Higher Order Improvements for Approximate Estimators*

Dennis Kristensen†
Columbia University and CREATES‡

Bernard Salanié§
Columbia University

February 25, 2009

Abstract

Many modern estimation methods in econometrics approximate an objective function, through simulation or discretization for instance. We here propose two methods to improve on the precision of such approximate estimators. Both of these methods only carry a small additional computation burden. The first method is targeted at estimators based on stochastic approximators, such as simulation-based estimators. It consists of a general formula that corrects the objective function and eliminates the leading order of the additional bias of the approximate estimator. Our second proposed improvement is a two-step method which applies quite generally. In the first step, we compute the approximate estimator, using an approximator that may be coarser than what is usually done; and in the second step we run one or several Newton-Raphson iterations based on the same objective function, but with a much finer degree of approximation. The second step removes some or all of the additional bias and variance of the initial approximate estimator.

VERY PRELIMINARY - DO NOT QUOTE WITHOUT PERMISSION

*We would like to thank participants at the CAM Research Workshop 2008, and seminars at UCL, Yale, and Northwestern for helpful comments and suggestions.

†E-mail: dk2313@columbia.edu

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

§Columbia University. E-mail: bs2237@columbia.edu.

1 Introduction

The complexity of econometric models has grown steadily over the past two decades. The increase in computer power contributed to this development in various ways, but in particular by allowing econometricians to estimate more complicated models using methods that rely on approximations. A leading example is simulation-based inference, where a function of the observables and the parameters is approximated using simulations. In this case, the function is an integral such as a moment, as in the simulated method of moments (McFadden (1989); Pakes and Pollard (1989)) and in simulated pseudo-maximum likelihood (Laroque and Salanié (1989, 1993, 1994)), or an integrated density, as in simulated maximum likelihood (Lee, 1992, 1995)).[1] Then the approximation technique often amounts to Monte Carlo integration. Other numerical integration techniques may be preferred for low-dimensional integrals, e.g. Gaussian quadrature, or both techniques can be mixed (see for example Lee (2001)). Within the class of simulation-based methods, some nonparametric alternatives rely on kernel sums instead of integration (e.g. Fermanian and Salanié (2004); Altissimo and Mele (2008); Creel and Kristensen (2008); Kristensen and Shin (2008)). Other estimation methods do not use simulations, but still involve numerical approximations, such as discretization of continuous processes, using a finite grid in the state space for dynamic programming models, and so on. Then the numerical approximation is essentially non-stochastic, unlike the case of simulation-based inference—this difference will play an important role in our paper.

In all of these cases, we call the "approximator" the numerical approximation that replaces the component of the objective function that we cannot evaluate exactly. Then the "exact estimator" is the infeasible estimator that reduces the approximation error to zero. E.g. in simulation-based inference, the exact estimator is based on an infinite number of simulations; in dynamic programming models it relies on an infinitely fine grid. We call "approximate estimator" the estimator that relies on a finite approximation. The use of approximations usually deteriorates the properties of the approximate estimator relative to those of the corresponding exact estimator: it is in general less efficient and may suffer from additional biases. These additional estimation errors can usually be controlled by choosing a sufficiently fine approximation, but this comes at the cost of increased computation time. In many applications this may be a seriously limiting factor; increased computer power helps, but it also motivates researchers to work on more complex models.

The contribution of this paper is twofold: First, we analyze the properties of the approximate estimator relative to the exact one in a very general setting that includes both M-estimators and GMM estimators. Our results encompass most known results in the literature on simulation-based estimators such as Lee (1995, 1999), Gouriéroux and Monfort

[1] Simulation-based inference is surveyed in Gouriéroux and Monfort (1996), van Dijk, Monfort and Brown (1995) and Mariano, Schuerman and Weeks (2001) among others.

(1996) and Laroque and Salanié (1989).

Second, we propose two methods to improve on the precision of approximate estimators. Both of these methods only carry a small additional computation burden. The first method is targeted at estimators based on stochastic approximators, such as simulation-based estimators. These approximators are usually unbiased (at least for a large number of simulations); but they have a variance that enters a nonlinear objective function. As a consequence, the variance component of the simulated approximation in general leads to an additional bias component relative to the exact estimator[2]. Our contribution to this is twofold. We first derive a general formula for the additional bias and variance of the approximate estimator. Then we show how to correct the objective function in a way that eliminates this additional bias term. Take for instance simulated maximum-likelihood on n observations, computed using S simulations. The resulting approximate estimator has a bias of order $1/S$, which dominates its efficiency loss in finite samples. Our corrected estimator only has a bias of order $1/S^{3/2}$, which can be a considerable improvement.

As we will show, our first method does not improve the properties of approximate estimators that rely on non-stochastic approximators. As noted above, our correction reduces the detrimental effect of the variance of the approximator on the approximate estimator. Therefore it works best when the approximator uses random draws to simulate an expectation, as then the bias of the approximator is zero. In contrast, if the approximator is non-stochastic then by definition it has zero variance, and our first method is of no help. Laffont et al. (1995) and Lee (1995) proposed a similar idea for SNLS estimators and SMLE of discrete choice models respectively. Our general method includes theirs as special cases.

On the other hand, our second proposed improvement is a two-step method which applies quite generally. In the first step, we compute the approximate estimator, using an approximator that may be coarser than what is usually done; and in the second step we run one or several Newton-Raphson iterations based on the same objective function, but with a much finer degree of approximation. The second step removes some or all of the additional bias and variance of the initial approximate estimator[3].

With simulation-based estimators or other stochastic approximation techniques, both approaches can be combined: the approximate objective function can be corrected so as to obtain an approximate estimator with a smaller bias, and this can be used in the first step of the Newton-Raphson method.

Our theoretical analysis is based on the insight that simulation-based estimators can be considered as a special case of a standard semiparametric estimation problem where the parameter of interest is computed using an infinite-dimensional nuisance parameter estimator,

[2] This is already well understood, and it is also known that the simulated method of moments is exempt from this additional bias as the objective function is linear in the approximator.

[3] Hajivassiliou (2000) considered a somewhat similar idea where an initial simulation-based estimator was adjusted so as to reduce its variance.

e.g. an expectation or a density. We use some of the tools that are applied in that setting; see for example Newey et al. (2004). Our analysis also shares some similarities with the recent literature on bias correction in the incidental parameters problem, see for example Newey and Hahn (2004) and Arellano and Hahn (2007) for results that are similar to ours in panel models with fixed effects.

Finally, our results are somewhat related to higher-order expansions of nonlinear fully parametric and semiparametric estimators as derived in, amongst others, Bao and Ullah (2007), Linton (1996) and Rilstone et al. (1996). However, in contrast to these papers, we carry out the expansion around the exact estimator, as opposed to doing it around the true parameter value. Thus, we only quantify biases and variances due to the approximation, and we set aside the sampling errors in the exact estimation problem.

Throughout, we take the criterion function that the estimator is based upon as given and we only discuss how the presence of additional biases and variances due to the approximation of some component in the criterion function can be dealt with. For results on higher order improvements through alternative specifications of the criterion function, we refer the reader to e.g. Newey and Smith (2004) and Newey et al. (2005).

We carry out a Monte Carlo study ... [MORE DETAILS LATER]

The paper is organised as follows: Section 2 presents our framework and the two methods we propose to improve the properties of approximate estimators. In Section 3, we derive a bias and variance expansion of the approximate estimator relative to the exact one. This expansion allows us to identify the leading terms; then in Section 4 we propose a bias adjustment that removes the leading bias term due to stochastic approximations. The properties of the Newton-Raphson method are derived in Section 5. Section 6 presents the results of a Monte Carlo study and section 7 concludes. Several proofs and lemmas have been relegated to Appendices.

2 Framework

At the most general level, our framework can be described as follows. Given a sample $\mathcal{Z}_n = \{z_1, ..., z_n\}$ of n observations, the econometrician proposes to estimate a parameter θ using some extremum estimator,

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\mathcal{Z}_n, \theta, \gamma(\cdot, \theta));$$

here γ is a function whose value may depend both on the data and on the parameters to be estimated. For notational simplicity, we suppress the dependence of γ on the data and we write

$$Q_n(\theta, \gamma) = Q_n(\mathcal{Z}_n, \theta, \gamma(\cdot, \theta));$$

note that Q_n has both a finite dimensional argument θ and a (usually) infinite-dimensional one (the function γ).

Our paper focuses on the common case when the true function γ_0 is not known to the econometrician, or its values have to be approximated numerically. In this case, a feasible estimator is obtained by minimizing the analog approximate criterion function

$$\hat{\theta}_{n,S} = \arg\min_{\theta\in\Theta} Q_n(\theta, \hat{\gamma}_S),$$

where $\hat{\gamma}_S$ depends on some approximation scheme of order S (e.g. S simulations, or a discretization on a grid of size S). We will refer to $\hat{\gamma}_S$ as an "approximator". We now present a few examples.

2.1 Examples of Approximate Estimators

Example 1: Simulated method of moments (SMM). The econometrician may just want to base estimation on a set of moment conditions

$$E\left[g(z, \theta_0)\right] = 0. \tag{1}$$

Given a weighting matrix W_n, the GMM estimator would minimize

$$Q_n(\theta) = G_n(\theta)' W_n G_n(\theta),$$

where

$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta).$$

Here, γ is simply the function g, which may be hard to evaluate, as in the multinomial probit example of McFadden (1989). If for instance the problematic component of g is itself an expectation, then it can easily be approximated as an average of simulated variables. In McFadden's example, g is the difference between choice dummy variables and their probabilities. Let $y_i = k$ if individual i choose the kth alternative conditional on observables x; then

$$g_k(z, \theta) = Z\left(x\right)\left[\mathbb{I}\left\{y = k\right\} - \Pr(y = k|x; \theta)\right],$$

where $\mathbb{I}\left\{\cdot\right\}$ denotes the indicator function, and $Z\left(x\right)$ are a set of instruments. The probability that an individual chooses k, $\gamma_k(z, \theta) = \Pr(y = k|x; \theta)$, can be approximated by drawing S choice errors for i and counting the proportion of times choice k brings highest utility, $\hat{\gamma}_{k,S}(z, \theta) = S^{-1}\sum_{s=1}^{S} \mathbb{I}\left\{y_s\left(x, \theta\right) = k\right\}$ where $y_s\left(x, \theta\right)$ are simulated choices (conditional on x).

In dynamic models, the above method is also applicable; but the simulations must be computed recursively from the time series model in question. Suppose for example that the observations come from a Markov model, $z_t = r(z_{t-1}, \varepsilon_t; \theta_0)$, and we wish to estimate θ_0 through the moment restriction $g(z_t, \theta) = w(z_t, z_{t-1}) - E_\theta[w(z_t, z_{t-1})]$ for some function w. Duffie and Singleton (1993) then propose to simulate a "long" trajectory from the model, $z_s(\theta) = r(z_{s-1}(\theta), \varepsilon_s; \theta)$, $s = 1, ...S$, and then approximate $E_\theta[w(z_t, z_{t-1})]$ by

$$\hat{\gamma}_S(\theta) = \frac{1}{S}\sum_{s=1}^{S} w(z_s(\theta), z_{s-1}(\theta)).$$

In certain situations, estimation based on conditional moment restrictions may be more attractive. These can in general still be estimated by simple sample averages in a cross-sectional setting, while this is normally not the case for dynamic latent variable models. Suppose for example that $z_t = (y_t, x_t)$, where only x_t has been observed, and we wish to compute $\gamma(x, \theta) = E_\theta[\phi(y_t) | x_{t-1} = x]$. Creel and Kristensen (2009) propose to approximate this conditional expectation by simulating a long string from the time series model as before and then using kernel regression techniques,

$$\hat{\gamma}_S(x; \theta) = \frac{1}{S}\sum_{s=1}^{S} \phi(y_s(\theta)) K_h(x - x_{s-1}(\theta)),$$

where $K_h(z) = K(z/h)/h^d$, $K : \mathbb{R}^d \mapsto \mathbb{R}$ is a kernel, $h > 0$ is a bandwidth, and $d = \dim(x_t)$. In contrast to the other approximators in this example, this approximator carries a bias due to the kernel smoothing.

Example 2: Parametric simulated M-estimators. Laroque and Salanié (1989) introduced a family of simulated pseudo-maximum likelihood (SPML) estimators. The simplest one is the simulated nonlinear least squares (SNLS) estimator. Suppose we want to estimate a nonlinear regression model,

$$y = g(x; \theta) + u,$$

where

$$g(x; \theta) = E[w(x, \varepsilon; \theta) | x],$$

for some function w. Defining $\gamma(x; \theta) = g(x; \theta)$, our exact criterion function takes the form

$$Q_n(\theta, \gamma_0) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \gamma(x_i; \theta))^2.$$

If the conditional expectation that defines γ cannot be evaluated analytically, we may use simulations instead. Draw i.i.d. random variables ε_s, $s = 1, ..., S$, and define $\hat{\gamma}_S(x; \theta) =$

$S^{-1}\sum_{s=1}^{S} w(x,\varepsilon_s;\theta)$. Then an SNLS estimator is obtained by minimizing

$$Q_n(\theta,\hat{\gamma}_S)=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{\gamma}_S(x_i;\theta)\right)^2.$$

It may be that in addition to the conditional mean, the econometrician wants to use the information in the conditional variance implied by the model. Now, $\gamma_0=(g,v)$ where $g(x;\theta)$ is the conditional mean and $v(x;\theta)$ is the conditional variance. Then we can define a pseudo-maximum likelihood estimator (PMLE) as the minimizer of:

$$Q_n(\theta,\gamma)=\frac{1}{n}\sum_{i=1}^{n}\left\{\log\left(v(x_i;\theta)\right)+\frac{\left[y_i-g(x_i;\theta)\right]^2}{v(x_i;\theta)}\right\}.$$

Again, in many situations $\gamma(x;\theta)$ cannot be written in closed form; but the conditional mean can be simulated as in the first part of this example, and obviously the conditional variance can be evaluated in the same way. The estimator based on the resulting approximate objective function is called an SPML estimator (of order 2).

Example 3: Simulated maximum likelihood. Simulated maximum-likelihood estimation (SML) is the second leading example of simulation-based M-estimation. As in Example 1, it comes in a parametric and a nonparametric variant.

The parametric version is well-known. Suppose we want to estimate a (conditional) distribution characterised by a parameter θ, $p(y|x;\theta)$. The natural choice is the maximum-likelihood estimator,

$$Q_n(\theta,\gamma)=-\frac{1}{n}\sum_{i=1}^{n}\log\left(\gamma(y_i,x_i;\theta)\right),$$

where $\gamma(y,x;\theta):=p(y|x;\theta)$. Sometimes the density γ cannot be written in analytical form. For example, in models with unobserved heterogeneity,

$$\gamma(z;\theta)=\int w(y|x,\varepsilon;\theta)\,f(\varepsilon)\,d\varepsilon,$$

for some densities w and f. In this example, we can draw $\varepsilon_{i,s}$, $s=1,...,S$, from the distribution of f and define $\hat{\gamma}_S(z;\theta)=S^{-1}\sum_{s=1}^{S} w(y|x,\varepsilon_s;\theta)$.

More recently, Fermanian and Salanié (2004) proposed using a kernel estimator as an approximator. The idea is simple, and it applies quite generally. Suppose data (y_i,x_i), $i=1,...,n$, has been generated by $y=r(x,\varepsilon;\theta_0)$, with implied conditional density $\gamma(y,x;\theta)=p(y|x,\theta_0)$. Then simulate the reduced form to generate samples $y_s(x,\theta)=r(x,\varepsilon_s;\theta)$ for $s=1,\ldots,S$, and approximate the density $f_{y|x}$ with a kernel density estimator based on the

y_s's:

$$\hat{\gamma}_S(y, x; \theta) = \frac{1}{S} \sum_{s=1}^{S} K_h(y - y_s(x, \theta)).$$

Maximizing the approximate likelihood in which $\hat{\gamma}_S$ replaces γ defines the nonparametric simulated maximum likelihood estimator (NPSML). It has somewhat different properties than other simulation based M-estimators, as the nonparametric approximator is biased for finite S. For a similar approach in time series models, see Altissimo and Mele (2008) and Kristensen and Shin (2008).

We now turn to two examples that involve non-stochastic approximation.

Example 4: Dynamic programming models. Dynamic programming models often have a multi-dimensional state space that forces analysts to resort to a finite grid and interpolation. Take a simple, stationary single-agent decision problem for instance:

$$V(s_t; \theta) = \max_{d_t} \{u(d_t, \theta) + \beta E[V(s_{t+1}; \theta)|s_t, d_t]\}.$$

Often the fixed point on the value function is computed by backwards induction, e.g. for use in maximum-likelihood estimation. This is infeasible in many cases, as the state space becomes too large. The fixed point of the value function may then be computed on a finite subset of S values of the state s_t by backwards induction. Let $(s_1, \ldots, s_S)$ be such a "grid". For $k = 1, \ldots, S$,

$$V_{S,t}^{(k)}(s_k, \theta) = \max_{d_t} \left\{ u(d_t, \theta) + \beta \hat{E}_S \left[V_{S,t+1}^{(k)}(s_{t+1}, \theta) | s_t = s_k, d_t \right] \right\}.$$

In this formula, the symbol $\hat{E}_S$ is meant to represent a numerical approximation of the conditional expectation of $V_{S,t+1}^{(k)}(s_{t+1}, \theta)$ based on its values at the points $(s_1, \ldots, s_S)$. Then the approximate estimator will match the policy function implied by the value function $V_{S,t}(\cdot, \theta)$ to the observed policy function. See Norets (2006) for an example of a specific approximation method for discrete choice models, and Fernández-Villaverde, Rubio-Ramirez and Santos (2006) for consequences of approximating policy functions in the estimation of DSGE models.

Example 5: Linearized models. Many models used in macroeconomics, for instance, have a very complex likelihood function, so that a limited information estimation method is used. But a large subclass cannot even be solved in a closed form. Then estimation is based on an approximate model, often by linearizing equations close to a steady state. For our purposes, this is quite similar to example 4 above: in both cases, the true model is replaced with one that is easier to work with. The quality of the approximation can be improved at a

larger computational cost by using a finer grid in example 4, or in example 5 by using more iterations of perturbations or projection methods for instance as advocated by Judd, Kubler and Schmedders (2003). Note one additional difficulty: approximation errors get magnified as the horizon is more remote. A recent paper by Fernández-Villaverde and Rubio-Ramirez (2005) shows that this may create severe difficulties for the approximate estimators.

2.2 A Summary of our Proposed Improvements

In all of the examples above, using approximation reduces the quality of the estimator. Start with our first three examples, which minimize objective functions where a mathematical expectation is replaced by a function of simulated draws. Of course the mean is an unbiased estimator of the expectation, but the objective function depends nonlinearly on the simulated mean in many simulation-based estimation methods, so that the approximate estimator based on S simulations has an additional bias, along with a loss of efficiency. In many cases both are of order $1/S$; this holds for example when simulated moments are employed. The efficiency loss may not be a concern in large samples; but the additional bias persists asymptotically.

On the other hand, the simulated method of moments (example 1) has nicer properties when the moment condition is linear in the simulated mean. Then the sampling errors from the simulations are averaged over observations, and the additional bias vanishes in large samples (the asymptotic efficiency loss is still of order $1/S$).

Similarly, non-stochastic approximations lead to deteriorations of the properties of the resulting estimators. Take simulated maximum likelihood for instance. If the dimensionality of the simulated mean is small, then integration may be done by an S point Gaussian quadrature. As demonstrated in the next section, the resulting approximate estimator will suffer from additional biases.

Thus in general the approximate estimator $\hat{\theta}_{n,S}$ can only be consistent if S goes to infinity as n goes to infinity; and $\sqrt{n}$-consistency requires that S go to infinity faster than n, in which case the asymptotic variance is the same as that of the exact estimator. In other words (Section 3 will give more precise statements and regularity conditions),

$$||\hat{\theta}_{n,S} - \hat{\theta}_n|| = o_P\left(1/\sqrt{n}\right)$$

as $n \to \infty$ for some sequence $S = S(n) \to \infty$, and there is no first order difference between the exact and the approximate estimator. However, to obtain this rate we need to choose S "large", which may be computationally costly. We here propose methods which yield estimators that are just as efficient as large-S approximate estimators, but are computationally much less burdensome.

We take as starting point the estimator

$$\hat{\theta}_{n,S} = \arg\min_{\theta\in\Theta} Q_n(\theta, \hat{\gamma}_S), \tag{2}$$

where S is "small" in the (admittedly loose) sense that the econometrician would dearly like to have enough computational power to increase S. In general the properties of $\hat{\theta}_{n,S}$ may not be very good. Our first method to improve its properties consists in changing the objective function: Instead of selecting $\hat{\theta}_{n,S}$ to minimize $Q_n(\theta, \hat{\gamma}_{n,S})$, we select

$$\tilde{\theta}_{n,S} = \arg\min_{\theta\in\Theta} \left\{ Q_n(\theta, \hat{\gamma}_S) - \Delta_{n,S}(\theta) \right\}, \tag{3}$$

where $\Delta_{n,S}(\theta)$ is a bias adjustment term. In section 4, we derive an expression of $\Delta_{n,S}(\theta)$ in the general case, and we show that it yields simple formulæ for examples 1 to 3.

This first approach was proposed in Laffont et al. (1995) in the context of SNLS (also see Laroque and Salanié (1989, 1993); Bierings and Sneek (1989).) They derived an unbiased and consistent estimator of the bias component due to simulations. SNLS is a quite special and favorable case, as the criterion function is only quadratic in the simulated mean; in general using $\Delta_{n,S}(\theta)$ will only correct for the leading term of the bias when using stochastic approximation. For instance, Lee (1995) derived a bias adjustment for SML that leaves higher-order bias terms. Moreover and more importantly, this first technique only works if the bias in the approximator itself is zero[4]. But almost by definition, non stochastic approximations are biased.

Our second proposed method works with non-stochastic approximations as well as with stochastic approximations; it extends the well-known idea that a consistent estimator can be made asymptotically efficient by applying one Newton-Raphson (NR) step of the log-likelihood function to it. E.g. if $\hat{\theta}_n$ is a consistent estimator of θ_0 in a model with log-likelihood $L_n(\theta)$, then $\hat{\theta}_n^{NR} = \hat{\theta}_n - \left[\partial^2 L(\hat{\theta}_n)/\partial\theta^2\right]^{-1} \partial L_n(\hat{\theta}_n)/\partial\theta$ is consistent and asymptotically efficient.

We apply this to our setting by starting from either the approximate estimator $\hat{\theta}_{n,S}$ obtained in (2), or the bias-corrected version $\tilde{\theta}_{n,S}$ of (3). We already know that both are consistent when both S and n go to infinity, and that when stochastic approximations are used, the finite-S bias of $\tilde{\theta}_{n,S}$ is smaller than that of $\hat{\theta}_{n,S}$. For notational simplicity, denote either of these two starting points as $\bar{\theta}_{n,S}$. We then define the corrected estimator through one or possibly several Newton-Raphson iterations of an approximate objective function that

[4] Or at least vanishes asymptotically at an appropriate rate.

uses a much finer approximation, $S^* \gg S$:

$$\hat{\theta}_{n,S}^{(k+1)} = \hat{\theta}_{n,S}^{(k)} - \left[\frac{\partial^2 Q_n(\hat{\theta}_{n,S}^{(k)}, \hat{\gamma}_{S^*})}{\partial\theta\partial\theta'}\right]^{-1} \frac{\partial Q_n(\hat{\theta}_{n,S}^{(k)}, \hat{\gamma}_{S^*})}{\partial\theta}, \quad k = 1, 2, 3, ... \tag{4}$$

where $\hat{\theta}_{n,S}^{(1)} = \bar{\theta}_{n,S}$ and we use the S^*th order approximator, $\hat{\gamma}_{S^*}$, in the iterations.

Note that the cost of computing this new estimator from the first one is (very) roughly S^*/S times the cost of one iteration in the maximization of $Q_n(\theta, \hat{\gamma}_{S^*})$. Since the maximization easily can require a hundred iterations or so, we can therefore take S^* ten or twenty times larger than S without significantly adding to the cost of the estimation procedure.[5] Also, one iteration is enough if S^* goes to infinity at least as fast as S. We discuss this method in more detail in Section 5.

3 Properties of Approximate Estimators

In order to be able to propose bias adjustments, we first derive an asymptotic expansion of the bias and variance of the unadjusted approximate estimator relative to the infeasible, exact estimator. This will also enable us to evaluate the improvements in terms of bias and variance that result from these adjustments. In order to compare the approximate estimator with the exact estimator, we first establish a very general lemma that may be useful in other contexts. It states that the stochastic difference between any two estimators is determined by the uniform stochastic difference between their respective criterion functions.

Lemma 1 *(i) Let* $\hat{\theta} = \arg\min_{\theta\in\Theta} L_n(\theta)$ *satisfy* $\hat{\theta} \to^P \theta_0$*, where* $\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)| \to^P 0$ *and* $L(\theta)$ *is twice differentiable with* $\partial L(\theta)/\partial\theta|_{\theta=\theta_0} = 0$ *and* $H(\theta) = \partial^2 L(\theta) / (\partial\theta\partial\theta')$ *satisfying* $H(\theta_0) > 0$*.*

For any alternative estimator $\tilde{\theta} = \arg\min_{\theta\in\Theta} \hat{L}_n(\theta)$*,*

$$||\hat{\theta} - \tilde{\theta}|| = O_P\left(\sup_{\theta\in\Theta} |L_n(\theta) - \hat{L}_n(\theta)|\right).$$

(ii) Let $\hat{\theta}$ *solving* $T_n(\hat{\theta}) = 0$ *be a consistent estimator, where* $\sup_{\theta\in\Theta} ||T_n(\theta) - T(\theta)|| = o_P(1)$ *and* $T(\theta)$ *is differentiable such that* $H(\theta) = \partial T(\theta)/\partial\theta$ *satisfies* $H(\theta_0) > 0$ *and* $T(\theta_0) = 0$*.*

[5] In many cases, a large part of the dimensionality of θ enters within some linear indexes $x\theta$; then the trade off would be even more favourable—each iteration in the maximization has to do some line search and therefore compute several values of $Q_n(\theta, \hat{\gamma}_S)$.

For any alternative estimator $\tilde{\theta}$ *solving* $\hat{T}_n(\tilde{\theta}) = 0$,

$$||\hat{\theta} - \tilde{\theta}|| = O_P\left(\sup_{\theta\in\Theta} ||T_n(\theta) - \hat{T}_n(\theta)||\right).$$

The first part of the lemma seems to be new, while a different version of the second part (requiring that $T_n(\theta)$ itself be differentiable, and not just its limit) appeared in Robinson (1988, Theorem 1). We will rely on part (i) of Lemma 1 to establish the stochastic difference between the actual and approximate extremum estimator by choosing $L_n(\theta) = Q_n(\theta, \gamma)$ and $\hat{L}_n(\theta) = Q_n(\theta, \hat{\gamma}_S)$. The second result is useful when estimators defined by a first-order condition are considered. Note that neither result imposes any smoothness condition on the finite-sample criterion function—only on its asymptotic limit. This is useful when working with approximate estimators, in particular ones based on simulations, since these in some cases lead to unsmooth criterion function, see e.g. Pakes and Pollard (1989).

We will restrict our attention to the class of MINPIN estimators as proposed in Andrews (1994). That is, we restrict $Q_n(\theta, \gamma)$ to be on the form

$$Q_n(\theta, \gamma) = d(\bar{m}_n(\theta, \gamma), \hat{\nu}),$$

where $\hat{\nu} \in \mathcal{W}$ is a preliminary estimator of nuisance parameters, $\bar{m}_n(\theta, \gamma)$ is a sample-average,

$$\bar{m}_n(\theta, \gamma) = \frac{1}{n}\sum_{i=1}^{n} m(z_i; \theta, \gamma) \in \mathbb{R}^M,$$

and the function $d : \mathbb{R}^M \times \mathcal{W} \mapsto \mathbb{R}$ is a pseudo-distance function. The MINPIN estimator includes most standard estimators, including M-estimators ($d(m, \nu) = m$) and GMM-estimators ($d(m, \nu) = m'\nu m$). We here restrict $\hat{\nu}$ to not depend on γ; this can in principle be accommodated for, but the results and proofs get more lengthy and cumbersome, so for simplicity we rule out this case.

Suppressing the dependence on γ, let $M(\theta)$ denote the moment function

$$M(\theta) := E[m(z; \theta, \gamma)],$$

which is assumed to be well-defined (see A.2 below). Also, let $G(\theta)$ and $H(\theta)$ denote the first and second order derivatives of the limit, $Q(\theta) = d(M(\theta), \nu)$, of $Q_n(\theta)$:

$$G(\theta) := \frac{\partial Q(\theta, \gamma)}{\partial \theta'} = D_1 \frac{\partial M(\theta, \gamma)}{\partial \theta'},$$

and

$$H(\theta) := \frac{\partial^2 Q(\theta, \gamma)}{\partial \theta \partial \theta'} = \frac{\partial M(\theta, \gamma)'}{\partial \theta} D_2 \frac{\partial M(\theta, \gamma)}{\partial \theta'} + D_1 \frac{\partial^2 M(\theta, \gamma)}{\partial \theta \partial \theta'},$$

where

$$D_1(\theta) = \frac{\partial d(M(\theta,\gamma),\nu)}{\partial m'}, \qquad D_2(\theta) = \frac{\partial d(M(\theta,\gamma),\nu)}{\partial m \partial m'}.$$

We then first impose conditions to ensure that the exact, but infeasible estimator satisfies the conditions of Lemma 1(i):

A.1 $\{z_i\}$ is stationary and α-mixing with the mixing coefficients satisfying $\sum_{i=1}^{\infty} \alpha_i^{\delta/(2+\delta)} < \infty$.

A.2 The parameter space Θ is compact and $E\left[\sup_{\theta\in\Theta} \|m(z_i;\theta,\gamma)\|\right] < \infty$. Furthermore, $M(\theta)$ is twice continuously differentiable in θ such that $G(\theta_0) = 0$ and $H(\theta_0) > 0$.

A.3 $\hat{\nu} \to^P \nu$.

Assumptions (A.1)-(A.2) imply the conditions of Lemma 1(i). Assumption A.1 rules out strongly persistent data, and is made to obtain standard rates of convergence of the resulting estimators. For results on simulation-baseed estimators with non-stationary data, we refer to Kristensen and Shin (2008). Primitive conditions for the uniform moment condition in A.2 to hold in a cross-sectional setting can be found in Newey and McFadden (1994).

We will consider two types of approximate estimators. For the class of MINPIN estimators, one can implement the approximate versions of them in two different ways: Either one single approximation is used across all observations, or a new approximation is used for each observation. In the first case, the approximate sample moment takes the form

$$\bar{m}_n(\theta,\hat{\gamma}_S) = \frac{1}{n}\sum_{i=1}^{n} m(z_i;\theta,\hat{\gamma}_S), \tag{5}$$

while in the second case,

$$\bar{m}_n(\theta,\hat{\gamma}_S) = \frac{1}{n}\sum_{i=1}^{n} m\left(z_i;\theta,\hat{\gamma}_{i,S}\right). \tag{6}$$

In the first case, one and the same approximation is used across all data points. In simulation-based estimation, this scheme was proposed by Lee (1992) for cross-sectional discrete choice models, and for Markov models in Kristensen and Shin (2008). The scheme has also been used in stationary time series models where one long trajectory of the model is simulated and used to compute simulated moments (see Example 1) or densities (see Fermanian and Salanié, 2004; Altissimo and Mele, 2008). In the second case, we have n approximators - one for each observation. Thus, the dimension of $\hat{\gamma}_S(x;\theta) = \left(\hat{\gamma}_{1,S}(x;\theta),...,\hat{\gamma}_{n,S}(x;\theta)\right)$ increases with sample size. For simulation-based estimators, this approach was taken in, amongst others, Laroque and Salanié (1989), McFadden (1989), and Fermanian and Salanié (2004), where the n approximations were chosen to be mutually independent.

We here note that the case where the dimension of $\hat{\gamma}_S$ increases with sample size is an incidental parameters problem. Some of our results for this situation are very similar to those found in the litterature on higher-order properties and bias-correction of estimators in an incidental parameters setting, see e.g. Arellano and Hahn (2007) and Hahn and Newey (2004). On the other hand, when the dimension remains fixed, the resulting approximate estimator is similar to semiparametric two-step estimators where in the first step a function is nonparametrically estimated, see e.g. Newey and McFadden (1994).

In the following we will use $\|\cdot\|$ to denote both the Euclidean norm and the function norm of γ; this should hopefully not lead to any confusion. In most cases, the norm will be the L_q-norm induced by the probability measure associated with our observations, $\|\gamma\| = E\left[\|\gamma(x)\|^q\right]^{1/q}$ for some $q \geq 1$. We then assume that the criterion functions are smooth functionals of γ:

A.4 There exists functions $\nabla m(z;\theta)[d\gamma]$ and $\nabla^2 m(z;\theta)[d\gamma, d\gamma]$ which are linear and quadratic respectively functionals of $d\gamma$ such that

$$E\left[\left|m(z;\theta,\gamma) - m\left(z;\theta,\gamma'\right) - \nabla m(z;\theta)[d\gamma] - \frac{1}{2}\nabla^2 m(z;\theta)[d\gamma, d\gamma]\right|\right] \leq \bar{M}_0 \|d\gamma\|^3, \tag{7}$$

where $d\gamma = \gamma - \gamma'$, and

$$E\left[\left|\nabla m(z;\theta)[d\gamma]\right|^2\right]^{1/2} \leq \bar{M}_1 \|d\gamma\|, \quad E\left[\left|\nabla^2 m(z;\theta)[d\gamma, d\gamma]\right|^{2+\delta}\right]^{1/(2+\delta)} \leq \bar{M}_2 \|d\gamma\|^2, \tag{8}$$

Note that assumption A.3 does not impose differentiability of $m(z;\theta,\gamma)$ as a function of θ, only that it is smooth as a function of γ. In particular, it applies to standard simulation-based estimators of limited-dependent variable and discrete choice models such as the moment estimators proposed in McFadden (1989) and Pakes and Pollard (1989): There, the simulated versions of the criterion functions are not differentiable w.r.t. θ, but they are smooth functionals of the probabilities being approximated by simulations.

Finally, we impose conditions on the approximators. To give conditions for both of the approximation schemes discussed above, we state our assumptions for J independent approximators: In the case of (5) $J = 1$, while in the case of (6) $J = n$.

A.5 Uniformly over θ, the approximator has the following properties:

(i) $\hat{\gamma}_{1,S}(x;\theta), ..., \hat{\gamma}_{J,S}(x;\theta)$ are mutually independent and are all independent of $\mathcal{Z}_n$.

(ii) The bias,

$$b_S(x;\theta) := E[\hat{\gamma}_{j,S}(x;\theta)|x] - \gamma(x;\theta),$$

is of order $\beta > 0$:

$$\|b_S(\cdot;\theta)\| = S^{-\beta}B(\theta) + o(S^{-\beta}).$$

(iii) The variance,

$$\psi_{j,S}(x;\theta) := \hat{\gamma}_{j,S}(x;\theta) - E\left[\hat{\gamma}_{j,S}(x;\theta)\,|x\right],$$

is of order $\alpha > 0$:

$$E\left[\left\|\psi_{j,S}(\cdot;\theta)\right\|^2\right] = S^{-\alpha}v(\theta) + o(S^{-\alpha}).$$

(iv) The third (centered) moment is of order $\eta > 0$:

$$E\left[\left\|\psi_{j,S}(\cdot;\theta)\right\|^3\right] = O(S^{-\eta}).$$

Assumption A.5 is sufficiently general to cover all of the examples in Section 2. When stochastic approximators are used, note that the independence assumption in A.5.i will be satisfied by drawing J independent batches of size S, $\{\varepsilon_{i,1},....,\varepsilon_{i,S}\}$, $i = 1,..,n$, and then use one batch per approximation. This does not rule out that the simulated values within each batch are dependent, as will for example be the case when drawing recursively from a time series models. For most simulation-based estimators in a dynamic setting, only $J = 1$ approximator is used for all observations, as is for example the case in Duffie and Singleton (1994), Creel and Kristensen (2009) and Kristensen and Shin (2008) and so A.4.i is automatically satisfied in these cases. We also note that $\hat{\gamma}_S(x;\theta)$ potentially can be a vector-valued function containing several different approximators. Thus, (A.5.i) only has bite when J goes to infinity. However, one situation where $J \to \infty$ and the independence assumption is violated is in the case of sequential approximation schemes used in dynamic latent variable models such as particle filters, see e.g. Brownlees, Kristensen and Shin (2009) and Olsson and Rydén (2008). In this case, we have a sequence of approximator where the approximator of the conditional density of the current observation depends on the one used for the previous observation, thereby not satisfying A.5.i.

In the case of stochastic approximators, they normally take the form of sample averages,

$$\hat{\gamma}_S(x;\theta) = \frac{1}{S}\sum_{s=1}^{S} w_S(x,\varepsilon_s;\theta). \tag{9}$$

For stochastic approximators on the form (9), the bias and variance directly result from those of the simulators w_S. Here,

$$\begin{aligned} \psi_S(x;\theta) &\equiv \frac{1}{S}\sum_{s=1}^{S}\left\{w_S(x,\varepsilon_s;\theta) - E\left[w_S(x,\varepsilon;\theta)\,|x\right]\right\}, \\ b_S(x;\theta) &\equiv E\left[w_S(x,\varepsilon;\theta)\,|x\right] - \gamma(x;\theta), \end{aligned} \tag{10}$$

Suppose that we work with the L_2-norm. If the simulated variables, $\{\varepsilon_{i,s}\}$, are i.i.d., then assumption A.5 holds if, for example

$$\frac{1}{S}E\left[||w_S(x,\varepsilon;\theta) - E\left[w_S(x,\varepsilon;\theta)\,|x\right]||^2\right] = O\left(S^{-\alpha}\right),$$

and

$$E\left[\left\|E\left[w_S(x,\varepsilon;\theta)\,|x\right] - \gamma(x;\theta)\right\|^2\right]^{1/2} = O(S^{-\beta}),$$

for some $\alpha, \beta > 0$.

In most cases, $w_S(x,\varepsilon;\theta) = w(x,\varepsilon;\theta)$ is independent of the number of simulations. The approximator will in this case have no bias ($b_S \equiv 0$) and so $\beta = \infty$, while, with i.i.d. simulations, its variance is constant so that $\alpha = 1$. In the more general case where the simulated variables are dependent (such as the case of the SMM of Duffie and Singleton (1994)), A.5 still holds with $\beta = \infty$ and $\alpha = 1$ assuming stationary and geometrically mixing and stronger moment conditions, c.f. Lemma 7. Furthermore, estimators based on sample averages will in general have $\eta \simeq 3/2$, c.f. Lemma 7.

Approximators on the form (9) also include simulation-based estimators that rely on kernel sums to approximate a density or a conditional mean, as in the NPSML method of Fermanian and Salanié (2004) and the NPSMM of Creel and Kristensen (2009). As an example, consider the NPSMLE: In this case, $w_S(y,x,\varepsilon_s;\theta) = K_h(y_s(x,\theta) - y)$ where the bandwidth $h = h(S) \to 0$ as $S \to \infty$. Let $d = \dim(y)$ and suppose that we use a kernel of order r. Then with a bandwidth of order $h \propto S^{-\delta}$ for some $\delta > 0$, the bias satisfies

$$b_S(y,x;\theta) = \int K_h(w-y)\,f(w|x;\theta)\,dw - f(y|x;\theta) = h^r\frac{\partial^r f(y|x;\theta)}{\partial y^r} + o(h^r) = O\left(S^{-r\delta}\right).$$

For both i.i.d. and stationary and mixing draws, the variance is of order $\log(S)/\left(Sh^d\right) \simeq S^{-(1-d\delta)}$ in the sup-norm, c.f. Kristensen (2009).[6] Moreover, it is easily seen that $\eta = \log(S)/\left(Sh^d\right)^{3/2}$. Thus, A.5 holds with $\beta = r\delta$, $\alpha \simeq 1 - d\delta$ and $\eta \simeq 3(1-d\delta)/2$.

As is well-known, the asymptotic mean integrated squared error is smallest when the bias and variance component are balanced. This occurs when $\delta = 1/(2r+d)$, leading to

[6] We have here left out a $\log(S)$ term in the order of the variance to simplify formulae, and assumed that the density is bounded away from zero.

$\beta = \alpha/2 = r/(2r+d)$. We recover of course the standard nonparametric rate of $S^{-2r/(2r+d)}$ for AMISE; for example in the textbook case $r = 2$ and $d = 1$, so that AMISE $= O\left(S^{-4/5}\right)$.

We should stress at this point that while the standard nonparametric rate is optimal for the approximation of the elements of the likelihood, this does not imply in any way that this rate yields the best NPSML estimators. In fact, we will see later that the bandwidth derived above oversmooths; the optimal rate for NPSML estimation turns out to be $1/(r+d)$.

Now consider an approximation that does not involve any randomness, as with numerical integration, discretization, or numerical solution of differential equations. Then by construction the conditional variance of the approximator is zero, so that $\alpha = +\infty$, but approximation imparts a bias, which in leading cases obeys assumption A.5 for some $\beta > 0$. We will see later that the bias adjustment technique based on correcting the objective function has no bite in this situation. On the other hand, the proposed Newton-Raphson procedure works for both stochastic and non-stochastic approximations.

Under (A.2) and (A.4), the first and second order differentials of $Q_n(\theta)$, denoted $\bigtriangledown Q_n(\theta)\,[d\gamma]$ and $\bigtriangledown^2 Q_n(\theta)\,[d\gamma, d\gamma']$, are well-defined (see Appendix A for their precise expressions). In particular, the difference between the approximate and exact criterion function can be written as

$$Q_n(\theta, \hat{\gamma}_S) - Q_n(\theta, \gamma) = \bigtriangledown Q_n(\theta)[\hat{\gamma}_S - \gamma] + \frac{1}{2} \bigtriangledown^2 Q_n(\theta)[\hat{\gamma}_S - \gamma, \hat{\gamma}_S - \gamma] + R_{n,S}, \tag{11}$$

where $R_{n,S} = O_P(\|\hat{\gamma}_S - \gamma\|^3)$. Each of the functional derivatives is a polynomial functional of $\hat{\gamma}_S - \gamma_0$ and can therefore be easily decomposed into bias and variance components:

$$\bigtriangledown Q_n(\theta)[\hat{\gamma}_S - \gamma] = \bigtriangledown Q_n(\theta)[\psi_S] + \bigtriangledown Q_n(\theta)[b_S], \tag{12}$$

$$\begin{aligned} \bigtriangledown^2 Q_n(\theta)[\hat{\gamma}_S - \gamma, \hat{\gamma}_S - \gamma] \;=\; & \bigtriangledown^2 Q_n(\theta)[\psi_S, \psi_S] + \bigtriangledown^2 Q_n(\theta)[b_S, b_S] \\ & + 2 \bigtriangledown^2 Q_n(\theta)[\psi_S, b_S], \end{aligned} \tag{13}$$

where ψ_S and b_S are the variance and bias components of $\hat{\gamma}_S$ as defined in A.5. In the Appendix, we then combine the Lipschitz conditions imposed on m together with the assumption A.5 to derive the order at which each of the terms on the right-hand sides of Eq. (12)-(13) vanishes. The following theorem states the rate with which the approximate criterion function converges towards the exact one. This translates into the rate of the approximate estimator by appealing to Lemma 1:

Theorem 2 *Assume that A.1-A.5 hold. Then:*

1. *The approximate criterion function based on eq. (5) satisfies:*

$$\bigtriangledown Q_n(\theta)[b_S] = O_P\left(S^{-\beta}\right) + O_P\left(n^{-1/2}S^{-\beta}\right),$$

$$\bigtriangledown Q_n(\theta)[\psi_S] = \bigtriangledown m_1\left(\psi_S;\theta\right) + O_P\left(n^{-1/2}S^{-\alpha/2}\right),$$

$$\bigtriangledown^2 Q_n(\theta)[b_S, b_S] = O_P\left(S^{-2\beta}\right) + O_P\left(n^{-1/2}S^{-2\beta}\right)$$

$$\bigtriangledown^2 Q_n(\theta)[\psi_S, \psi_S] = O_P\left(S^{-\alpha}\right) + O_P\left(n^{-1/2}S^{-\alpha}\right)$$

where

$$\bigtriangledown m_1\left(\psi_S;\theta\right) = \int \bigtriangledown m\left(z;\theta\right)[\psi_S]\, dF\left(z\right).$$

As a consequence,

$$\begin{aligned} ||\hat{\theta}_{n,S} - \hat{\theta}_n|| &= \bigtriangledown m_1\left(\psi_S;\theta\right) + O_P\left(S^{-\alpha}\right) + O_P\left(S^{-\beta}\right) + O_P\left(S^{-\eta}\right) \\ &\quad + O_P\left(n^{-1/2}S^{-\alpha/2}\right) + O_P\left(n^{-1/2}S^{-\beta}\right). \end{aligned}$$

2. *The approximate criterion function based on eq. (6) satisfies:*

$$\bigtriangledown Q_n(\theta)[b_S] = O_P\left(S^{-\beta}\right) + O_P\left(n^{-1/2}S^{-\beta}\right),$$

$$\bigtriangledown Q_n(\theta)[\psi_S] = O_P\left(n^{-1/2}S^{-\alpha/2}\right)$$

$$\bigtriangledown^2 Q_n(\theta)[b_S, b_S] = O_P\left(S^{-2\beta}\right) + O_P\left(n^{-1/2}S^{-2\beta}\right),$$

$$\bigtriangledown^2 Q_n(\theta)[\psi_S, \psi_S] = O_P\left(S^{-\alpha}\right) + O_P\left(n^{-1/2}S^{-\alpha}\right).$$

As a consequence,

$$\begin{aligned} ||\hat{\theta}_{n,S} - \hat{\theta}_n|| &= O_P\left(S^{-\alpha}\right) + O_P\left(S^{-\beta}\right) + O_P\left(S^{-\eta}\right) \\ &\quad + O_P\left(n^{-1/2}S^{-\alpha/2}\right) + O_P\left(n^{-1/2}S^{-\beta}\right). \end{aligned} \tag{14}$$

In the first part of the theorem, the term $\bigtriangledown m_1\left(\psi_S;\theta\right)$ appears. It is easily seen that due to the Lipschitz condition imposed on $\bigtriangledown m\left(z;\theta\right)[\cdot]$, we have

$$\left|\bigtriangledown m_1\left(\psi_S;\theta\right)\right| = O_P\left(S^{-\alpha/2}\right).$$

However, in some cases this bound is not sharp. In particular, when $\hat{\gamma}_S$ is a kernel estimator $\left|\bigtriangledown m_1\left(\psi_S;\theta\right)\right| = O_P\left(S^{-1/2}\right)$ which is faster than $O_P\left(S^{-\alpha/2}\right) = O_P(\left(Sh^d\right)^{-1/2})$, c.f. Example

3.2 below.

We have seen that for a large class of simulation-based estimators, the bias and variance of the approximator is of order $\alpha = 1$, $\beta = \infty$ and $\eta = 3/2$, c.f. Assumption A.5 and the following discussion. With weakly dependent data, the above corrollary states that the leading term of $||\hat{\theta}_{n,S} - \hat{\theta}_n||$ is $O_P(1/S)$ which is due to the conditional variance of each simulator. This is a well-known result for specific simulation-based estimators, see e.g. Laffont et al. (1993) and Lee (1995). We demonstrate here that this result holds more generally under weak regularity conditions.

In some cases, the rate of the approximate estimator obtained in Theorem 2 is not sharp. For example, if $q(z;\theta,\gamma)$ is three times pathwise differentiable and the approximator is a (simulated) sample average, the remainder term $R_{n,S}$ will vanish at the faster rate of $R_{n,S} = O_P\left(S^{-2}\right) + O_P\left(n^{-1/2}S^{-3/2}\right)$. This follows as a straightforward application of Lee (1995, Propositions A.3-A.4). It appears difficult to obtain precise rates at our level of generality though.

To illustrate the use of our results, we return to Examples 2-3 of Section 2. In both examples, $d(m,\nu) = m$. We first consider the case where unbiased simulators are employed:

Example 2.1 (SNLS). In this example,

$$Q_n(\theta,\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - g(x_i;\theta)\right)^2,$$

and $w_S(x,\varepsilon;\theta) = w(x,\varepsilon;\theta)$ satisfies $E[w(x,\varepsilon;\theta)] = g(x;\theta) = E_\theta[y|x]$. With $\xi_i(\theta) := y_i - g(x_i;\theta)$, the functional differentials are

$$\bigtriangledown m(z_i;\theta)[dg] = -2\xi_i(\theta)\,dg(x_i;\theta), \qquad \bigtriangledown^2 m(z_i;\theta)[dg,dg] = 2dg(x_i;\theta)\,dg(x_i;\theta),$$

such that (using a single approximation for all observations)

$$\bigtriangledown Q_n(\theta,\gamma)[dg] = -\frac{2}{n}\sum_{i=1}^{n}\xi_i(\theta)\,dg(x_i;\theta), \qquad \bigtriangledown^2 Q_n(\theta,\gamma)[dg,dg] = \frac{2}{n}\sum_{i=1}^{n}dg(x_i;\theta)\,dg(x_i;\theta).$$

Since Q_n is quadratic in g, (7) holds with $\bar{M}_0 = 0$ such that the remainder term R_n in eq. (11) is zero.

Assuming that $E\left[\sup_{\theta\in\Theta}\xi_i^2(\theta)\right] < \infty$, it is easily seen that Eq. (8) also holds when using the L_2 norm, $\|m\| = \sqrt{E[m^2(x)]}$.

Also note that functional differentials are well-defined without $\hat{\gamma}_S(x;\theta)$ being differentiable in θ; therefore it also applies to limited dependent variable models. Take for instance the binomial choice model, with $y = \mathbb{I}\{y^* > 0\}$,and $y^* = r(x,\varepsilon;\theta)$. Then $m(x;\theta) =$

$E[y|x] = P(y=1|x)$ which can be simulated using $w(x_i, \varepsilon_{is}; \theta) = \mathbb{I}\{y_{is}^*(\theta) > 0\}$, where $y_{is}^*(\theta) = r(x_i, \varepsilon_{is}; \theta)$.

Example 2.2 (SPML). Here, with $\xi_i(\theta) = y_i - g(x_i; \theta)$,

$$m(z; \theta, \gamma) = \log(v(x_i; \theta)) + \frac{\xi_i^2(\theta)}{v(x_i; \theta)}.$$

Thus, with dg and dv denoting mean and variance directions,

$$\begin{aligned} \bigtriangledown m(z;\theta)[d\gamma] &= \bigtriangledown_g m(\theta,\gamma)[dg] + \bigtriangledown_v q(z;\theta)[dv] \\ \bigtriangledown^2 m(z;\theta)[d\gamma, d\gamma] &= \bigtriangledown^2_{g,g} m(z;\theta)[dg, dg] + 2\bigtriangledown^2_{g,v} m(z;\theta)[dg, dv] + \bigtriangledown^2_{v,v} m(z;\theta)[dv, dv], \end{aligned}$$

where,

$$\bigtriangledown_g m(z_i;\theta)[dm] = -2\left\{\frac{\xi_i(\theta)}{v(x_i;\theta)} dg(x_i;\theta)\right\}, \quad \bigtriangledown_v m(z;\theta)[dv] = \left\{1 - \frac{\xi_i^2(\theta)}{v(x_i;\theta)}\right\}\frac{dv(x_i;\theta)}{v(x_i;\theta)},$$

$$\begin{aligned} \bigtriangledown^2_{g,g} m(z_i;\theta)[dg, dg] &= -2\frac{dg(x_i;\theta)^2}{v(x_i;\theta)}, \quad \bigtriangledown^2_{g,v} m(z_i;\theta)[dg, dv] = 2\frac{\xi_i(\theta)}{v^2(x_i;\theta)} dg(x_i;\theta)\, dv(x_i;\theta), \\ \bigtriangledown^2_{v,v} m(z;\theta)[dv, dv] &= \left\{\frac{\xi_i^2(\theta)}{v(x_i;\theta)} - 1\right\}\frac{dv(x_i;\theta)^2}{v^2(x_i;\theta)}. \end{aligned}$$

In contrast to Example 2.1, the third order differential is non-zero. It can still easily be checked that Eqs. (7)-(8) hold with $\bar{M}_k = E[\bar{m}_k(z_i)]$, $k = 1, 2, 3$, where

$$\bar{m}_0(z_i) := \sup_{\theta\in\Theta}\left[\frac{1}{v^4(x_i;\theta)} + \frac{\xi_i^2(\theta)}{v^6(x_i;\theta)} + \frac{\xi_i^4(\theta)}{v^8(x_i;\theta)}\right],$$

$$\bar{m}_1(z_i) := \sup_{\theta\in\Theta}\left[\frac{1}{v^4(x_i;\theta)} + \frac{\xi_i^4(\theta)}{v^4(x_i;\theta)}\right], \quad \bar{m}_2(z_i) = \sup_{\theta\in\Theta}\left[\frac{1}{v^4(x_i;\theta)} + \frac{\xi_i^2(\theta)}{v^4(x_i;\theta)} + \frac{\xi_i^4(\theta)}{v^6(x_i;\theta)}\right].$$

Example 3.1 (SML). Here,

$$\bigtriangledown m(z_i;\theta)[dp] = -\frac{dp(z_i;\theta)}{p(z_i;\theta)}, \quad \bigtriangledown^2 m(z_i;\theta)\left[dp, dp'\right] = \frac{dp(z_i;\theta)\, dp(z_i;\theta)'}{p^2(z_i;\theta)}.$$

Since $\bigtriangledown^3 m(z_i;\theta)[dp] = -2dp(z_i;\theta)^3/p^3(z_i;\theta)$, it is easily seen that Eq. (7) holds with $\bar{m}_0 := E\left[\sup_{\theta\in\Theta} p^{-3}(z_i;\theta)\right]$. Similarly, Eq. (8) holds with $\bar{M}_k := E\left[\sup_{\theta\in\Theta} p^{-k}(z_i;\theta)\right]$, $k = 1, 2$. Thus, for the moment conditions to hold, it is necessary that $\int p^{-2}(z;\theta)\, dz < \infty$. This only holds if the density is bounded away from zero (this is for example imposed in Lee,

1995), which is a very strong requirement.

In general, we use trimming to circumvent such assumptions (see Fermanian and Salanié, 2004; Kristensen and Shin, 2008) However, trimming imparts a bias component to the approximator, so that $\beta > 0$. Fortunately, our bias-correction does not require that β be exactly, as we will see later.

In the above three cases, $\alpha = 1$ and $\beta = \infty$. More generally, assume only that $\alpha > 0$ and $\beta > 0$ (and that n approximators are used as in eq. (5)). Then $||\hat{\theta}_{n,S} - \hat{\theta}_n|| = o_P\left(n^{-1/2}\right)$ if $n/S^{\alpha} \to 0$, $n/S^{2\beta} \to 0$; that is, if the approximation improves fast enough as the sample size grows, then the approximate estimator is first-order equivalent to the exact estimator $\hat{\theta}_n$ (assuming that it is $\sqrt{n}$-consistent). When nonparametric kernel methods are used to approximate γ and the bandwidth is chosen to be of order $h = O\left(S^{-\delta}\right)$, we saw that $\beta = r\delta$ and $\alpha \simeq 1 - d\delta$. Thus, Theorem 2 yields

$$||\hat{\theta}_{n,S} - \hat{\theta}_n|| = O_P\left(S^{-\delta r}\right) + O_P\left(S^{-(1-\delta d)}\right) + O_P\left(n^{-1/2}S^{-(1-\delta d)/2}\right) + O_P\left(n^{-1/2}S^{-\delta r}\right).$$

We then require that $\sqrt{n}S^{-\delta r} \to 0$ and $\sqrt{n}/S^{-(1-\delta d)} \to 0$. The optimal rate is obtained with $\delta^* = 1/(r+d)$ such that $h^* = O\left(S^{-1/(r+d)}\right)$ which is a non-standard choice compared to what is normally used in kernel smoothing, namely $h = O\left(S^{-1/(2r+d)}\right)$. Thus, it is here recommendable to undersmooth since the bias component plays a more dominant role. Using h^* as our bandwidth, we get

$$||\hat{\theta}^*_{n,S} - \hat{\theta}_n|| = O_P\left(S^{-r/(r+d)}\right) + O_P\left(n^{-1/2}S^{-r/(2r+2d)}\right).$$

Thus, with $\sqrt{n}S^{-r/(r+d)} \to 0$, we obtain first-order equivalence.

However, if only one batch is used in the simulations (such that the approximator is given by eq. (6)), it will hold in great generality that

$$\bigtriangledown Q_n(\theta)[\psi_S] = O_P\left(S^{-1/2}\right), \quad \bigtriangledown^2 Q_n(\theta)[\psi_S, \psi_S] = O_P\left(1/\left(Sh^d\right)\right).$$

In the case of the NPSMLE, this is demonstrated in Example 3.2 below. That $\bigtriangledown Q_n(\theta)[\psi_S]$ exhibits $\sqrt{S}$-convergence is akin to the well-known result from two-step semiparametric estimators where first-step kernel estimation does not reduce the convergence rate of the parametric estimator under regularity conditions. As a consequence, we obtain

$$||\hat{\theta}_{n,S} - \hat{\theta}_n|| = O_P\left(S^{-\delta r}\right) + O_P\left(S^{-(1-\delta d)}\right) + O_P\left(S^{-1/2}\right).$$

As before, the optimal bandwidth is $h^* = O\left(S^{-1/(r+d)}\right)$ yielding

$$||\hat{\theta}^*_{n,S} - \hat{\theta}_n|| = O_P\left(S^{-r/(r+d)}\right) + O_P\left(S^{-1/2}\right),$$

where, for sufficiently smooth densities relative to the dimension, $r/(r+d) > 1/2$ such that the kernel smoothing has no asymptotic influence on the performance.

Example 3.2 (NPSML). We here derive the precise orders of the first order bias and variance components. Regarding the bias component,

$$\begin{aligned}\bigtriangledown Q_n(\theta,\gamma)\left[b_S\right] &= -\frac{1}{n}\sum_{i=1}^{n}\frac{b_S\left(z_i;\theta\right)}{p\left(z_i;\theta\right)} = -h^r\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p\left(y_i|x_i;\theta\right)}\frac{\partial^r p\left(y|x;\theta\right)}{\partial y^r} + o\left(h^r\right)\\ &= -h^r\int\frac{\partial^{r-1}p\left(y|x;\theta\right)}{\partial y^{r-1}}p\left(x\right)dx + o_P\left(h^r\right),\end{aligned}$$

as $n\to\infty$, where $p\left(x\right)$ is the marginal density of x. This holds irrespectively of whether a single or n simulation batches are used.

Next, we derive the rate of the variance component. First, consider the case where n independent batches are used. Then the first order variance component is given by:

$$\bigtriangledown Q_n(\theta)[\psi_S] = -\frac{1}{nS}\sum_{i=1}^{n}\sum_{s=1}^{S}\frac{1}{p\left(z_i;\theta\right)}\left\{K_h\left(y_i - y\left(x_i,\varepsilon_{is};\theta\right)\right) - E\left[K_h\left(y_i - y\left(x_i,\varepsilon_{is};\theta\right)\right)|z_i\right]\right\}.$$

For simplicity suppose that observations are i.i.d. Then, ignoring the cross term,

$$\begin{aligned}\mathrm{Var}\left(\bigtriangledown Q_n(\theta)[\psi_S]|Z_n\right) &= \frac{1}{nS}\int\int\int\frac{p\left(y|x;\theta_0\right)}{p^2\left(y|x;\theta\right)}p\left(x\right)p\left(w|x;\theta\right)\frac{1}{h^{2d}}K\left(\frac{y-w}{h}\right)^2 dydwdx\\ &\quad + \frac{h^{2r}}{nS}\sum_{i=1}^{n}\sum_{s=1}^{S}\frac{1}{p\left(y_i|x_i;\theta\right)}\left[\frac{\partial^r p\left(y_i|x_i;\theta\right)}{\partial y^r}\right]^2\end{aligned}$$

where

$$\begin{aligned}&\frac{1}{nS}\int\int\frac{p\left(y|x;\theta_0\right)}{p^2\left(y|x;\theta\right)}p\left(x\right)\left[\int p\left(w|x;\theta\right)\frac{1}{h^{2d}}K\left(\frac{y-w}{h}\right)^2 dw\right]dydx\\ = \;&\frac{1}{nSh^d}\left[\int\int\frac{p\left(y|x;\theta_0\right)}{p\left(y|x;\theta\right)}p\left(x\right)dydx\right]\left[\int K\left(z\right)^2 dz\right].\end{aligned}$$

Thus, $\bigtriangledown Q_n(\theta)[\psi_S] = O_P(1/\sqrt{nSh^d}) + O_P(h^r/\sqrt{nS})$ which is the rate obtained in the corollary.

However, if we use only one batch of simulations, a better rate is obtained: Employing standard U-statistics results, see e.g. Lee (1990), we easily obtain

$$\begin{aligned}\nabla Q_n(\theta)[\psi_S] &= -\frac{1}{nS}\sum_{i=1}^{n}\sum_{s=1}^{S}\frac{1}{p\left(z_i;\theta\right)}\left\{K_h\left(y_i - y\left(x_i,\varepsilon_s;\theta\right)\right) - E\left[K_h\left(y_i - y\left(x_i,\varepsilon_s;\theta\right)\right)|z_i\right]\right\} \\ &= \frac{1}{S}\sum_{s=1}^{S}E\left[\frac{1}{p\left(z_i;\theta\right)}\left\{K_h\left(y_i - y\left(x_i,\varepsilon_s;\theta\right)\right) - E\left[K_h\left(y_i - y\left(x_i,\varepsilon_s;\theta\right)\right)|z_i\right]\right\}|\varepsilon_s\right] \\ &\quad + o_P(1/\sqrt{S}),\end{aligned}$$

where the leading term satisfies

$$\begin{aligned}&E\left[\frac{1}{p\left(y_i|x_i;\theta\right)}\left\{K_h\left(y_i - y_i\left(x_i,\varepsilon_s;\theta\right)\right) - E\left[K_h\left(y_i - y\left(x_i,\varepsilon_s;\theta\right)\right)|z_i\right]\right\}|\varepsilon_s\right] \\ \simeq\ &\int\int\frac{p\left(y|x;\theta_0\right)}{p\left(y|x;\theta\right)}p\left(x\right)\left[\frac{1}{h^d}K\left(\frac{y - y\left(x,\varepsilon_s;\theta\right)}{h}\right) - h^r\frac{\partial^r p\left(y|x;\theta\right)}{\partial y^r}\right]dydx + o\left(h^r\right) \\ =\ &\int\frac{p\left(y\left(x,\varepsilon_s;\theta\right)|x;\theta_0\right)}{p\left(y\left(x,\varepsilon_s;\theta\right)|x;\theta\right)}p\left(x\right)dydx - h^r\int\int\frac{p\left(y|x;\theta_0\right)}{p\left(y|x;\theta\right)}p\left(x\right)\frac{\partial^r p\left(y|x;\theta\right)}{\partial y^r}dydx + o\left(h^r\right).\end{aligned}$$

Thus, $\nabla Q_n(\theta)[\psi_S] = O\left(S^{-1}\right) + O\left(h^r S^{-1}\right)$ exhibits parametric rate.

4 Analytical Bias Adjustment

We here propose an adjustment of the criterion function $Q_n\left(\theta,\hat{\gamma}_S\right)$ which will remove the leading term of the bias incurred by using $\hat{\gamma}_S$ if the variance component is of a larger order than the bias component, $\alpha < \beta$. This is clearly the case for the core simulation-based estimation methods as $\alpha = 1$ and $\beta = \infty$. In the previous section, we derived the order of the bias and variance of the second order expansion of $Q_n\left(\theta,\hat{\gamma}_S\right)$ in terms of $\hat{\gamma}_S$, and translated these into an error bound for the approximate estimator as stated in Eq. (14). The two leading terms are of the following orders:

$$E\left[\nabla^2 Q_n(\theta,\gamma)[\psi_S,\psi_S]\right] = \sum_{j=1}^{J}E\left[\nabla^2_{jj}Q_n(\theta,\gamma)[\psi_{j,S},\psi_{j,S}]\right] = O\left(S^{-\alpha}\right),$$

$$E\left[\nabla Q_n(\theta)[b_S]\right] = O\left(S^{-\beta}\right).$$

We discuss in turn how these can be adjusted for.

To adjust for the bias in $\hat{\theta}_{n,S}$ due to $E\left[\nabla^2 Q_n(\theta,\gamma)[\psi_S,\psi_S]\right]/2$, we here propose an estimator of this term which we then include in the criterion function. This adjustment will work under the following additional assumption regarding the the approximator:

A.5' Assume that $\hat{\gamma}_{j,S}(x)$ takes the form

$$\hat{\gamma}_{j,S}(x;\theta) = \frac{1}{S}\sum_{s=1}^{S} w_{j,S}(x,\varepsilon_{js};\theta),$$

for some function w_S such that A.5 holds.

Under this assumption, we can write the variance component of $\hat{\gamma}_{j,S}$ as

$$\psi_{j,S}(x;\theta) = \frac{1}{S}\sum_{s=1}^{S} w_{j,S}(x,\varepsilon_{js};\theta) - E\left[w_{j,S}(x,\varepsilon;\theta)\,|x\right],$$

such that, with $w_{js}(x) = w_{j,S}(x,\varepsilon_{is};\theta)$ and $\bar{w}_j(x) = E\left[w_{j,S}(x,\varepsilon_{i,s};\theta)\,|x\right]$ where we have suppressed the dependence on S,

$$\begin{aligned}\sum_{j=1}^{J}\nabla^2_{jj}Q_n(\theta)[\psi_{j,S},\psi_{j,S}] &= \frac{1}{S^2}\sum_{j=1}^{J}\sum_{s=1}^{S}\sum_{t=1}^{S}\nabla^2_{jj}Q_n(\theta)[w_{js}-\bar{w}_j, w_{jt}-\bar{w}_j] \\ &= \frac{1}{S^2}\sum_{j=1}^{J}\sum_{s=1}^{S}\nabla^2_{jj}Q_n(\theta)[w_{js}-\bar{w}_j, w_{js}-\bar{w}_j] + o_P\left(S^{-\alpha}\right),\end{aligned}$$

where the second equality follows from the independence between the batches. Since $\hat{\gamma}_{j,S}$ is an unbiased and consistent estimator of $\bar{w}_j$, we define the adjustment term $\Delta_{n,S}(\theta)$ as:

$$\Delta_{n,S}(\theta) = \frac{1}{2S(S-1)}\sum_{j=1}^{J}\sum_{s=1}^{S}\nabla^2_{jj}Q_n(\theta)[w_{js}-\hat{\gamma}_{j,S}, w_{js}-\hat{\gamma}_{j,S}]. \tag{15}$$

Under regularity conditions, we show that this is an unbiased and consistent estimator of $E\left[\nabla^2 Q_n(\theta)[\psi_S,\psi_S]\right]/2$, $E\left[\Delta_{n,S}(\theta)\right] = E\left[\nabla^2 Q_n(\theta)[\psi_S,\psi_S]\right]/2$ for all $S \geq 1$, and $|\Delta_{n,S}(\theta) - \nabla^2 Q_n(\theta,\gamma_0)[\psi_S,\psi_S]/2| \to^P 0$ as $S \to \infty$. Thus, by including $\Delta_{n,S}(\theta)$ in the computation of $\tilde{\theta}_{n,S}$ as given in Eq. (3), the bias term $E\left[\nabla^2 Q_n(\theta)[\psi_S,\psi_S]\right]/2$ drops out of the expansion of the adjusted criterion function which in turn improves on the rate of convergence. We state this result in the following Theorem:

Theorem 3 *Assume that A.1-A.4 and A.5' hold. Then $\tilde{\theta}_{n,S}$ defined in (3) with $\Delta_{n,S}(\theta)$ given in (15) satisfies:*

1. *Using the approximate estimator in eq. (5):*

$$\begin{aligned}||\hat{\theta}_{n,S} - \hat{\theta}_n|| &= \nabla m_1(\psi_S;\theta) + O_P\left(S^{-(1+\alpha)}\right) + O_P\left(S^{-\beta}\right) \\ &\quad + O_P\left(S^{-\eta}\right) + O_P\left(n^{-1/2}S^{-\alpha/2}\right) + O_P\left(n^{-1/2}S^{-\beta}\right).\end{aligned}$$

2. Using the approximate estimator in eq. (6):

$$\begin{aligned} ||\hat{\theta}_{n,S} - \hat{\theta}_n|| &= O_P\left(S^{-(1+\alpha)}\right) + O_P\left(S^{-\beta}\right) + O_P\left(S^{-\eta}\right) \\ &\quad + O_P\left(n^{-1/2}S^{-\alpha/2}\right) + O_P\left(n^{-1/2}S^{-\beta}\right). \end{aligned}$$

Note that, compared to the unadjusted estimator, the error term $O_P\left(S^{-\alpha}\right)$ has been replaced by one of order $O_P\left(S^{-(1+\alpha)}\right)$. This is due to the fact that $|\Delta_{n,S}(\theta) - \nabla^2 Q_n(\theta, \gamma_0)[\psi_S, \psi_S]/2| = O_P\left(S^{-(1+\alpha)}\right)$. In the leading case where $\alpha = 1$, $O_P\left(S^{-(1+\alpha)}\right)$ is of smaller order than $O_P\left(S^{-3\alpha/2}\right)$, but for other approximators, e.g. NPSML with n independent batches being used, the relationship may be reverse.

For M-estimators ($d(m, \nu) = m$), including SMM, SPML, SNLS or SML estimators, the adjustment term takes the form

$$\Delta_{n,S}(\theta) = \frac{1}{2nS(S-1)} \sum_{i=1}^{n} \sum_{s=1}^{S} \nabla^2 m(z_i; \theta)[w_{is} - \hat{\gamma}_{i,S}, w_{is} - \hat{\gamma}_{i,S}],$$

where $w_{is} = w_S(x_i, \varepsilon_{is}; \theta)$ if n independent batches are used in the simulations, or $w_{is} = w_S(x_i, \varepsilon_s; \theta)$ if only a single batch is used.

With unbiased simulators, we have $\alpha = 1$ and $\beta = \infty$, and the leading bias term of the approximation error of the unadjusted estimator is of order $O_P\left(S^{-1}\right)$. The above theorem shows that for the adjusted estimator the leading term is of order $O_P\left(S^{-3/2}\right)$. The improvement is by a factor $\sqrt{S}$ and so may be very significant. More generally, the proposed adjustment will remove the largest bias component as long as $\alpha < \beta$. Otherwise the bias term $O_P\left(S^{-\beta}\right)$ is of a larger order than $O_P\left(S^{-\alpha}\right)$ in which case the proposed bias adjustment is still valid, but we are evidently not removing the leading term anymore.

Finally, note that when non-stochastic approximations are employed, the above adjustment is irrelevant. With non-stochastic approximations the leading term of the approximation error is not $\nabla^2 Q_n(\theta)[\psi_S, \psi_S]$, which the $\Delta_{n,S}$ correction is aimed at: in fact this term is identically zero as we saw earlier. To phrase things differently, with non-stochastic approximations, $\alpha = -\infty$ and so $\alpha < \beta$.

We now return to the examples introduced in Section 2, and derive the bias adjustments for the ones where stochastic approximators are employed.

Example 2.1 (SNLS). We saw in the previous section that

$$\nabla^2 Q_n(\theta, \gamma)\left[d\gamma, \gamma'\right] = \frac{2}{n} \sum_{i=1}^{n} d\gamma'(x_i; \theta)\, d\gamma(x_i; \theta).$$

Let $r_{is}(\theta) = w_S(x_i, \varepsilon_{is}; \theta) - \hat{\gamma}_{i,S}(x_i; \theta)$ denote the difference of a given simulator from the

mean simulation for the same observation. Then the adjustment term becomes

$$\Delta_{n,S}(\theta) = \frac{1}{nS(S-1)} \sum_{i=1}^{n} \sum_{s=1}^{S} r_{is}^2(\theta).$$

This is exactly the correction proposed in Laffont et al. (1995). Take for instance the binomial choice model discussed earlier of this example, $y = \mathbb{I}\{y^* > 0\}$ and $y^* = m(x, \varepsilon; \theta)$. For this model, the adjustment term is:

$$\Delta_{n,S}(\theta) = \frac{1}{nS(S-1)} \sum_{i=1}^{n} \sum_{s=1}^{S} \left(\mathbb{I}\{m_{is}(\theta) > 0\} - \frac{1}{S} \sum_{t=1}^{S} \mathbb{I}\{m_{it}(\theta) > 0\} \right)^2.$$

The adjustment term is discontinuous in θ, which may make minimization difficult; but this can in principle be solved by replacing the indicator functions above with integrated kernels such as the (scaled) cdf of the logistic distribution, c.f. Fermanian and Salanié (2004).

Example 2.2 (SPML). In this example,

$$\bigtriangledown^2 Q_n(\theta, \gamma)[d\gamma, d\gamma] = \bigtriangledown^2 Q_n(\theta, \gamma)[dg, dg] + 2 \bigtriangledown^2 Q_n(\theta, \gamma)[dg, dv] + \bigtriangledown^2 Q_n(\theta, \gamma)[dv, dv],$$

where, with $\xi_i(\theta) = y_i - m(x_i; \theta)$,

$$\begin{aligned}
\bigtriangledown^2 Q_n(\theta, \gamma)[dg, dg] &= \frac{2}{n} \sum_{i=1}^{n} \frac{dg(x_i; \theta)^2}{v_0(x_i; \theta)}, \\
\bigtriangledown^2 Q_n(\theta, \gamma)[dg, dv] &= \frac{2}{n} \sum_{i=1}^{n} \frac{\xi_i(\theta)}{v_0^2(x_i; \theta)} dg(x_i; \theta)\, dv(x_i; \theta), \\
\bigtriangledown^2 Q_n(\theta, \gamma)[dv, dv] &= \frac{2}{n} \sum_{i=1}^{n} \left\{ \frac{\xi_i(\theta)}{v_0(x_i; \theta)} - \frac{1}{2} \right\} \frac{dv(x_i; \theta)^2}{v_0^2(x_i; \theta)}.
\end{aligned}$$

Let the simulated versions of the conditional mean and variance be on the form

$$\hat{m}_i(x_i; \theta) = \frac{1}{S} \sum_{s=1}^{S} w^{[m]}(x_i, \varepsilon_{is}; \theta), \quad \hat{v}_i(x_i; \theta) = \frac{1}{S} \sum_{s=1}^{S} w^{[v]}(x_i, \varepsilon_{is}; \theta).$$

Then we obtain the following expressison for the bias adjustment: Write $\hat{\xi}_i(\theta) = y_i - \hat{m}_i(x_i; \theta)$ and

$$r_{is}(\theta) = w^{[m]}(x_i, \varepsilon_{is}; \theta) - \hat{m}_i(x_i; \theta), \quad d_{is}(\theta) = w^{[v]}(x_i, \varepsilon_{is}; \theta) - \hat{v}(x_i; \theta).$$

Then we obtain

$$\Delta_{n,S}(\theta) = \Delta_{n,S}^{(1)}(\theta) + \Delta_{n,S}^{(2)}(\theta) + \Delta_{n,S}^{(3)}(\theta),$$

where,

$$\begin{aligned}
\Delta_{n,S}^{(1)}(\theta) &= \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S}\frac{r_{is}^2(\theta)}{\hat{v}(x_i;\theta)}, \\
\Delta_{n,S}^{(2)}(\theta) &= \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S}\frac{\hat{\xi}_i(\theta)}{\hat{v}^2(x_i;\theta)}r_{is}(\theta)d_{is}(\theta), \\
\Delta_{n,S}^{(3)}(\theta) &= \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S}\left\{\frac{\hat{\xi}_i(\theta)}{\hat{v}(x_i;\theta)}-\frac{1}{2}\right\}\frac{d_{is}^2(\theta)}{\hat{v}^2(x_i;\theta)}.
\end{aligned}$$

If two independent batches of simulated draws are used to compute $\hat{g}$ and $\hat{v}$ respectively, then $\Delta_{n,S}^{(2)}(\theta)$ has mean zero and can be left out in the computation of $\Delta_{n,S}(\theta)$. In this case, $\bigtriangledown^3 Q_n(\theta,\gamma)[d\gamma] \neq 0$ and so the bias adjustment does not ensure consistency for fixed S.

Example 3.1 (SML). Here,

$$\bigtriangledown^2 Q_n(\theta,\gamma)[d\gamma] = \frac{1}{n}\sum_{i=1}^{n}\frac{d\gamma^2(z_i;\theta)}{\gamma^2(z_i;\theta)},$$

and the adjustment term becomes

$$\Delta_{n,S}(\theta) = \frac{1}{2nS(S-1)}\sum_{i=1}^{n}\sum_{s=1}^{S}\left[\frac{w_S(z_i,\varepsilon_{is};\theta)-\hat{\gamma}(z_i;\theta)}{\hat{\gamma}(z_i;\theta)}\right]^2.$$

As in the previous example, $\bigtriangledown^3 Q_n(\theta,\gamma)[d\gamma] \neq 0$ and the bias adjustment does not ensure consistency for fixed S. On the other hand, we note that

$$\bigtriangledown^2 Q_n(\theta)[\psi,\psi,\psi] = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p^3(z_i;\theta)}\psi^3(z_i;\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p^3(z_i;\theta)}\left[\frac{1}{S}\sum_{s=1}^{S}r(z_i,\varepsilon_{i,s};\theta)\right]^3,$$

where $r(z_i,\varepsilon_{i,s};\theta) = w(z_i,\varepsilon_{i,s};\theta) - E[w(z_i,\varepsilon_{i,s};\theta)]$. It is easily shown for i.i.d. simulations that $\bigtriangledown^2 Q_n(\theta)[\psi,\psi,\psi] = O_P\left(S^{-2}\right) + O_P\left(n^{-1/2}S^{-3/2}\right)$ by Lee (1995, Propositions A.3-A.4). Thus, the bias adjustment leads to a significant improvement in the rate of the approximate estimator.

Example 3.2 (NPSML). In the case where n batches of simulations are used, bias corrected estimator satisfies:

$$||\hat{\theta}_{n,S} - \hat{\theta}_n|| = O_P\left(S^{-\delta r}\right) + O_P\left(S^{-(2-\delta d)}\right) + O_P\left(n^{-1/2}S^{-(1-\delta d)/2}\right) + O_P\left(n^{-1/2}S^{-\delta r}\right),$$

and the optimal bandwidth is of order $h^* = O\left(S^{-2/(r+d)}\right)$ such that

$$||\hat{\theta}^*_{n,S} - \hat{\theta}_n|| = O_P\left(S^{-2r/(r+d)}\right) + O_P\left(n^{-1/2}S^{-r/(r+d)}\right).$$

A similar result is found for the adjusted estimator based on a single batch of simulations.

Instead of adjusting the objective function, one could think of correcting the approximate estimator instead. The former can be seen as a preventive bias adjustment while the latter is a corrective one. To illustrate how this is could be done, consider the case where $Q_n(\theta, \gamma)$ is differentiable in θ with $S_n(\theta, \gamma) = \partial Q_n(\theta, \gamma)/\partial\theta$. Then the unadjusted approximate estimator satisfies $S_n(\hat{\theta}_{n,S}, \hat{\gamma}_S) = 0$ while the estimator based on the adjusted criterion function solves $S_n(\tilde{\theta}_{n,S}, \hat{\gamma}_S) - \dot{\Delta}_{n,S}(\tilde{\theta}_{n,S}) = 0$, where

$$\dot{\Delta}_{n,S}(\theta) = \frac{1}{2S(S-1)} \sum_{j=1}^{J} \sum_{s=1}^{S} \nabla^2_{jj} S_n(\theta, \gamma)[w_{js} - \hat{\gamma}_{j,S}, w_{js} - \hat{\gamma}_{j,S}].$$

is an estimator of the second order pathwise derivative of the score function $S_n(\theta, \gamma)$. That is, $\dot{\Delta}_{n,S}(\theta) = \partial\Delta_{n,S}/\partial\theta$.

Taking a first-order expansion in θ around $\hat{\theta}_{n,S}$, the formula leads to the following bias-adjustment:

$$\tilde{\theta}_{n,S} = \hat{\theta}_{n,S} - H_n(\hat{\theta}_{n,S}, \hat{\gamma}_{n,S})^{-1}\dot{\Delta}_{n,S}(\hat{\theta}_{n,S}),$$

where $H_n(\theta, \gamma) = \partial^2 Q_n(\theta, \gamma) / \left(\partial\theta\partial\theta'\right)$. So here, we adjust the approximate estimator after it is computed, instead of correcting the objective function and obtaining an adjusted approximate estimator in one step. This two-step procedure was proposed in Lee (1995) for the special case of SMLE and SNLS of limited dependent variable models.

As an illustration, in the SNLS example, the adjustment term takes the following form:

$$\dot{\Delta}_{n,S}(\theta) = \frac{2}{nS^2} \sum_{i=1}^{n} \sum_{k=1}^{m} \left(\dot{w}(x_i, \varepsilon_{is}; \theta) - \dot{\hat{\gamma}}_S(x_i; \theta)\right)\left(\dot{w}(x_i, \varepsilon_{is}; \theta) - \hat{\gamma}_S(x_i; \theta)\right),$$

where $\dot{g}$ denotes the derivative w.r.t. θ. For the SML example it is

$$\dot{\Delta}_{n,S}(\theta) = -\frac{1}{nS^2} \sum_{i=1}^{n} \sum_{k=1}^{m} \frac{\left(\dot{w}(z_i, \varepsilon_{is}; \theta) - \dot{\hat{\gamma}}_S(z_i; \theta)\right)\left(\dot{w}(z_i, \varepsilon_{is}; \theta) - \hat{\gamma}_S(z_i; \theta)\right)}{\hat{\gamma}^2_S(z_i; \theta)}.$$

One complication of this corrective procedure relative to the preventive one is that we here need to be able to compute the derivatives of the simulators. We refer to Arellano and Hahn (2007) for a further discussion of corrective and preventive bias correction in a panel data setting.

Next, we here discuss the issue of adjustment for the bias term, b_S. One would ideally

also like to adjust for the bias due to $b_S(z;\theta) \neq 0$, which drives the term in $E[\bigtriangledown Q_n(\theta)[b_S]] = O\left(S^{-\beta}\right)$. If one is able to obtain (an estimator of) b_S, the associated adjustment term can straightforwardly be estimated by

$$\Delta_{n,S}^{(B)}(\theta) = \bigtriangledown Q_n(\theta)[b_S].$$

However, in most cases, only approximate expressions of b_S are available, and these expressions involve unknown components that need to be estimated, so this estimator is not easily computed.

Instead of trying to estimate $\bigtriangledown Q_n(\theta)[b_S]$, one may try to improve the order of $Q_n(\theta)[b_S]$ by adjusting the estimator $\hat{\gamma}_S$ itself. Lee (2001) demonstrates how combining numerical approximations and simulations can improve the order of the estimator. When kernel-based estimators are used, so-called twicing kernels can also be used to decrease the bias component. Suppose for example that $\gamma(z;\theta) = f(y|x;\theta)$ and $\hat{\gamma}_S(y|x;\theta) = S^{-1}\sum_{s=1}^{S} K_h(Y_s(\theta,x) - y)$, where K is a r-order kernel function. Then $b_S(z;\theta) = h^r \partial^r f(y|x;\theta)/\partial y^r + o(h^r)$ which is not easily estimated. Instead, we here propose to reduce this bias component by using twicing kernels as advocated by Newey et al. (2004) in a different context: For a given kernel function K, define the associated twicing kernel $\bar{K}$ by

$$\bar{K}(z) = 2K(z) - \int K(z-w)K(w)\,dw.$$

Suppose now that the first order pathwise derivative takes the form

$$\bigtriangledown Q_n(\theta,\gamma)[d\gamma] = \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown q(z_i;\theta)\,d\gamma(z_i;\theta);$$

the order of the variance is then the same whether twicing kernels or standard kernels are used. On the other hand, with regard to the bias component the use of a standard kernel function yields:

$$\bigtriangledown Q_n(\theta,\gamma)\left[b_S^K\right] = O_P(h^r) + O_P\left(n^{-1/2}h^{-d/2}\right),$$

where d is the dimension of y, while the use of a twicing kernel estimator yields

$$\bigtriangledown Q_n(\theta,\gamma)[b_S^{\bar{K}}] = O_P\left(h^{2r}\right) + O_P\left(n^{-1/2}h^{-d/2}\right),$$

cf. Newey et al. (2004, Theorem 1). Again, the improvement obtained here is not through an adjustment term added to the criterion function since the adjustment takes place in the construction of $\hat{\gamma}_S(y|x;\theta)$ itself.

5 Other Bias Correction Methods

In addition to the analytical bias corrrection proposed in the previous section, other methods for bias correction can be found in the literature. We here give a brief overview and describe how these potentially can be used in our setting.

One prominent alternative is Jackknifing. The advantage of this is that it will in general handle the biases due to the stochastic, but also due to the non-stochastic component of the approximator. It will on the other hand be computationally more demanding than the analytical bias correction proposed in the previous section. See Hahn and Newey (2004) for similar applications of the Jackknife in the context of panel models, while we refer to Phillips and Yu (2005) for a time series application.

In our setting, the Jackknife involves computing m approximate estimators of order $s = S/m$: Let $\hat{\theta}_{n,s}^{[j]}$ be the estimator based on using the jth sub-approximator, $j = 1, ...m$. We then propose the following jackknife estimator:

$$\tilde{\theta}_{n,S}^{\text{jack}} = \frac{m}{m-1}\hat{\theta}_{n,S} - \frac{1}{m\,(m-1)}\sum_{j=1}^{m}\hat{\theta}_{n,s}^{[j]}.$$

In order to derive the asymptotic properties, we assume that $Q_n\,(\theta, \gamma)$ is twice differentiable w.r.t. θ and that satisfies the same conditions as a functional of γ as we imposed on $Q_n\,(\theta, \gamma)$. We then first Taylor-expand at $\hat{\theta}_n$ to obtain

$$0 = S_n(\hat{\theta}_{n,S}, \hat{\gamma}_S) = S_n(\hat{\theta}_n, \hat{\gamma}_S) + H_n(\hat{\theta}_n, \hat{\gamma}_S)(\hat{\theta}_{n,S} - \hat{\theta}_n),$$

where in great generality $H_n(\hat{\theta}_n, \hat{\gamma}_S) \to^P H_0$ while

$$S_n(\hat{\theta}_n, \hat{\gamma}_S) = \bigtriangledown S_n(\hat{\theta}_{n,S})\,[\hat{\gamma}_S - \gamma] + \frac{1}{2}\bigtriangledown S_n(\hat{\theta}_{n,S})\,[\hat{\gamma}_S - \gamma, \hat{\gamma}_S - \gamma] + O\left(\|\hat{\gamma}_S - \gamma\|^3\right).$$

Here, following the same arguments as in Lemmas 5-6, the biases can, under suitable regularity conditions, be written as:

$$\begin{aligned} B_{n,S}^{[1]} &: = E\left[\bigtriangledown S_n(\hat{\theta}_{n,S})\,[\hat{\gamma}_S - \gamma]\right] = B_1 S^{-\beta} + o\left(S^{-\beta}\right) \\ B_{n,S}^{[2]} &: = E\left[\bigtriangledown S_n(\hat{\theta}_{n,S})\,[\hat{\gamma}_S - \gamma, \hat{\gamma}_S - \gamma]\right] = B_2 S^{-\alpha} + o\left(S^{-\alpha}\right). \end{aligned}$$

In total,

$$E\left[\hat{\theta}_{n,S} - \hat{\theta}_n\right] = H_0^{-1}\left[B_1 S^{-\beta} + B_2 S^{-\alpha}\right] + O\left(S^{-2\beta}\right) + O\left(S^{-3\alpha/2}\right),$$

and

$$E\left[\hat{\theta}_{n,s}-\hat{\theta}_n\right]=H_0^{-1}\left[\frac{B_1}{m}S^{-\beta}+\frac{B_2}{m}S^{-\alpha}\right]+O\left(S^{-2\beta}\right)+O\left(S^{-3\alpha/2}\right),$$

This yields, ignoring the higher order terms,

$$E\left[\tilde{\theta}_{n,S}^{\text{jack}}-\hat{\theta}_n\right]\simeq\frac{m}{m-1}H_0^{-1}\left[B_1S^{-\beta}+B_2S^{-\alpha}\right]-\frac{1}{m\,(m-1)}H_0^{-1}\sum_{j=1}^{m}\left[\frac{B_1}{m}S^{-\beta}+\frac{B_2}{m}S^{-\alpha}\right]\simeq 0.$$

6 Newton-Raphson Adjustment

We here propose an additional adjustment that works for general estimators, including ones that involve no simulations. We show that starting from either $\bar{\theta}_{n,S}=\tilde{\theta}_{n,S}$ or even $\bar{\theta}_{n,S}=\hat{\theta}_{n,S}$, one or more Newton-Raphson iterations based on the approximate objective function with a finer approximation $S^*>>S$ produce an estimator that has the (presumably) higher precision of $\hat{\theta}_{n,S^*}$. The resulting estimator based on k iterations, $\hat{\theta}_{n,S}^{(k+1)}$, is defined in eq. (4). In order for this estimator to be well-defined, we have to assume that $Q_n(\theta,\gamma)$ is twice differentiable w.r.t. θ:

A.6 For all γ' in a neighbourhood of γ, $Q_n(\theta,\gamma')$ is twice differentiable w.r.t. θ.

This assumption does rule out certain types of approximate estimators such as the simulated method of moment estimators for discrete choice models proposed in McFadden (1989) and Pakes and Pollard (1989). However, by using kernel smoothers instead of indicator functions, their estimators can easily be adjusted to satisfy A.6.

To evaluate the performance of $\hat{\theta}_{n,S}^{(k+1)}$ relative to $\bar{\theta}_{n,S^*}$, we first note that

$$||\hat{\theta}_{n,S}^{(k+1)}-\hat{\theta}_n||\leq||\hat{\theta}_{n,S}^{(k+1)}-\bar{\theta}_{n,S^*}||+||\bar{\theta}_{n,S}-\hat{\theta}_n||.$$

We then apply Robinson (1988, Theorem 2) to obtain that

$$\begin{aligned}||\hat{\theta}_{n,S}^{(k+1)}-\bar{\theta}_{n,S^*}||&=O_P\left(||\bar{\theta}_{n,S}-\bar{\theta}_{n,S^*}||^{2^k}\right)\\&=O_P\left(||\bar{\theta}_{n,S}-\hat{\theta}_n||^{2^k}\right)+O_P\left(||\bar{\theta}_{n,S^*}-\hat{\theta}_n||^{2^k}\right),\end{aligned}$$

which in turn implies

$$||\hat{\theta}_{n,S}^{(k+1)}-\hat{\theta}_n||=O_P\left(||\bar{\theta}_{n,S}-\hat{\theta}_n||^{2^k}\right)+O_P\left(||\bar{\theta}_{n,S^*}-\hat{\theta}_n||\right).\tag{16}$$

We then simply choose the number of iterations, k, large enough so that the first term is of smaller order than the second, and $\hat{\theta}_{n,S}^{(k+1)}$ is first-order equivalent to $\bar{\theta}_{n,S^*}$. This is stated in the following theorem:

Theorem 4 *Assume that A.1-A.6 hold. Assume that the initial estimator is computed with* $S = O\left(n^{\underline{r}}\right)$ *and the 2nd step NR-iterations are done with* $S^* = O(n^{\bar{r}})$ *where* $0 < \underline{r} \leq \bar{r}$. *Then with* $k > \left[\log\left(\bar{r}/\underline{r}\right)/\log\left(2\right)\right]$,

$$||\hat{\theta}_{n,S}^{(k+1)} - \hat{\theta}_n|| = O_P(||\bar{\theta}_{n,S^*} - \hat{\theta}_n||).$$

Note here that we allow for $\underline{r} < XXXX$ and so we do not require the initial estimator to be $\sqrt{n}$-consistent, merely that it be consistent. Moreover, if S^* goes to infinity with n at the same speed as S, then $\bar{r} = \underline{r}$ and the formula shows that one iteration is enough.

The above iterative estimator requires computation of the Hessian, $H_n\left(\theta, \hat{\gamma}_S\right)$. If this is not feasible or computationally burdensome, an approximation can be employed, e.g. numerical derivatives. This however will slow down the convergence rate and the result of Theorem 4 has to be adjusted, cf. Robinson (1988, Theorem 5). In particular, more iterations are required to obtain a given level of precision.

7 A (Preliminary) Simulation Study

To explore the performance of our proposed approaches, we set up a small Monte Carlo study of an autoregressive Tobit time series model: the econometrician observes

$$y_t = \max(z_t, 0), \quad \text{for } t = 1, \ldots, T,$$

where the latent variable z_t is an AR(1) process:

$$z_t = a + bz_{t-1} + u_t,$$

and z_0 is drawn from the ergodic distribution of the z_t process for the current parameter values.

We assume that the true model has $a = 0, b = 0.5$ and u_t is iid $N(0, 1)$.

We use various sample sizes T and number of draws S. For a time series model on macroeconomic data, $T = 150$ seems to be a natural benchmark, and we start from this case with $S = 50$ simulations. We ran 200 simulations, starting from initial values of the parameters drawn randomly as

$$a \sim U[-0.2, 0.2], \quad b \sim U[0.25, 0.75], \quad u_1 \sim U[0.5, 1.5].$$

The estimation method we use is NPSML. As the model has no covariates, we need to approximate the likelihood of $(y_t, \ldots, y_{t-K})$ for some $K > 0$ in order to identify such a model, as explained in Fermanian-Salanié (2004). We choose $K = 1$ here.

The initial bias adjustment (correcting the objective function with the $\Delta_{n,S}$ term) has somewhat surprising properties. Even though we start the minimization for the corrected objective function with a new set of initial parameter values, the corrected estimates are often strikingly close to the uncorrected estimates. In fact, for about one third of the sample the Euclidean distance between the corrected and the uncorrected vector of estimates is smaller than 0.01. We document this in Table XXX.

This feature makes standard statistical summaries much less informative than usual, so we resort to plots. Since the coefficient of most interest to the researcher is likely to be the autoregressive parameter b, we focus on it; but the other two parameters tell a similar story. On each graph we plotted on the x-axis $\hat{b}_{n,S} - 0.5$, the deviation of the uncorrected NPSML1 estimate from the true value of b. The y-axis measure the impact of the correction, $\tilde{b}_{n,S} - \hat{b}_{n,S}$.

The first thing apparent from the graphs is the cluster of points on the $y = 0$ line, where the correction is very small. For those roughly 65% of samples in which the correction has a sizable impact, the cloud of points slopes downwards around the $x + y = 0$ line; since a point on the $x + y = 0$ line signals that the correction fully brought $\tilde{b}_n$ to the true value 0.5, this is rather comforting.

References

Altissimo, F. and A. Mele (2008) Simulated Nonparametric Estimation of Dynamic Models. Forthcoming in *Review of Economic Studies*.

Arellano, M. and J. Hahn (2007) Understanding Bias in Nonlinear Panel Models: Some Recent Developments. In *Advances in Economics and Econometrics*, Volume III (eds. R. Blundell, W.K. Newey and T. Persson). Cambridge: Cambridge University Press.

Andrews, D.W.K. (1994) Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity. *Econometrica* 62, 43-72.

Bao, Y. and A. Ullah (2007) The Second-order Bias and Mean Squared Error of Estimators in Time-Series Models. *Journal of Econometrics* 140, 650–669.

Bierings, H. and K. Sneek (1989) Pseudo Maximum Likelihood Techniques in a Simple Rationing Model of the Dutch Labour Market. Research Memoranda No. 1989-63, Faculty of Economics, Business Administration and Econometrics, Free University Amsterdam.

Brownlees, C.T., D. Kristensen and Y. Shin (2009) Nonparametric Simulated Maximum Likelihood Estimation of Dynamic Latent Variable Models. Manuscript, Department of Economics, Columbia University.

Creel, M. and D. Kristensen (2008) Estimation of Dynamic Latent Variable Models Using Simulated Nonparametric Moments. Manuscript, Department of Economics, Universitat Autònoma de Barcelona.

van Dijk, H., A. Monfort and B. Brown (1995) *Econometric Inference Using Simulation Techniques*. John Wiley.

Duffie, D. and K. J. Singleton (1993) Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica* 61, 929–952.

Fermanian, J.-D. and B. Salanié (2004) A Nonparametric Simulated Maximum Likelihood Estimation Method. *Econometric Theory* 20, 701-734.

Fernández-Villaverde, J. and J.F. Rubio-Ramirez (2005) Estimating Dynamic Equilibrium Economies: Linear versus Nonlinear Likelihood. *Journal of Applied Econometrics* 20, 891–910.

Fernández-Villaverde, J., J.F. Rubio-Ramirez and M. Santos (2006) Convergence Properties of the Likelihood of Computed Dynamic ModelsÓ . *Econometrica,* 74, 93-119.

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

Hahn, J. and G. Kuersteiner (2004) Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects. Manuscript, Department of Economics, MIT.

Hahn, J. and W.K. Newey (2004) Jackknife and Analytical Bias Reduction for Nonlinear Panel Models. *Econometrica* 72, 1295-1319.

Hajivassiliou, V.A. (2000) Some Practical Issues in Maximum Simulated Likelihood. In *Simulation-based Inference in Econometrics* (eds. R. Mariano, T. Schuermann and M.J. Weeks), 71-99. Cambridge: Cambridge University Press.

Hall, P. and C.C. Heyde (1980) *Martingale Limit Theory and its Applications*. New York: Academic Press.

Judd, K., F. F. Kubler and K. Schmedder (2003) Computational Methods for Dynamic Equilibria with Heterogeneous Agents. In *Advances in Economics and Econometrics* (eds. M. Dewatripont, L.P. Hansen, and S. Turnovsky). Cambridge University Press.

Kristensen, D. (2008) Uniform Convergence Rates of Kernel Estimators with Heterogeneous, Dependent Data. Forthcoming in *Econometric Theory*.

Kristensen, D. and Y. Shin (2008) Estimation of Dynamic Models with Nonparametric Simulated Maximum Likelihood. CREATES Research Papers 2008-58, University of Aarhus.

Laffont, J.-J., H. Ossard and Q. Vuong (1995) Econometrics of First-Price Auctions. *Econometrica* 63, 953-980.

Laroque, G. and B. Salanié (1989) Estimation of Multimarket Fix-Price Models: An Application of Pseudo-maximum Likelihood Methods. *Econometrica* 57, 831–860.

Laroque, G. and B. Salanié (1993) Simulation-based Estimation of Models with Lagged latent Variables. *Journal of Applied Econometrics* 8, 119–133.

Lee, A.J. (1990) *U-Statistics: Theory and Practice*. Marcel Dekker, Inc.

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, 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. (2001) Interpolation, Quadrature, and Stochastic Integration. *Econometric Theory* 17, 933-961.

Linton, O. (1996) Edgeworth Approximation for MINPIN Estimators in Semiparametric Regressions Models. *Econometric Theory* 12, 30-60.

R. Mariano, T. Schuerman and M. Weeks (2000) *Simulation-based Inference in Econometrics*. Cambridge University Press.

McFadden, D. (1989) A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration. *Econometrica* 57, 995-1026.

Newey, W.K., F. Hsieh and J.M. Robins (2004) Twicing Kernels and A Small Bias Property of Semiparametric Estimators. *Econometrica* 72, 947-962.

Newey, W.K., J.J.S. Ramalho and R. Smith (2005) Asymptotic Bias for GMM and GEL Estimators with Estimated Nuisance Parameters. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (eds. D.W.K. Andrews and J.H. Stock). Cambridge University Press

Newey, W.K. and R. Smith (2004) Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica* 72, 219-255.

Norets, A. (2006) Inference in Dynamic Discrete Choice Models with Serially Correlated Unobserved State Variables. Manuscript, Department of Economics, Princeton University.

Olsson, J. and T. Rydén (2008) Asymptotic Properties of Particle Filter-Based Maximum Likelihood Estimators for State Space Models. *Stochastic Processes and their Applications* 118, 649-680.

Pakes, A. and D. Pollard (1989) Simulation and the Asymptotics of Optimization Estimators. *Econometrica* 57, 1027-57.

Rilstone, P., V.K. Srivastavab and A. Ullah (1996) The Second Order Bias and MSE of nonlinear Estimators. *Journal of Econometrics* 75, 369-395.

Robinson, P.M. (1988) The Stochastic Difference Between Econometric Statistics. *Econometrica* 56, 531-548.

Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley.

A Proofs

Proof of Lemma 1. To show the first part, we combine the uniform convergence of $L_n(\theta)$ with a 2nd order Taylor expansion of $L(\theta)$ to obtain that for some $\bar{\theta} \in [\tilde{\theta}, \hat{\theta}]$,

$$\begin{aligned} L_n(\tilde{\theta}) - L_n(\hat{\theta}) &= \left\{L_n(\tilde{\theta}) - L(\tilde{\theta})\right\} - \{L_n(\hat{\theta}) - L(\hat{\theta})\} + \{L(\tilde{\theta}) - L(\hat{\theta})\} \\ &= \{L(\tilde{\theta}) - L(\hat{\theta})\} + o_P(1) \\ &= \frac{\partial L(\hat{\theta})}{\partial \theta} + \frac{1}{2}(\tilde{\theta} - \hat{\theta})' H(\bar{\theta})(\tilde{\theta} - \hat{\theta}) + o_P(1) \\ &= \frac{1}{2}(\tilde{\theta} - \hat{\theta})' H(\theta_0)(\tilde{\theta} - \hat{\theta}) + o_P(1), \end{aligned}$$

where the last equality follows from the fact that $\partial L(\hat{\theta})/\partial\theta = L(\theta_0)/\partial\theta + o_P(1) = o_P(1)$ and $H(\bar{\theta}) \to^P H(\theta_0)$ since $\bar{\theta} \to^P \theta_0$. Next, since $\hat{\theta}$ and $\tilde{\theta}$ are the minimizers of $L_n(\theta)$ and $Q_n(\theta)$ respectively, we have

$$0 \leq L_n(\tilde{\theta}) - L_n(\hat{\theta}) = \{L_n(\tilde{\theta}) - \hat{L}_n(\hat{\theta})\} + \{\hat{L}_n(\hat{\theta}) - L_n(\hat{\theta})\} \leq \{L_n(\tilde{\theta}) - \hat{L}_n(\tilde{\theta})\} + \{\hat{L}_n(\hat{\theta}) - L_n(\hat{\theta})\},$$

where both the first and second term on the RHS of the last inequality is bounded by $\sup_{\theta\in\Theta} |L_n(\theta) - \hat{L}_n(\theta)|$. Thus,

$$||\hat{\theta} - \tilde{\theta}||^2 \leq \left\|H^{-1}(\theta_0)\right\| (\tilde{\theta} - \hat{\theta})' H(\theta_0)(\tilde{\theta} - \hat{\theta}) \leq 4 \sup_{\theta\in\Theta} |L_n(\theta) - \hat{L}_n(\theta)| + o_P(1).$$

The second part follows along the same lines: First,

$$\begin{aligned} T_n(\tilde{\theta}) - T_n(\hat{\theta}) &= T(\tilde{\theta}) - T(\hat{\theta}) + o_P(1) \\ &= T(\tilde{\theta}) + o_P(1) \\ &= (\tilde{\theta} - \hat{\theta})' H(\hat{\theta})(\tilde{\theta} - \hat{\theta}) \\ &= (\tilde{\theta} - \hat{\theta})' H(\theta_0)(\tilde{\theta} - \hat{\theta}) + o_P(1), \end{aligned}$$

where we have used the uniform convergence condition and that $\hat{\theta} \to^P \theta_0$. The rest of the proof proceeds just as before. ■

Proof of Theorem 2. The functional differentials are given by

$$\bigtriangledown Q_n(\theta)[d\gamma] = \frac{\partial d(\bar{m}_n(\theta, \gamma), \hat{\nu})}{\partial m} \bigtriangledown \bar{m}_n(\theta)[d\gamma],$$

and

$$\begin{aligned}\bigtriangledown^2 Q_n(\theta)\left[d\gamma\right] &= \bigtriangledown \bar{m}_n\left(\theta\right)\left[d\gamma'\right]' \frac{\partial^2 d\left(\bar{m}_n\left(\theta,\gamma\right),\hat{\nu}\right)}{\partial m^2} \bigtriangledown \bar{m}_n\left(\theta\right)\left[d\gamma\right] \\ &\quad + \frac{\partial d\left(\bar{m}_n\left(\theta,\gamma\right),\hat{\nu}\right)}{\partial m} \bigtriangledown^2 \bar{m}_n\left(\theta\right)\left[d\gamma, d\gamma'\right],\end{aligned}$$

where $\bigtriangledown \bar{m}_n\left(\theta\right)\left[d\gamma\right]$ and $\bigtriangledown^2 \bar{m}_n\left(\theta\right)\left[d\gamma, d\gamma'\right]$ are given in the Proof of Lemmas 6 and 5. Thus, with $\bar{m} = \bar{m}_n\left(\theta,\gamma\right)$, $\hat{m} = \bar{m}_n\left(\theta,\hat{\gamma}_S\right)$ and $d\hat{\gamma} = \hat{\gamma}_S - \gamma$,

$$\begin{aligned}& Q_n(\theta,\hat{\gamma}_S) - Q_n(\theta,\gamma) - \bigtriangledown Q_n(\theta)\left[d\hat{\gamma}\right] - \frac{1}{2}\bigtriangledown^2 Q_n(\theta)\left[d\hat{\gamma}\right] \\ = & \left\{ d\left(\bar{m},\hat{\nu}\right) - d\left(\hat{m},\hat{\nu}\right) - \frac{\partial d\left(\bar{m},\hat{\nu}\right)}{\partial m}\left(\bar{m}-\hat{m}\right) - \frac{1}{2}\left(\bar{m}-\hat{m}\right)' \frac{\partial d\left(\bar{m},\hat{\nu}\right)}{\partial m^2}\left(\bar{m}-\hat{m}\right)\right\} \\ & + \frac{\partial d\left(\bar{m},\hat{\nu}\right)}{\partial m}\left\{\bar{m}-\hat{m} - \bigtriangledown \bar{m}_n\left(\theta\right)\left[d\hat{\gamma}\right] - \bigtriangledown^2 \bar{m}_n\left(\theta\right)\left[d\gamma, d\hat{\gamma}'\right]\right\} \\ & + \frac{1}{2}\left\{\bar{m}-\hat{m} - \bigtriangledown \bar{m}_n\left(\theta\right)\left[d\hat{\gamma}\right]\right\}' \frac{\partial^2 d\left(\bar{m},\hat{\nu}\right)}{\partial m^2}\left\{\bar{m}-\hat{m} - \bigtriangledown \bar{m}_n\left(\theta\right)\left[d\hat{\gamma}\right]\right\} \\ = & : A_1 + A_2 + A_3.\end{aligned}$$

Using the stationarity and mixing assumption in conjunction with the uniformly boundedness of $\bar{m}_n\left(\theta,\gamma\right)$ and the consistency of $\hat{\nu}$ as assumed in (A.1)-(A.3), the following two convergence results hold uniformly over θ:

$$\begin{aligned}\frac{\partial d\left(\bar{m}_n\left(\theta,\gamma\right),\hat{\nu}\right)}{\partial m} &\rightarrow {}_P \frac{\partial d\left(M\left(\theta,\gamma\right),\nu\right)}{\partial m}, \\ \frac{\partial^2 d\left(\bar{m}_n\left(\theta,\gamma\right),\hat{\nu}\right)}{\partial m^2} &\rightarrow {}_P \frac{\partial^2 d\left(M\left(\theta,\gamma\right),\nu\right)}{\partial m^2}.\end{aligned} \tag{17}$$

Since d is twice continuously differentiable, it now follows that uniformly over θ,

$$\left|A_1\right| = O\left(\left\|\bar{m}-\hat{m}\right\|^3\right) = O_P\left(\left\|\hat{\gamma}_S-\gamma\right\|^3\right).$$

Using (17) in conjunction with A.4,

$$\left|A_2\right| = O\left(\left\|\bar{m}-\hat{m} - \bigtriangledown \bar{m}_n\left(\theta\right)\left[d\gamma\right] - \bigtriangledown^2 \bar{m}_n\left(\theta\right)\left[d\gamma, d\gamma'\right]\right\|\right) = O_P\left(\left\|\hat{\gamma}_S-\gamma\right\|^3\right),$$

$$\left|A_3\right| = O\left(\left\|\bar{m}-\hat{m} - \bigtriangledown \bar{m}_n\left(\theta\right)\left[d\gamma\right]\right\|^2\right) = O_P\left(\left\|\hat{\gamma}_S-\gamma\right\|^4\right).$$

This shows that the remainder term in the expansion of eq. (11) satisfies $R_{n,S} = O_P\left(\left\|\hat{\gamma}_S-\gamma\right\|^3\right)$.

The claimed rates for the first and second order differentials of $Q_n(\theta,\gamma)$ now follows by combining the results of Lemmas 5 and 6 with (17). Substituting the orders of the functional

derivatives just derived into eq. (11) we obtain

$$\sup_{\theta\in\Theta}|Q_n(\theta,\hat{\gamma}_S)-Q_n(\theta,\gamma)| = O_P\left(S^{-\alpha}\right)+O_P\left(S^{-\beta}\right)+O_P\left(S^{-\eta}\right)+O_P\left(n^{-\lambda/2}S^{-\alpha/2}\right)+O_P\left(n^{-\lambda/2}S^{-\beta}\right),$$

where we've left out the higher order terms. The claimed result now follows as a consequence of Lemma 1(i). ■

Proof of Theorem 3. Use Eq. (11) to write

$$\begin{aligned}\left\{Q_n\left(\theta,\hat{\gamma}\right)-\Delta_{n,S}\left(\theta\right)\right\}-Q_n\left(\theta,\gamma\right) &= \left\{\frac{1}{2}\bigtriangledown^2 Q_n(\theta)[\hat{\gamma}-\gamma,\hat{\gamma}-\gamma]-\bar{\Delta}_{n,S}\left(\theta\right)\right\} \\ &\quad +\left\{\bar{\Delta}_{n,S}\left(\theta\right)-\Delta_{n,S}\left(\theta\right)\right\} \\ &\quad +\bigtriangledown Q_n(\theta)[\hat{\gamma}-\gamma_0]+o_P\left(\left\|\hat{\gamma}-\gamma_0\right\|^2\right),\end{aligned} \tag{18}$$

where, with $w_{is}:=w_S\left(x_i,\varepsilon_{js};\theta\right)$, $\bar{w}_i:=E\left[w_S\left(x_i,\varepsilon;\theta\right)|x_i\right]$,

$$\bar{\Delta}_{n,S}\left(\theta\right)=\frac{1}{2S\left(S-1\right)}\sum_{j=1}^{J}\sum_{s=1}^{S}\bigtriangledown_{jj}^2 Q_n(\theta)[w_{js}-\bar{w}_j,w_{js}-\bar{w}_j].$$

The first term of eq. (18) satisfies

$$\begin{aligned}&\frac{1}{2}\bigtriangledown^2 Q_n(\theta)[\hat{\gamma}-\gamma_0,\hat{\gamma}-\gamma_0]-\bar{\Delta}_{n,S}\left(\theta\right) \\ =\;&\frac{1}{2}\bigtriangledown^2 Q_n(\theta)[b,b]+\frac{1}{2}\bigtriangledown^2 Q_n(\theta)[b,\psi]+\frac{1}{2}\sum_{j\neq k}\bigtriangledown_{j,k}^2 Q_n(\theta)[\psi_j,\psi_k] \\ =\;&O_P\left(S^{-2\beta}\right)+O_P\left(n^{-\lambda/2}S^{-2\beta}\right)+O_P\left(n^{-\lambda/2}S^{-\alpha}\right),\end{aligned}$$

where the last equality follows from the proof of Theorem 2.

Consider the second term of eq. (18): Use that $(d\gamma,d\gamma')\to\bigtriangledown_{jj}^2 Q_n(\theta)\left[d\gamma,d\gamma\right]$ is linear in

both $d\gamma$ and $d\gamma'$, and that $\hat{\gamma}_j = \sum_{s=1}^{S} w_{js}/S$, $i = 1, ..., n$, to obtain

$$\begin{aligned}
& \Delta_{n,S}(\theta) - \bar{\Delta}_{n,S}(\theta) \\
= & \frac{1}{2S(S-1)} \sum_{j=1}^{J} \sum_{s=1}^{S} \{2 \bigtriangledown_{jj}^2 Q_n(\theta,\gamma)[w_{js}, \hat{\gamma}_j] - 2 \bigtriangledown_{jj}^2 Q_n(\theta,\gamma)[w_{js}, \bar{w}_j] \\
& \qquad + \bigtriangledown_{jj}^2 Q_n(\theta,\gamma)[\bar{w}_j, \bar{w}_j] - \bigtriangledown_{jj}^2 Q_n(\theta,\gamma)[\hat{\gamma}_j, \hat{\gamma}_j]\} \\
= & \frac{1}{2(S-1)} \sum_{j=1}^{J} \{\bigtriangledown_{jj}^2 Q_n(\theta,\gamma)[\hat{\gamma}_j, \hat{\gamma}_j] - 2 \bigtriangledown_{jj}^2 Q_n(\theta,\gamma)[\hat{\gamma}_j, \bar{w}_j] \\
& \qquad + \bigtriangledown_{jj}^2 Q_n(\theta,\gamma)[\bar{w}_j, \bar{w}_j]\} \\
= & \frac{1}{2(S-1)} \sum_{j=1}^{J} \bigtriangledown_{jj}^2 Q_n(\theta,\gamma)[\hat{\gamma}_j - \bar{w}_j, \hat{\gamma}_j - \bar{w}_{ji}].
\end{aligned}$$

By A.4, the resulting expression on the right hand side satisfies

$$\begin{aligned}
E\left[\left|\frac{1}{2(S-1)} \sum_{j=1}^{J} \bigtriangledown_{jj}^2 Q_n(\theta,\gamma)[\hat{\gamma}_j - \bar{w}_j, \hat{\gamma}_j - \bar{w}_j]\right|\right] & \leq \left(\frac{1}{2(S-1)} \sum_{j=1}^{J} \bar{Q}_{n,ii}^{[2]}\right) \times \|\hat{\gamma} - \bar{w}\|^2 \\
& = O_P\left(S^{-(1-\alpha)}\right).
\end{aligned}$$

The rates of the third and fourth term of eq. (18) were derived in Theorem 2, and we obtain in total that:

$$\begin{aligned}
& |Q_n(\theta, \hat{\gamma}) - Q_n(\theta, \gamma) - \Delta_{n,S}(\theta)| \\
= & O_P\left(S^{-\beta}\right) + O_P\left(S^{-(1+\alpha)}\right) + O_P\left(S^{-3\alpha/2}\right) + O_P\left(n^{-\lambda/2} S^{-\alpha/2}\right) + O_P\left(n^{-\lambda/2} S^{-\beta}\right).
\end{aligned}$$

Applying Lemma 1, the claimed result follows. ■

Proof of Theorem 4. This is an immediate consequence of Theorems 2 and 3 in conjunction with Eq. (16). ■

B Lemmas

Lemma 5 *Under A.1-A.2 and A.4-A.5, the leading terms of the first and second order differentials of $\bar{m}_n$ are given below for approximators on the form (6):*

$$\bigtriangledown \bar{m}_n(\theta)[b_S] = S^{-\beta} E\left[\bigtriangledown m(z;\theta)\left[\bar{b}\right]\right] + n^{-1/2} S^{-\beta} \sqrt{\text{LRVar}\left(\bigtriangledown m(z;\theta)\left[\bar{b}\right]\right)},$$

$$\bigtriangledown \bar{m}_n(\theta)[\psi_S] \propto n^{-1/2} S^{-\alpha/2} \bar{M}_1 v(\theta),$$

$$\bigtriangledown^2 \bar{m}_n(\theta)[b_S, b_S] = S^{-2\beta} E\left[\bigtriangledown^2 m\left(z_i;\theta\right)\left[\bar{b},\bar{b}\right]\right] + n^{-1/2} S^{-2\beta}\sqrt{\text{LRVar}\left(\bigtriangledown^2 m\left(z;\theta\right)\left[\bar{b},\bar{b}\right]\right)},$$

$$\bigtriangledown^2 \bar{m}_n(\theta)[\psi_S, \psi_S] \propto S^{-\alpha}\bar{M}_2 v\left(\theta\right) + n^{-1/2} S^{-\alpha}\bar{M}_2 v^2\left(\theta\right).$$

Proof. First, consider the case where the approximator of $\bar{m}_n\left(\theta,\gamma\right)$ is on the form (6). Define

$$\begin{aligned}
\bigtriangledown \bar{m}_n(\theta)\left[d\gamma\right] &= \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown m\left(z_i;\theta\right)\left[d\gamma_i\right], \\
\bigtriangledown^2 \bar{m}_n(\theta)\left[d\gamma, d\gamma'\right] &= \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown^2 m\left(z_i;\theta\right)\left[d\gamma_i, d\gamma_i'\right],
\end{aligned}$$

for any $d\gamma = (d\gamma_1, ..., d\gamma_n)$ and $d\gamma' = (d\gamma_1', ..., d\gamma_n')$.

$$\bigtriangledown \bar{m}_n(\theta)[b_S] = S^{-\beta}\frac{1}{n}\sum_{i=1}^{n} \bigtriangledown m\left(z_i;\theta\right)\left[\bar{b}\right] + o\left(S^{-\beta}\right) \times \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown m\left(z_i;\theta\right)\left[1\right],$$

such that, using the stationarity and mixing assumption in (A.1),

$$E\left[\bigtriangledown \bar{m}_n(\theta)[b_S]\right] = S^{-\beta} E\left[\bigtriangledown m\left(z;\theta\right)\left[\bar{b}\right]\right] + o\left(S^{-\beta}\right),$$

and

$$\text{Var}\left(\bigtriangledown \bar{m}_n(\theta)[b_S]\right) = \frac{S^{-2\beta}}{n}\text{LRVar}\left(\bigtriangledown m\left(z;\theta\right)\left[\bar{b}\right]\right) + o\left(S^{-2\beta} n^{-1}\right).$$

Next,

$$E\left[\bigtriangledown \bar{m}_n(\theta)[\psi_S]|\mathcal{Z}_n\right] = \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown m\left(z_i;\theta\right)\left[E\left[\psi_{i,S}|z_i\right]\right] = 0,$$

and, using the independence assumption,

$$\begin{aligned}
\text{Var}\left(\bigtriangledown \bar{m}_n(\theta)[\psi_S]\right) &= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\left(\bigtriangledown m\left(z;\theta\right)\left[\psi_{i,S}\right]\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} E\left[\bigtriangledown m\left(z;\theta\right)\left[\psi_{i,S}\right]^2\right] \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n} \bar{M}_1 E\left[\left\|\psi_{i,S}\left(x\right)\right\|^2\right] \\
&= \frac{S^{-\alpha}}{n}\bar{M}_1 v\left(\theta\right) + o(n^{-1}S^{-\alpha}).
\end{aligned}$$

Moving on to the second order differential,

$$\bigtriangledown^2\bar{m}_n(\theta)[b_S,b_S]=S^{-2\beta}\frac{1}{n}\sum_{i=1}^{n}\bigtriangledown^2 m\left(z_i;\theta\right)\left[\bar{b},\bar{b}\right]+o\left(S^{-2\beta}\right)\times\frac{1}{n}\sum_{i=1}^{n}\bigtriangledown^2 m\left(z_i;\theta\right)[1,1],$$

where

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\bigtriangledown^2 m\left(z_i;\theta\right)\left[\bar{b},\bar{b}\right]\right]=E\left[\bigtriangledown^2 m\left(z_i;\theta\right)\left[\bar{b},\bar{b}\right]\right]$$

$$\begin{aligned}\mathrm{Var}\left(\bigtriangledown^2\bar{m}_n(\theta)[b_S,b_S]\right) &= S^{-4\beta}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\bigtriangledown^2 m\left(z_i;\theta\right)\left[\bar{b},\bar{b}\right]\right)+o\left(S^{-2\beta}n^{-1}\right)\\ &= \frac{S^{-4\beta}}{n}\mathrm{LRVar}\left(\bigtriangledown m\left(z;\theta\right)\left[\bar{b},\bar{b}\right]\right)+o\left(S^{-2\beta}n^{-1}\right)\end{aligned}$$

Similarly,

$$E\left[\bigtriangledown^2\bar{m}_n(\theta)[\psi_S,\psi_S]\right]=\frac{1}{n}\sum_{i=1}^{n}E\left[\bigtriangledown^2 m\left(z_i;\theta\right)\left[\psi_{i,S},\psi_{i,S}\right]\right],$$

where

$$\left|E\left[\bigtriangledown^2 m\left(z_i;\theta\right)\left[\psi_{i,S},\psi_{i,S}\right]\right]\right|\leq S^{-\alpha}\bar{M}_2 v\left(\theta\right)+o\left(S^{-\alpha}\right),$$

and

$$\begin{aligned}\mathrm{Var}\left[\bigtriangledown^2\bar{m}_n(\theta)[\psi_S,\psi_S]\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(\bigtriangledown^2 m\left(z_i;\theta\right)\left[\psi_{i,S},\psi_{i,S}\right]\right)\\ &\leq \frac{1}{n}E\left[\bigtriangledown m\left(z;\theta\right)\left[\psi_S,\psi_S\right]^2\right]\\ &\leq \frac{1}{n}\bar{M}_1 E\left[\left\|\psi_S\left(x\right)\right\|^2\right]^2\\ &= \frac{S^{-2\alpha}}{n}\bar{M}_1 v^2\left(\theta\right)+o(n^{-1}S^{-2\alpha}).\end{aligned}$$

■

Lemma 6 *Under A.1-A.2 and A.4-A.5, the leading terms of the first and second order differentials of $\bar{m}_n$ are given below for approximators on the form (5):*

$$\bigtriangledown\bar{m}_n(\theta)[b_S]=S^{-\beta}E\left[\bigtriangledown m\left(z;\theta\right)\left[\bar{b}\right]\right]+n^{-1/2}S^{-\beta}\sqrt{\mathrm{LRVar}\left(\bigtriangledown m\left(z;\theta\right)\left[\bar{b}\right]\right)},$$

$$\bigtriangledown\bar{m}_n(\theta)[\psi_S]=\bigtriangledown m_1\left(\psi_S;\theta\right)+O_P\left(n^{-1/2}S^{-\alpha/2}\right),$$

$$\bigtriangledown^2\bar{m}_n(\theta)[b_S,b_S]=S^{-2\beta}E\left[\bigtriangledown^2 m\left(z_i;\theta\right)\left[\bar{b},\bar{b}\right]\right]+n^{-1/2}S^{-2\beta}\sqrt{\mathrm{LRVar}\left(\bigtriangledown^2 m\left(z;\theta\right)\left[\bar{b},\bar{b}\right]\right)},$$

$$\bigtriangledown^2 \bar{m}_n(\theta)[\psi_S, \psi_S] \propto S^{-\alpha} \bar{M}_2 v(\theta) + O_P\left(n^{-1/2} S^{-\alpha}\right)$$

where

$$\bigtriangledown m_1(\psi_S; \theta) = \int \bigtriangledown m(z; \theta)[\psi_S] \, dF(z). \tag{19}$$

Proof. When the approximation of $\bar{m}_n(\theta, \gamma)$ is on the form (6), the functional differentials are given by

$$\begin{aligned} \bigtriangledown \bar{m}_n(\theta)[d\gamma] &= \frac{1}{n} \sum_{i=1}^{n} \bigtriangledown m(z_i; \theta)[d\gamma], \\ \bigtriangledown^2 \bar{m}_n(\theta)\left[d\gamma, d\gamma'\right] &= \frac{1}{n} \sum_{i=1}^{n} \bigtriangledown^2 m(z_i; \theta)\left[d\gamma, d\gamma'\right], \end{aligned}$$

with $d\gamma$ and $d\gamma'$ only having one component. It is easily seen that the bias components are the same as in the multi-batch case, and so we only consider the variance components. Here,

$$E\left[\bigtriangledown \bar{m}_n(\theta)[\psi_S] | \mathcal{Z}_n\right] = \frac{1}{n} \sum_{i=1}^{n} \bigtriangledown m(z_i; \theta)\left[E\left[\psi_S | z_i\right]\right] = 0.$$

With $\bigtriangledown m_1(\psi_S; \theta)$ given in eq. (19), write

$$\bigtriangledown \bar{m}_n(\theta)[\psi_S] = \bigtriangledown m_1(\psi_S; \theta) + \frac{1}{n} \sum_{i=1}^{n} \left[\bigtriangledown m(z_i; \theta)[\psi_S] - \bigtriangledown m_1(\psi_S; \theta)\right],$$

where

$$E\left[\frac{1}{n} \sum_{i=1}^{n} \left[\bigtriangledown m(z_i; \theta)[\psi_S] - \bigtriangledown m_1(\psi_S; \theta)\right]\right] = 0,$$

and

$$\begin{aligned} & \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} \left[\bigtriangledown m(z_i; \theta)[\psi_S] - \bigtriangledown m_1(\psi_S; \theta)\right]\right) \\ = \; & E\left[\mathrm{Var}\left[\frac{1}{n} \sum_{i=1}^{n} \left[\bigtriangledown m(z_i; \theta)[\psi_S] - \bigtriangledown m_1(\psi_S; \theta)\right] \middle| \mathcal{E}_S\right]\right] \\ = \; & \frac{1}{n} E\left[\mathrm{LRVar}\left(\bigtriangledown m(z_i; \theta)[\psi_S] - \bigtriangledown m_1(\psi_S; \theta) | \, \mathcal{E}_S\right)\right] \\ = \; & O\left(\frac{S^{-\alpha}}{n}\right). \end{aligned}$$

In total,

$$\bigtriangledown \bar{m}_n(\theta)[\psi_S] = \bigtriangledown m_1(\psi_S; \theta) + O_P\left(\sqrt{\frac{S^{-\alpha}}{n}}\right).$$

Regarding the second order differential, define

$$\begin{aligned}\bigtriangledown^2 m_1(z;\theta) &= \int \bigtriangledown^2 m(z;\theta)[\psi_S,\psi_S]\,dF(\psi_S),\\ \bigtriangledown^2 m_2(\psi_S;\theta) &= \int \bigtriangledown^2 m(z;\theta)[\psi_S,\psi_S]\,dF(z),\end{aligned}$$

and

$$\bigtriangledown^2 \bar{m}(\theta) = \int\int \bigtriangledown^2 m(z;\theta)[\psi_S,\psi_S]\,dF(z)\,dF(\psi_S) = E\left[\bigtriangledown^2 m(z;\theta)[\psi_S,\psi_S]\right].$$

Then write

$$\begin{aligned}\bigtriangledown^2 \bar{m}_n(\theta)[\psi_S,\psi_S] &= \frac{1}{n}\sum_{i=1}^{n}\bigtriangledown^2 m(z_i;\theta)[\psi_S,\psi_S]\\ &= \bigtriangledown^2\bar{m}(\theta) + \frac{1}{n}\sum_{i=1}^{n}\left[\bigtriangledown^2 m_1(z_i;\theta) - \bigtriangledown^2\bar{m}(\theta)\right] + \left[\bigtriangledown^2 m_2(\psi_S;\theta) - \bigtriangledown^2\bar{m}(\theta)\right]\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left[\bigtriangledown^2 m(z_i;\theta)[\psi_S,\psi_S] - \bigtriangledown^2 m_1(z_i;\theta) - \bigtriangledown^2 m_2(\psi_S;\theta) + \bigtriangledown^2\bar{m}(\theta)\right]\end{aligned}$$

where, by the same arguments as before, $\left|\bigtriangledown^2\bar{m}(\theta)\right| = O\left(S^{-2\alpha}\right)$,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\bigtriangledown^2 m_1(z_i;\theta) - \bigtriangledown^2\bar{m}(\theta)\right] = O_P\left(\frac{S^{-2\alpha}}{\sqrt{n}}\right),$$

$$\bigtriangledown^2 m_2(\psi_S;\theta) - \bigtriangledown^2\bar{m}(\theta) = O_P\left(S^{-2\alpha}\right),$$

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left[\bigtriangledown^2 m(z_i;\theta)[\psi_S,\psi_S] - \bigtriangledown^2 m_1(z_i;\theta) - \bigtriangledown^2 m_2(\psi_S;\theta) + \bigtriangledown^2\bar{m}(\theta)\right]\\ =\ &\frac{1}{\sqrt{n}}\left\{\text{LRVar}\left(\bigtriangledown^2 m(z_i;\theta)[\psi_S,\psi_S] - \bigtriangledown^2 m_1(z_i;\theta) - \bigtriangledown^2 m_2(\psi_S;\theta) + \bigtriangledown^2\bar{m}(\theta)\right) + o_P(1)\right\}\\ =\ &O_P\left(\frac{S^{-2\alpha}}{\sqrt{n}}\right).\end{aligned}$$

■

Lemma 7 *Assume that $\{W_t\}$ is a stationary and α-mixing sequence with mixing coefficients $\alpha(m)$, $m = 1,2,...$, that satisfy $\alpha(m) \leq Am^{-\beta}$ for some $A,\beta > 0$. Also, assume that $E[W_t] = 0$, $E\left[\|W_t\|^{2r+\delta}\right] < \infty$ for some $r,\delta > 0$. Then, there exists a constant $C(r) < \infty$*

such that:

$$E\left[\left\|\frac{1}{T}\sum\nolimits_{t=1}^{T} W_t\right\|^{2r}\right] \leq C(r)\, E\left[\|W_t\|^{2r+\delta}\right] T^{-q},$$

where

$$q = \frac{r}{2r/a+1}, \quad a = \frac{\beta\delta}{2r+\delta}.$$

Proof. From Hahn and Kuersteiner (2004, Lemma 7), there exists a constant $C(r) < \infty$ such that for any $1 \leq m \leq C(r)\, T$:

$$E\left[\left\|\frac{1}{T}\sum\nolimits_{t=1}^{T} W_t\right\|^{2r}\right] \leq C(r)\, E\left[\|W_t\|^{2r+\delta}\right]\left[T^{-r}m^{2r} + Am^{-a}\right].$$

The claimed result now follows by choosing $m = T^{-\gamma}$ with $\gamma = ra/(2r+a)$. ■