Columbia University

Department of Economics
Discussion Paper Series

Higher Order Improvements for Approximate Estimators

Dennis Kristensen
Bernard Salanié

Discussion Paper No.: 0910-15

Department of Economics
Columbia University
New York, NY 10027

May 2010

Higher Order Improvements for Approximate Estimators[*]

Dennis Kristensen[†]
Columbia University and CREATES[‡]

Bernard Salanié[§]
Columbia University

April 20, 2010

Abstract

Many modern estimation methods in econometrics approximate an objective function, through simulation or discretization for instance. The resulting "approximate" estimator is often biased; and it always incurs an efficiency loss. We here propose three methods to improve the properties of such approximate estimators at a low computational cost. The first two methods correct the objective function so as to remove the leading term of the bias due to the approximation. One variant provides an analytical bias adjustment, but it only works for estimators based on stochastic approximators, such as simulation-based estimators. Our second bias correction is based on ideas from the resampling literature; it eliminates the leading bias term for non-stochastic as well as stochastic approximators. Finally, we propose an iterative procedure where we use Newton-Raphson (NR) iterations based on a much finer degree of approximation. The NR step removes some or all of the additional bias and variance of the initial approximate estimator. A Monte Carlo simulation on the mixed logit model shows that noticeable improvements can be obtained rather cheaply.

[*] We would like to thank participants at the CAM Research Workshop 2008, the Econometric Society Latin-American Meeting 2009, and seminars at Columbia, Northwestern, Ohio State, Toulouse, UCL and Yale for helpful comments and suggestions.

[†] E-mail: dk2313@columbia.edu. Kristensen acknowledges financial support from the Danish Social Sciences Research Council (through CREATES), the National Science Foundation (grant no. SES-0961596), and the Velux Foundation. Parts of this research was conducted while Kristensen visited Princeton University and University of Copenhagen whose hospitality is gratefully acknowledged.

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

[§] E-mail: bs2237@columbia.edu. Parts of this paper were written while Salanié was visiting the Toulouse School of Economics; he gratefully acknowledges the Georges Meyer endowment for its support.

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, and 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), Duffie and Singleton (1993)) and in simulated pseudo-maximum likelihood (Laroque and Salanié (1989, 1993, 1994)). It may also be an integrated density/cdf, as in simulated maximum likelihood (Lee (1992, 1995)) and in some testing procedures (Corradi and Swanson (2007)).[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 (2009); Creel and Kristensen (2009); 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 would be obtained with an infinite number of simulations; in dynamic programming models it would rely 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. When the approximation error is unbiased and the objective function is linear in the approximation error, then using approximations does not create additional bias, although it deteriorates efficiency: a case in point is the simulated method of moments (SMM). In all other cases, approximation creates both a bias and a loss of efficiency. These 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

[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.

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 findings encompass most known results in the literature on simulation-based estimators such as Lee (1995, 1999), Gouriéroux and Monfort (1996) and Laroque and Salanié (1989).

Second, we propose three methods to improve on the precision of approximate estimators. Each of these methods only carries a small additional computation burden. The first method is targeted at a class of estimators that includes most 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 approximator in general leads to an additional bias component in the approximate estimator relative to the exact one[2]. This point is well-known; our contribution is mainly to derive a general formula for the additional bias and variance of the approximate estimator, and to build upon our asymptotic expansions in order to correct the objective function and eliminate the leading term of the additional bias. 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 $1/S^{3/2}$ at most, which can be a considerable improvement (applications typically use $S = 50$ to 500 or even more simulations, so that the bias should be reduced by a factor of ten at least.)

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.

The second method is a more general bias correction procedure. We show that the leading term of the additional bias in an approximate estimator based an an approximator of quality S (say, S simulations) can also be removed by subtracting from the objective function an average of similar objective functions computed with smaller values of S. This is in the spirit of the parametric bootstrap and the jackknife. It applies equally well to stochastic and non-stochastic approximators, although the terms to be subtracted differ.

Finally, our third proposed improvement is a two-step method which applies quite generally. In the first step, we compute the approximate estimator, using an approximator that

[2] As explained above, the simulated method of moments is exempt from this additional bias.

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.

We should stress that as our first and second method aim at reducing the bias that approximation imparts on estimators, they are not meant to be useful for SMM estimators. On the other hand, our third method can be used to improve efficiency as well as to reduce bias, and so it is applicable to SMM estimation.

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, e.g. an expectation or a density. We use some of the tools that are applied in that setting; see for example Andrews (1994) and Chen et al. (2003). 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 in panel models with fixed effects.

Our results are also 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.

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

The paper is organised as follows: Section 2 presents our framework and informally introduces the 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 approxima-

[3] Hajivassiliou (2000) considered a somewhat similar idea, where Newton-Raphson step based on the exact likelihood function were used to improve the efficiency of a first-step simulated method of moments estimator.

tions. The properties of the Newton-Raphson method are derived in Section 5. Section 6 discusses our third approach, based on the resampling literature. Finally, section 7 presents the results of a Monte Carlo simulation study, using the mixed logit model as an example. Several examples and proofs have been relegated to appendices A and B.

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_0 using some extremum estimator,

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta, \gamma_0), \tag{1}$$

for some objective function

$$Q_n(\theta, \gamma) = Q_n(\mathcal{Z}_n, \theta, \gamma_0(\cdot, \theta)).$$

The objective function depends on data, a finite dimensional parameter θ and a (usually) infinite-dimensional one, some function $\gamma_0(z, \theta)$.

Our paper focuses on the common case when the true function γ is not known on closed form to the econometrician, and instead it has to be approximated numerically. In this case, a feasible estimator is obtained by minimizing the analog approximate objective function

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

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"of γ_0. 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

$$E\left[g(z, \theta_0)\right] = 0. \tag{3}$$

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_0 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

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

where $\mathbb{I}\left\{\cdot\right\}$ denotes the indicator function, and $Z(x)$ are a set of instruments. The probability that an individual chooses k, $\gamma_{0,k}(z, \theta) = \Pr(y = k|x; \theta)$, can be approximated by drawing S choice errors for i and counting the proportion of draws for which choice k brings highest utility,

$$\hat{\gamma}_{k,S}(z, \theta) = S^{-1}\sum_{s=1}^{S} \mathbb{I}\left\{y_s(x, \theta) = k\right\}$$

where the $y_s(x, \theta)$ 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 $g(z_t, z_{t-1}, \theta_0) = 0$, with

$$g(z_t, z_{t-1}, \theta) = w(z_t, z_{t-1}) - E_\theta\left[w(z_t, z_{t-1})\right]$$

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 $\gamma_0(\theta) = E_\theta\left[w(z_t, z_{t-1})\right]$ 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 $\gamma_0(x, \theta) = E_\theta\left[\phi(y_t)|x_{t-1} = x\right]$. 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,

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

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,

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

where, for some function w and some unobserved error ε,

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

Defining $\gamma_0(x;\theta) = m(x;\theta)$, our exact objective function takes the form

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

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

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

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 $\gamma_0 = (m,v)$ where $m(x;\theta)$ 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:

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

Again, in many situations $\gamma_0(x;\theta)$ 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 another 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, $p\left(y|x;\theta\right)$. The natural choice is the maximum-likelihood estimator,

$$Q_n(\theta,\gamma_0) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\gamma_0\left(y_i,x_i;\theta\right)\right),$$

where $\gamma_0\left(y,x;\theta\right) := p\left(y|x;\theta\right)$. Sometimes the density γ_0 cannot be written in closed form. For example, in models with unobserved heterogeneity,

$$\gamma_0\left(z;\theta\right) = \int w\left(y|x,\varepsilon;\theta\right) f\left(\varepsilon\right) d\varepsilon,$$

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 $\hat{\gamma}_S\left(z;\theta\right) = S^{-1}\sum_{s=1}^{S} w\left(y|x,\varepsilon_s;\theta\right)$.

More recently, Fermanian and Salanié (2004) proposed using a kernel estimator as an approximator. The idea is simple, and it applies quite generally. Suppose that data (y_i,x_i), $i=1,...,n$, has been generated by $y=r(x,\varepsilon;\theta_0)$, with implied conditional density $\gamma_0\left(y,x;\theta\right) = p(y|x,\theta_0)$. 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:

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

Maximizing the approximate likelihood in which $\hat{\gamma}_S$ replaces γ defines the nonparametric simulated maximum likelihood estimator (NPSML). It has 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 (2009), Brownless, Kristensen and Shin (2019) and Kristensen and Shin (2008).

We now turn to three 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:

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

Here the function γ_0 is the unknown value function V. 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", and assume that $V_{S,t+1}(., \theta)$ has been evaluated on this grid. Then the backward recursion evaluates for $k = 1, \ldots, S$,

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

In this formula, the symbol $\hat{E}_S$ is meant to represent a numerical approximation of the conditional expectation of $V_{S,t+1}(s_{t+1}, \theta)$ 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 (2009) for an example of a specific approximation method for discrete choice models.

Example 5: Nested fixed-point algorithms. Fixed-point algorithms have found many applications in the estimation of structural IO models after Berry, Levinsohn and Pakes (1995). Here market shares are modelled as functions of unobserved and observed characteristics, $share = s(\xi, z; \theta)$ for some function s where ξ and z respectively denote unobserved and observed characteristics. The BLP procedure requires that the econometrician compute the unobserved product characteristics given observed market shares; this involves inverting the market share function in its first argument, $\xi(share, z; \theta) = s^{-1}(share, z; \theta)$. Since s^{-1} is normally not available on closed form, this is usually performed using a numerical fixed-point algorithm. It leads to an approximate solution, $\xi_S(share, z; \theta)$, where S captures the number of iterations and/or the tolerance level used in the algorithm[4].

Example 6: 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, as shown by Fernández-Villaverde, Rubio-Ramirez and Santos

[4] Some more recent implementations use mathematical programming under equilibrium constraints, as advocated by Judd and Su (2007).

(2006).

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. The mean of course is an unbiased estimator of the expectation; but in many simulation-based estimation methods the objective function depends nonlinearly on the simulated mean, 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 the approximator simulates an expectation through a simple average. 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 still is of order $1/S$.

Similarly, non-stochastic approximations lead to deteriorations of the properties of the resulting estimators. Take the problem of computing the density $p\left(y|x;\theta\right)$ in Example 3 for instance. If the dimensionality of the integration variable (ε) is small, then instead of simulations the numerical integration may be done by an S point Gaussian quadrature, as in Lee (2001). As demonstrated in the next section, the resulting approximate estimator will suffer from additional biases relative to the exact one.

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 fast enough, 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 $S = S\left(n\right) \to \infty$, and there is no first-order difference between the exact and the approximate estimator. However, in practice S may need to be quite large before this result can apply; and the resulting computations may become prohibitively costly. Our proposed methods yield estimators that may be just as efficient as large-S approximate estimators, and yet are computationally much less burdensome.

We take as starting point the approximate estimator defined in eq. (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 two methods correct the objective function so as to obtain an estimator with better bias properties. Instead of selecting $\hat{\theta}_{n,S}$ to minimize $Q_n(\theta, \hat{\gamma}_{n,S})$, we select

$$\hat{\theta}^{\mathrm{b}}_{n,S} = \arg\min_{\theta \in \Theta} \left\{ Q_n(\theta, \hat{\gamma}_S) - \Delta_{n,S}(\theta) \right\}, \tag{4}$$

where $\Delta_{n,S}(\theta)$ corrects for at least the leading term of the approximation bias. Sections 4 and 5 present two approaches to computing this Δ term.

This first approach is an analytical bias adjustment that works for all known simulation-based estimators. In the context of SNLS, it boils down to the adjustment proposed in Laffont et al. (1995) (also see Laroque and Salanié (1989, 1993); Bierings and Sneek (1989)); and for SML of discrete choice models, it yields the adjustment in Lee (1995). These papers derived an unbiased and consistent estimator of the leading bias component due to simulations. We extend their result to general simulation-based estimators and show how to compute $\Delta_{n,S}(\theta)$. We note that SNLS is a quite special and favorable case, as the objective function is only quadratic in the simulated mean such that $\Delta_{n,S}(\theta)$ adjusts for all biases due to simulations. In general using $\Delta_{n,S}(\theta)$ will only correct for the leading term of the bias when using stochastic approximation. This is for example the case in SML.

Our second proposal is an alternative to the analytic bias adjustment and works for both stochastic and non-stochastic approximators. The corrected estimator is defined as in equation (4), but the adjustment term $\Delta_{n,S}(\theta)$ is constructed in a different manner, more closely related to the jackknife bias adjustment. To illustrate, suppose that

$$E\left[Q_n(\theta, \hat{\gamma}_S) - Q_n(\theta, \gamma)\right] = \frac{B(\theta)}{S} + o\left(S^{-1}\right).$$

Now take two independent approximators $\hat{\gamma}^{[1]}_{S/2}$ and $\hat{\gamma}^{[2]}_{S/2}$ of order $S/2$. For each approximator $m = 1, 2$, we can define a corresponding objective function based on data and on the approximator, $Q_n(\theta, \hat{\gamma}^{[m]}_{S/2})$. We then define the adjustment as

$$\Delta_{n,S}(\theta) = \frac{1}{4}\left[Q_n(\theta, \hat{\gamma}^{[1]}_{S/2}) + Q_n(\theta, \hat{\gamma}^{[2]}_{S/2})\right].$$

Then the adjusted objective function satisfies

$$E\left[\left\{Q_n(\theta, \hat{\gamma}_S) - \Delta_{n,S}(\theta)\right\} - Q_n(\theta, \gamma)\right] = \frac{B(\theta)}{S} - \frac{1}{4}\left[\frac{2B(\theta)}{S} + \frac{2B(\theta)}{S}\right] + o\left(S^{-1}\right) = o\left(S^{-1}\right),$$

so that the leading term cancels out. We provide details in section 5. Note that the above argument does not require the approximators to be independent. However, for the variance of the bias corrected estimator not to increase too much, it is desirable to choose $\hat{\gamma}^{[m]}_{S_m}$ to be independent.

Our third 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 $\hat{\theta}_n^{NR} = \hat{\theta}_n - \left[\partial^2 L(\hat{\theta}_n)/\partial\theta\partial\theta'\right]^{-1} \partial L_n(\hat{\theta}_n)/\partial\theta$ 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 $\hat{\theta}_{n,S}^b$ of (4). 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 $\hat{\theta}_{n,S}^b$ 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 uses a much finer approximation, $S^* \gg S$. Denote

$$G_n(\theta, \gamma) = \frac{\partial Q_n}{\partial\theta}(\theta, \gamma) \text{ and } H_n(\theta, \gamma) = \frac{\partial^2 Q_n}{\partial\theta\partial\theta'}(\theta, \gamma);$$

and define

$$\hat{\theta}_{n,S}^{(k+1)} = \hat{\theta}_{n,S}^{(k)} - H_n^{-1}(\hat{\theta}_{n,S}^{(k)}, \hat{\gamma}_{S^*}) G_n(\hat{\theta}_{n,S}^{(k)}, \hat{\gamma}_{S^*}), \quad k = 1, 2, 3, ... \tag{5}$$

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 minimization of $Q_n(\theta, \hat{\gamma}_{S^*})$. Since the minimization 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 6.

3 Properties of Approximate Estimators

Before we come to our proposed 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 enable us to identify the leading bias and viarance terms that we wish to adjust for, and evaluate the improvements from these adjustments. In order to establish the establish the expansion formally, we need to make assumptions both on the estimating equation and on the approximators.

[5] In many cases, a large part of the dimensionality of θ only comes into play within some linear indexes $\theta' x$; then the trade off is even more favourable since the computation of the second derivative H_n is much simplified.

3.1 The Estimating Equation

We restrict our attention to estimators $\hat{\theta}_n$ that (asymptotically) satisfy a first order condition of the form

$$G_n(\hat{\theta}_n, \gamma_0) = o_P\left(1/\sqrt{n}\right),$$

while the approximate estimator, $\tilde{\theta}_n = \hat{\theta}_{n,S}$, satisfies

$$G_n(\hat{\theta}_{n,S}, \hat{\gamma}_S) = o_P\left(1/\sqrt{n}\right).$$

Furthermore, we assume that $G_n\left(\theta, \gamma\right)$ takes the form of a sample average,

$$G_n\left(\theta, \gamma\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(z_i; \theta, \gamma\right). \tag{6}$$

Our setup allows for two-step GMM estimators where the weight matrix has been estimated. In the following we shall assume that $G_n\left(\theta, \gamma\right)$ is a smooth function in θ and γ which rules out estimators minimizing non-differentiable objective functions. We conjecture that our results could be generalized to this class of estimators by combining our approach with the results of, for example, Newey and McFadden (1994, Section 7) and Pollard (1985).

The above framework includes all of the examples described in Section 2. When the estimator is defined by (1) we may choose

$$G_n\left(\theta, \gamma\right) = \frac{\partial Q_n}{\partial \theta}\left(\theta, \gamma\right).$$

For example, with $Q_n\left(\theta, \gamma\right) = n^{-1}\sum_{i=1}^{n} q\left(z_i; \theta, \gamma\right)$, we have $g\left(z_i; \theta, \gamma\right) = \partial q\left(z_i; \theta, \gamma\right) / \left(\partial \theta\right)$. In the case of GMM estimators where $Q_n\left(\theta, \gamma\right) = M_n(\theta, \gamma) W_n M_n(\theta, \gamma)$ with $W_n \to^P W$ and $M_n(\theta, \gamma) = \frac{1}{n}\sum_{i=1}^{n} m\left(z_i; \theta, \gamma\right)$, we may choose $g\left(z_i; \theta, \gamma\right) = Wm\left(z_i; \theta, \gamma\right)$ since this is (asymptotically) equivalent to $g_n\left(z_i; \theta, \gamma\right) = W_n m\left(z_i; \theta, \gamma\right)$.

Our estimation problem is very similar to two-step semiparametric estimation where in the first step a (possibly infinite-dimensional) nuisance parameter (γ_0) is replaced by its estimator (the approximator γ_S), which in turn is used to obtain an estimator $\hat{\theta}_S$ of θ_0; see, for example, Andrews (1994) and Chen et al (2003).

We assume that the function of interest $\gamma_0 : \mathcal{Z} \times \Theta \mapsto \mathbb{R}^p$ belongs to a function space Γ equipped with a norm $\|\cdot\|$. In most cases, the norm will be the L_q-norm induced by the probability measure associated with our observations, $\|\gamma\| = E\left[\|\gamma\left(z\right)\|^q\right]^{1/q}$ for some $q \geq 1$. We also assume that the objective functions are smooth functionals of θ and γ, and introduce the first-order derivative of $G_n(\theta,\gamma)$ w.r.t. θ,

$$H_n(\theta,\gamma)=\frac{1}{n}\sum_{i=1}^{n}h(z_i;\theta,\gamma), \quad \text{with} \quad h(z_i;\theta,\gamma)=\frac{\partial g}{\partial\theta}(z_i;\theta,\gamma),$$

and their corresponding population versions,

$$G(\theta,\gamma)=E[g(z_i;\theta,\gamma)], \quad H(\theta,\gamma)=E\left[\frac{\partial g(z_i;\theta,\gamma)}{\partial\theta}\right].$$

We first impose conditions to ensure that the exact, but infeasible estimator and its approximate version are both well-behaved:

A.1 $\{z_i\}$ is stationary and geometrically α-mixing.

A.2 The parameter space Θ is compact and θ_0 is in its interior.

A.3 (i) The function $g(z;\theta,\gamma)$ is continuous in $\theta\in\Theta$, $E[\sup_{\theta\in\Theta}\|g(z_i;\theta,\gamma_0)\|]<\infty$ and (ii) $G(\theta,\gamma_0)=0$ if and only if $\theta=\theta_0$.

A.4 For all (θ,γ) in a neighbourhood of (θ_0,γ_0):

(a) $g(z;\theta,\gamma)$ is continuously differentiable w.r.t. θ, and its derivative, $h(z;\theta,\gamma)$, is continuous in $\theta\in\Theta$,

(b) For some $\delta>0$,

$$E\left[\sup_{\|\theta-\theta_0\|<\delta}\|h(z_i;\theta,\gamma_0)\|\right]<\infty$$

(c) $H_0:=H(\theta_0,\gamma_0)$ is positive definite,

(d) for some $\delta,\lambda,\bar{H}>0$,

$$E\left[\sup_{\|\theta-\theta_0\|<\delta}\|h(z_i;\theta,\gamma)-h(z_i;\theta,\gamma_0)\|\right]\leq\bar{H}\|\gamma-\gamma_0\|^{\lambda}.$$

Assumption A.1 rules out strongly persistent data, and allows us to obtain standard rates of convergence for the resulting estimators. The geometric mixing condition could be weakened, but this would lead to more complicated results; we refer the reader to Kristensen and Shin (2008) for results on strongly persistent and/or non-stationary data (and thereby estimators with non-standard rates.)

The second assumption, A.2, is standard in the asymptotic analysis of extremum estimators, while A.3 ensures that a uniform law of large numbers hold for $G_n(\theta,\gamma)$ and that θ_0 is identified. Primitive conditions for the uniform moment condition in A.3 to hold in a cross-sectional setting can be found in Newey and McFadden (1994).

Finally, A.4 imposes additional smoothness conditions on $g(z;\theta,\gamma)$ for $\gamma \neq \gamma_0$. In particular, when γ depends on θ (as is the case for all of our examples), it requires the approximator to be a smooth function of θ. Therefore A.4 rules out discontinuous and non-differentiable approximators such as the simulated method of moment estimators for discrete choice models proposed in McFadden (1989) and Pakes and Pollard (1989), as the approximate moment conditions for these models involve indicator functions.[6] The Lipschitz condition imposed on $h(z;\theta,\gamma')$ is used to ensure that $H_n(\theta,\hat{\gamma}_S) \to^P H(\theta,\gamma)$ uniformly in θ as $\hat{\gamma}_S \to^P \gamma$.

Under the additional assumption that $E\left[\|g(z_i;\theta_0,\gamma_0)\|^2\right] < \infty$, conditions A.1-A.4 imply that $\hat{\theta}_n$ has standard "sandwich" asymptotics,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to^d N\left(0, H_0^{-1}\Omega H_0^{-1}\right).$$

Our higher-order results will rely on a functional expansion of $G_n(\theta,\gamma)$ w.r.t. γ. To take a finite-dimensional analogy, we would like to be able to use a Taylor expansion,

$$G_n(\theta,\hat{\gamma}_S) = G_n(\theta,\gamma_0) + \frac{\partial G_n(\theta,\gamma_0)}{\partial\gamma'}(\hat{\gamma}_S - \gamma) + \frac{1}{2}(\hat{\gamma}_S - \gamma)'\frac{\partial^2 G_n(\theta,\gamma_0)}{\partial\gamma\partial\gamma'}(\hat{\gamma}_S - \gamma) + o_P\left(\|\hat{\gamma}_S - \gamma\|^2\right).$$

Then we can use our knowledge of the properties of the approximators $\hat{\gamma}_S$ to bound the difference between approximate and exact estimating equation, and finally to characterize the difference between approximate and exact estimators. For such an expansion to be well-defined and for the individual terms in the expansion to be well-behaved, we need to impose some further regularity conditions on $g(z_i;\theta_0,\gamma)$ as a functional of γ; and since our γ's are not vectors but functions, the notation will be somewhat more involved.

In all of the following, $\Delta\gamma \in \Gamma$ denotes a small change around γ_0.

A.5(m) Assume that for some $0 < \delta, \lambda, \bar{G}_0 < \infty$, the following hold:

(i)

$$E\left[\sup_{\|\Delta\gamma\|\leq\delta}\sup_{\theta\in\Theta}\|g(z;\theta,\gamma_0+\Delta\gamma) - g(z;\theta,\gamma_0)\|\right] \leq \bar{G}_0\|\Delta\gamma\|^{\lambda}. \tag{7}$$

(ii) There exist functionals $\nabla^k g(z;\theta)[d\gamma_1,...,d\gamma_k]$, $k = 1,...,m$, which are linear in each component $d\gamma_k \in \Gamma$ such that for some $K_0 < \infty$:

$$E\left[\sup_{\|\theta-\theta_0\|\leq\delta}\left\|g(z;\theta,\gamma_0+\Delta\gamma) - g(z;\theta,\gamma_0) - \sum_{k=1}^{m}\frac{1}{k!}\nabla^k g(z;\theta)[\Delta\gamma,...,\Delta\gamma]\right\|\right] \leq \bar{G}_0\|\Delta\gamma\|^{m+1},. \tag{8}$$

[6] These cases can be handled by introducing a smoothed version of the approximators in the spirit of Fermanian and Salanié (2004).

Furthermore, for $k = 2, ..., m$,

$$\begin{aligned} E\left[\sup_{\|\theta-\theta_0\|\leq\delta} \|\nabla g\left(z;\theta\right)\left[\Delta\gamma\right]\|^2\right] &\leq \bar{G}_1 \|\Delta\gamma\|^2, & (9) \\ E\left[\sup_{\|\theta-\theta_0\|\leq\delta} \left\|\nabla^k g\left(z;\theta\right)\left[\Delta\gamma_1, ..., \Delta\gamma_k\right]\right\|^{2+\delta}\right] &\leq \bar{G}_2 \left(\|\Delta\gamma_1\| \cdots \|\Delta\gamma_k\|\right)^{(2+\delta)}. & (10) \end{aligned}$$

Assumption A.5 restricts $g\left(z;\theta,\gamma\right)$ to be m times pathwise differentiable w.r.t. γ with differentials $\nabla^k g\left(z;\theta\right)\left[d\gamma_1, ..., d\gamma_k\right]$, $k = 1, ..., m$. These differentials are required to be Lipchitz in $d\gamma_1, ..., d\gamma_k$. For a given choice of m, this allows us to use an mth order expansion of $G_n\left(\theta,\gamma\right)$ w.r.t. γ to evaluate the impact of $\hat{\gamma}_S$. In particular, the difference between the approximate and exact objective function can be written as

$$G_n(\theta, \hat{\gamma}_S) - G_n(\theta, \gamma_0) = \sum_{k=1}^{m} \frac{1}{k!} \triangledown^k\, G_n(\theta)[\hat{\gamma}_S - \gamma, ..., \hat{\gamma}_S - \gamma] + R_{n,S}, \qquad (11)$$

where $R_{n,S} = O_P(\|\hat{\gamma}_S - \gamma_0\|^{m+1})$ is the remainder term, and

$$\triangledown^k G_n(\theta)\left[d\gamma_1, ..., d\gamma_m\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla^k g\left(z_i;\theta_0\right)\left[d\gamma_1, ..., d\gamma_k\right].$$

To evaluate the higher-order errors due to the approximation, we will derive (the order of) the mean and variance of each of the terms in the sum on the right hand side of Eq. (11).

3.2 The Approximators

We now impose regularity conditions on the approximation method. Let us first introduce two alternative ways of implementing the approximation: Either one common approximator is used across all observations, or a new approximator is used for each observation. In the first case, the approximate sample moment takes the form

$$G_n\left(\theta, \hat{\gamma}_S\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(z_i; \theta, \hat{\gamma}_S\right), \qquad (12)$$

while in the second case,

$$G_n\left(\theta, \hat{\gamma}_S\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(z_i; \theta, \hat{\gamma}_{i,S}\right). \qquad (13)$$

We will refer to the approximate estimator based on eq. (12) as an *estimator based on common approximators* (ECA) and to (13) as an *estimator based on individual approximators* (EIA).

Thus ECAs use one and the same approximator 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 Altissimo and Mele, 2009; Fermanian and Salanié, 2004). When the number of approximators 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. Andrews (1994) and Chen et al (2003).

In contrast, EIAs employ 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 note that EIAs, where the dimension of $\hat{\gamma}_S$ increases with sample size, give rise to an incidental parameters problem. Some of our results for this situation are similar to those found in the literature 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).

Finally, we impose conditions on the approximators. In order to give conditions that apply to both of the approximation schemes discussed above (ECA and EIA), we state our assumptions for J independent approximators: $J = 1$ for the ECA in (12), while $J = n$ for the EIA in (13). In what follows, it is crucial to separate assumptions on the bias of the approximator

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

from assumptions on its stochastic component

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

A.6(p) The approximator has the following properties:

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

(ii) The bias b_S is of order $\beta > 0$:

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

(iii) The stochastic component of the approximator satisfies

$$E\left[\|\psi_{j,S}(\cdot;\theta)\|^p\right] = S^{-\alpha_p}v_p(\theta) + o(S^{-\alpha_p}),$$

for some constant $\alpha_p > 0$ and some $p \geq 1$.

Assumption A.6 is sufficiently general to cover all of the examples in Section 2. A.6.i clearly has no bite when non-stochastic approximators are used, or in an ECA setting. For most simulation-based estimators in a dynamic setting for instance, only one approximator is used for all observations[7]; and so A.6.i is automatically satisfied in these cases.

For stochastic approximators in an ECA, A.6.i will be satisfied by drawing J independent batches of size S, and then using one batch per approximation. This does not rule out dependence between the simulated values within each batch, as will for example be the case when drawing recursively from a time series models.

There is one situation where $J = n \to \infty$ and the independence assumption is violated: 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 approximators where the approximator of the conditional density of the current observation depends on the one used for the previous observation, thereby not satisfying A.6.i.

For parametric approximators in simulation-based inference, the bias b_S is typically zero and so A.6.ii holds with $\beta = \infty$. We discuss this and other cases in more detail below.

A.6.iii requires that the approximator have p moments and that each of these be suitably bounded as a function of S. Note that, by Jensen's inequality, the individual rates are ordered, $\alpha_p/p \leq \alpha_q/q$ for $1 \leq p \leq q$.[8] We will choose $p \geq 1$ in conjunction with the order of the expansion $m \geq 1$ of Eq. (11), since we wish to evaluate the mean and variance of each of the higher-order terms. For example, in order to ensure that the variance of $\nabla^k G_n(\theta_0)[\hat{\gamma}_S, ..., \hat{\gamma}_S]$ exists and to evaluate its rate of convergence, we will require A.6.iii to hold with $p = 2k$.

One particular class of stochastic approximators that we consider in more detail is the following:

A.6'(p) Assume that $\hat{\gamma}_{j,S}(z;\theta)$ takes the form

$$\hat{\gamma}_{j,S}(z;\theta) = \frac{1}{S}\sum_{s=1}^{S} w_S(z, \varepsilon_{j,s};\theta). \tag{14}$$

For each $j = 1, ..., J$, $\{\varepsilon_{js}\}_{s=1}^{S}$ is stationary and geometrically β-mixing; $\{\varepsilon_{js}\}_{s=1}^{S}$ and $\{\varepsilon_{ks}\}_{s=1}^{S}$ are independent for $j \neq k$, and they are all independent of the sample; the

[7] See e.g. Duffie and Singleton (1994), Creel and Kristensen (2009) and Kristensen and Shin (2008).

[8] We have $E\left[\left\|\psi_{j,S}(\cdot;\theta)\right\|^p\right] = c_p S^{-\alpha_p}$ for any $p \geq 1$. Then by Jensen's inequality, since $q/p \geq 1$,

$$c_p^{q/p} S^{-\alpha_p q/p} = E\left[\left\|\psi_{j,S}(\cdot;\theta)\right\|^p\right]^{q/p} \leq E\left[\left\|\psi_{j,S}(\cdot;\theta)\right\|^q\right] = c_q S^{-\alpha_q}.$$

This inequality can only hold for all $S \geq 1$ if $\alpha_p q/p \geq \alpha_q$.

function $w_S(z, \varepsilon_{js}; \theta)$ satisfies

$$\bar{w}_S(z;\theta) := E[w_S(z, \varepsilon_{js}; \theta)|x] = \gamma(z;\theta) + S^{-\beta}\bar{b}(z;\theta),$$

while

$$\psi_{j,S}(z;\theta) = \frac{1}{S}\sum_{s=1}^{S} e_S(z, \varepsilon_{js}; \theta), \quad e_S(z, \varepsilon_{js}; \theta) = w_S(z, \varepsilon_{js}; \theta) - \bar{w}_S(z;\theta),$$

satisfies the conditions in A.6(p).iii.

To our knowledge, the above class of approximators includes all simulation-based approximators proposed in the literature. The requirement that $\{\varepsilon_{js}\}_{s=1}^{S}$ be geometrically β-mixing is only needed in the proof of Theorem 2 and could be weakened to strongly mixing elsewhere, but we maintain the assumption of β-mixing throughout to streamline the assumptions. The bias and variance of approximators on the form given in eq. (14) follow directly result from those of the simulators w_S. Suppose that we work with the L_2-norm; then assumption A.6(p).iii holds in great generality if $E[||w_S(x, \varepsilon; \theta)||^p|x] = O\left(S^{p/2-\mu}\right)$for some $\mu > 0$; this is proved in Lemma 5 in the Appendix.

In most cases, the simulating function $w_S \equiv w$ is actually independent of the number of simulations, and the approximator has no bias: $b_S \equiv 0$ and so $\beta = \infty$. Moreover, $E[||w(x, \varepsilon; \theta)||^p|x]$ then is constant and A.6(p).iii typically holds with $\alpha_p = p/2$.

Approximators of the form (14) 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 NPSML estimator: 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. The bias component satisfies

$$\bar{w}_S(y, x; \theta) = f(y|x;\theta) + h^r \frac{\partial^r f(y|x;\theta)}{\partial y^r} + o(h^r),$$

Furthermore, it is easily checked that $E[|K_h(y_s(x, \theta) - x)|^p|x] = O\left(1/\left(h^{d(p-1)}\right)\right)$ for all $p \geq 2$ under suitable regularity conditions. Thus, with a bandwidth of order $h \propto S^{-\delta}$ for some $\delta > 0$, A.6(p) holds with $\beta = r\delta$ and $\alpha_p = p/2 - \delta d(p-1)$, $p \geq 2$.

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 $\beta = \alpha_1/2 = r/(2r + d)$. We recover of course the standard nonparametric rate of $S^{-2r/(2r+d)}$ for AMISE; for example in the textbook case with $r = 2$ and $d = 1$, we obtain 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 individual densities that make up the 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 is not necessarily optimal when the goal is to minimize the MSE of $\hat{\theta}_{n,S}$. This is akin to results for semiparametric two-step estimators where undersmoothing of the first-step nonparametric estmator is normally required for the parametric estimator to be $\sqrt{n}$-consistent. For example, we show that the optimal rate for NPSML estimation turns out to be $\delta^{**} = 1/(r+d+2)$, see Section 3.2. Interestingly, in the case where standard second-order kernels are employed ($r=2$), the optimal rate minimizing the MSE of the kernel estimator is also optimal w.r.t. the MSE of $\hat{\theta}_{n,S}$, $\delta^{*} = \delta^{**} = 1/(4+d)$.

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 $\alpha_p = +\infty$, $p \geq 2$, but approximation imparts a bias, which in leading cases obeys assumption A.6 for some $\beta > 0$. We will see later that the analytical bias adjustment technique based on correcting the objective function has no bite in this situation. On the other hand, the proposed Jackknife-type bias adjustment and Newton-Raphson procedure work for both stochastic and non-stochastic approximations.

3.3 The Effect of Approximators

The following theorem states the rate at which the approximate objective function converges towards the exact one; and shows how it translates directly into a bound on the difference between the approximate estimator and the exact estimator. To state the asymptotic expansion in a compact manner, we introduce some moments which will make up the bias terms:

$$\begin{aligned} B_1 &= H_0^{-1} E\left[\bigtriangledown G_n(\theta_0)[b_S]\right], \quad B_{S,2}(\theta) = \frac{1}{2} H_0^{-1} E\left[\bigtriangledown^2 G_n(\theta_0)[\psi_S, \psi_S]\right], \qquad (15) \\ B_{S,3}(\theta) &= \frac{1}{2} H_0^{-1} E\left[\bigtriangledown^2 G_n(\theta_0)[b_S, b_S]\right]. \end{aligned}$$

Theorem 1 *Assume that A.1-A.4, A.5*(2) *and A.6*(4) *hold. Then for both the ECA and EIA the approximate objective function satisfies:*

$$E\left[\bigtriangledown G_n(\theta_0)[\psi_S]\right] = 0, \quad B_{S,1} = O\left(S^{-\beta}\right), \quad B_{S,2} = O\left(S^{-\alpha_2}\right), \quad B_{S,3} = O\left(S^{-2\beta}\right), \qquad (16)$$

$$\mathrm{Var}\left(\bigtriangledown G_n(\theta_0)[b_S]\right) = O\left(n^{-1} S^{-\beta}\right), \quad \mathrm{Var}\left(\bigtriangledown^2 G_n(\theta_0)[b_S, b_S]\right) = O\left(n^{-1} S^{-2\beta}\right) \qquad (17)$$

and the remainder term in equation (11) satisfies:

$$R_{n,S}(\theta) = O\left(S^{-3\beta}\right) + O\left(S^{-\alpha_3}\right).$$

Furthermore:

(ECA) *The approximate objective function based on eq. (12) satisfies:*

$$\mathrm{Var}\left(\bigtriangledown G_n(\theta_0)[\psi_S]\right) = \mathrm{Var}\left(\bigtriangledown \bar{g}\left(\psi_S;\theta_0\right)\right) + O\left(n^{-1}S^{-\alpha_2}\right), \tag{18}$$

$$\mathrm{Var}\left(\bigtriangledown^2 G_n(\theta_0)[\psi_S,\psi_S]\right) = O\left(S^{-\alpha_2}\right) + O\left(n^{-1}S^{-\alpha_4}\right), \tag{19}$$

where $\bigtriangledown \bar{g}\left(d\gamma;\theta_0\right) := E\left[\bigtriangledown g\left(z;\theta_0\right)\left[d\gamma\right]\right]$ *satisfies* $\mathrm{Var}\left(\bigtriangledown \bar{g}\left(\psi_S;\theta_0\right)\right) = O\left(S^{-\alpha_2}\right)$*.*

As a consequence, the ECA satisfies:

$$||\hat{\theta}_{n,S}-\hat{\theta}_n|| = B_{S,1}+B_{S,2}+\left\|H_0^{-1}\bigtriangledown \bar{g}\left(\psi_S;\theta_0\right)\right\|+O_P\left(n^{-1/2}S^{-\alpha_2/2}\right)+O_P\left(n^{-1/2}S^{-\beta}\right).$$

(EIA) *The approximate objective function based on eq. (13) satisfies*

$$\mathrm{Var}\left(\bigtriangledown G_n(\theta_0)[\psi_S]\right) = O\left(n^{-1}S^{-\alpha_2}\right) \tag{20}$$

$$\mathrm{Var}\left(\bigtriangledown^2 G_n(\theta_0)[\psi_S,\psi_S]\right) = O\left(n^{-1}S^{-\alpha_4}\right), \tag{21}$$

As a consequence, the EIA satisfies:

$$||\hat{\theta}_{n,S}-\hat{\theta}_n|| = B_{S,1}+B_{S,2}+O_P\left(n^{-1/2}S^{-\alpha_2/2}\right)+O_P\left(n^{-1/2}S^{-\beta}\right).$$

Under our assumptions, the term $\bigtriangledown \bar{g}\left(\psi_S;\theta\right)$ that appears in the expansion of the ECA is at least of order $O_P\left(S^{-\alpha_2/2}\right)$; but in some important cases this rate is not sharp. For example, when $\hat{\gamma}_S$ is a kernel estimator, we can show that $\mathrm{Var}\left(\bigtriangledown \bar{g}\left(\psi_S;\theta_0\right)\right) = O\left(S^{-1}\right)$ which is faster than $O\left(S^{-\alpha_2}\right) = O(1/\left(Sh^d\right))$, c.f. Example 3.2 below.

We have seen that for a large class of simulation-based estimators, the bias and the stochastic component of the approximator are of order $\alpha_p = p/2$ and $\beta = \infty$, c.f. Assumption A.6(p) and the subsequent discussion. With weakly dependent data, the above corollary states that for the EIA's, 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 ECA's in a cross-sectional setting, see e.g. Laffont et al. (1993) and Lee (1992). Our theorem shows that this result holds more generally under weak regularity conditions. In the case of ECA's, the leading term is $O_P(1/\sqrt{S})$; again, this is consistent with the findings of, for example, Duffie and Singleton (1993) and Corradi and Swanson (2007).

Comparing the results for the two approximate estimators, we see that the only difference appears in the variances of $\bigtriangledown^2 G_n(\theta_0)[\psi_S]$ and $\bigtriangledown^2 G_n(\theta_0)[\psi_S,\psi_S]$, which both have an additional term when common approximators are employed. This is due to the additional correlations across observation, which vanish when independent approximators are employed.

This is why ECA's are of order $O_P(1/\sqrt{S})$ while EIA's are $O_P(1/S)$. This does not imply that the EIA is preferable to the ECA: Note that we generate nS draws in total to compute the EIA, but only S draws for the ECA. Thus, for a fair comparison, one should replace S with nS in the case of ECA, in which case the ECA is in fact more precise for a given number of draws.

In some cases, the rate of the approximate estimator obtained in Theorem 1 is not sharp. By imposing more structure on the problem, better rates can be obtained for the remainder term $R_{n,S}$ of equation (11). Strengthening A.6(p) to A.6'(p) and combining it with the assumption that $g\left(z;\theta,\gamma\right)$ is three (instead of two) times differentiable w.r.t. γ, we can obtain slightly sharper rates for the estimator. Lemma 9 in the appendix delivers this refinement of Theorem 1. The sharper rate stated there can in turn be used to establish better rates for the bias adjusted estimators considered in the next section.

3.4 Applications to Standard Approximate Estimators

To illustrate the use of our results, we return to Examples 2-3 of Section 2. We will throughout only consider the first two functional derivatives; the third order term is easily derived but we leave it out to save space. In the following, the notation $\dot{f}(x,\theta)$ stands for $\frac{\partial f}{\partial \theta}(x,\theta)$.

Example 2.1 (SNLS). In this example,

$$g_i(\theta,m) = \frac{\partial q_i(\theta,m)}{\partial \theta} = 2\left(y_i - m\left(x_i;\theta\right)\right)\dot{m}\left(x_i;\theta\right),$$

where $m_i\left(\theta\right) = m\left(x_i;\theta\right)$ and $\dot{m}_i\left(\theta\right) = \partial m\left(x_i;\theta\right)/\left(\partial\theta\right)$. The approximator is of the form (14) where $w_S\left(x,\varepsilon;\theta\right) = w\left(x,\varepsilon;\theta\right)$ satisfies $E\left[w\left(x,\varepsilon;\theta\right)\right] = m\left(x;\theta\right) = E_\theta\left[y|x\right]$.

Denote $\xi_i\left(\theta\right) := y_i - m_i\left(\theta\right)$; then the functional differentials are

$$\triangledown g(z_i;\theta)\left[dm\right] = 2\dot{m}_i\left(\theta\right)d\gamma\left(x_i;\theta\right) + 2\xi_i\left(\theta\right)d\dot{\gamma}\left(x_i;\theta\right),$$

$$\triangledown^2 g(z_i;\theta)\left[dm,dm\right] = 4d\dot{m}\left(x_i;\theta\right)dm\left(x_i;\theta\right),$$

so that (using a single approximation for all observations)

$$\begin{aligned}\triangledown G_n(\theta,m)\left[dm\right] &= -\frac{2}{n}\sum_{i=1}^{n}\left\{\dot{m}_i\left(\theta\right)dm\left(x_i;\theta\right) + \xi_i\left(\theta\right)d\dot{m}\left(x_i;\theta\right)\right\},\\ \triangledown^2 G_n(\theta,m)\left[dm,dm\right] &= \frac{4}{n}\sum_{i=1}^{n}d\dot{m}\left(x_i;\theta\right)dm\left(x_i;\theta\right).\end{aligned}$$

Since $\triangledown^3 g(z;\theta)\left[dm,dm,dm\right] = 0$, (8) holds with $\bar{G}_0 = 0$ and the remainder term $R_{S,n}$ in eq. (11) is zero.

Assuming that $E\left[y^2\right] < \infty$, $E\left[\sup_{\theta\in\Theta}\left\|m_i\left(\theta\right)\right\|^2\right] < \infty$ and $E\left[\sup_{\theta\in\Theta}\left\|\dot{m}_i\left(\theta\right)\right\|^2\right] < \infty$ it is easily seen that Eqs. (9)-(10) also hold when using an appropriate L_2-norm. Depending on how the simulated estimator has been implemented, different norms should be used. If two independent batches have been used for the conditional mean and its derivative respectively, we use $\left\|\gamma\right\|^2 = E\left[\left\|\gamma\left(x_i;\theta\right)\right\|^2\right]$. If on the other hand the same simulations have been used for both, we need to use $\left\|\gamma\right\|^2 = E\left[\left\|\gamma\left(x_i;\theta\right)\right\|^2\right] + E\left[\left\|\dot{\gamma}\left(x_i;\theta\right)\right\|^2\right]$.

Example 2.2 (SPML). Since the derivations for this estimator follows along the same lines as the SNLS, we have relegated them to Appendix A.

Example 3.1 (SML in discrete choice models). Consider a discrete choice model where $P\left(y = d_l|x\right) = P_l\left(x;\theta\right)$ for $l = 1, ..., L$, so that given observations $\left(d_{l,i}\right)$, the log-likelihood is given by:

$$\log p_i\left(\theta\right) = \sum_{l=1}^{L} d_{l,i}\log P_{l,i}\left(\theta\right).$$

Let unbiased simulations be used to approximate $P\left(x;\theta\right) = \left(P_1\left(x;\theta\right), ..., P_L\left(x;\theta\right)\right)$. Then

$$g_i(\theta) = \frac{\partial\log p_i\left(\theta\right)}{\partial\theta} = \sum_{l=1}^{L} d_{l,i}\frac{\dot{P}_{l,i}\left(\theta\right)}{P_{l,i}\left(\theta\right)}$$

and

$$\begin{aligned}
\bigtriangledown g_i(\theta)\left[dP\right] &= \sum_{l=1}^{L} d_{l,i}\left[\frac{1}{P_{l,i}\left(\theta\right)}d\dot{P}_{l,i}\left(\theta\right) - \frac{\dot{P}_{l,i}\left(\theta\right)}{P_l^2\left(x_i;\theta\right)}dP_{l,i}\left(\theta\right)\right], \\
\bigtriangledown^2 g_i(\theta)\left[dP, dP\right] &= \sum_{l=1}^{L} d_{l,i}\left[-\frac{2}{P_{l,i}^2\left(\theta\right)}d\dot{P}_{l,i}\left(\theta\right)dP_{l,i}\left(\theta\right) + \frac{2\dot{P}_{l,i}\left(\theta\right)}{P_l^3\left(x_i;\theta\right)}dP_{l,i}^2\left(\theta\right)\right], \\
\bigtriangledown^3 g_i(\theta)\left[dP, dP\right] &= \sum_{l=1}^{L} d_{l,i}\left[\frac{4}{P_{l,i}^3\left(\theta\right)}d\dot{P}_{l,i}\left(\theta\right)dP_{l,i}^2\left(\theta\right) - \frac{6\dot{P}_{l,i}\left(\theta\right)}{P^4\left(x_i;\theta\right)}dP_{l,i}^3\left(\theta\right)\right]
\end{aligned}$$

Comparing with the expansion of the SMLE in Lee (1995, Theorem 1), we recognize his first and second order terms, L_n and Q_n in his notation, as the first and second order differentials respectively: $L_n = \bigtriangledown G_n(\theta_0)[\hat{P}_S - P]$ and $Q_n = \bigtriangledown^2 G_n(\theta_0)[\hat{P}_S - P, \hat{P}_S - P]$. By standard arguments, we see that eq. (8) holds with $m = 2$ if

$$\bar{G}_0 := \sum_{l=1}^{L} E\left[\left\{\frac{6\left\|\dot{P}_{l,i}\left(\theta\right)\right\|}{P_{l,i}^3\left(\theta_0\right)} + \frac{4}{P_{l,i}^2\left(\theta_0\right)}\right\}\right] < \infty.$$

Thus $\bar{G}_0$ cannot be finite unless $EP_l^{-2-k}\left(x;\theta\right) < \infty$ for $k = 1, 2$. This will typically not hold when covariates have unbounded support. We could impose that the density of the covariates be bounded away from zero as in Lee (1995), but this is a very strong requirement. To circumvent such assumptions, one can instead use trimming techniques (see Fermanian and Salanié, 2004; Kristensen and Shin, 2008). This imparts an additional bias component to the approximator, but the bias in general is of smaller order than the simulation component however, and then it can be ignored.

Example 3.2 (NPSML). Here,

$$g_i(\theta, p) = -\frac{\dot{p}_i(\theta)}{p_i(\theta)},$$

so that

$$\bigtriangledown g_i(\theta)[dp] = \frac{\dot{p}_i(\theta)}{p_i^2(\theta)} dp_i(\theta) - \frac{1}{p_i(\theta)} d\dot{p}_i(\theta),$$

$$\bigtriangledown^2 g_i(\theta)[dp, dp] = \frac{2}{p_i^2(\theta)} d\dot{p}_i(\theta)\, dp_i(\theta) - \frac{2\dot{p}_i(\theta)}{p_i^3(\theta)} dp_i(\theta)^2.$$

It is easily seen that Eq. (8) holds for $m = 2$ with

$$\bar{G}_0 := E\left[\sup_{\theta\in\Theta}\left\{\frac{6\,\|\dot{p}_i(\theta_0)\|}{p_i^3(\theta_0)} + \frac{2}{p_i^2(\theta_0)}\right\}\right].$$

The discussion of $\bar{G}_0 < \infty$ in Example 3.1 applies here too: we either have to assume that the density of covariates is bounded away from zero, or to resort to trimming.

Since kernel estimators are used in the approximation, the first order bias and stochastic components for the approximator have non-standard rates. Assume for simplicity that the density is bounded away from zero. Then the bias component of the first order term is

$$\bigtriangledown g_i(\theta)[b_S] = h^r\left\{\frac{\dot{p}(y_i|x_i;\theta)}{p^2(y_i|x_i;\theta)}\frac{\partial^r p(y_i|x_i;\theta)}{\partial y_i^r} - \frac{1}{p(y_i|x_i;\theta)}\frac{\partial^r \dot{p}(y_i|x_i;\theta)}{\partial y_i^r}\right\} + o(h^r).$$

This holds irrespectively of whether a single simulation batch (ECA) or n (EIA) simulation batches are used.

Next, we derive the rate of the variance component of the first order term. First, consider the EIA: By Lemma 6, we obtain that $\mathrm{Var}(\bigtriangledown G_n(\theta)[\psi_S]) = O\left(1/\left(nSh^{d+2}\right)\right)$. Note that the $(d+2)$ term comes from the fact that we need to approximate the derivative of the loglikelihood as well as the function itself. Next consider the ECA: we show in Appendix A that

$$\bigtriangledown G_n(\theta)[\psi_S] = \frac{1}{S}\sum_{s=1}^{S}\bigtriangledown \bar{g}(\theta)[e_s] + O_P\left(\frac{1}{\sqrt{nSh^{d+2}}}\right) = O_P\left(\frac{1}{\sqrt{S}}\right) + O_P\left(\frac{1}{\sqrt{nSh^{d+2}}}\right).$$

As a consequence, for EIA's we have

$$\begin{aligned}||\hat{\theta}_{n,S}-\hat{\theta}_n|| &= B\times h^r + V_1\times\frac{1}{Sh^{d+2}} + V_2\times\frac{1}{\sqrt{nSh^{d+2}}} \\ &\quad + o_P\left(h^r\right) + o_P\left(\frac{1}{Sh^{d+2}}\right) + o_P\left(\frac{1}{\sqrt{nSh^{d+2}}}\right),\end{aligned}$$

while for ECA's an additional $O_P\left(1/\sqrt{S}\right)$ appears.

3.5 Asymptotic First-Order Equivalence and Variance Estimation

Our results allow us to state precisely when the approximate estimator is asymptotically equivalent to the exact estimator; that is, which sequences $\{S_n\}$ guarantee that $||\hat{\theta}_{n,S_n}-\hat{\theta}_n|| = o_P\left(n^{-1/2}\right)$.

In general, asymptotic equivalence for ECA's are obtained if $S_n^{\min(\alpha_2,2\beta)}$ goes to infinity faster than n; for EIA's we have a weaker condition, replacing α_2 with $2\alpha_2$.

For parametric simulation-based estimators ($\beta=0$, $\alpha_2=1$), this gives the standard result that n/S_n should go to zero for ECA's (Duffie and Singleton, 1993; Lee, 1995, Theorem 1), while $\sqrt{n}/S_n$ should go to zero for EIA's (Laroque and Salanié, 1989; Lee, 1995, Theorem 4).

When nonparametric kernel methods are used, we have to choose both S and h. Assume that y is d-dimensional, and we use an r-order kernel. Given the calculations made in the previous section for both EIA and ECA, we need $\sqrt{n}h^r\to 0$ and $\sqrt{n}/\left(Sh^{d+2}\right)\to 0$ for the NPSMLE to be equivalent to the MLE. As usual, the optimal bandwidth makes these two terms go to zero at the same rate; this yields $h^* = O\left(S^{1/(r+d+2)}\right)$. In general, this is a non-standard bandwidth rate which is due to the fact that we here try to balance the bias and variance of the kernel estimator, while in standard problems one tries to balance the *squared* bias and variance in order to minimize the MSE. Yet with kernels of order $r=2$ the rate becomes standard, a somewhat surprising result

Even when the approximate estimator is asymptotically equivalent to the exact estimator, in finite samples it may be useful to adjust computed standard errors to account for the additional variance due to the approximation. This turns out to be quite straightforward in some cases. The exact estimator has variance $H_0^{-1}\Omega_n H_0^{-1}$ where $\Omega_n = \mathrm{Var}\left(G_n\left(\theta_0,\gamma\right)\right)$. For the approximate estimator,

$$\mathrm{Var}(\hat{\theta}_{n,S}) \approx H_0^{-1}\Sigma_{n,S}H_0^{-1}, \quad \Sigma_{n,S} = \mathrm{Var}\left(G_n\left(\theta_0,\gamma\right) + \triangledown G_n(\theta_0)[\psi_S]\right).$$

To approximate $\Sigma_{n,S}$, suppose for simplicity that the observations, z_i, $i=1,...,n$, are independent (otherwise HAC-type estimators should be employed). Then, we compute

$$\hat{\Sigma}_{n,S} = \frac{1}{n}\sum_{i=1}^{n} \hat{s}_i \hat{s}_i', \quad \hat{s}_i := g(z_i, \hat{\theta}_{n,S}) + \hat{\delta}_i,$$

where $\hat{\delta}_i$ is an estimator of $\bigtriangledown g(z_i, \hat{\theta}_{n,S})[\psi_{i,S}]$ and thereby accounting for the additional variance due to the simulations. In the leading example where $\hat{\gamma}_{i,S}$ satisfies A.6',

$$\bigtriangledown g(z_i, \hat{\theta}_{n,S})[\psi_{i,S}] = \frac{1}{S}\sum_{s=1}^{s} \bigtriangledown g(z_i, \hat{\theta}_{n,S})[w_{i,S} - \bar{w}_{i,S}],$$

and so a natural choice for the estimator $\hat{\delta}_i$ is

$$\hat{\delta}_i = \frac{1}{S}\sum_{s=1}^{s} \bigtriangledown g(z_i, \hat{\theta}_{n,S})[w_{i,S} - \hat{\gamma}_S].$$

This estimator is similar to the one proposed in Newey (1994) for semiparametric two-step estimators.

4 Analytical Bias Adjustment

We here propose an analytical bias adjustment of the objective function $G_n(\theta, \hat{\gamma}_S)$ which removes the leading term of the bias incurred by using $\hat{\gamma}_S$ if the stochastic component of the approximator is of a larger order than its bias component: $\alpha_2 < \beta$. This is clearly the case for the parametric simulation-based estimation methods, as $\alpha_2 = 1$ and $\beta = \infty$. In the previous section, we derived the order of the bias and variance of the second order expansion of $G_n(\theta, \hat{\gamma}_S)$ in terms of $\hat{\gamma}_S$, and translated these into an error bound for the approximate estimator as stated in Theorem 1. The two leading bias terms are $B_{S,1}$ and $B_{S,2}$ as defined in eq. (15). We discuss in turn how these can be adjusted for.

Start with $B_{S,2} = H_0^{-1} E\left[\bigtriangledown^2 G_n(\theta_0, \gamma)[\psi_S, \psi_S]\right]/2$. To adjust for the bias in $\hat{\theta}_{n,S}$ due to $B_{S,2}$, we propose an estimator of $E\left[\bigtriangledown^2 G_n(\theta_0, \gamma)[\psi_S, \psi_S]\right]/2$ which we then include in the objective function. When the approximator belongs to the class defined in A.6', we can write the bias component in terms of e defined in A.6',

$$e_S(z, \varepsilon_s; \theta) = w_S(z, \varepsilon_s; \theta) - \bar{w}_S(z; \theta),$$

which represents the deviation of simulation s from its expected value for a given observation z.

In the following, we suppress the dependence of the functions w and e on S. We first note that

$$E\left[\triangledown^2 G_n(\theta_0,\gamma)[\psi_S,\psi_S]\right] = \lim_{n\to\infty}\frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S}\triangledown^2 g_i(\theta_0)[e_{i,s},e_{i,s}],$$

where $\triangledown^2 g_i(\theta)[e_{i,s},e_{i,s}] = \triangledown^2 g(z_i;\theta)[e_{i,s},e_{i,s}]$, and $e_{i,s} = e\left(z,\varepsilon_{i,s};\theta\right)$, $i=1,...,n$. Note that in the case of ECA's, the same simulations are used across all observations such that $e_{i,s} = e_s = e\left(z,\varepsilon_s;\theta\right)$ for $i=1,...,n$.

We wish to obtain an estimator of this term and use it to remove the bias. Ideally, we would like to compute $e = w - \bar{w}$, but since in general $\bar{w}$ is unknown, this is not feasible. On the other hand, we can compute $\hat{\gamma}_S$ which is an unbiased and consistent estimator of $\bar{w}$. Thus, a natural estimator of $E\left[\triangledown^2 G_n(\theta_0,\gamma)[\psi_S,\psi_S]\right]/2$ is:

$$\dot{\Delta}_{n,S}\left(\theta\right) = \frac{1}{2S\left(S-1\right)}\sum_{i=1}^{n}\sum_{s=1}^{S}\triangledown^2 g_i(\theta)[w_{is}-\hat{\gamma}_i, w_{is}-\hat{\gamma}_i], \tag{22}$$

Under regularity conditions, $||H_0^{-1}\dot{\Delta}_{n,S}\left(\theta_0\right) - B_{S,2}|| \to^P 0$ as $n\to\infty$. This motivates our definition of an analytically bias-adjusted estimator $\hat{\theta}_{n,S}^{\text{AB}}$ as the solution to:

$$o_P\left(n^{-1/2}\right) = G_n(\hat{\theta}_{n,S}^{\text{AB}},\hat{\gamma}_S) - \dot{\Delta}_{n,S}(\hat{\theta}_{n,S}^{\text{AB}}). \tag{23}$$

When using an extremum estimator, $\hat{\theta}_{n,S} = \arg\max_\theta Q_n(\theta,\hat{\gamma}_S)$ where $Q_n(\theta,\gamma) = \sum_{i=1}^n q(z_i;\theta,\gamma)/n$, the above adjustment corresponds to

$$\hat{\theta}_{n,S}^{\text{AB}} = \arg\min_{\theta\in\Theta}\left\{Q_n(\theta,\hat{\gamma}_S) - \Delta_{n,S}\left(\theta\right)\right\}, \tag{24}$$

where

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

After such an adjustment, the bias component $B_{S,2}$ drops out of the expansion of the resulting adjusted estimator, which improves on the rate of convergence. To state the theoretical result, we introduce the moment

$$M_{S,p}\left(z_i\right) := \sup_{s<t} E\left[\left\|\triangledown^2 g(z_i)[e_{is},e_{it}]\right\|^{p+\epsilon}|z_i\right]^{p/(p+\epsilon)}, \text{ for some } \epsilon > 0. \tag{25}$$

Theorem 2 *Assume that A.1-A.4, A.5*(2) *and A.6'*(4) *hold. Then with* $\hat{\theta}_{n,S}^{\text{AB}}$ *defined in (23) with* $\dot{\Delta}_{n,S}\left(\theta\right)$ *given in (22), the following holds:*

1. *The bias adjusted version of the ECA in eq. (12) satisfies:*

$$\begin{aligned}||\hat{\theta}_{n,S}^{\text{AB}}-\hat{\theta}_{n}|| &= \|\triangledown g_{1}\left(\psi_{S};\theta\right)\|+O_{P}\left(S^{-(1+\alpha_{2})}\right)+O_{P}\left(S^{-\beta}\right)+O_{P}\left(S^{-\alpha_{3}}\right)\\ &\quad +O_{P}\left(n^{-1/2}S^{-1}\sqrt{E\left[M_{S,4}\left(z_{i}\right)\right]}\right)+O_{P}\left(n^{-1/2}S^{-(1+\alpha_{4}/2)}\right)+O_{P}\left(n^{-1/2}S^{-\beta}\right).\end{aligned}$$

2. *The bias adjusted version of the EIA in eq. (13) satisfies:*

$$\begin{aligned}||\hat{\theta}_{n,S}^{\text{AB}}-\hat{\theta}_{n}|| &= O_{P}\left(S^{-(1+\alpha_{2})}\right)+O_{P}\left(S^{-\beta}\right)+O_{P}\left(S^{-\alpha_{3}}\right)\\ &\quad +O_{P}\left(n^{-1/2}S^{-1}\sqrt{E\left[M_{S,4}\left(z_{i}\right)\right]}\right)+O_{P}\left(n^{-1/2}S^{-(1+\alpha_{4}/2)}\right)+O_{P}\left(n^{-1/2}S^{-\beta}\right).\end{aligned}$$

Remark 3 *In the proof, we employ moment bounds for U-statistics with mixing variables (Yoshihara, 1976). In the case where* $\{e_{is}:s=1,...,S\}$ *are i.i.d., we can instead employ results for i.i.d. variables (Ferger, 1996) and exchange* $M_{S,4}\left(z_{i}\right)$ *for* $M_{S,2}\left(z_{i}\right)$ *in the theorem. The rate of the moment* $E\left[M_{S,p}\left(z_{i}\right)\right]$*, with* $p=2$ *or* 4*, can be derived in most applications. For example, if* $\left\|\triangledown^{2}g(z_{i})[e_{is},e_{it}]\right\|\leq b\left(z_{i}\right)\left\|e_{is}\left(z_{i}\right)\right\|\left\|e_{is}\left(z_{i}\right)\right\|$*, which holds in all our examples, then with dependent simulations,*

$$\begin{aligned}E\left[M_{S,4}\left(z_{i}\right)\right] &\leq E\left[b^{4}\left(z_{i}\right)\sup_{s<t}E\left[\left\|e_{is}\left(z_{i}\right)\right\|^{4+\epsilon}\left\|e_{it}\left(z_{i}\right)\right\|^{4+\epsilon}|z_{i}\right]^{4/(4+\epsilon)}\right]\\ &\leq E\left[b^{4}\left(z_{i}\right)E\left[\left\|e_{is}\left(z_{i}\right)\right\|^{8+2\epsilon}|z_{i}\right]^{4/(4+\epsilon)}\right]\end{aligned}$$

while with independent simulations,

$$E\left[M_{S,2}\left(z_{i}\right)\right]\leq E\left[b^{2}\left(z_{i}\right)E\left[\left\|e_{is}\left(z_{i}\right)\right\|^{4+\epsilon}|z_{i}\right]^{2/(4+\epsilon)}\right].$$

Thus, with standard simulators, $\sqrt{E\left[M_{S,p}\left(z_{i}\right)\right]}=O\left(1\right)$*.*

Comparing with Theorem 1 on the unadjusted estimators shows that the bias term $B_{S,2}=O_{P}\left(S^{-\alpha_{2}}\right)$ has been replaced by a term of order $O_{P}\left(S^{-(1+\alpha_{2})}\right)$. This is due to the fact that $|\dot{\Delta}_{n,S}\left(\theta\right)-\triangledown^{2}G_{n}(\theta,\gamma_{0})[\psi_{S},\psi_{S}]/2|=O_{P}\left(S^{-(1+\alpha_{2})}\right)$. In the leading case where $\beta=\infty$ and $\alpha_{p}=p/2$, $O_{P}\left(S^{-(1+\alpha_{2})}\right)$ is of smaller order than the next term, which is $O_{P}\left(S^{-\alpha_{3}}\right)$. For other approximators, e.g. NPSML, the relationship may be reversed; but in either case, the adjusted estimator is (asymptotically) superior to the unadjusted one since the bias term of order $O_{P}\left(S^{-\alpha_{2}}\right)$ has been removed.

With unbiased simulators, we have $\alpha_{2}=1$ 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.

If we strengthen A.5(2) and A.6'(4) to A.5(3) and A.6'(6), then we know from Lemma 9 that the leading terms after adjustment are $O_P\left(S^{-\alpha_4}\right)$ and $O\left(S^{-(1+\alpha_3)}\right)$; this follows by the same arguments we used to prove Lemma 9. With unbiased simulators $\alpha_4 = 2$, so that the proposed adjustment in fact takes the bias from S^{-1} down to S^{-2}.

More generally, the proposed adjustment will remove the largest bias component as long as $\alpha_2 < \beta$. Otherwise the bias term $O_P\left(S^{-\beta}\right)$ is of a larger order than $O_P\left(S^{-\alpha_2}\right)$ and the proposed bias adjustment does not remove the leading term anymore. In particular, when non-stochastic approximations are employed the above adjustment does not help. With non-stochastic approximations the leading term of the approximation error is not $\nabla^2 Q_n(\theta)[\psi_S, \psi_S]$, which the $\Delta_{n,S}(\theta)$ correction is aimed at: in fact this term is identically zero as we saw earlier. To phrase things differently, with non-stochastic approximations, for every $p, \alpha_p = \infty$ and so $\alpha_p > \beta$.

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

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

$$\nabla^2 G_n(\theta, \gamma)\left[dm, dm\right] = \frac{4}{n}\sum_{i=1}^{n} d\dot{m}_i\left(\theta\right) dm_i\left(\theta\right).$$

Let $r_{is}(\theta) = w_S\left(x_i, \varepsilon_{is}; \theta\right) - \hat{m}_{i,S}\left(x_i; \theta\right)$ denote the difference of a given simulator from the mean simulation for the same observation. Then the adjustment term becomes

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

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

$$\Delta_{n,S}\left(\theta\right) = \frac{1}{nS\left(S-1\right)}\sum_{i=1}^{n}\sum_{s=1}^{S}\left(F\left(\frac{m_{is}\left(\theta\right)}{h}\right) - \frac{1}{S}\sum_{t=1}^{S} F\left(\frac{m_{it}\left(\theta\right)}{h}\right)\right)^2,$$

where we have replaced the indicator function $\mathbb{I}\left\{y > 0\right\}$ by a symmetric cdf, $F\left(y/h\right)$ as proposed in Fermanian and Salanié (2004). As $h \to 0$, $F\left(y/h\right) = \mathbb{I}\left\{y > 0\right\} + o\left(h^2\right)$, and so we only pay a small price in terms of bias to obtain a smooth criterion function.

Example 3.1 (Discrete choice). Here, $q_i(\theta, P) = -\sum_{l=1}^{L} d_{l,i} \log P_{l,i}(\theta)$ and so

$$\bigtriangledown^2 q_i(\theta)\left[dP, dP\right] = \sum_{l=1}^{L} d_{l,i} \frac{dP_{l,i}^2(\theta)}{P_{l,i}^2(\theta)}.$$

Thus, the adjustment term becomes

$$\Delta_{n,S}(\theta) = \frac{1}{2nS(S-1)} \sum_{i=1}^{n} \sum_{l=1}^{L} d_{l,i} \sum_{s=1}^{S} \left[\frac{w_l(x_i, \varepsilon_{is}; \theta) - \hat{P}_l(x_i; \theta)}{\hat{P}_l(x_i; \theta)} \right]^2.$$

In contrast to the previous example, $\bigtriangledown^3 q_i(\theta)\left[dP, dP, dP\right] \neq 0$ and so the bias adjustment does not ensure consistency for fixed S.

Example 3.2 (NPSML). Here, $Q_n(\theta, p) = \frac{1}{n}\sum_{i=1}^{n} \log p(z_i; \theta)$ and so

$$\bigtriangledown^2 Q_n(\theta)\left[dp\right] = \frac{1}{n} \sum_{i=1}^{n} \frac{dp^2(z_i; \theta)}{p^2(z_i; \theta)}.$$

Thus, the adjustment term becomes

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

When n batches of simulations are used as in EIA, the bias corrected estimator satisfies:

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

A similar result can be derived for the adjusted ECA estimator.

Instead of adjusting the objective function("preventive bias adjustment"), we could first compute the unadjusted estimator, $\hat{\theta}_{n,S}$, and then directly correct its bias ("corrective bias adjustment"): Taking a first-order expansion in θ around $\hat{\theta}_{n,S}$ in eq. (23), we obtain

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

where $H_n(\theta, \gamma) = \partial G_n(\theta, \gamma) / (\partial \theta)$. Such a two-step procedure was proposed in Lee (1995) for the special case of SMLE and SNLS in limited dependent variable models.

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

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

where as, before, $\dot{f}$ denotes the derivative of f w.r.t. θ. For the SML example it is

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

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.

Can we find a simple adjustment for the bias term, b_S which in turn leads to the bias term $E[\bigtriangledown G_n(\theta)[b_S]] = O\left(S^{-\beta}\right)$? If we were able to obtain (an estimator of) b_S, the associated adjustment term could straightforwardly be estimated by $\dot{\Delta}_{n,S}^{(B)}(\theta) = \bigtriangledown G_n(\theta)[b_S]$. 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 $\bigtriangledown G_n(\theta)[b_S]$, one may try to improve the order of $G_n(\theta)[b_S]$ 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 higher-order kernels can also be used to decrease the bias component. Suppose for example that $\gamma(z;\theta) = p(y|x;\theta)$ 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 the bias takes the form $b_S(z;\theta) = h^r\partial^r p(y|x;\theta)/\partial y^r + o(h^r)$, and for large r the bias is of small order. Removing the leading bias component requires knowledge of $\partial^r p(y|x;\theta)/\partial y^r$ which is not easily estimated.

An alternative way to reduce this bias component for kernel-based approximators is to use so-called 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 $\bigtriangledown g(z;\theta,\gamma)[d\gamma] = \delta(z;\theta)\,d\gamma(z;\theta)$ for some function δ as is the case in all of our examples. 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 $E\left[\bigtriangledown G_n(\theta,\gamma)\left[b_S^K\right]\right] = O(h^r)$, while the use of a twicing kernel estimator yields $E\left[\bigtriangledown G_n(\theta,\gamma)[b_S^{\bar{K}}]\right] = O\left(h^{2r}\right)$, cf. Newey et al. (2004, Theorem 1). Again, the improvement obtained here is not through an adjustment term added to the objective function since the adjustment takes place in the construction of $\hat{\gamma}_S(y|x;\theta)$ itself.

5 Bias Adjustment by Resampling

As an alternative to analytical bias corrections, resampling methods for bias correction can be used[9]. They will in general handle the biases due to both the stochastic and the non-stochastic component of the approximator; and the researcher is not required to derive an expression of the bias. On the other hand, they are computationally more demanding than the analytical bias correction proposed in the previous section.

To motivate the bias adjustment, we first note that, according to Lemmas 7-8,

$$\begin{aligned} E\left[\bigtriangledown G_n(\theta_0,\gamma)\left[\hat{\gamma}_S-\gamma\right]\right] &= H_0B_1S^{-\beta}+o\left(S^{-\beta}\right), \\ \frac{1}{2}E\left[\bigtriangledown G_n(\theta_0,\gamma)\left[\hat{\gamma}_S-\gamma,\hat{\gamma}_S-\gamma\right]\right] &= H_0B_2S^{-\alpha_2}+o\left(S^{-\alpha_2}\right), \end{aligned}$$

where $B_1=E\left[\bigtriangledown g(z;\theta_0)\left[\bar{b}\right]\right]$ and $B_2=\lim_{S\to\infty}S^{\alpha_2}E\left[\bigtriangledown g(z;\theta_0)\left[\psi_S,\psi_S\right]\right]$. Thus, the leading biases due to the approximation can be written as:

$$E\left[G_n(\theta_0,\hat{\gamma}_S)-G_n(\theta_0,\gamma)\right]=H_0B_1S^{-\beta}+H_0B_2S^{-\alpha_2}+o\left(S^{-\beta}\right)+o\left(S^{-\alpha_2}\right).$$

From the proof of Theorem 1 it therefore follows

$$E\left[\hat{\theta}_{n,S}-\hat{\theta}_n\right]\simeq B_1S^{-\beta}+B_2S^{-\alpha_2}+o\left(S^{-\beta}\right)+o\left(S^{-\alpha_2}\right).$$

We then wish to obtain an estimator of (parts of) the bias $B_1S^{-\beta}+B_2S^{-\alpha_2}$, and use this for bias correction. We here propose to do this by resampling methods: First, compute two approximators of order S^* which we denote $\hat{\gamma}_{S^*}^{[1]}$ and $\hat{\gamma}_{S^*}^{[2]}$. Let $\hat{\theta}_{n,S^*}^{[m]}$ be the estimator based on the same data sample $\mathcal{Z}_n$ but using the mth approximator $\hat{\gamma}_{S^*}^{[m]}$, $m=1,2$. This has the following bias:

$$E\left[\hat{\theta}_{n,S^*}^{[m]}-\hat{\theta}_n\right]\simeq B_1\left(S^*\right)^{-\beta}+B_2\left(S^*\right)^{-\alpha_2}+o\left(S^{-\beta}\right)+o\left(S^{-\alpha_2}\right).$$

We then propose the following jackknife (JK) type estimator:

$$\hat{\theta}_{n,S}^{\text{JK}}:=2\hat{\theta}_{n,S}-\frac{1}{2}\left(\hat{\theta}_{n,S^*}^{[1]}+\hat{\theta}_{n,S^*}^{[2]}\right), \tag{26}$$

[9] See Hahn and Newey (2004), Dhaene and Jochmans (2010), Gouriéroux, Phillips and Yu (2007) for bias correction using Jackknife in the context of panel models, while we refer to Phillips and Yu (2005) for a time series application.

and we easily see that

$$\begin{aligned} E\left[\hat{\theta}^{\mathrm{JK}}_{n,S}-\hat{\theta}_n\right] &= 2E\left[\hat{\theta}_{n,S}-\hat{\theta}_n\right]-\frac{1}{2}\left(E\left[\hat{\theta}^{[1]}_{n,S^*}-\hat{\theta}_n\right]+E\left[\hat{\theta}^{[2]}_{n,S^*}-\hat{\theta}_n\right]\right) \\ &\simeq B_1\left[2S^{-\beta}-(S^*)^{-\beta}\right]+B_2\left[2S^{-\alpha}-(S^*)^{-\alpha_2}\right], \end{aligned}$$

where higher-order terms have been ignored. We would now ideally choose S^* such that both of the above bias terms cancel out. However, we can only remove either of the two: By choosing either

$$S^*=\frac{S}{2^{1/\beta}} \text{ or } S^*=\frac{S}{2^{1/\alpha_2}}, \tag{27}$$

we will remove the first or the second term respectively. Obviously, S^* should be chosen so as to remove the bias component that dominates in the expansion.

One can generalize the above and compute M approximators, $\hat{\gamma}^{[m]}_{S_m}$, $m=1,...,M$, of order $S_m<S$, and for each of those the corresponding approximate estimator, $\hat{\theta}^{[m]}_{n,S_m}$. For a given set of weights p_m, $m=1,...,M$, we then define the adjusted estimator as

$$\hat{\theta}^{\mathrm{JK}}_{n,S}=2\hat{\theta}_{n,S}-\frac{1}{2}\sum_{m=1}^{M}p_m\hat{\theta}^{[m]}_{n,S_m}. \tag{28}$$

However, if the main objective is to remove the first-order bias term, Dhaene and Jochmans (2010, Section 3.1) demonstrate in a panel data context that the optimal procedure in terms of minimum bias is $M=2$, $p_m=1/2$ and $S_m=S/2$. We expect that this result carries over to our setting as well. On the other hand, the generalized adjustment as given in eq. (28) can be used to remove further higher-order bias components by appropriate choice of weights and appproximation orders, c.f. Dhaene and Jochmans (2010, Section 3.2). While we do not pursue this here, we conjecture that the generalized adjustment would enable us to remove both B_1 and B_2.

The implementation of the above Jackknife procedure can be computationally time-consuming. In particular, one has to carry out additional two minimization routines. This can be bypassed by using a Newton-Raphson procedure, leading to a Jackknife version of the k-step bootstrap of Andrews (2002a): For each $m=1,2$, compute

$$\hat{\theta}^{[m,k+1]}_{n,S^*}=\hat{\theta}^{[m,k]}_{n,S^*}-\left[\frac{\partial G_n(\hat{\theta}^{[m,k]}_{n,S^*},\hat{\gamma}^{[m]}_{S^*})}{\partial\theta}\right]^{-1}G(\hat{\theta}^{[m,k]}_{n,S^*},\hat{\gamma}^{[m]}_{S^*}),\quad k=1,2,3,... \tag{29}$$

with starting value $\hat{\theta}^{[m,1]}_{n,S^*}=\hat{\theta}_{n,S}$, and compute $\hat{\theta}^{\mathrm{JK}}_{n,S}$ with $\hat{\theta}^{[m,k+1]}_{n,S^*}$ replacing $\hat{\theta}^{[m]}_{n,S^*}$.

An alternative way to reduce the computational cost is to jackknife the objective function directly: Define

$$G_n^*(\theta, \hat{\gamma}_S) = 2G_n(\theta, \hat{\gamma}_S) - \frac{1}{2}\left[G_n(\theta, \hat{\gamma}_{S^*}^{[1]}) + G_n(\theta, \hat{\gamma}_{S^*}^{[2]})\right],$$

By the same computations as before,

$$E\left[G_n^*(\theta, \hat{\gamma}_S) - G_n(\theta, \gamma)\right] \simeq H_0 B_1 \left[2S^{-\beta} - (S^*)^{-\beta}\right] + H_0 B_2 \left[2S^{-\alpha_2} - (S^*)^{-\alpha_2}\right]$$

By choosing S^* as in eq. (27), we remove either of the two dominating bias terms. Thus, the estimator defined by

$$G_n^*(\tilde{\theta}_{n,S}^{\mathrm{JK}}, \hat{\gamma}_S) = 0 \tag{30}$$

is equivalent to $\hat{\theta}_{n,S}^{\mathrm{JK}}$ given in eq. (26) in terms of bias.

In contrast to the analytical bias correction, the resampling-based correction can remove the leading term of the bias for both stochastic and non-stochastic approximation schemes. Another advantage of this alternative bias adjustment method is that we expect it to remove finite-sample biases. Since we are here focusing on biases due to approximation errors, we will merely give the intuition. Suppose that the approximate estimator suffers from biases of order $n^{-\nu}$ *relative to the true value* due to finite samples. That is,

$$E\left[\hat{\theta}_{n,S} - \theta_0\right] \simeq B_1 S^{-\beta} + B_2 S^{-\alpha_2} + B_3 n^{-\nu},$$

where we have suppressed any higher-order terms. Note that we here consider $E\left[\hat{\theta}_{n,S} - \theta_0\right]$ instead of $E\left[\hat{\theta}_{n,S} - \hat{\theta}_n\right]$. Then, by the same arguments as before, it is easily seen that $\hat{\theta}_{n,S}^{\mathrm{JK}}$ also removes the third term, $B_3 n^{-\nu}$, for any choice of S^*.

6 Newton-Raphson Adjustment

We here propose an additional adjustment that also works for general approximation-based estimators. We show that starting from either $\bar{\theta}_{n,S} = \hat{\theta}_{n,S}^{\mathrm{AB}}$, $\hat{\theta}_{n,S}^{\mathrm{JK}}$ 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. (5).

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

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

which in turn implies

$$||\hat{\theta}_{n,S}^{(k+1)} - \hat{\theta}_n|| = O_P\left(||\bar{\theta}_{n,S} - \hat{\theta}_n||^{2^k}\right) + O_P\left(||\bar{\theta}_{n,S^*} - \hat{\theta}_n||\right). \tag{31}$$

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 $\bar{\theta}_{n,S^*}$. In order to give a general result covering the different choices of the initial estimator $\bar{\theta}_{n,S}$, we assume that

Theorem 4 *Assume that the initial estimator is computed with S such that $||\bar{\theta}_{n,S} - \hat{\theta}_n|| = O_P(n^{-r})$ and the 2nd step NR-iterations are done with S^* such that $||\bar{\theta}_{n,S^*} - \hat{\theta}_n|| = O_P(n^{-r^*})$ where $0 < r \leq r^*$. Then with $k > [\log(r^*/r)/\log(2)]$, the NR-estimator $\hat{\theta}_{n,S}^{(k+1)}$ defined in eq. (5) satisfies:*

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

Note here that we only require that the initial estimator converges at some rate $r > 0$. Thus, we do not require the initial estimator to be $\sqrt{n}$-consistent, merely consistent. Moreover, if the NR-estimator goes to infinity with n at the same speed as the initial one, then $r = r^*$ and the formula shows that one iteration is enough.

The above result holds under very general, but rather high-level assumptions regarding the first-step estimator and the target estimator, $\bar{\theta}_{n,S}$ and $\bar{\theta}_{n,S^*}$. We can employ Theorems 1 and 2 and the results for the jackknife estimator to verify these high-level conditions for specific choices of $\bar{\theta}_{n,S}$. For example, let $\bar{\theta}_{n,S}$ be the unadjusted simulation-based ECA in Examples 1-3 where unbiased simulators are employed, an. In this case, the initial estimator satisfies $||\bar{\theta}_{n,S} - \hat{\theta}_n|| = O_P\left(S^{-1/2}\right) + O_P\left(n^{-1/2}S^{-1}\right)$, where only the leading terms have been included. Thus, with $S = Cn^{-a}$ and $S^* = C^*n^{-a^*}$, we obtain $||\bar{\theta}_{n,S} - \hat{\theta}_n|| = O_P\left(n^{-a/2}\right) + O_P\left(n^{-1/2-a}\right)$ such that $r = \min\{a/2, 1/2 + a\}$ and $r^* = \min\{a^*/2, 1/2 + a^*\}$.

The above iterative estimator requires computation of the Hessian, $H_n(\theta, \hat{\gamma}_S)$. 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.

Finally, we conjecture that the above theoretical result can be extended along the lines of Andrews (2002b) to demonstrate improvements in terms of convex variational distance, thereby establishing higher order asymptotic efficiency.

7 A Simulation Study

To explore the performance of our proposed approaches, we set up a small Monte Carlo study of a mixed logit model: the econometrician observes i.i.d. draws of (x_i, y_i) for $i = 1, \ldots, n$, with x_i a centered normal of variance τ^2 and

$$y_i = \mathbb{1}(b + (a + su_i)x_i + e_i > 0)$$

where e_i is standardized type I extreme value and u_i is a centered normal with unit variance, independent of e_i.

The mixed logit, in its multinomial form, has become a workhorse in studies of consumer demand (see e.g. the book by Train (2009)); it also figures prominently on the demand side of models of empirical industrial organization. It is usually estimated by simulation-based methods, or by Monte Carlo Markov Chains techniques. In empirical IO, the simulated method of moments is more commonly used because of endogeneity concerns; but it is not a useful benchmark for us as the approximate estimator in SMM inherits no additional bias from the simulations. We focus here on SML, which is perhaps the most popular method outside of empirical IO.

We ran experiments for several sets of parameter values; since the results are similar, we only present here those we obtained when the true model has $a = 1, s = 1, b = 0$ and the covariate has a standard error $\tau = 1$ or 2.

In these two specifications, the mean probability of $y = 1$ is close to 0.5; and the generalized R^2 is respectively 0.21 and 0.11. In the corresponding simple logit model (which has $s = 0$) the R^2 would be 0.39 and 0.17. Thus these two choices of parameters yield models that have low to fairly high explanatory power.

The mixed logit is still a very simple model; thus we can use Gaussian quadrature to compute the integral

$$\Pr(y = 1|x) = \int \frac{\phi(u)}{1 + \exp(-(b + (a + su)x)} du. \tag{32}$$

Since Gaussian quadrature achieves almost correct numerical integration in such a regular, one-dimensional case, we can rely on it to do (almost) exact maximum likelihood estimation. By the same token, it is easy to compute the asymptotic variance of the exact ML estimator $\hat{\theta}_n$, and the leading term of the bias of the SML estimator. Simple calculations give the numbers in Table 1.

The columns labeled $\sqrt{n}\hat{\sigma}$ give the square roots of the diagonal terms of the inverse of the Fisher information matrix. As can be seen from the values of $\sqrt{n}\hat{\sigma}$, it takes a large number of observations to estimate this model reliably. To take an example, assume that the econometrician would be happy with a modestly precise 95% confidence interval of half-

τ	$\sqrt{n}\hat{\sigma}$			S times bias		
	a	s	b	a	s	b
1	7.2	17.2	2.4	-9.0	-23.5	-0.0
2	6.7	10.8	2.8	-8.3	-13.5	-0.0

Table 1: Rescaled asymptotic standard errors and simulation biases

diameter 0.2 for the mean slope a. With $\tau = 1$ it would take about $(7.2/0.2)^2 \simeq 5,000$ observations; and still about $4,300$ for $\tau = 2$, even though the model has a generalized R^2 that is larger by half. With such sample sizes, the estimate of the size of the heterogeneity s would still be very noisy: the 95% confidence intervals would have half-diameters 0.48 and 0.32, respectively. We also found that the correlation between the estimators of a and of s is always large and positive—of the order of 0.8. Thus the confidence region for the pair (a, s) is in fact a rather elongated ellipsoid. On the other hand, the estimates of b are reasonably precise, which is not very surprising as b shifts the mean probability of $y = 1$ strongly.

The figures in the columns labeled "S times bias" refer to the expansions of $\hat{\theta}_{nS} - \hat{\theta}$ in our theorems. We will be using SML under the EIA scheme (independent draws). Then we know that the leading term of the bias due to the simulations is $B_{S,2}$ and is of order $1/S$. The figures give our numerical evaluation of $SB_{S,2}$, using our formulæ and Gaussian quadrature again. As appears clearly from Table 1, once again the heterogeneity coefficient s is the harder to estimate, followed by a, while there is hardly any bias on b. With $S = 200$ simulations for instance, the biases on s are -0.12 for $\tau = 1$ and -0.07 for $\tau = 2$. For sample sizes of a couple thousand observations, they are actually much smaller than the dispersion of the estimates implied by the parametric efficiency bounds; but they become more relevant in larger samples.

We used various sample sizes n and number of draws S. We ran 1,000 simulations in each case, starting from initial values of the parameters drawn randomly from uniform distributions:

$$a \sim U[0.5, 1.5], \quad b \sim U[-0.5, 0.5], \quad s \sim U[0.5, 1.5].$$

For each simulated sample, we estimated the model using both analytic bias correction (ABA), jackknife (JK) and Newton-Raphson. The ABA was done on the objective funtion, while the JK on the estimator itself.

1. exact ML (using adaptive Gaussian quadrature as in equation 32)

2. SML with S independent draws of u_i for each observations

3. SML with S draws + one NR step with $S^* = 10 \times S$ draws

4. SML with S draws + ABA.

5. SML with S draws + ABA + one NR step with $S^* = 10 \times S$ draws
6. SML with S draws + JK with $S^* = S/2$ draws.

We present below the results for $n = 5,000$ and $n = 25,000$, using $S = 200$ simulations. Given our discussion of the Fisher bounds, there is little point in considering smaller samples as the dispersion of the MLE would swamp the bias. As for S, we obtained very similar results for $S = 100$, with larger biases due to the approximation of course.

We faced very few numerical difficulties. The optimization algorithm sometimes stopped very close to the bounds we had imposed for the heterogeneity parameter, $0.1 \leq s \leq 5$. In even fewer cases it failed to find an optimum. Finally, the second derivative of the simulated log-likelihood was not invertible in a very small number of samples. Altogether, we had to discard 7 to 18 of the 1,000 samples, depending on the run. The tables and graphs below only refer to the remaining samples. We focus on a and s since there is little bias to correct for in b. We report (Huber) robust means, standard errors and RMSEs; the robustness correction only matters in a few cells of the table where the Newton correction generates estimates of s that are unusually large.

Tables 2 and 3 report our results for the smaller and the larger sample size, both when covariates have little explanatory power ($\tau = 1$) and when they have more power ($\tau = 2$). All numbers in the last five rows of these tables pertain to the bias due to the approximation; that is, we compute the "error terms" $\hat{\theta}_{n,S} - \hat{\theta}_n$, and we average them over the 1,000 samples (minus the small number that were eliminated due to numerical issues). The standard error of these averages is about 0.002, so that many of the biases from the corrected estimates are not only small, but actually insignificant.

The "SML" rows in the tables report the bias of the uncorrected SML estimator. They are very similar in both tables, as they should be. Building on Table 1, it is easy to see that the theoretical values of the leading term of the bias are

- for $\tau = 1$: -0.045 for a and -0.117 for s
- for $\tau = 2$: -0.042 for a and -0.068 for s.

Therefore the leading term is a very good approximation to the actual size of the bias in these simulations; and the two methods that focus on correcting it, our analytical bias adjustment (ABA in the tables) and the resampling method, should work very well. ABA in fact does eliminate most of the bias; resampling also works quite well, with the exception of the s estimator for $\tau = 1$ for which there is still a small bias. The Newton step with $2,000$ simulations reduces the bias, as expected; but it does not do quite as well as ABA and resampling.

The discussion above only bears on bias, but one may legitimately be concerned about the possibility that our adjustment procedures introduce more noise into the estimates. Figures 1

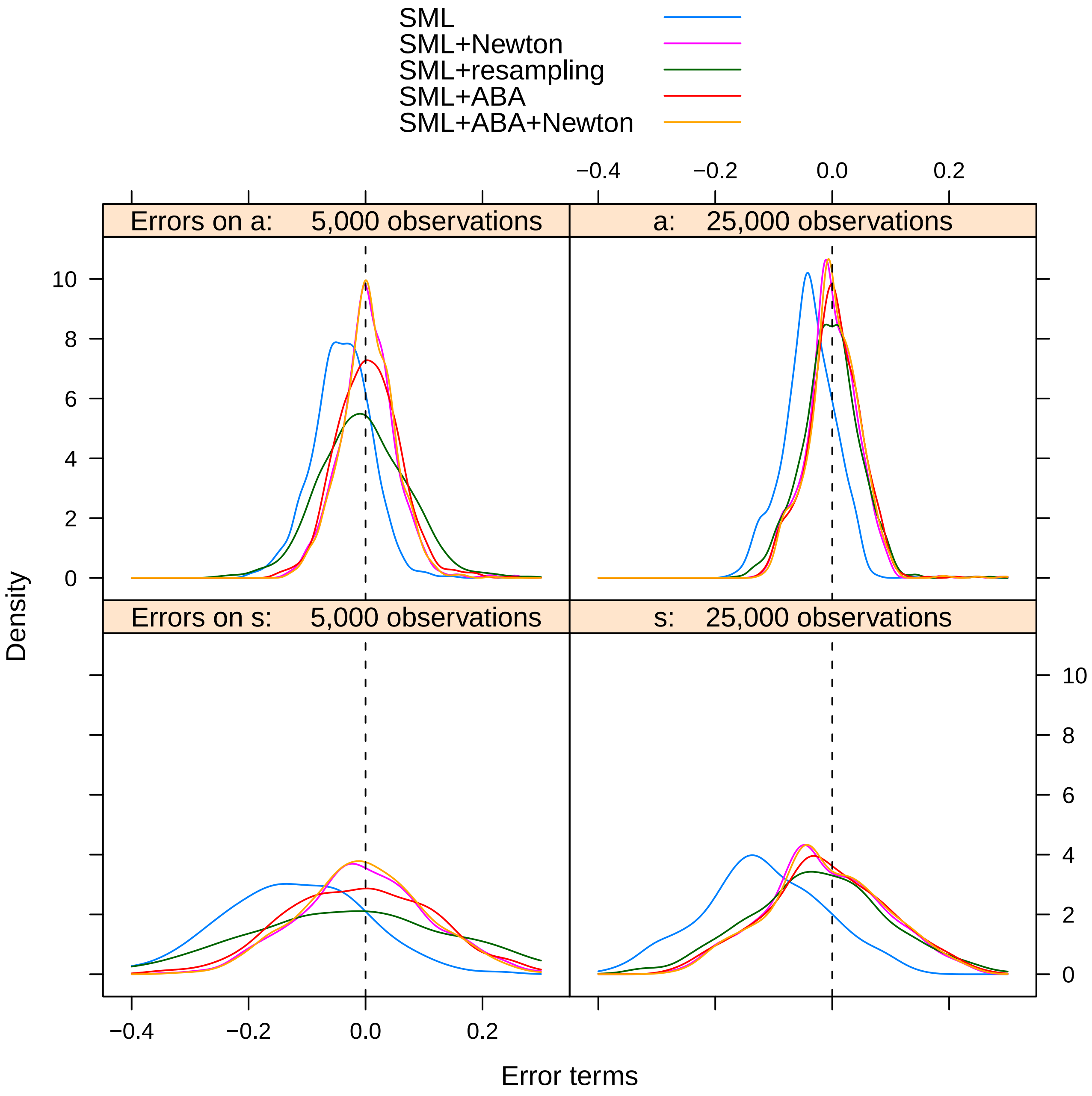

Figure 1: Density of $\hat{\theta}_{nS} - \hat{\theta}_n$ when $\tau = 1$

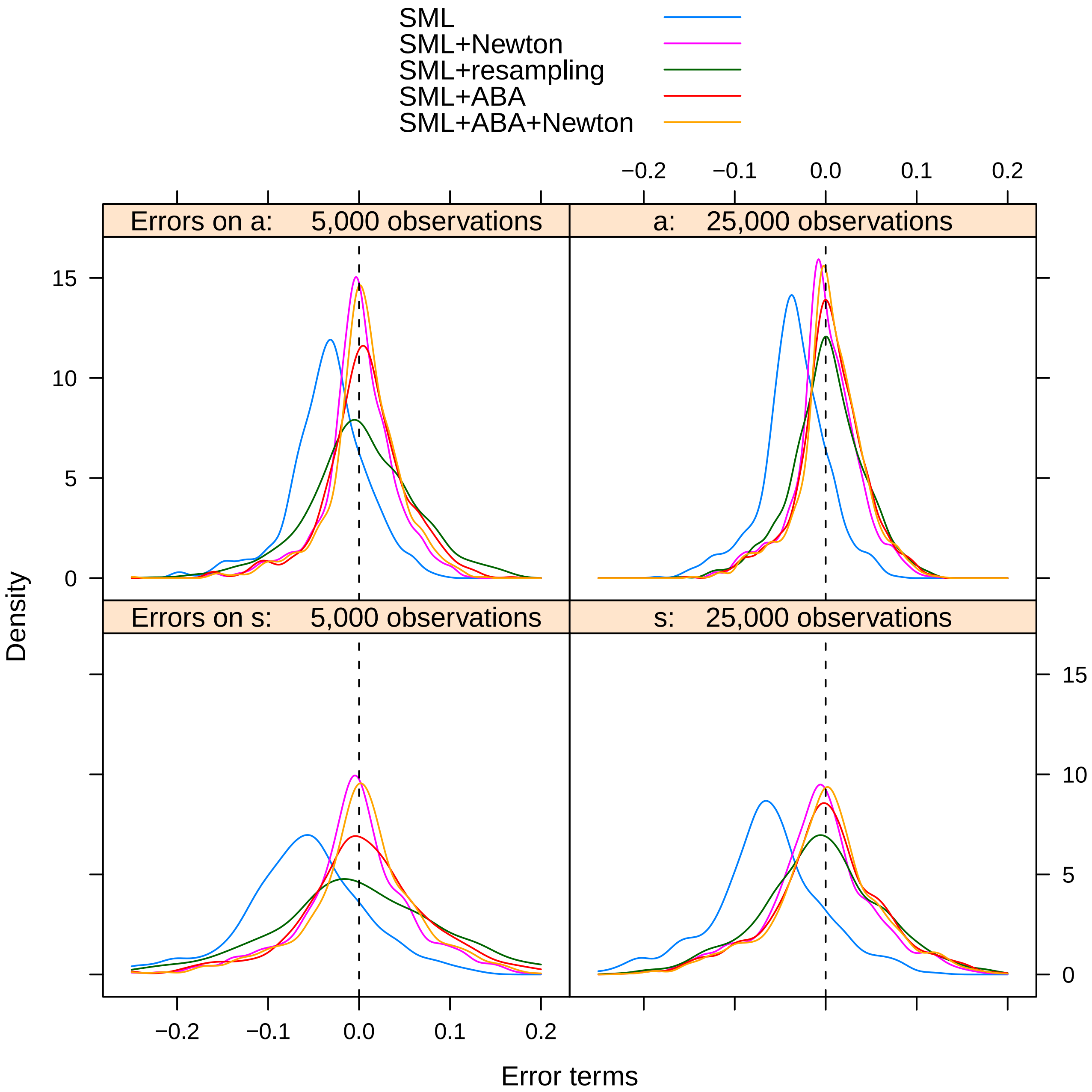

Figure 2: Density of $\hat{\theta}_{nS} - \hat{\theta}_n$ when $\tau = 2$

Method	$\tau = 1$		$\tau = 2$	
	a	s	a	s
SML	-0.040	-0.121	-0.033	-0.059
SML+Newton	0.003	0.003	0.001	-0.004
SML+resampling	-0.004	-0.022	0.002	-0.002
SML+ABA	0.004	-0.007	0.006	0.006
SML+ABA+Newton	0.004	-0.002	0.006	0.004

Table 2: Biases and their corrections, $n = 5,000$ observations

Method	$\tau = 1$		$\tau = 2$	
	a	s	a	s
SML	-0.038	-0.118	-0.032	-0.063
SML+Newton	-0.001	-0.014	0.001	-0.006
SML+resampling	-0.003	-0.024	0.003	-0.005
SML+ABA	0.004	-0.009	0.008	0.001
SML+ABA+Newton	0.003	-0.010	0.007	0.001

Table 3: Biases and their corrections, $n = 25,000$ observations

and 2 plot the estimated densities of the error terms $\hat{\theta}_{n,S} - \hat{\theta}_n$ when $\tau = 1$ and $\tau = 2$. The improvements in the biases are obvious. More interesting is the contrasting performance of the methods when it comes to the dispersion of the errors. While our analytical bias adjustment hardly changes the dispersion, the Newton procedure reduces it; and the resampling procedure increases it. Since the Newton adjustment aims at giving the estimator the asymptotic properties of one with 10 times more simulations, it reduces the efficiency loss relative to the MLE. On the other hand, resampling corrects the $S = 200$ estimator by using an average of estimators with $S = 100$, and so it introduces more noise.

These trade-offs are reflected in the RMSEs of the error terms, as collected in tables 4 and 5. Two other considerations are worth mentioning:

- *Ease of implementation:* The resampling method wins on that count; the analytical bias adjustment is not far behind, since it is usually easy to get a formula for the Δ term and to program it. The Newton method may be more troublesome in model with more than a few parameters, as it requires a reasonably accurate evaluation of the matrix of second derivatives.

- *Computer time:* Here, the analytical bias adjustment wins hands down. For SML for instance, the evaluation of the corrected objective function requires the variance of the simulated p's in addition to their mean—-a very small computational cost. Resampling, as implemented in this study, roughly doubles the cost of the uncorrected estimator;

and Newton can be more costly still, depending on the structure of the model and the care needed to estimate the Hessian.

Like all Monte Carlo studies, ours can only be illustrative; yet our results suggest that the resampling method is dominated by the other two. If the Hessian is easy to compute with enough accuracy, then the Newton method is probably the best choice; otherwise, the analytical bias adjustment seems to be a good choice, at least if the bias induced by the approximations is the main concern.

Method	$\tau = 1$		$\tau = 2$	
	a	s	a	s
SML	0.062	0.173	0.046	0.086
SML+Newton	0.045	0.116	0.031	0.053
SML+resampling	0.073	0.189	0.038	0.092
SML+ABA	0.052	0.129	0.033	0.065
SML+ABA+Newton	0.044	0.109	0.032	0.054

Table 4: Root mean squared errors, $n = 5,000$ observations

Method	$\tau = 1$		$\tau = 2$	
	a	s	a	s
SML	0.058	0.159	0.049	0.084
SML+Newton	0.042	0.105	0.032	0.053
SML+resampling	0.047	0.121	0.055	0.064
SML+ABA	0.044	0.109	0.040	0.056
SML+ABA+Newton	0.042	0.105	0.034	0.053

Table 5: Root mean squared errors, $n = 25,000$ observations

8 Conclusion

We developed in this paper a unifying framework for the analysis of approximate estimators. We derived bias and variance rates of the approximate estimator relative to the exact estimator, and used them to propose three methods for reducing the bias and the efficiency loss that result from the approximation. Simulations on the mixed logit model confirm that the proposed methods work well in finite samples.

We restricted ourselves to estimators solving a first-order condition given in eq. (6). It would be of interest to extend our results to a more general setting. Consider the case of non-smooth objective functions and non-smooth approximators (as functions of θ). In principle, one could import the arguments of Chen et al (2003) for semiparametric estimators in order to handle this complication. Another approach would be to employ a slight generalization of Robinson (1988, Theorem 1) which in our setting would yield

$$||\hat{\theta}_{n,S} - \tilde{\theta}_n|| = O_P\left(\sup_{\|\theta-\theta_0\|\leq\delta} \|G_n(\theta,\hat{\gamma}_S) - G_n(\theta,\gamma)\|\right) + o_P\left(1/\sqrt{n}\right),$$

for some $\delta > 0$. If one could then strengthen the pointwise bias and variance results derived here to hold uniformly over $\|\theta - \theta_0\| \leq \delta$, all our results would remain valid. To extend our results to hold uniformly, one could rely on standard uniform convergence results as developed in, e.g. van der Vaart and Wellner (1996).

Also, we require the approximators to be mutually independent, which rules out certain recursive approximation schemes such as particle filtering. Establishing results for this more complicated case would be highly useful. One could here try to use the results of Chen and White (1998, 2002) who analyze random dynamic function systems.

Finally, we only allowed for one source of approximation in γ. More general situations could have several such terms, possibly with quite different properties. As an example, we could have evaluate a quantity γ_1 using simulations, and another term γ_2 by discretizing over a grid and interpolating. We could still write a Taylor expansion as in section 3.1, and evaluate the corresponding terms. While we have not formally explored this extension, we feel that we can venture some conjectures. The Newton method would still work, using here both a larger number of simulations and a more precise grid in computing the Newton correction. The analytical bias-adjustment method would only work if all sources of approximations were "stochastic" (unlike γ_2 in our example); and then one would focus on the approximation whose size goes to zero most slowly. As for the resampling method, we would need to use different choices of m and S^* along the various dimensions of approximation.

References

Altissimo, F. and A. Mele (2009) Simulated Nonparametric Estimation of Dynamic Models. *Review of Economic Studies* 76, 413-450.

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.

Andrews, D.W.K. (2002a) Higher-order Improvements of a Computationally Attractive k-step Bootstrap for Extremum Estimators. *Econometrica* 70, 119-162.

Andrews, D.W.K. (2002b) Equivalence of the Higher-order Asymptotic Efficiency of k-step and Extremum Statistics. *Econometric Theory* 18, 1040-1065.

Berry, S., Levinsohn, J., and Pakes, A. (1995) Automobile Prices in Market Equilibrium. *Econometrica* 63, 841-890.

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.

Chen, X., O. Linton and I. Van Keilegom (2003) Estimation of Semiparametric Models When the Criterion Function Is Not Smooth. *Econometrica* 71, 1591-1608.

Chen, X. and H. White (1998) Nonparametric Adaptive Learning with Feedback. *Journal of Economic Theory* 82, 190 222.

Chen, X. and H. White (2002) Asymptotic Properties of Some Projection-Based Robbins-Monro Procedures in a Hilbert Space. *Studies in Nonlinear Dynamics & Econometrics* 6(1), Article 1.

Creel, M. and D. Kristensen (2009) Estimation of Dynamic Latent Variable Models Using Simulated Nonparametric Moments. Manuscript, Department of Economics, Universitat Autònoma de Barcelona.

Corradi, V. and N.R. Swanson (2007) Evaluation of Dynamic Stochastic General Equilibrium Models Based on Distributional Comparison of Simulated and Historical Data. *Journal of Econometrics* 136, 699-723.

Dhaene, G. and K. Jochmans (2010) Split-panel Jackknife Estimation of Fixed-effect Models. Manuscript, Katholieke Universiteit Leuven.

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.

Ferger, D. (1996) Moment Inequalities for U-Statistics with Degeneracy of Higher Order. *Sankhyā: The Indian Journal of Statistics, Series A*, 58, 142-148.

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 Models. *Econometrica* 74, 93-119.

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

Gouriéroux, C., P.C.B. Phillips and J. Yu (2007) Indirect Inference for Dynamic Panel Models. Forthcoming in *Journal of Econometrics.*

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.

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.

Judd, K. and C. Su (2007) Constrained Optimization Approaches to Estimation of Structural models. Working paper, CMS-EMS.

Kristensen, D. (2009) Uniform Convergence Rates of Kernel Estimators with Heterogeneous, Dependent Data. *Econometric Theory* 25, 1433-1445.

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

Kristensen, D. and A. Rahbek (2005) Asymptotics of the QMLE for a Class of ARCH(q) Models. *Econometric Theory* 21, 946-961

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

Newey, W.K. (1991) Uniform Convergence in Probability and Stochastic Equicontinuity, *Econometrica* 59, 1161-1167.

Newey, W.K. (1991) Kernel Estimation of Partial Means and a General Variance Estimator. *Econometric Theory* 10, 233-253.

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. and D. McFadden (1994) Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics*, Vol. 4 (eds. R.F. Engle and D.L. McFadden), Chapter 36. Elsevier Science B.V.

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.

Nze, P.A. and P. Doukhan (2004) Weak Dependence: Models and Applications to Econometrics. *Econometric Theory* 20, 995-1045.

Norets, A. (2009) Inference in Dynamic Discrete Choice Models with Serially Correlated Unobserved State Variables. *Econometrica* 77, 1665–1682.

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.

Pollard, D. (1985) New Ways to Prove Central Limit Theorems. *Econometric Theory* 1, 295-314.

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

Rio, E. (1994) Inégalités de Moments pour les Suites Stationnaires et Fortement Mélangeantes. *Comptes Rendus de l'Académie des Sciences* 318, 355–360.

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

Train, K. (2009), *Discrete Choice Methods with Simulation*, Cambridge University Press.

van der Vaart, A & J. Wellner (1996) *Weak Convergence and Empirical Processes.* Springer-Verlag.

Yoshihara, K. (1976) Limiting Behaviour of U-Statistics for Stationary, Absolutely Regular Processes. *Zeitschrift für Wahrenscheinlichkeittheorie und verwandte Gebeite* 35, 237-252.

A Examples

In this appendix, we give more details for the SPML (Example 2.2) and the NPSMLE (Example 3.2).

A.1 Example 2.2 (SPML)

We here derive the first and second order differentials for the SPML estimator, and obtain an expression of the analytical bias adjustment.

With $m_i(\theta) = m(x_i;\theta)$, $v_i(\theta) = v(x_i;\theta)$ and $\xi_i(\theta) = y_i - m_i(\theta)$,

$$g_i(\theta,\gamma) = \frac{\partial}{\partial\theta}\left\{\log\left(v_i(\theta)\right) + \frac{\xi_i^2(\theta)}{v_i(\theta)}\right\} = \frac{\dot{v}_i(\theta)}{v_i(\theta)} - \frac{2\xi_i(\theta)\,\dot{m}_i(\theta)}{v_i(\theta)} - \frac{\xi_i^2(\theta)\,\dot{v}_i(\theta)}{v_i^2(\theta)}.$$

Thus, with $d\gamma_i = (dm_i, dv_i)$ and $dm_i = dm(x_i)$ and $dv_i = dv(x_i)$ denoting mean and variance directions,

$$\begin{aligned}\triangledown g_i(\theta)\,[d\gamma] &= \triangledown_m g_i(\theta)\,[dm] + \triangledown_v g_i(\theta)\,[dv] \\ \triangledown^2 g_i(\theta)\,[d\gamma, d\gamma] &= \triangledown^2_{m,m} g_i(\theta)\,[dm, dm] + 2\,\triangledown^2_{g,m}\, g_i(\theta)\,[dm, dv] + \triangledown^2_{v,v} g_i(\theta)\,[dv, dv]\,,\end{aligned}$$

where, by easy but tedious calculations,

$$\triangledown_m g_i(\theta)\,[dm] = \frac{2}{v_i(\theta)}\left\{\dot{m}_i(\theta) + \frac{\xi_i(\theta)\,\dot{v}_i(\theta)}{v_i(\theta)}\right\} dm_i(\theta) - \frac{2\xi_i(\theta)}{v_i(\theta)} d\dot{m}_i(\theta)\,,$$

$$\triangledown_v g_i(\theta)\,[dv] = \frac{1}{v_i(\theta)}\left\{1 - \frac{\xi_i^2(\theta)}{v_i(\theta)}\right\} d\dot{v}_i(\theta) + \frac{1}{v_i^2(\theta)}\left\{2\xi_i(\theta)\,\dot{m}_i(\theta) + \frac{2\xi_i^2(\theta)\,\dot{v}_i(\theta)}{v_i(\theta)} - \dot{v}_i(\theta)\right\} dv_i(\theta)$$

$$\triangledown^2_{m,m} g_i(\theta)\,[dm, dm] = \frac{1}{v_i(\theta)}\left\{4d\dot{m}_i(\theta) - \frac{2\dot{v}_i(\theta)}{v_i(\theta)} dm_i(\theta)\right\} dm_i(\theta)\,,$$

$$\begin{aligned}\triangledown^2_{m,v} g(z_i;\theta)\,[dm, dv] &= \frac{2\xi_i(\theta)}{v_i^2(\theta)} dm_i(\theta)\,d\dot{v}_i(\theta) + \frac{2\xi_i(\theta)}{v_i^2(\theta)} d\dot{m}_i(\theta)\,dv_i(\theta) \\ &\quad - \frac{1}{v_i^2(\theta)}\left\{2\dot{m}_i(\theta) + \frac{4\xi_i(\theta)\,\dot{v}_i(\theta)}{v_i(\theta)}\right\} dm_i(\theta)\,dv_i(\theta)\end{aligned}$$

$$\begin{aligned}\triangledown^2_{v,v} g(z;\theta)\,[dv, dv] &= \frac{2}{v_i^2(\theta)}\left\{\frac{2\xi_i^2(\theta)}{v_i(\theta)} - 1\right\} dv_i(\theta)\,d\dot{v}_i(\theta) \\ &\quad - \frac{2}{v_i^3(\theta)}\left\{2\xi_i(\theta)\,\dot{m}_i(\theta) + \left(3\frac{\xi_i^2(\theta)}{v_i(\theta)} - 1\right)\dot{v}_i(\theta)\right\} dv_i(\theta)^2\end{aligned}$$

In contrast to Example 2.1, the third order differential is non-zero. It can still easily be checked that Eqs. (8)-(10) hold with $\bar{G}_k = E\left[\bar{g}_k(z_i)\right]$, $k = 1, 2, 3$, where

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

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

Given the above differentials, we can derive an expression of the analytical bias adjustment. Suppose the simulated versions of the conditional mean and variance are of the form

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

We then obtain the following expression for the analytical bias adjustment[10]:

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

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

where $\hat{\xi}_i(\theta) = y_i - \hat{m}_i(x_i;\theta)$ and

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

In this case, $\bigtriangledown^3 G_n(\theta,\gamma)[d\gamma] \neq 0$ and so the bias adjustment does not ensure consistency for fixed S.

[10] If two independent batches of simulated draws are used to compute $\hat{m}$ and $\hat{v}$, then $\Delta_{n,S}^{(2)}(\theta)$ has mean zero and can be left out in the computation of $\Delta_{n,S}(\theta)$.

A.2 Example 3.2 (NPSMLE)

We here derive the optimal rate for the bandwidth used in NPSMLE: The bias component of the first order term is

$$\begin{aligned} \triangledown g_i(\theta)\left[b_S\right] &= \frac{\dot{p}\left(y_i|x_i;\theta\right)}{p^2\left(y_i|x_i;\theta\right)} b_S\left(y_i|x_i;\theta\right) - \frac{1}{p\left(y_i|x_i;\theta\right)} \dot{b}_S\left(y_i|x_i;\theta\right) \\ &= h^r\left\{\frac{\dot{p}\left(y_i|x_i;\theta\right)}{p^2\left(y_i|x_i;\theta\right)} \frac{\partial^r p\left(y_i|x_i;\theta\right)}{\partial y_i^r} - \frac{1}{p\left(y_i|x_i;\theta\right)} \frac{\partial^r \dot{p}\left(y_i|x_i;\theta\right)}{\partial y_i^r}\right\} + o\left(h^r\right). \end{aligned}$$

This holds irrespectively of whether a single simulation batch (ECA) or n (EIA) simulation batches are used.

Next, we derive the rate of the variance component of the first order term. First, consider the EIA: By Lemma 6, we obtain that

$$\begin{aligned} \mathrm{Var}\left(\triangledown G_n(\theta)[\psi_S]\right) &\leq \frac{C}{n}\left\{E\left[\left\|\hat{p}_S\left(y|x;\theta\right) - p\left(y|x;\theta\right)\right\|^2\right] + E\left[\left\|\widehat{\dot{p}}_S\left(y|x;\theta\right) - \dot{p}\left(y|x;\theta\right)\right\|^2\right]\right\} \\ &= O\left(\frac{1}{nSh^{d+2}}\right). \end{aligned}$$

Note that the $(d+2)$ term comes from the fact that we need to approximate the derivative of the loglikelihood as well as the function itself.

Next consider the ECA: we claim that

$$\triangledown G_n(\theta)[\psi_S] = \frac{1}{S}\sum_{s=1}^{S} \triangledown \bar{g}(\theta)[e_s] + O_P\left(\frac{1}{\sqrt{nSh^{d+1}}}\right) = O_P\left(\frac{1}{\sqrt{S}}\right) + O_P\left(\frac{1}{\sqrt{nSh^{d+1}}}\right).$$

The first equality follows from Lemma 6, while to show the second one we write $\triangledown \bar{g}(\theta)[e_s] = \triangledown \bar{g}_1(\theta)[e_s] + \triangledown \bar{g}_2(\theta)[e_s]$ where

$$\triangledown \bar{g}_1(\theta)[e] = E\left[\frac{\dot{p}\left(y_i|x_i;\theta\right)}{p^2\left(y_i|x_i;\theta\right)} e\left(y_i|x_i;\theta\right)\right], \quad \triangledown\ \bar{g}_1(\theta)[e] = -E\left[\frac{1}{p\left(y_i|x_i;\theta\right)} \dot{e}\left(y_i|x_i;\theta\right)\right]$$

and

$$e_s\left(y_i|x_i;\theta\right) = K\left(\frac{y - y\left(x,\varepsilon_s;\theta\right)}{h}\right) - E\left[K\left(\frac{y - y\left(x,\varepsilon_s;\theta\right)}{h}\right)\right]$$

By standard arguments,

$$\begin{aligned} \triangledown \bar{g}_1(\theta)[e_S] &= \int\int \frac{\dot{p}\left(y|x;\theta_0\right)}{p\left(y|x;\theta\right)} p\left(x\right)\left[\frac{1}{h^d} K\left(\frac{y - y\left(x,\varepsilon_s;\theta\right)}{h}\right) - h^r \frac{\partial^r p\left(y|x;\theta\right)}{\partial y^r}\right] dydx + o\left(h^r\right) \\ &= \int \frac{\dot{p}\left(y\left(x,\varepsilon_s;\theta\right)|x;\theta_0\right)}{p\left(y\left(x,\varepsilon_s;\theta\right)|x;\theta\right)} p\left(x\right) dydx + O_P\left(h^r\right), \end{aligned}$$

where

$$\int \frac{\dot{p}\left(y\left(x,\varepsilon_s;\theta_0\right)|x;\theta_0\right)}{p\left(y\left(x,\varepsilon_s;\theta_0\right)|x;\theta_0\right)} p\left(x\right) p\left(\varepsilon\right) dydxd\varepsilon = 0.$$

Thus, $S^{-1/2}\sum_{s=1}^{S}\bigtriangledown\bar{g}_1(\theta)[e_s] = O_P\left(1\right)$. Similarly, we find that $S^{-1/2}\sum_{s=1}^{S}\bigtriangledown\bar{g}_2(\theta)[e_s] = O_P\left(1\right)$.

As a consequence, for both ECA and EIA we have

$$||\hat{\theta}_{n,S}-\hat{\theta}_n|| = O_P\left(h^r\right)+O_P\left(\frac{1}{Sh^{d+2}}\right)+O_P\left(\frac{1}{\sqrt{nSh^{d+2}}}\right).$$

B Proofs

Proof of Theorem 1. We first note that under (A.1)-(A.2) and (A.3.i),

$$\sup_{\theta\in\Theta}\left\|G_n\left(\theta,\gamma_0\right)-G\left(\theta,\gamma_0\right)\right\| \to^P 0, \tag{33}$$

as $n\to\infty$; see e.g. Kristensen and Rahbek (2005, Proposition 1). This together with (A.3.ii) implies that $\hat{\theta}_n$ is consistent, see e.g. Newey and McFadden (1994, Theorem 2.1).

In the case of ECA's, we will in the following write $\hat{\gamma}_{i,S} := \hat{\gamma}_S$, $i=1,...,n$, so we do not have to treat the two approximation schemes separately. Then, by part (i)-(ii) of (A.6), for any $\lambda\leq 2$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}E\left[\left\|\hat{\gamma}_{i,S}-\gamma_0\right\|^{\lambda}\right] &\leq \frac{1}{n}\sum_{i=1}^{n}E\left[\left\|\hat{\gamma}_{i,S}-\gamma_0\right\|^{2}\right]^{\lambda/2}\\ &= \left[O\left(S^{-2\beta}\right)+O\left(S^{-\alpha_2}\right)\right]^{\lambda/2}\\ &= o\left(1\right),\end{aligned}$$

as $S\to\infty$. Thus, by (A.1), part (i) of (A.5), and part (i) of (A.6), where without loss of generality we assume $\lambda\leq 2$,

$$\begin{aligned}E\left[\sup_{\theta\in\Theta}\left\|G_n\left(\theta,\hat{\gamma}_S\right)-G_n\left(\theta,\gamma_0\right)\right\|\right] &\leq \frac{1}{n}\sum_{i=1}^{n}E\left[\sup_{\theta\in\Theta}\left\|g\left(z_i;\theta,\hat{\gamma}_S\right)-g\left(z_i;\theta,\gamma_0\right)\right\|\right]\\ &\leq \bar{G}_0\frac{1}{n}\sum_{i=1}^{n}E\left[\left\|\hat{\gamma}_{i,S}-\gamma_0\right\|^{\lambda}\right]\\ &= o_P\left(1\right) \qquad (34)\end{aligned}$$

Combining this result with eq. (33), we obtain $\sup_{\theta\in\Theta}\left\|G_n\left(\theta,\hat{\gamma}_S\right)-G\left(\theta,\gamma_0\right)\right\| \to^P 0$. Together with (A.3), this proves that $\hat{\theta}_{n,S}$ is consistent as $n,S\to\infty$; see Newey and McFadden (1994, Theorem 2.1).

To derive more precise rates of the approximate estimator, we first take a Taylor expansion of $G_n(\theta, \hat{\gamma}_S)$ w.r.t. θ:

$$o_P\left(n^{-1/2}\right) = G_n(\hat{\theta}_{n,S}, \hat{\gamma}_S) = G_n(\theta_0, \hat{\gamma}_S) + H_n(\bar{\theta}_{n,S}, \hat{\gamma}_S)(\hat{\theta}_{n,S} - \theta_0), \tag{35}$$

for some $\bar{\theta}_{n,S}$ between $\hat{\theta}_{n,S}$ and θ_0. Since $\hat{\theta}_{n,S}$ is consistent, $\bar{\theta}_{n,S} \to^P \theta_0$. By the same arguments used to establish eqs. (33)-(34), Assumption A.4 then ensures that,

$$\begin{aligned} \left\|H_n\left(\bar{\theta}_{n,S}, \hat{\gamma}_S\right) - H_0\right\| &\leq \left\|H_n\left(\bar{\theta}_{n,S}, \hat{\gamma}_S\right) - H_n\left(\bar{\theta}_{n,S}, \gamma_0\right)\right\| + \left\|H_n\left(\bar{\theta}_{n,S}, \gamma_0\right) - H\left(\bar{\theta}_{n,S}, \gamma_0\right)\right\| \\ &\quad + \left\|H\left(\bar{\theta}_{n,S}, \gamma_0\right) - H\left(\theta_0, \gamma_0\right)\right\| \\ &\leq \sup_{\|\theta-\theta_0\|\leq\delta} \left\|H_n\left(\theta, \hat{\gamma}_S\right) - H_n\left(\theta, \gamma_0\right)\right\| + \sup_{\|\theta-\theta_0\|\leq\delta} \left\|H_n\left(\theta, \gamma_0\right) - H\left(\theta, \gamma_0\right)\right\| \\ &\quad + \left\|H\left(\bar{\theta}_{n,S}, \gamma_0\right) - H\left(\theta_0, \gamma_0\right)\right\| \\ &= o_P\left(1\right). \end{aligned}$$

Going back to eq. (35), we have now shown that

$$\hat{\theta}_{n,S} - \theta_0 = -H_0^{-1} G_n(\theta_0, \hat{\gamma}_S) + o_P\left(n^{-1/2}\right),$$

while

$$\hat{\theta}_n - \theta_0 = -H_0^{-1} G_n(\theta_0, \gamma_0) + o_P\left(n^{-1/2}\right).$$

Subtracting gives

$$\hat{\theta}_{n,S} - \hat{\theta}_n = -H_0^{-1}\left\{G_n(\theta_0, \hat{\gamma}_S) - G_n(\theta_0, \gamma_0)\right\} + o_P\left(n^{-1/2}\right).$$

We now use the expansion given in eq. (11) with $m = 2$ and $\theta = \theta_0$, to get

$$\left\|\hat{\theta}_{n,S} - \hat{\theta}_n\right\| = O_P\left(\left\|\triangledown G_n(\theta_0)\left[\Delta\hat{\gamma}_S\right] + \frac{1}{2}\triangledown^2 G_n(\theta_0)\left[\Delta\hat{\gamma}_S, \Delta\hat{\gamma}_S + R_{n,S}\right]\right\|\right) + o_P(n^{-1/2}), \tag{36}$$

where $\Delta\hat{\gamma}_{i,S} = \hat{\gamma}_{i,S} - \gamma_0$. We first derive the rate of the remainder term $R_{n,S}$:

$$\begin{aligned} E\left[\|R_{n,S}\|\right] &= E\left\|G_n(\theta_0, \hat{\gamma}_S) - G_n(\theta_0, \gamma_0) - \triangledown G_n(\theta_0)\left[\Delta\hat{\gamma}_S\right] - \frac{1}{2}\triangledown^2 G_n(\theta_0)\left[\Delta\hat{\gamma}_S, \Delta\hat{\gamma}_S\right]\right\| \\ &\leq \frac{1}{n}\sum_{i=1}^n E\left\|g_i(\theta_0, \hat{\gamma}_{i,S}) - g_i(\theta_0, \gamma_0) - \triangledown g_i(\theta_0)\left[\Delta\hat{\gamma}_{i,S}\right] - \frac{1}{2}\triangledown^2 g_i(\theta_0)\left[\Delta\hat{\gamma}_{i,S}, \Delta\hat{\gamma}_{i,S}\right]\right\| \\ &\leq \frac{\bar{G}_0}{n}\sum_{i=1}^n E\left\|\Delta\hat{\gamma}_{i,S}\right\|^3, \end{aligned}$$

where we have used A.5(2).

Applying first Minkowski's inequality and then the inequality

$$(a+b)^p \leq 2^{p-1}a^p + 2^{p-1}b^p$$

(which holds for all $a, b > 0$ and $p \geq 1$), we obtain—dropping the i index:

$$\begin{aligned}
E\left\|\Delta\hat{\gamma}_S\right\|^3 &= E\left[\left\|\psi_S + (E\left[\hat{\gamma}_S\right] - \gamma_0)\right\|^3\right] \\
&\leq \left(E\left[\left\|\psi_S\right\|^3\right]^{1/3} + \left\|E\left[\hat{\gamma}_S\right] - \gamma_0\right\|\right)^3 \\
&\leq 4E\left[\left\|\psi_S\right\|^3\right] + 4\left\|E\hat{\gamma}_S - \gamma_0\right\|^3 \\
&= O\left(S^{-\alpha_3}\right) + O\left(S^{-3\beta}\right),
\end{aligned}$$

The rates of the first and second order functional differentials of $G_n(\theta_0, \gamma)$ are given in Lemmas 7 and 8 depending on whether the ECA approximator of (12) or the EIA approximator of eq. (13) is used. By plugging those into eq. (36) together with the rate of $R_{n,S}$, we obtain the desired result. ■

Proof of Theorem 2. We only give a proof for the case of EIA's; the proof for ECA's follows along the same lines. One can easily show that $\sup_{\theta\in\Theta}||\dot{\Delta}_{n,S}(\theta)|| = o_P(1)$ as $n, S \to \infty$, and it now follows by the same arguments as in the proof of Theorem 1 that $\hat{\theta}^{\text{AB}}_{n,S}$ is consistent.

Next, we make a Taylor expansion of eq. (23),

$$o_P\left(n^{-1/2}\right) = \left\{G_n(\theta_0, \hat{\gamma}_S) - \dot{\Delta}_{n,S}(\theta_0)\right\} + \left\{H_n(\bar{\theta}_{n,S}, \hat{\gamma}_S) - \ddot{\Delta}_{n,S}\left(\bar{\theta}_{n,S}\right)\right\}(\hat{\theta}^{\text{AB}}_{n,S} - \theta_0),$$

where $\ddot{\Delta}_{n,S}(\theta) = \partial\dot{\Delta}_{n,S}(\theta)/\partial\theta$. From the proof of Theorem 1, $H_n(\bar{\theta}_{n,S}, \hat{\gamma}_S) = H_0 + o_P(1)$, while it is easily shown that $\ddot{\Delta}_{n,S}\left(\bar{\theta}_{n,S}\right) = o_P(1)$ as $n, S \to 0$, so that, by the same arguments as in the proof of Theorem 1,

$$\hat{\theta}^{\text{AB}}_{n,S} - \hat{\theta}_n = O_P\left(\left\|G_n(\theta_0, \hat{\gamma}_S) - \dot{\Delta}_{n,S}(\theta_0) - G_n(\theta_0, \gamma)\right\|\right).$$

Suppressing any dependence on θ_0, use eq. (11) to write

$$\begin{aligned}
G_n\left(\hat{\gamma}_S\right) - \dot{\Delta}_{n,S} - G_n\left(\gamma\right) &= \left\{\frac{1}{2}\bigtriangledown^2 G_n[\psi_{n,S}, \psi_{n,S}] - \dot{\Delta}_{n,S}\right\} + \bigtriangledown G_n[\hat{\gamma}_S - \gamma] \\
&\quad + \frac{1}{2}\left\{\bigtriangledown^2 G_n[\hat{\gamma}_S - \gamma, \hat{\gamma}_S - \gamma] - \bigtriangledown^2 G_n[\psi_{n,S}, \psi_{n,S}]\right\} + R_{n,S}.
\end{aligned} \tag{37}$$

The rates of the second, third and fourth term of eq. (37) are derived in Lemma 8. The crucial term is the first term of eq. (37). Now, recall that $\hat{\gamma}_i = S^{-1}\sum_{s=1}^{S} w_{is}$, and that

$$\Delta_{n,S} = \frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S} \nabla g_i[w_{is} - \hat{\gamma}_i, w_{is} - \hat{\gamma}_i].$$

Thus, using the bilinearity of $(d\gamma, d\gamma') \mapsto \nabla^2 g_i\left[d\gamma, d\gamma'\right]$, and denoting $\bar{w}_i = E\left[w_{i,s}\right]$ and $e_{is} = w_{is} - \bar{w}_i$, the first term of eq. (37) can be rewritten as

$$\begin{aligned}
& \frac{1}{2}\nabla^2\, G_n[\psi_{n,S}, \psi_{n,S}] - \dot{\Delta}_{n,S} \\
= \; & \frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s\neq t} \nabla^2 g_i[e_{is}, e_{it}] + \frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S} \nabla^2 g_i[e_{is}, e_{is}] - \frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S} \nabla g_i[w_{is} - \hat{\gamma}_i, w_{is} - \hat{\gamma}_i] \\
= \; & \frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s\neq t} \nabla^2 g_i[e_{is}, e_{it}] + \frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S} \left\{\nabla^2 g_i[e_{is}, e_{is}] - \nabla g_i[w_{is} - \hat{\gamma}_i, w_{is} - \hat{\gamma}_i]\right\} \\
= \; & \frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s\neq t} \nabla^2 g_i[e_{is}, e_{it}] + \frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S} \left\{\nabla^2 g_i[\hat{\gamma}_i - \bar{w}_i, e_{is}] + \nabla^2 g_i[e_{is}, \hat{\gamma}_i - \bar{w}_i]\right\} \\
= \; & \frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s\neq t} \nabla^2 g_i[e_{is}, e_{it}] + \frac{1}{nS}\sum_{i=1}^{n} \nabla^2 g_i[\hat{\gamma}_i - \bar{w}_i, \hat{\gamma}_i - \bar{w}_i]
\end{aligned}$$

where the last equality uses the fact that $S^{-1}\sum_{s=1}^{S} e_{is} = \hat{\gamma}_i - \bar{w}_i$.

Start with the first term, and note that $E\left[\nabla^2 g_i[e_{is}, e_{it}]\right] = 0$ when $s \neq t$. Then apply Lemma 5 with $r = 1$ to $W_{i,S} := S^{-2}\sum_{s\neq t}\nabla^2 g_i[e_{is}, e_{it}]$, getting

$$\operatorname{Var}\left(\frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s\neq t}\nabla^2 g_i[e_{is}, e_{it}]\right) \leq \frac{C}{n} E\left[\left\|W_{i,S}\right\|^{2+\delta}\right]^{2/(2+\delta)}.$$

Now $W_{i,S}$ is a degenerate U-statistic since

$$E\left[\nabla^2 g(z_i)[e_{is}, e_{it}] | z_i, e_{it}\right] = E\left[\nabla^2 g(z_i)[e_{is}, e_{it}] | z_i, e_{is}\right] = 0.$$

Given the conditions imposed on $\{e_{i,s} : 1 \leq s \leq S\}$ in (A.6'), we can employ U-statistic results for absolutely regular sequences: Yoshihara (1976, Lemma 3) states that $E\left[\left\|W_{i,S}\right\|^4 | z_i\right] = O\left(S^{-4}\right)$. By inspection of the proof of Yoshihara (1976, Lemma 3), it is easily checked that in fact, for some constant $C > 0$ and with $M_{S,4}\left(z_i\right)$ defined in eq. (25), $E\left[\left\|W_{i,S}\right\|^4 | z_i\right] \leq CS^{-4}M_{S,4}\left(z_i\right)$. Thus, with $\delta = 2$, we obtain

$$E\left[\left\|W_{i,S}\right\|^4\right]^{1/2} \leq \sqrt{C}S^{-2}\sqrt{E\left[M_{S,4}\left(z_i\right)\right]}.$$

It follows that:

$$\frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s\neq t}\triangledown^2 g_i[e_{is},e_{it}] = O_P(n^{-1/2}S^{-1}\sqrt{E\left[M_{S,4}\left(z_i\right)\right]}).$$

As for the second term, by definition $\hat{\gamma}_i - \bar{w}_i = \psi_{i,S}$; and it follows from Lemma 6 that $E\left[\triangledown^2 g_i[\psi_{i,S},\psi_{i,S}]\right] = O\left(S^{-\alpha_2}\right)$ and

$$\frac{1}{n}\sum_{i=1}^{n}\left(\triangledown^2 g_i[\psi_{i,S},\psi_{i,S}] - E\left[\triangledown^2 g_i[\psi_{i,S},\psi_{i,S}]\right]\right) = O_P\left(n^{-1/2}S^{-\alpha_4/2}\right).$$

Summing up, we have shown that

$$\frac{1}{2}\triangledown^2 G_n[\psi_{n,S},\psi_{n,S}] - \dot{\Delta}_{n,S} = O_P\left(S^{-(1+\alpha_2)}\right) + O_P(n^{-1/2}S^{-1}\sqrt{E\left[M_{S,4}\left(z_i\right)\right]}) + O_P\left(n^{-1/2}S^{-(1+\alpha_4/2)}\right).$$

■

Proof of Theorem 4. We only need to check that conditions A.1-A.3 of Robinson (1988) hold. First, note that in his notation, our estimators are given by $\hat{\theta}_T = \hat{\theta}_{n,S^*}$ and $\tilde{\theta}_T = \hat{\theta}^{(k)}_{n,S}$. His Assumption A.1 is satisfied under our assumptions since in the proof of Theorem 1 we showed that $\hat{\theta}_{n,S^*} \rightarrow^P 0$ as n and $S^* \rightarrow \infty$. This also shows that Robinson's Assumption 3 is satisfied. Thus, we can appeal to his Theorem 2, which in conjunction with eq. (31) yields the desired result. ■

C Lemmas

To establish the rates for the first and second order differentials, we first establish some useful auxiliary results:

Lemma 5 *Assume that $\{W_i\}$ is an sequence α-mixing satisfying $E\left[W_i\right] = 0$, $E\left[\left\|W_i\right\|^{2r+\delta}\right] < \infty$ for some $r \geq 1$ and $\delta > 0$, and with its mixing coefficients α_i, $i = 1, 2, ...$, satisfying $\alpha_i \leq Ai^{-a}$ for some $A > 0$, and $a > 2r + 4r\left(r-1\right)/\delta - 2$. Then there exists a constant $C = C\left(r, a, A\right) < \infty$ such that:*

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n} W_i\right\|^{2r}\right] \leq CE\left[\left\|W_i\right\|^{2r+\delta}\right]^{2/(2r+\delta)} n^{-r}.$$

Proof. From Rio (1994), we obtain that

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n} W_i\right\|^{2r}\right] \leq C_r\left[n^{-r}M_{2,\alpha}^{2r} + n^{1-2r}M_{2r,\alpha}\right],$$

where $M_{p,\alpha}$, $p \geq 2$, is defined in Rio (1994) and $n^{-r} \geq n^{1-2r}$ for $r \geq 1$. By Nze and Doukhan (2004, p. 1040),

$$M_{p,\alpha} \leq E\left[\|W_i\|^{p+\delta}\right]^{p/(p+\delta)} \times \frac{(p+\delta)(p-1)}{\delta} \sum_{n=0}^{\infty} (n+1)^{p-2+p(p-1)/\delta} \alpha_n, \quad p \geq 1,$$

where, given the bound imposed on the mixing coefficients,

$$\sum_{n=0}^{\infty} (n+1)^{p+p(p-1)/\delta-2} \alpha_n \leq C(A,a) \sum_{n=0}^{\infty} (n+1)^{p+p(p-1)/\delta-2-a} < \infty.$$

■

Lemma 6 *Assume that $\{z_i\}$ satisfies (A.1), and that for ECA or EIA, the $\hat{\gamma}_{j,S}$ satisfy (A.6(4)) for $j = 1, ..., J$. Let $m(z; d\gamma)$ be a functional satisfying:*

$$E\left[\|m(z; d\gamma)\|^{2r+\delta}\right] \leq \bar{M} \|d\gamma\|^{k(2r+\delta)}, \tag{38}$$

for some $r, k \geq 1$ and $\delta > 0$.

Then, with b_S and ψ_S given in A.5 and with

$$M_S(\psi) = E[m(z; \psi_S)], \quad M_S(b) = E[m(z; b_S)]$$

the following hold:

(i) For EIA's,

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n} \{m(z_i; b_{i,S}) - M_S(b)\}\right\|^{2r}\right] \leq C(r,A)\bar{M}E\left[\|b_S\|^{k(2r+\delta)}\right] n^{-r},$$

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n} \{m(z_i; \psi_{i,S}) - M_S(\psi)\}\right\|^{2r}\right] \leq C(r,A)\bar{M}E\left[\|\psi_S\|^{k(2r+\delta)}\right] n^{-r}.$$

(ii) For ECA's, with $\bar{m}(\gamma) = E[m(z;\gamma)]$,

$$E\left[\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n} \{m(z_i; b_S) - \bar{m}(\theta, b_S)\}\right\|^{2r}\right] = C(r,A)\bar{M}E\left[\|\psi_S\|^{k(2r+\delta)}\right] n^{-r}.$$

$$E\left[\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n} \{m(z_i; \psi_S) - \bar{m}(\theta, \psi_S)\}\right\|^{2r}\right] \leq C(r,A)\bar{M}E\left[\|\psi_S\|^{k(2r+\delta)}\right] n^{-r},$$

where

$$E\left[\|\bar{m}(\psi_S)\|^{2r}\right] \leq \bar{M}E\left[\|\psi_S\|^{2kr}\right].$$

(iii) The means satisfy:

$$\|M_S(b)\| \leq \bar{M}E\left[\|b_S\|^k\right], \quad \|M_S(\psi)\| \leq \bar{M}E\left[\|\psi_S\|^k\right].$$

Proof. Define $W_{i,S} = m\left(z_i; \psi_{i,S}\right) - M_S$. By assumptions (A.1) and (A.5), for any given value of $S \geq 1$, this is a mixing process. Furthermore, eq. (38) implies that $E\left[\|W_{i,S}\|^{2r+\delta}\right] < \infty$. We can therefore apply Lemma 5,

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^n \left\{m\left(z_i; \psi_{i,S}\right) - M_S(\psi)\right\}\right\|^{2r}\right] \leq CE\left[\left\|m\left(z; \psi_{i,S}\right) - M_S(\psi)\right\|^{2r+\delta}\right]^{2/(2r+\delta)} n^{-r},$$

where, by eq. (38),

$$E\left[\left\|m\left(z; \psi_{i,S}\right)\right\|^{2r+\delta}\right] \leq CE\left[\left\|\psi_{i,S}\right\|^{k(2r+\delta)}\right] n^{-r},$$

and

$$\|M_S(\psi)\| \leq E\left[\left\|m\left(z_i; \psi_{i,S}\right)\right\|\right] \leq \bar{M}E\left[\left\|\psi_{i,S}\right\|^k\right].$$

It is easily seen that the above arguments still go through when replacing $\psi_{i,S}$ with $b_{i,S}$. This shows (i) and (iii).

To show the second inequality of (ii), redefine $W_{S,i}$ as $W_{S,i} = m(z_i; \psi_S) - \bar{m}(\psi_S)$. Conditional on ψ_S, it is easily seen that $W_{S,i}$ satisfies the conditions of Lemma 5 such that

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^n W_{S,i}\right\|^{2r} |\psi_S\right] \leq CE\left[\|W_{S,i}\|^{2r+\delta} |\psi_S\right] n^{-r}.$$

Next, observe that

$$E\left[\|W_{S,i}\|^{2r+\delta}\right] \leq CE\left[\|m(z; \psi_S)\|^{2r+\delta}\right] \leq C\bar{M}E\left[\|\psi_S\|^{k(2r+\delta)}\right],$$

and we conclude that

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^n W_{S,i}\right\|^{2r}\right] = E\left[E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^n W_{S,i}\right\|^{2r} |\psi_S\right]\right] \leq CE\left[\|\psi_S\|^{k(2r+\delta)}\right] n^{-r}.$$

Finally,

$$E\left[\|\bar{m}(\psi_S)\|^{2r}\right] \leq E\left[\|m(z; \psi_S)\|^{2r}\right] \leq \bar{M}E\left[\|\psi_S\|^{2rk}\right].$$

The proof of the first inequality of (ii) follows along the same lines. ■

Lemma 7 *Under A.1-A.4, A.5(2) and A.6(4), the first and second order differentials of G_n for the ECA in (12) satisfy equations (16)-(17) and (18)-(19).*

Proof. In the following we suppress the dependence on θ_0 since this is kept fixed. When the approximation of $G_n(\gamma)$ is on the form (13), the functional differentials are given by

$$\bigtriangledown G_n[d\gamma] = \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown g_i[d\gamma], \quad \bigtriangledown^2 G_n[d\gamma, d\gamma'] = \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown^2 g_i[d\gamma, d\gamma'],$$

and $d\gamma$ and $d\gamma'$ are the same for all observations $i = 1, \ldots, n$.

Given A.6(4), the application of the first-order differential to the bias component can be rewritten as

$$\bigtriangledown G_n[b_S] = S^{-\beta}\frac{1}{n}\sum_{i=1}^{n} \bigtriangledown g_i\left[\bar{b}\right] + \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown g_i\left[b_S - S^{-\beta}\bar{b}\right].$$

Now,

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \bigtriangledown g_i\left[\bar{b}\right]\right] = E\left[\bigtriangledown g_i\left[\bar{b}\right]\right], \text{ and}$$

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \left\|\bigtriangledown g_i\left[b_S - S^{-\beta}\bar{b}\right]\right\|\right] \leq G_1\left\|b_S - S^{-\beta}\bar{b}\right\| = o\left(S^{-\beta}\right).$$

By Lemma 6(i) with $m(z; d\gamma) = \bigtriangledown g(z)[d\gamma]$, $k = 1$ and $r = 1$,

$$\mathrm{Var}\left(\bigtriangledown G_n[b_S]\right) \leq \frac{1}{n} C\left\|b_S\right\|^2 = O\left(\frac{S^{-2\beta}}{n}\right).$$

Since $d\gamma \mapsto \bigtriangledown g_i[d\gamma]$ is linear, the conditional mean of the stochastic component of the first-order term is

$$E\left[\bigtriangledown G_n[\psi_S]|\mathcal{Z}_n\right] = \frac{1}{n}\sum_{i=1}^{n} \bigtriangledown g_i\left[E\left[\psi_S|z_i\right]\right] = 0.$$

Moreover, with $\bigtriangledown\bar{g}(\psi_S; \theta_0)$ given in Theorem 1,

$$\bigtriangledown G_n[\psi_S] = \bigtriangledown\bar{g}(\psi_S; \theta_0) + \frac{1}{n}\sum_{i=1}^{n}\left\{\bigtriangledown g_i[\psi_S] - \bigtriangledown\bar{g}(\psi_S; \theta_0)\right\}.$$

Recalling the definition of $\bigtriangledown\bar{g}(\psi_S; \theta_0)$, it follows from Lemma 6(ii) with $m(z; d\gamma) = \bigtriangledown g(z)[d\gamma, \gamma]$ and $k = 2$ that the second term is $O_P(n^{-1/2}S^{-\alpha_2})$.

Regarding the second order differential, its application to the bias component satisfies

$$\bigtriangledown^2 G_n[b_S, b_S] = S^{-2\beta}\frac{1}{n}\sum_{i=1}^{n} \bigtriangledown^2 g_i\left[\bar{b}, \bar{b}\right] + o_P\left(S^{-2\beta}\right);$$

moreover,

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \bigtriangledown^2 g_i\left[\bar{b}, \bar{b}\right]\right] = E\left[\bigtriangledown^2 g_i\left[\bar{b}, \bar{b}\right]\right],$$

and, applying Lemma 6(i) with $m\left(z;d\gamma\right)=\bigtriangledown^{2}g\left(z\right)\left[d\gamma,d\gamma\right]$, $k=2$ and $r=1$,

$$\operatorname{Var}\left(\bigtriangledown^{2}G_{n}[b_{S},b_{S}]\right)\leq\frac{1}{n}C\left\|b_{S}\right\|^{4}=O\left(n^{-1}S^{-4\beta}\right).$$

To bound the variance component, define

$$\bigtriangledown^{2}\bar{g}\left[\gamma,\gamma\right]=E\left[\bigtriangledown^{2}g_{i}\left[\gamma,\gamma\right]\right],$$

and write

$$\bigtriangledown^{2}G_{n}[\psi_{S},\psi_{S}]=\bigtriangledown^{2}\bar{g}\left[\psi_{S},\psi_{S}\right]+\frac{1}{n}\sum_{i=1}^{n}\left(\bigtriangledown^{2}g_{i}\left[\psi_{S},\psi_{S}\right]-\bigtriangledown^{2}\bar{g}\left[\psi_{S},\psi_{S}\right]\right).$$

Applying Lemma 6(ii) with $m\left(z;d\gamma\right)=\bigtriangledown^{2}g\left(z\right)\left[d\gamma,d\gamma\right]$ and $k=2$, we obtain that $\left\|\bigtriangledown^{2}\bar{g}\left[\psi_{S},\psi_{S}\right]\right\|=O_{P}\left(S^{-2\alpha_{2}}\right)$ and that the second term is $O_{P}(n^{-1/2}S^{-\alpha_{4}})$.

Finally, ´by the same arguments as before, $E\left[\bigtriangledown^{2}G_{n}[\psi_{S},b_{S}]\right]=0$ while $\operatorname{Var}\left(\bigtriangledown^{2}G_{n}[\psi_{S},b_{S}]\right)=O(n^{-1}S^{-\alpha_{2}}S^{-2\beta})$, and so we can ignore the cross term since it is of lower order. ■

Lemma 8 *Under A.1-A.4, A.5(2) and A.6(4), the first and second order differentials of $G_{n}(\theta_{0},\gamma)$ for the EIA in (12) satisfy equations (16)-(17) and (20)-(21).*

Proof. Again, we suppress dependence on θ_{0}. For the EIA, the first and second order differentials are

$$\begin{aligned}\bigtriangledown G_{n}\left[d\gamma\right]&=\frac{1}{n}\sum_{i=1}^{n}\bigtriangledown g_{i}\left[d\gamma_{i}\right],\\ \bigtriangledown^{2}G_{n})\left[d\gamma,d\gamma'\right]&=\frac{1}{n}\sum_{i=1}^{n}\bigtriangledown^{2}g_{i}\left[d\gamma_{i},d\gamma_{i}'\right],\end{aligned}$$

for any $d\gamma=(d\gamma_{1},...,d\gamma_{n})$ and $d\gamma'=(d\gamma_{1}',...,d\gamma_{n}')$. It is easily seen that the bias components are the same as those we derived for the ECA in Lemma 7, and so we only consider the variance components. With $\mathcal{Z}_{n}=(z_{1},...,z_{n})$, the mean of the first-order variance component is zero,

$$E\left[\bigtriangledown G_{n}[\psi_{S}]|\mathcal{Z}_{n}\right]=\frac{1}{n}\sum_{i=1}^{n}\bigtriangledown g_{i}\left[E\left[\psi_{i,S}|z_{i}\right]\right]=0,$$

while its variance satisfies, using Lemma 6(i),

$$\operatorname{Var}\left(\bigtriangledown G_{n}[\psi_{S}]\right)\leq\frac{1}{n}CE\left[\left\|\psi_{S}\right\|^{2}\right]=O\left(n^{-1}S^{-\alpha_{2}}\right).$$

Applying Lemma 6(i) and (iii) with $m\left(z;d\gamma\right)=\triangledown^{2}g\left(z\right)\left[d\gamma,d\gamma\right]$ and $k=2$, the mean and the variance of the second order differential satisfy

$$E\left[\triangledown^{2}G_{n}[\psi_{S},\psi_{S}]\right]=E\left[\triangledown^{2}g_{i}\left[\psi_{i,S},\psi_{i,S}\right]\right]\leq CE\left[\left\Vert\psi_{i,S}\right\Vert^{2}\right]=O\left(S^{-\alpha_{2}}\right),$$

$$\mathrm{Var}\left[\triangledown^{2}G_{n}[\psi_{S},\psi_{S}]\right]=O(n^{-1}S^{-\alpha_{4}}).$$

The cross term satisfies $E\left[\triangledown^{2}G_{n}[\psi_{S},b_{S}]\right]=0$ while $\mathrm{Var}\left(\triangledown^{2}G_{n}[\psi_{S},b_{S}]\right)=O(n^{-1}S^{-\alpha_{2}}S^{-2\beta})$, and so we can ignore this term since it is of lower order. ■

Lemma 9 *Assume that A.1-A.4, A.5*(3) *and A.6'*(6) *hold. Then, the results of Theorem 1 still hold with*

$$B_{S,1}=S^{-\beta}H_{0}^{-1}E\left[\triangledown g(z;\theta)[\bar{b}]\right]+o\left(S^{-\beta}\right),\quad B_{S,3}=S^{-2\beta}H_{0}^{-1}E\left[\triangledown g(z;\theta)[\bar{b},\bar{b}]\right]+o\left(S^{-2\beta}\right),$$

where $\bar{b}$ is defined in (A.6'), and the rate of the remainder term $R_{n,S}$ can be sharpened to:

$$R_{n,S}=O_{P}\left(S^{-3\beta}\right)+O_{P}\left(S^{-\alpha_{4}}\right)+O\left(S^{-(1+a_{3})}\right)+O\left(n^{-1/2}S^{-\alpha_{6}/2}\right).$$

Proof. The results for the first and second order derivatives derived in Theorem 1 are still valid. The bias expressions stated in the theorem follow as a simple consequence of A.5'. The only difference is that the remainder term in eq. (11) now takes the form

$$R_{n,S}=\frac{1}{6}\triangledown^{3}G_{n}\left[\Delta\hat{\gamma}_{S},\Delta\hat{\gamma}_{S},\Delta\hat{\gamma}_{S}\right]+\bar{R}_{n,S},$$

where, by A.4(3) and the same arguments used in the proof of Theorem 1, $\bar{R}_{n,S}=O_{P}\left(S^{-4\beta}\right)+O_{P}\left(S^{-\alpha_{4}}\right)$. Regarding the third order term, it is easily checked that the bias component is of order $O_{P}\left(S^{-3\beta}\right)+O_{P}\left(n^{-1/2}S^{-3\beta}\right)$ by the same arguments employed in Lemma 7, so what remains is the variance component: In the case of EIA, the variance component can be written as

$$\triangledown^{3}G_{n}\left[\psi_{S},\psi_{S},\psi_{S}\right]=\frac{1}{n}\sum_{i=1}^{n}\triangledown^{3}g_{i}\left[\psi_{S},\psi_{S},\psi_{S}\right].$$

By Lemma 6, we obtain

$$\triangledown^{3}G_{n}\left[\psi_{S},\psi_{S},\psi_{S}\right]-E\left[\triangledown^{3}G_{n}\left[\psi_{S},\psi_{S},\psi_{S}\right]\right]=O\left(n^{-1/2}S^{-\alpha_{6}/2}\right),$$

while, due to the independence,

$$\begin{aligned}\left|E\left[\triangledown^3 G_n\left[\psi_S,\psi_S,\psi_S\right]\right]\right| &\leq \frac{1}{S^3}\sum_{s,t,u=1}^{S}\left|E\left[\triangledown^3 g_i\left[e_{i,s},e_{i,t},e_{i,u}\right]\right]\right| \\ &= \frac{\left|E\left[\triangledown^3 g_i\left[e_{i,s},e_{i,s},e_{i,s}\right]\right]\right|}{S^2} \\ &\leq CE\left[\left\|\psi_{i,S}\right\|^3\right] \\ &= O\left(S^{-\alpha_3}\right).\end{aligned}$$

In the case of ECA, define $\triangledown^3\bar{g}\left[\gamma,\gamma,\gamma\right]=E\left[\triangledown^2 g_i\left[\gamma,\gamma,\gamma\right]\right]$ and write

$$\triangledown^3 G_n[\psi_S,\psi_S,\psi_S]=\triangledown^3\bar{g}\left[\psi_S,\psi_S,\psi_S\right]+\frac{1}{n}\sum_{i=1}^{n}\left\{\triangledown^3 g_i\left[\psi_S,\psi_S,\psi_S\right]-\triangledown^3\bar{g}\left[\psi_S,\psi_S,\psi_S\right]\right\}.$$

Applying Lemma 6(ii) with $m\left(z;d\gamma\right)=\triangledown^3 g\left(z\right)\left[d\gamma,d\gamma,d\gamma\right]$, the two terms are $O_P\left(S^{-\alpha_3}\right)$ and $O_P(n^{-1/2}S^{-\alpha_6/2})$ respectively. ■