-statistic $W_{2,T,N}$ is second-order degenerate, implying $W_{2,T,N} = O_P(1/(N^{3/2}h^{k+1}))$. This establishes Eq. (21).

Finally, appealing to Lemma B.2 and the LLN, the remainder term of the expansion of $\hat{S}_T(\theta_0)$ satisfies

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

Next, we consider the simulated Hessian:

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

We write

$$\begin{aligned}\|\hat{H}_T(\bar{\theta}) - i(\theta_0)\| &\le \sup_{\theta\in\Theta}\|\hat{H}_T(\theta) - H_T(\theta)\| \\ &\quad + \sup_{\theta\in\Theta}\|H_T(\theta) - i(\theta)\| + \|i(\bar{\theta}) - i(\theta_0)\|.\end{aligned}$$

As shown in the proof of Corollary 3.2, the second and third terms are both $o_P(1)$. Regarding the first term, write

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

The first term, $A_1(\theta)$, satisfies uniformly in $\theta \in \Theta$:

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

where we have used the LLN together with $|\hat{p}_t(\theta) - p_t(\theta)| = o_P(1)$, $\|\partial^2_{\theta\theta}\hat{p}_t(\theta) - \partial^2_{\theta\theta}p_t(\theta)\| = o_P(1)$ and $p_t(\theta) > 0$ uniformly over $1 \le t \le T$ and $\theta \in \Theta$. We can show that $\sup_{\theta\in\Theta}\|A_2(\theta)\| = o_P(1)$ by the same arguments. The claimed result now follows by combining the results obtained for the simulated score and Hessian. □

Proof of Theorem 3.4. We follow the exact same arguments as in the proof of Theorem 3.3, except that we now have

$$\begin{aligned}&|\hat{p}_{M,t}(y_t|x_t;\theta) - p_t(y_t|x_t;\theta)| \\ &\quad \le |\hat{p}_{M,t}(y_t|x_t;\theta) - \hat{p}_t(y_t|x_t;\theta)| + |\hat{p}_t(y_t|x_t;\theta) - p_t(y_t|x_t;\theta)|\end{aligned}$$

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

$$\begin{aligned}&|\hat{p}_{M,t}(y_t|x_t;\theta) - \hat{p}_t(y_t|x_t;\theta)| \\ &\le \frac{1}{N}\sum_{i=1}^{N}|K_h(\hat{Y}^\theta_{1t,i} - y_{1t})\hat{Y}^\theta_{2t,i} - K_h(Y^\theta_{1t,i} - y_{1t})Y^\theta_{2t,i}| \\ &\le \frac{1}{Nh}\sum_{i=1}^{N}|K_h^{(1)}(\bar{Y}^\theta_{1t,i} - y_{1t})|\,|\hat{Y}^\theta_{2t,i}|\,\|\hat{Y}^\theta_{1t,i} - Y^\theta_{1t,i}\| \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}|K_h(Y^\theta_{1t,i} - y_{1t})|\,|\hat{Y}^\theta_{2t,i} - Y^\theta_{2t,i}| \\ &\le \left\{\frac{\delta_{M,1}}{h}B_1(1 + \|x_t\|^{\lambda_{3,1}} + t^{\lambda_{3,2}}) + \delta_{M,2}B_2(1 + \|x_t\|^{\lambda_{4,1}} + t^{\lambda_{4,2}})\right\} \\ &\quad \times O_P(1),\end{aligned}$$

where $\bar{Y}^\theta_{1t,i}$ lies on the line segment between $\hat{Y}^\theta_{1t,i}$ and $Y^\theta_{1t,i}$. Thus, uniformly over $y_t \in \mathbb{R}^k$, $\|x_t\| \le N^\gamma$ and $\theta \in \Theta$,

$$\begin{aligned}&\frac{T}{a\nu_T} \times |\hat{p}_{M,t}(y_t|x_t;\theta) - \hat{p}_t(y_t|x_t;\theta)| \\ &\quad = O_P\left(\frac{T}{a\nu_T}\frac{\delta_{M,1}}{h}[N^{\gamma\lambda_{3,1}} + T^{\lambda_{3,2}}]\right) \\ &\qquad + O_P\left(\frac{T}{a\nu_T}\delta_{M,2}[N^{\gamma\lambda_{4,1}} + T^{\lambda_{4,2}}]\right),\end{aligned}$$

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., 2008. Closed-form likelihood expansions for multivariate diffusions. Annals of Statistics 36, 906–937.

Altissimo, F., Mele, A., 2009. Simulated nonparametric estimation of dynamic models. Review of Economic Studies 76, 413–450.

Amemiya, T., 1983. Advanced Econometrics. Harvard University Press, Massachusets.

Andersen, T.G., Benzoni, L., Lund, J., 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., Doukhan, P., 2004. Weak dependence: models and applications to econometrics. Econometric Theory 20, 995–1045.

Aït-Sahalia, Y., 2002. Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. Econometrica 70, 223–262.

Aït-Sahalia, Y., Kimmel, R., 2007. Maximum likelihood estimation of stochastic volatility models. Journal of Financial Economics 83, 413–452.

Bandi, F.M., Phillips, P.C.B., 2003. Fully nonparametric estimation of scalar diffusion models. Econometrica 71, 241–283.

Basawa, I.V., Scott, D.J., 1983. Asymptotic Optimal Inference for Non-Ergodic Models. Springer-Verlag, New York.

Bichteler, K., Graveraux, J.-B., Jacod, J., 1987. Mallivin Calculus for Processes with Jumps. Gordon and Breach Science Publishers.

Brandt, M.W., Santa-Clara, P., 2002. Simulated likelihood estimation of diffusions with an application to exchange rates dynamics in incomplete markets. Journal of Financial Economics 63, 161–210.

Brownlees, C., Kristensen, D., Shin, Y., 2011. Smooth filtering and maximum-likelihood estimation of dynamic latent variable models. Manuscript, Department of Economics. Columbia University.

Bruti-Liberati, N., Platen, E., 2007. Approximation of jump diffusions in finance and economics. Computational Economics 29, 283–312.

Cai, Z., Yao, Q., Zhang, W., 2001. Smoothing for discrete-valued time series. Journal of the Royal Statistical Society (B) 63, 357–375.
Carrasco, M., Chernov, M., Florens, J.-P., Ghysels, E., 2007. Efficient estimation of general dynamic models with a continuum of moment conditions. Journal of Econometrics 140, 529–573.
Cox, J.C., Ingersoll Jr., J.E., Ross, S.A., 1985. A theory of the term structure of interest rates. Econometrica 53, 385–408.
Creel, M., Kristensen, D., 2011. Indirect likelihood inference. Manuscript, Department of Economics. Columbia University.
Creel, M., Kristensen, D., 2009. Estimation of Dynamic Latent Variable Models Using Simulated Nonparametric Moments, Working Paper 792.09, UFAE and IAE.
Detemple, J., Garcia, R., Rindisbacher, M., 2006. Asymptotic properties of monte carlo estimators of diffusion processes. Journal of Econometrics 134, 1–68.
Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D., 1996. Density estimation by wavelet tresholding. Annals of Statistics 24, 508–539.
Doraszelski, U., Judd, K., 2008. Avoiding the curse of dimensionality in dynamic stochastic games. Manuscript. Stanford University.
Doraszelski, U., Pakes, A., 2007. In: Armstrong, M., Porter, R. (Eds.), A Framework for Applied Dynamic Anaysis in IO. In: Handbook of Industrial Organization, vol. 3. Elsevier B.V., Amsterdam, pp. 1887–1966.
Duffie, D., Singleton, K.J., 1993. Simulated moments estimation of markov models of asset prices. Econometrica 61, 929–952.
Durham, G.B., Gallant, A.R., 2002. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. Journal of Business and Economic Statistics 20, 297–316.
Elerian, O., Chib, S., Shephard, N., 2001. Likelihood inference for discretely observed non-linear diffusions. Econometrica 69, 959–993.
Ericson, R., Pakes, A., 1995. Markov-perfect industry dynamics: a framework for empirical work. Review of Economic Studies 62, 53–82.
Fenton, V.M., Gallant, A.R., 1996. Notes and comments: convergence rates of SNP density estimators. Econometrica 64, 719–727.
Fermanian, J.-D., Salanié, B., 2004. A nonparametric simulated maximum likelihood estimation method. Econometric Theory 20, 701–734.
Gallant, A.R., Nychka, D.W., 1987. Semi-nonparametric maximum likelihood estimation. Econometrica 55, 363–390.
Gallant, A.R., Tauchen, G., 1996. Which moments to match? Econometric Theory 12, 657–681.
Gouriéroux, C., Monfort, A., Renault, É, 1993. Indirect inference. Journal of Applied Econometrics 8, S85–S118.
Hajivassiliou, V.A., McFadden, D.L., 1998. The method of simulated scores for the estimation of LDV models. Econometrica 66, 863–896.
Ichimura, H., Todd, P.E., 2007. In: Heckman, J.J., Leamer, E.E. (Eds.), Implementing Nonparametric and Semiparametric Estimators. In: Handbook of Econometrics, vol. 6. North Holland, Amsterdam, pp. 5369–5468.
Jeganathan, P., 1995. Some aspects of asymptotic theory with applications to time series models. Econometric Theory 11, 818–887.
Jones, M., Marron, J., Sheather, S., 1996. A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association 91, 401–407.
Karlsen, H.A., Tjøstheim, D., 2001. Nonparametric estimation in null recurrent time series. Annals of Statistics 29, 372–416.
Kloeden, P.E., Platen, E., 1992. Numerical Solution of Stochastic Differential Equations. Springer-Verlag, New York.
Kristensen, D., 2010. Pseudo-maximum-likelihood estimation in two classes of semiparametric diffusion models. Journal of Econometrics 156, 239–259.
Kristensen, D., Rahbek, A., 2010. Likelihood-based inference for cointegration with nonlinear error-correction. Journal of Econometrics 158, 78–94.
Kristensen, D., 2009. Uniform convergence rates of kernel estimators with heterogeneous, dependent data. Econometric Theory 25, 1433–1445.
Kristensen, D., Rahbek, A., 2005. Asymptotics of the QMLE for a class of ARCH(q) models. Econometric Theory 21, 946–961.
Kristensen, D., Salanie, B., 2010. Higher Order Improvements for Approximate Estimators, CAM Working Paper 2010-04, Department of Economics, University of Copenhagen.
Kristensen, D., Schjerning, B., 2011. Implementation and Estimation of Discrete Markov Decision Models By Sieve Approximations, Manuscript, University of Copenhagen.
Lee, L.-F., 1995. Asymptotic bias in simulated maximum likelihood estimation of discrete choice models. Econometric Theory 11, 437–483.
Lee, L.-F., 1999. Statistical inference with simulated likelihood functions. Econometric Theory 15, 337–360.
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.
Lee, S.-W., Hansen, B.E., 1994. Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. Econometric Theory 10, 29–52.
Lee, B.-S., Ingram, B.F., 1991. Simulation estimation of time-series models. Journal of Econometrics 47, 197–205.
Lee, D., Song, K., 2009. Simulated MLE for Discrete Choices using Transformed Simulated Frequencies, Manuscript, Department of Economics, University of Pennsylvania.
Lehmann, E., 1951. Consistency and unbiasedness of certain nonparametric tests. Annals of Matematical Statistics 22, 165–179.
Li, Q., Racine, J., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press, New Jersey.
Lo, A.W., 1988. Maximum likelihood estimation of generalized itô processes with discretely sampled data. Econometric Theory 4, 231–247.
Manrique, A., Shephard, N., 1998. Simulation-based likelihood inference for limited dependent processes. Econometrics Journal 1, 174–202.
Park, J.Y., Phillips, P.C.B., 2001. Nonlinear regressions with integrated time series. Econometrica 68, 1249–1280.
Park, J.Y., Phillips, P.C.B., 2000. Nonstationary binary choice models. Econometrica 69, 117–161.
Pedersen, A.R., 1995b. A new approach to maximum-likelihood estimation for stochastic differential equations based on discrete observations. Scandinavian Journal of Statistics 22, 55–71.
Pedersen, A.R., 1995a. Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. Bernoulli 1, 257–279.
Phillips, P.C.B., 1983. ERA's: a new approach to small sample theory. Econometrica 51, 1505–1525.
Rust, J., 1994. In: Engle, R.F., McFadden, D.L. (Eds.), Structural Estimation of Markov Decision Processes. In: Handbook of Econometrics, vol. 4. North Holland, Amsterdam, pp. 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., Koopman, S.J., 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. Wiley, New York.
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., 1997. New minimum chi-square methods in empirical finance. In: Kreps, D.M., Wallis, K.F. (Eds.), Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress. Cambridge University Press, Cambridge, pp. 279–317.
Wahba, G., 1981. Data-based optimal smoothing of orthogonal series density estimates. Annals of Statistics 9, 146–156.
White Jr., H.L., 1994. Estimation, Inference and Specification Analysis. Cambridge University Press, Cambridge.
White Jr., H.L., 1984. Maximum likelihood estimation of misspecified dynamic models. In: Dijkstra, T.K. (Ed.), Misspecification Analysis. Springer-Verlag, New York, pp. 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 141, 1245–1280.