Higher Order Improvements for Approximate Estimators[*]

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

Bernard Salanié[§]
Columbia University

June 14, 2011

Abstract

Many modern estimation methods in econometrics approximate an objective function, through simulation or discretization for instance. Approximations typically impart additional bias and variance to the resulting estimator. We here propose three methods to improve the properties of such "approximate" estimators at a low computational cost. The first method provides an analytical bias adjustment for estimators based on stochastic approximators, such as simulation-based estimators. Our second proposal 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 much of the additional bias and variance of the initial approximate estimator. A Monte Carlo simulation on the mixed logit model shows that combining these approaches can yield spectacular improvements at a low cost.

[*] We would like to thank participants at the CAM Research Workshop 2008 and at the "Conference on Dynamic Aspects in Economic Decision Making" (University of Copenhagen), the Econometric Society Latin-American Meeting 2009 in Buenos Aires, and seminars at 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 the 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), 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)), Kolmogorov-Smirnov type statistics (Corradi and Swanson (2007)), or an integrated value function (Rust (1997)).[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); Creel and Kristensen (2009); Kristensen and Shin (2008)), or on sieve methods (Kristensen and Schjerning (2011); Norets (2011)). Other estimation methods 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 the estimator that relies on a finite approximation an "approximate estimator".

The use of approximations usually deteriorates the properties of the approximate estimator relative to those of the corresponding exact estimator: it is often 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 reduces efficiency: a case in point is the simulated method of moments. In all other cases, approximation creates a bias and potentially (in case of simulations) 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

[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.

a seriously limiting factor; increased computer power helps, but it also motivates researchers to work on more complex models.

The contribution of this paper is twofold: First, we analyze the higher-order properties of the approximate estimator relative to the exact one in a very general setting that includes both M-estimators and GMM estimators, and allows for a wide range of approximation schemes—both stochastic and non-stochastic. This higher-order expansion can be used to choose the degree of approximation, and to quantify the additional estimation error due to approximation. Our findings encompass and extend results in the literature on simulation-based estimators, such as Lee (1995, 1999), Gouriéroux and Monfort (1996) and Laroque and Salanié (1989). Moreover, they can be used to analyze the behavior of approximate estimators based on non-stochastic numerical approximation, which are often used in structural econometric models[2].

Second, based on the higher-order expansion, 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[3]. We derive a general formula for the additional bias and variance of the approximate estimator, and we build upon our asymptotic expansions in order to correct the objective function and eliminate the leading term of the additional bias[4]. 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 has a much lower bias: the leading term is of order $1/S^2$ for parametric simulation-based inference.

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] To list just a few examples: asset pricing models (Tauchen and Hussey (1991)), DSGE models (Fernández-Villaverde, Rubio-Ramirez and Santos (2006)), and dynamic discrete choice models (Rust (1997)).

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

[4] Laffont et al. (1995) and Lee (1995) proposed a similar idea for SNLS estimators and SMLE of discrete choice models respectively.

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 much of the additional bias and variance of the initial approximate estimator.[5] The Newton-Raphson adjustment can be combined with either of the two aforementioned bias correction methods: the approximate objective function can first be corrected so as to obtain an approximate estimator with a smaller bias that in turn is used as the initial estimator of Newton-Raphson algorithm.

To test the practical performance of our proposed methods, we run a simulation study on a mixed logit model. The mixed logit is a building block of much work in demand analysis; and it is simple enough that we can use Gaussian quadrature to obtain a quasi-exact maximum-likelihood estimator, which we then compare to basic and improved simulated maximum-likelihood estimators. We show that uncorrected SML has non-negligible bias, even for large sample sizes. Our analytical bias-correction removes most of it, at almost no additional computational cost; and it does not add more variance to the estimators. In addition, the Newton-Raphson correction markedly reduces the bias, and especially the variance of the SML estimator. Combining both approaches therefore brings the SML estimator much closer to the MLE. Taking one Newton-Raphson step increases the cost of SML since we need to compute a Hessian matrix; but it is still a much more practical proposition than increasing the number of simulations during estimation.

It bears repeating that we carry out our expansions around the exact estimator, as opposed to doing them 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. To obtain a higher-order expansion around the true parameter value, one could simply combine the results derived in this paper with those obtained for exact estimators. In principle, correcting for the approximation error might actually make the corrected estimator more biased or less precise than the original approximate estimator. There is no reason to believe that such quirky behavior is the norm, however; and that has not been our experience in the simulation studies reported here.

Neither do we deal in this paper with difficulties incurred in numerical optimization or in solving for a fixed-point, as often arise in estimating structural models. This is a separate issue; and we refer to Judd and Su (2010) and to Dubé, Fox and Su (2009) for some recent results.

The paper is organized 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 Sections 4 and

[5] 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.

5, we analyze in turn the two proposed bias adjustments. The properties of the Newton-Raphson method are derived in Section 6. Finally, section 7 presents the results of a Monte Carlo simulation study using the mixed logit model as an example. All proofs and lemmas 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 $\theta_0 \in \Theta \subseteq \mathbb{R}^k$ using some extremum estimator,

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

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

Note that in many cases of interest the value of Q_n only depends on θ through the function γ_0—this will in fact hold in our examples below. However, we do not require this, and we allow for

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

Our paper focuses on situations where the true function γ_0 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; to $\hat{\theta}_n$ as the "exact" estimator; and to $\hat{\theta}_{n,S}$ as the "approximate" estimator. We now present a few examples.

Example 1: Simulated maximum likelihood (SML). 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 $\gamma_0\left(z;\theta\right) := p\left(y|x;\theta\right)$, $z = (y, x)$. 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 obtain a simulated version by $\hat{\gamma}_S(z;\theta) = S^{-1}\sum_{s=1}^{S} w(y|x,\varepsilon_s;\theta)$. The resulting estimator is a SMLE.

More recently, Fermanian and Salanié (2004) proposed using kernel estimators as approximators. Suppose that $y = r(x, \varepsilon; \theta_0)$, with implied conditional density $\gamma_0(z;\theta) = p(y|x,\theta)$. Then generate samples, $y_s(x,\theta) = r(x, \varepsilon_s; \theta)$ for $s = 1, \ldots, S$, and approximate the density γ_0 with a kernel density estimator based on the y_s's: $\hat{\gamma}_S(z;\theta) = \sum_{s=1}^{S} K_h(y - y_s(x,\theta))/S$. For a similar approach in time series models, see Altissimo and Mele (2009), Brownlees, Kristensen and Shin (2011) and Kristensen and Shin (2008).

Example 2: Simulated pseudo-maximum likelihood (SPML) method of moments. Suppose that we have the following conditional moment restriction, $E[y|x] = m(x;\theta)$, 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 Gaussian pseudo-log likelihood 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 γ_0 cannot be evaluated analytically, Laroque and Salanié (1989) proposed simulated pseudo-maximum likelihood (SPML) estimators: 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 replacing γ_0 with $\hat{\gamma}_S$. The above idea can be extended to incorporate information regarding the conditional variance of y.

Example 3: Simulated method of moments (SMM). The parameter of interest is identified through a set of moment conditions $E[g(z, \theta_0)] = 0$. Given a weighting matrix W_n, the GMM estimator would minimize

$$Q_n(\theta, \gamma_0) = G_n(\theta)' W_n G_n(\theta)$$

where $G_n(\theta) = \sum_{i=1}^{n} g(z_i, \theta)/n$. Here, γ_0 is simply the function g, which may be hard to evaluate, as in the multinomial probit example of McFadden (1989). Another example is the simulated method of moments (SMM) proposed by Duffie and Singleton (1993) to estimate dynamic models where a long string of simulations from the model, say $\{y_s(\theta) : s = 1, ..., S\}$, are used to approximate unconditional moments of the model. The resulting estimator is of the minimum-distance type. Creel and Kristensen (2009) generalize the approach of Duffie and Singleton (1993) and propose to approximate conditional expectations by combining simulations with kernel regression techniques.

Many other examples fall within the above general framework: Evaluating the value function in dynamic programming models most often requires numerical approximations that involve simulations, interpolation or sieve methods (also referred to as parametric approximations); see Rust (1997), and more recently Kristensen and Schjerning (2011) and Norets (2009, 2011). Here the approximated value function plays the role of γ_0.

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, $share = s\left(\xi, z; \theta\right)$ 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, $\xi\left(share, z; \theta\right) = s^{-1}\left(share, z; \theta\right)$. 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, $\xi_S\left(share, z; \theta\right)$, where S captures the number of iterations and/or the tolerance level used in the algorithm[6].

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. The quality of the model approximation can be improved at a larger computational cost by using a finer grid or by using, for example, more iterations of perturbations or projection methods as advocated by Judd, Kubler and Schmedders (2003). For a first-order theoretical analysis of the impact on the resulting approximate MLE, see Fernández-Villaverde, Rubio-Ramirez and Santos (2006) and Ackerberg, Geweke and Hahn (2009).

In all of the examples above, approximations reduce the quality of the estimator. Start with our first three examples where stochastic approximations (i.e. simulations) are used to evaluate a mathematical expectation. 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. When using nonparametric techniques such as kernel smoothers or sieve methods in the approximation, the approximator itself is biased, and the objective function will be biased even if the approximator enters linearly.

One exception from the above is simulated method of moments (Example 3). This ap-

[6] Some more recent implementations use mathematical programming under equilibrium constraints, as advocated by Judd and Su (2010).

proximate estimator has nicer properties since the objective function is linear in the simulated mean. Then no additional biases due to simulations appear. The asymptotic efficiency loss still is of order $1/S$ though.

Similarly, non-stochastic approximations lead to deteriorations of the properties of the resulting estimators. Take the problem of computing the density $p\,(y|x;\theta)$ in Example 1 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. On the other hand, no efficiency loss will be incurred.

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 diverges fast enough. In other words (Section 3 will give more precise statements and regularity conditions), $||\hat{\theta}_{n,S} - \hat{\theta}_n|| = o_P\,(1/\sqrt{n})$ as $n \to \infty$ for some sequence $S = S\,(n) \to \infty$, in which case there is no first-order difference between the exact and approximate estimator. However, in practice, S is finite and so it is desirable to quantify the discrepancy between the exact and infeasible estimator. Our higher-order expansion allows the researcher to gauge the degree of inaccuracy (relative to the exact estimator) that the chosen approximate estimator suffers from.

Moreover, to reach a given level of tolerance for the approximation error, S may have to be chosen very large. This motivates our proposed methods that yield adjusted estimators that may perform just as well as large-S approximate estimators, and yet are computationally much less burdensome: 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. 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

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

where $\Delta_{n,S}\,(\theta)$ corrects for at least the leading term of the approximation bias.

The 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)); 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 results to general simulation-based estimators such as, for example, dynamic program models where the Bellman operator is evaluated through simulations (Rust (1997)). We show how to compute $\Delta_{n,S}\,(\theta)$ for a broad class of simulation-based estimators, and analyze the theoretical properties of the resulting estimator.

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 (3), 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 $E\left[Q_n(\theta,\hat{\gamma}_S)-Q_n(\theta,\gamma)\right]=B(\theta)/S+o\left(1/S\right)$. Now take two approximators $\hat{\gamma}_{S/2}^{[1]}$ and $\hat{\gamma}_{S/2}^{[2]}$ of order $S/2$. For each approximator $m=1,2$, we can compute the corresponding objective function, $Q_n(\theta,\hat{\gamma}_{S/2}^{[m]})$. We then choose $\Delta_{n,S}(\theta)=\frac{1}{4}\left[Q_n(\theta,\hat{\gamma}_{S/2}^{[1]})+Q_n(\theta,\hat{\gamma}_{S/2}^{[2]})\right]$ so that the adjusted objective function satisfies

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

and the leading bias terms cancel out. We provide details in section 5.

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 $\sqrt{n}$-consistent estimator of θ_0 in a model with log-likelihood $L_n(\theta)$, then a single NR-step yields a consistent and asymptotically efficient estimator. We apply this idea to our setting by starting from some initial approximate estimator based on a small degree of approximation S, say $\bar{\theta}_{n,S}$. This can for example arrive from eq. (2) or (3). 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 $G_n\left(\theta,\gamma\right)=\partial Q_n(\theta,\gamma)/\partial\theta$ and $H_n\left(\theta,\gamma\right)=\partial^2 Q_n\left(\theta,\gamma\right)/(\partial\theta\partial\theta')$, and define

$$\hat{\theta}_{n,S}^{(k+1)}=\hat{\theta}_{n,S}^{(k)}-H_n^{-1}(\hat{\theta}_{n,S}^{(k)},\hat{\gamma}_{S^*})G_n(\hat{\theta}_{n,S}^{(k)},\hat{\gamma}_{S^*}),\quad k=1,2,3,... \tag{4}$$

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

Note that the cost of computing this new estimator from the first one is (very) roughly S^*/S times the cost of one iteration in the 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 adding much to the cost of the estimation procedure.[7] 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.

[7] 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 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. To do so, we need to introduce assumptions both on the estimating equation and on the approximators.

3.1 The Estimating Equation

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

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

for some random functional $G_n(\theta, \gamma)$. The corresponding approximate estimator, $\hat{\theta}_{n,S}$, is implicitly defined by

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

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

$$G_n(\theta, \gamma) = \frac{1}{n}\sum_{i=1}^{n} g(z_i; \theta, \gamma). \tag{5}$$

Our setup allows for two-step GMM estimators where the weight matrix has been estimated. In the following we shall assume that $G_n(\theta, \gamma)$ is a smooth function in both θ and γ. Remember that the dependence on θ comes partly through the function γ, so that smoothness in θ also requires that γ be smooth in θ. We conjecture that our results could be generalized to estimators minimizing non-differentiable objective functions 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 $G_n(\theta, \gamma) = \partial Q_n(\theta, \gamma)/\partial\theta$. In the case of GMM estimators, $Q_n(\theta, \gamma) = M_n(\theta, \gamma) W_n M_n(\theta, \gamma)$ with $W_n \to^P W$ and $M_n(\theta, \gamma) = \sum_{i=1}^{n} m(z_i; \theta, \gamma)/n$. We may then choose $g(z_i; \theta, \gamma) = H_0 W m(z_i; \theta, \gamma)$, where $H_0 = E[\partial m(z_i; \theta, \gamma)/\partial\theta]$, since this is (asymptotically) first-order equivalent to the first-order condition of the GMM estimator.

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 $\hat{\gamma}_S$), 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 linear 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, $\|\gamma\| = E\left[\|\gamma(z)\|^q\right]^{1/q}$ for some $q \geq 1$. We introduce the first-order derivative of $G_n(\theta, \gamma)$ w.r.t. θ,

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

and the corresponding population versions,

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

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 $\theta \in \Theta$, $E\left[\sup_{\theta \in \Theta} \|g(z_i; \theta, \gamma_0)\|\right] < \infty$ and (ii) $G(\theta, \gamma_0) = 0$ if and only if $\theta = \theta_0$.

A.4 For all γ in a neighbourhood $\mathcal{N}$ of γ_0, $g(z; \theta, \gamma)$ and its derivative, $h(z; \theta, \gamma)$, satisfy:

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

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

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

(c) for some $\delta, \lambda, \bar{H} > 0$, and for all $\gamma \in \mathcal{N}$,

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

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.[8] 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 $\hat{\gamma}_S \to^P \gamma$.

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,

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

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. To do this, we start from

$$G_n(\hat{\theta}_{n,S},\hat{\gamma}_S) - G_n(\hat{\theta}_n,\gamma_0) = o_P(1/\sqrt{n});$$

and we break down the left hand-side into

$$G_n(\hat{\theta}_{n,S},\hat{\gamma}_S) - G_n(\hat{\theta}_{n,S},\gamma_0) = \frac{\partial G_n}{\partial \theta}(\hat{\theta}_n,\gamma_0)(\hat{\theta}_{n,S} - \hat{\theta}_n) + O_P(\|\hat{\theta}_{n,S} - \hat{\theta}_n\|^2).$$

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) There exist functionals $\nabla^k g(z;\theta)[d\gamma_1,...,d\gamma_k]$, $k = 1,...,m$, which are linear in each component $d\gamma_i \in \Gamma$, $i = 1,...,k$, and positive constants δ, λ, and $\bar{G}_i$, $i = 0,1,2$, such that:

$$E\left[\sup_{\|\theta-\theta_0\|\leq\delta}\left\|g(z;\theta,\gamma_0+\Delta\gamma) - g(z;\theta,\gamma_0) - \sum_{k=1}^{m}\frac{1}{k!}\nabla^k g(z;\theta)[\Delta\gamma,...,\Delta\gamma]\right\|\right] \leq \bar{G}_0\|\Delta\gamma\|^{m+\lambda}. \tag{6}$$

Furthermore,

$$E\left[\sup_{\|\theta-\theta_0\|\leq\delta}\|\nabla g(z;\theta)[\Delta\gamma]\|^2\right] \leq \bar{G}_1\|\Delta\gamma\|^2, \tag{7}$$

[8] These cases can be handled by introducing a smoothed version of the approximators as discussed in McFadden (1989); see also Fermanian and Salanié (2004).

and for $k = 2, ..., m$ and for some $\nu > 0$,

$$E\left[\sup_{\|\theta-\theta_0\|\leq\delta}\left\|\nabla^k g\left(z;\theta\right)\left[\Delta\gamma_1,...,\Delta\gamma_k\right]\right\|^{2+\nu}\right] \leq \bar{G}_k\left(\|\Delta\gamma_1\|\cdots\|\Delta\gamma_k\|\right)^{2+\nu}. \tag{8}$$

Assumption A.5(m) restricts $g(z;\theta,\gamma)$ to be m times pathwise differentiable w.r.t. γ with differentials $\nabla^k g(z;\theta)[d\gamma_1,...,d\gamma_k]$, $k = 1,...,m$. 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 $G_n(\theta,\gamma)$ 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

$$G_n(\theta,\hat{\gamma}_S) - G_n(\theta,\gamma_0) = \sum_{k=1}^{m}\frac{1}{k!}\nabla^k G_n(\theta)[\hat{\gamma}_S-\gamma_0,...,\hat{\gamma}_S-\gamma_0] + R_{n,S}, \tag{9}$$

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

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

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. (9).

3.2 The Approximators

We also 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

$$G_n\left(\theta,\hat{\gamma}_S\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(z_i;\theta,\hat{\gamma}_S\right), \tag{10}$$

and one single approximator, $\hat{\gamma}_S$, is used in the computation of the moment conditions. In the second case,

$$G_n\left(\theta,\hat{\gamma}_S\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(z_i;\theta,\hat{\gamma}_{i,S}\right), \tag{11}$$

and n approximators, $\hat{\gamma}_{1,S},....\hat{\gamma}_{n,S}$, are used in the computation. To differentiate between the two approximation schemes, we will refer to the approximate estimator based on eq. (10) as an *estimator based on common approximators* (ECA) and to (11) as an *estimator based on individual approximators* (EIA). We stress that the ECA and EIA are both targeting the same infeasible estimator; the only difference lies in how the approximators are used in the computation of the objective function.

In simulation-based estimation, ECAs were 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 3) 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 (10), while $J = n$ for the EIA in (11). In what follows, it is crucial to separate assumptions on the bias of the approximator

$$b_S(z;\theta) := E[\hat{\gamma}_{j,S}(z;\theta)|x] - \gamma_0(z;\theta)$$

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(s) lies in Γ and satisfies:

(i) for any fixed value z, the J random variables $\hat{\gamma}_{1,S}(z;\theta), ..., \hat{\gamma}_{J,S}(z;\theta)$ are mutually independent and are all independent of $\mathcal{Z}_n$.

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

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

(iii) For $2 \leq q \leq p$, the stochastic component of the approximator satisfies:

$$E\left[\left\|\psi_{j,S}\left(z;\theta\right)\right\|^{q}\right] = S^{-\alpha_q} v_q\left(z;\theta\right) + o(S^{-\alpha_q}),$$

for some constant $\alpha_q > 0$.

Note that the $o\left(\cdot\right)$ terms in (ii)-(iii) are w.r.t. the function norm on Γ. Assumption A.6 is sufficiently general to cover all of the examples in Section 2 under suitable regularity conditions. First consider A.6.iii. It 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 $1 \leq p \leq q$.[9] We will choose $p \geq 1$ in conjunction with the order of the expansion $m \geq 1$ of Assumption A.5, 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 $\nabla^k G_n\left(\theta_0\right)\left[\hat{\gamma}_S, ..., \hat{\gamma}_S\right]$ exists and to evaluate its rate of convergence, we will require A.6.iii to hold with $p = 2k$.

3.2.1 Non-stochastic approximators

First, consider an approximation that does not involve any randomness, as with numerical integration, discretization, or numerical solution of differential equations. A.6.i clearly has no bite when non-stochastic approximators are used. Then by construction the conditional variance of the approximator is zero, so that $\alpha_p = +\infty$ for all $p \geq 2$. Non-stochastic approximation imparts a bias, which in leading cases obeys assumption A.6.ii 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.

3.2.2 Stochastic approximators

Next, let us examine stochastic approximation schemes, which encompass simulation-based inference methods. Most simulation-based estimators in a dynamic setting use the ECA scheme: only one approximator is used for all observations, c.f. the discussion in Example 3, and so A.6.i is automatically satisfied. The typical EIA scheme draws J independent batches of size S and then uses one batch per approximation; this again satisfies A6.i. It does not rule out dependence between the simulated values within each batch, as will for example be

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

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

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

the case when drawing recursively from a time series models. Note that A.6.i is stated for some *fixed* value of z; the requirement that the simulations be independent of data is satisfied by most standard simulation schemes[10]. For parametric approximators in simulation-based inference, the bias b_S is typically zero and so A.6.ii holds with $\beta = \infty$.

Monte Carlo schemes are of course the most prominent example of stochastic approximators; and they have specific properties that allows for a more precise analysis of the approximation error appearing in the resulting estimator. We will therefore specialize some of our results to the following class of Monte Carlo approximators:

A.7(p) Assume that $\hat{\gamma}_{j,S}(z;\theta)$ takes the form

$$\hat{\gamma}_{j,S}(z;\theta) = \frac{1}{S}\sum_{s=1}^{S} w_S(z, \varepsilon_{j,s}; \theta). \tag{12}$$

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

$$\bar{w}_S(z;\theta) := E\left[w_S(z, \varepsilon_{js}; \theta)\,|x\right] = \gamma_0(z;\theta) + S^{-\beta}\bar{b}(z;\theta) + o\left(S^{-\beta}\right).$$

Define $e_{jS} \equiv w_S(z, \varepsilon_{js}; \theta) - \bar{w}_S(z;\theta)$; then

$$E\|e_{jS}\|^p = O(S^{\mu_p}) \text{ for some } \mu_p < p/2.$$

To our knowledge, the class of approximators that satisfies A7 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 (12) follow directly from those of the simulators w_S: it is easy to see that under A7(p), A6(p) holds with the same rate β for the bias term in A6(p).ii and with $\alpha_p = p/2 - \mu_p > 0$ in A6(p).iii.

In standard simulation-based estimation, the simulating function $w_S \equiv w$ in A7 is actually independent of the number of simulations, and the approximator has no bias: $b_S \equiv 0$ and so $\beta = \infty$. Moreover, $E\|e_{jS}\|^p$ then is constant; A.7(p).iii typically holds with $\mu_p = 0$, and A6(p).iii with $\alpha_p = p/2$.

The class A.7 also include approximators that combine simulations and nonparametric

[10] There is one situation where 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 (2011) and Olsson and Rydén (2008). Then 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.

techniques such as the methods proposed in Fermanian and Salanié (2004), Creel and Kristensen (2009), Kristensen and Scherning (2011) and Norets (2009, 2011). These will incur both a bias and variance component, but A7 still applies. 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

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

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.7(p) holds with $\beta = r\delta$ and $\mu_p = \delta d(p-1)$ for $p \geq 2$.

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 $\beta = \alpha_2/2 = r/(2r + d)$. 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, 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 estimator is normally required for the parametric estimator to be $\sqrt{n}$-consistent[11]. For example, the optimal rate for NPSML estimation turns out to be $\delta^{**} = 1/(r + d + 2)$. Interestingly, when 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 $\hat{\theta}_{n,S}$: $\delta^* = \delta^{**} = 1/(4 + d)$.

3.3 The Effect of Approximators

The following theorem states the rate at which the approximate objective function converges towards the exact one; and it provides a bound on the difference between the approximate estimator and the exact estimator. In the following, when we discuss biases and variances and, for example, write $E[\hat{\theta}_{n,S}]$, we refer to the (well-defined!) means and variances of the leading terms of a valid stochastic expansion of the estimators. This is a standard approach in the higher-order analysis of estimators since Rothenberg (1984); see also Newey-Smith (2004, section 3.)

[11] See Kristensen-Salanié (2010) for details.

Since we are concerned with the error due to the approximation, we let $\mathrm{EBias}(\hat{\theta}_{n,S})$ denote the leading term of the error-bias, $E[\hat{\theta}_{n,S} - \hat{\theta}_n]$; and $\mathrm{EVar}(\hat{\theta}_{n,S})$ for the leading term of the error-variance, $\mathrm{Var}(\hat{\theta}_{n,S} - \hat{\theta}_n)$. To state the asymptotic expansion in a compact manner, we introduce some moments which will make up the leading bias terms:

$$\begin{aligned} B_1 &= -H_0^{-1} E\left[\nabla g(z_i;\theta_0)[b_S]\right] && (13) \\ B_2 &= -\frac{1}{2} H_0^{-1} E\left[\nabla^2 g(z_i;\theta_0)[\psi_S,\psi_S]\right]. && (14) \end{aligned}$$

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

$$\mathrm{EBias}(\hat{\theta}_{n,S}) = B_1 + B_2,$$

with

$$B_1 = O(S^{-\beta}),\ B_2 = O(S^{-\alpha_2}).$$

For EIA,

$$\mathrm{EVar}(\hat{\theta}_{n,S}) = O(n^{-1}S^{-\beta}) + O(n^{-1}S^{-\alpha_2})$$

while for ECA,

$$\mathrm{EVar}(\hat{\theta}_{n,S}) = O(n^{-1}S^{-\beta}) + O(S^{-\alpha_2}).$$

The bias and the variance of the approximator enter the two leading bias terms separately: the bias b_S drives B_1, and the stochastic component ψ_S drives B_2. When the approximator takes the form of a simple unbiased simulated average, $B_1 = 0$ and the leading bias term $B_2 = O(1/S)$; this is a well-known result for specific simulation-based estimators in cross-sectional settings, see e.g. Gouriéroux-Monfort (1996). Our theorem shows that this result holds more generally under weak regularity conditions.

EIA's and ECA's differ regarding the variance due to the approximator: First, common approximators introduce additional correlations across observations, which drive an additional term in the variances of $\nabla^2 G_n(\theta_0)[\psi_S]$ and $\nabla^2 G_n(\theta_0)[\psi_S,\psi_S]$. Second, in contrast to the ECA's, the EIA's asymptotically have no additional variance as $n \to \infty$. These two points do not imply however EIA's are preferable to ECA's: we need to 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 variances of the two approximate estimators have the same rate in the leading case $\alpha_2 = 1$.

Finally, we should mention that the rate we give for the variance of ECA's is not always sharp. For example, when $\hat{\gamma}_S$ is a kernel estimator, we can show that $\mathrm{EVar}(\hat{\theta}_{n,S}) = O(n^{-1}S^{-\beta}) + O(S^{-1})$ which is sharper since $O(S^{-\alpha_2}) = O(1/(Sh^d))$; see Creel and Kristensen (2009) and Kristensen and Shin (2008).

To illustrate the use of our results, we return to Examples 1 and 2 of Section 2; more examples are given in the working paper version of the paper, Kristensen and Salanié (2010). In the following, the notation $\dot{f}(x,\theta)$ stands for $\partial f(x,\theta)/(\partial\theta)$.

Example 1 (SML in discrete choice models). Consider a discrete choice model where $d \in \{1,...,L\}$ is the decision variable and x a set of covariates. Let $P_l(x;\theta) := P_\theta(d = l|x)$, $l = 1,...,L$, denote the choice probabilities and suppose these are not available on closed form. Then $\gamma_0(x;\theta) := (P_1(x;\theta),...,P_L(x;\theta))$ is a vector function. Given observations of $z = (y,x)$, where $y = (d_1,....,d_L)$ with $d_l = 1$ if $d = l$, the individual log-likelihood is given by $\log p(z;\theta) = \sum_{l=1}^L d_{l,i}\log\gamma_{0,l}(x;\theta)$. In this case,

$$g(z;\theta,\gamma) = \frac{\partial\log p(z;\theta)}{\partial\theta} = \sum_{l=1}^L d_{l,i}\frac{\dot{\gamma}_l(x_i;\theta)}{\gamma_l(x_i;\theta)}.$$

The pathwise derivatives take the form

$$\begin{aligned}
\nabla g(z;\theta)[d\gamma] &= \sum_{l=1}^L d_{l,i}\left[\frac{1}{\gamma_{0,l}(x;\theta)}d\dot{\gamma}_l(x;\theta) - \frac{\dot{\gamma}_{0,l}(x;\theta)}{\gamma_{0,l}^2(x;\theta)}d\gamma_l(x;\theta)\right], \\
\nabla^2 g(z;\theta)[d\gamma,d\gamma] &= \sum_{l=1}^L d_{l,i}\left[-\frac{2}{\gamma_{0,l}^2(x;\theta)}d\gamma_l(x;\theta)\,d\dot{\gamma}_l(x;\theta) + \frac{2\dot{\gamma}_{0,l}(x;\theta)}{\gamma_{0,l}^3(x;\theta)}d\gamma_l(x;\theta)^2\right], \\
\nabla^3 g(z;\theta)[d\gamma,d\gamma,d\gamma] &= \sum_{l=1}^L d_{l,i}\left[\frac{4}{\gamma_{0,l}^2(x;\theta)}d\gamma_l(x;\theta)^2\,d\dot{\gamma}_l(x;\theta) - \frac{6\dot{\gamma}_{0,l}(x;\theta)}{\gamma_{0,l}^4(x;\theta)}d\gamma_l(x;\theta)^3\right]
\end{aligned}$$

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: $L_n = \nabla G_n(\theta_0)[\hat{\gamma}_S - \gamma_0]$ and $Q_n = \nabla^2 G_n(\theta_0)[\hat{\gamma}_S - \gamma_0, \hat{\gamma}_S - \gamma_0]$. By standard arguments, we see that eq. (6) holds with $m = 2$ if

$$\bar{G}_0 := \sum_{l=1}^L E\left[\left\{\frac{6\left\|\dot{\gamma}_{0,l}(x;\theta_0)\right\|}{\gamma_{0,l}^3(x;\theta_0)} + \frac{4}{\gamma_{0,l}^2(x;\theta_0)}\right\}\right] < \infty.$$

Thus $\bar{G}_0$ cannot be finite unless $E\left[\gamma_{0,l}^{-2-k}(x;\theta)\right] < \infty$ for $k = 0,1$. 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 2 (SNLS). For the PMLE proposed in Laroque and Salanié (1989), $\gamma_0(x;\theta) =$

$E\left[y|x,\theta\right]$. The first-order condition takes the form

$$g(z;\theta,\gamma)=2\left(y-\gamma\left(x;\theta\right)\right)\dot{\gamma}\left(x;\theta\right)$$

Define $\xi\left(z;\theta\right):=y-\gamma_0\left(x;\theta\right)$; then the functional differentials are

$$\nabla g(z;\theta)\left[d\gamma\right]=2\dot{\gamma}_0\left(x;\theta\right)d\gamma\left(x_i;\theta\right)+2\xi\left(z;\theta\right)d\dot{\gamma}\left(x;\theta\right),$$

$$\nabla^2 g(z;\theta)\left[d\gamma,d\gamma\right]=4d\dot{\gamma}\left(x;\theta\right)d\gamma_0\left(x;\theta\right).$$

Since $\nabla^3 g(z;\theta)\left[d\gamma,d\gamma,d\gamma\right]=0$, eq. (6) holds with $\bar{G}_0=0$ and the remainder term $R_{S,n}$ in eq. (9) is zero.

Assuming that $E\left[y^2\right]<\infty$, $E[\sup_{\theta\in\Theta}\left\|\gamma_0\left(x,\theta\right)\right\|^2]<\infty$ and $E[\sup_{\theta\in\Theta}\left\|\dot{\gamma}_0\left(x;\theta\right)\right\|^2]<\infty$, it is easily seen that eqs. (7)-(8) 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 $\left\|\gamma\right\|^2=E[\left\|\gamma\left(x_i;\theta\right)\right\|^2]$. If on the other hand the same simulations have been used for both, we need to use $\left\|\gamma\right\|^2=E[\left\|\gamma\left(x;\theta\right)\right\|^2]+E[\left\|\dot{\gamma}\left(x;\theta\right)\right\|^2]$.

3.4 Asymptotic First-Order Equivalence

Our results allow us to state precisely when the approximate estimator is asymptotically first-order equivalent to the exact estimator; that is, which choices of the sequence $S=S_n$ guarantee $||\hat{\theta}_{n,S_n}-\hat{\theta}_n||=o_P\left(n^{-1/2}\right)$. In general, asymptotic equivalence for ECA's are obtained if $n/S^{\min(\alpha_2,2\beta)}\rightarrow 0$; 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. One can show (see Kristensen and Shin, 2008) that for the NPSMLE based on ECA's to be equivalent to the MLE, we need $\sqrt{n}h^r\rightarrow 0$, $n/S\rightarrow 0$ and $\sqrt{n}/\left(Sh^d\right)\rightarrow 0$.

3.5 Estimating the Variance

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 many cases.

Under the additional assumption that $E\left[\left\|g\left(z_i;\theta_0,\gamma_0\right)\right\|^2\right]<\infty$, conditions A.1-A.4 imply that $\hat{\theta}_n$ has standard "sandwich" asymptotics, $\sqrt{n}(\hat{\theta}_n - \theta_0) \to^d N\left(0, H_0^{-1}\Omega H_0^{-1}\right)$ where $\Omega = \lim_{n\to\infty} \mathrm{Var}\left(G_n\left(\theta_0, \gamma_0\right)\right)$.

For the approximate estimator,

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

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 $\nabla g(z_i, \hat{\theta}_{n,S})[\psi_{i,S}]$; it accounts for the additional variance due to the simulations. In the leading example where $\hat{\gamma}_{i,S}$ satisfies A.7,

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

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

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

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

4 Analytical Bias Adjustment

We now propose an analytical bias adjustment of the objective function $G_n\left(\theta, \hat{\gamma}_S\right)$ which removes the term B_2 in the formula for $\mathrm{EBias}(\hat{\theta}_{n,S})$. For the EIA scheme, B_2 is in fact the leading term of the approximation bias if the approximator's variance is of a larger order than its bias: $\alpha_2 < \beta$. This is clearly the case for the parametric simulation-based estimation methods ($\alpha_2 = 1$, $\beta = \infty$), and so for this class of approximators are proposed method will remove the leading bias term. We would like to stress that the proposed bias adjustment method requires the approximator to satisfy A.7. Thus, the method is not applicable to approximation schemes that cannot be expressed as an average.

Our adjustment is based on an estimator of the bias term $B_2 = -H_0^{-1}E\left[\nabla^2 g(z_i; \theta_0)[\psi_S, \psi_S]\right]/2$ which we then include in the objective function. For approximators satisfying A.7, first note that

$$\nabla^2 G_n(\theta_0)[\psi_S,\psi_S] = \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S}\nabla^2 g(z_i;\theta_0)[e_{S,is},e_{S,is}],$$

Here, we write, for EIA's, $e_{S,is} := e_S\left(z,\varepsilon_{i,s};\theta\right)$ and, for ECA's, $e_{S,is} = e_S\left(z,\varepsilon_s;\theta\right)$ for $i = 1,...,n$ where e_S is defined in A.7.

To obtain an estimator of the right hand side in the above equation, we would ideally compute $e_S = w_S - \bar{w}_S$; but since in general $\bar{w}_S$ 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 $E\left[\nabla^2 g(z_i;\theta)[\psi_S,\psi_S]\right]/2$ is:

$$\dot{\Delta}_{n,S}\left(\theta\right) = \frac{1}{2S\left(S-1\right)}\sum_{i=1}^{n}\sum_{s=1}^{S}\nabla^2 g(z_i;\theta)[w_{S,is}-\hat{\gamma}_{S,i},w_{S,is}-\hat{\gamma}_{S,i}]. \tag{15}$$

Under regularity conditions, $||H_0^{-1}\dot{\Delta}_{n,S}\left(\theta_0\right) - B_2|| \to^P 0$ as $n\to\infty$. This motivates our definition of an analytically bias-adjusted estimator $\hat{\theta}_{n,S}^{\mathrm{AB}}$ as the solution to:

$$o_P\left(n^{-1/2}\right) = G_n(\hat{\theta}_{n,S}^{\mathrm{AB}},\hat{\gamma}_S) - \dot{\Delta}_{n,S}(\hat{\theta}_{n,S}^{\mathrm{AB}}). \tag{16}$$

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

$$\hat{\theta}_{n,S}^{\mathrm{AB}} = \arg\min_{\theta\in\Theta}\left\{Q_n(\theta,\hat{\gamma}_S) - \Delta_{n,S}\left(\theta\right)\right\}, \tag{17}$$

where

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

After such an adjustment, the bias component B_2 changes to

$$\tilde{B}_2 := H_0^{-1}E\left[\frac{1}{2}\nabla^2 G_n(\theta_0)[\psi_S,\psi_S] - \dot{\Delta}_{n,S}\right].$$

The following Theorem shows that under slightly stronger conditions[12] than in Theorem 1, $\tilde{B}_2$ has a faster rate of convergence than B_2, while the rate of the other leading terms is unchanged:

Theorem 2 *Assume that A.1-A.4, A.5*(3) *and A.7*(8) *hold together with*

$$\left\|\nabla^2 g(z)[e_{is},e_{it}]\right\| \leq b\left(z\right)\left\|e_{is}\left(z\right)\right\|\left\|e_{it}\left(z\right)\right\|,$$

[12] Note the different orders on A.5 and A.7; they are required to ensure that the remainder term, $R_{n,S}$, in the asymptotic expansion is still dominated.

where $E\left[b^{8}(z)\right]<\infty$*. Then, with* $\dot{\Delta}_{n,S}(\theta)$ *and* $\hat{\theta}_{n,S}^{\mathrm{AB}}$ *defined in eqs.(15)-(16):*

$$\mathrm{EBias}(\hat{\theta}_{n,S})=B_1+\tilde{B}_2$$

where $\tilde{B}_2=O(S^{-2+\mu_2})$*. The rates of* B_1 *and* $\mathrm{EVar}(\hat{\theta}_{n,S})$ *are the same as in Theorem 1.*

Comparing with Theorem 1, the bias term $B_2=O\left(S^{-\alpha_2}\right)=O\left(S^{-1+\mu_2}\right)$ has been replaced by $\tilde{B}_2=O(S^{-2+\mu_2})$. With unbiased simulators, we have $\mu_2=0$ and $\beta=\infty$, and the leading bias term of the approximation error of the unadjusted estimator is of order $O\left(S^{-1}\right)$. The above theorem shows that for the adjusted estimator the leading term of the bias is of order $O\left(S^{-2}\right)$. The improvement is by a factor S and so may be very significant.

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[13]. With non-stochastic approximations, the leading term of the approximation error is not driven by $\nabla^2 G_n(\theta)[\psi_S,\psi_S]$, which the $\dot{\Delta}_{n,S}(\theta)$ 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 1 (SML on discrete choice models). Here, $q(z;\theta,\gamma)=-\sum_{l=1}^{L}d_{l,i}\log\gamma_l(x;\theta)$ and so $\nabla^2 q(z;\theta)[d\gamma,d\gamma]=\sum_{l=1}^{L}d_{l,i}d\gamma_l^2(x;\theta)/\gamma_{0,l}^2(x;\theta)$. .Thus, the adjustment term becomes

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

Example 2 (SNLS). Using the results obtained for the SNLS in the previous section,

$$\nabla^2 G_n(\theta)[d\gamma,d\gamma]=\frac{4}{n}\sum_{i=1}^{n}d\dot{\gamma}(x_i;\theta)\,d\gamma(x_i;\theta).$$

Then the adjustment term becomes

$$\Delta_{n,S}(\theta)=\frac{1}{nS(S-1)}\sum_{i=1}^{n}\sum_{s=1}^{S}r^2(x_i,\varepsilon_s;\theta),\quad r(x_i,\varepsilon_s;\theta):=w_S(x_i,\varepsilon_s;\theta)-\hat{\gamma}_S(x_i;\theta)$$

[13] If we could estimate b_S, then B_1 could also be adjusted straightforwardly with $\dot{\Delta}_{n,S}^{(B)}(\theta)=\nabla G_n(\theta)[b_S]$. However, estimating b_S is usually a difficult task. Lee (2001) demonstrates how combining numerical approximations and simulations can improve the order of the estimator. When kernel-based estimators are used, higher-order kernels can also be used to reduce the bias component.

This is exactly the correction proposed in Laffont et al. (1995); and as $\nabla^3 q_i \equiv 0$ in SNLS, all approximation biases are removed.

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. (16), we obtain

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

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:

$$\dot{\Delta}_{n,S}(\theta) = \frac{2}{nS(S-1)} \sum_{i=1}^{n} \sum_{s=1}^{S} r(x_i, \varepsilon_s; \theta) \dot{r}(x_i, \varepsilon_s; \theta),$$

where as, before, $\dot{f}$ denotes the derivative of f w.r.t. θ. One complication of this corrective procedure relative to the preventive one is that the derivatives of the simulators must be computed. We refer to Arellano and Hahn (2007) for a further discussion of corrective and preventive bias correction in a panel data setting.

5 Bias Adjustment by Resampling

As an alternative to analytical bias corrections, resampling methods can be used[14]. 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, and may lead to an increase in finite-sample variance.

To motivate the bias adjustment, recall from Theorem 1 that $\mathrm{EBias}(\hat{\theta}_{n,S}) \simeq b_1 S^{-\beta} + b_2 S^{-\alpha_2}$. As before, the goal is to obtain an estimator of (parts of) the leading bias terms 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 $\hat{\gamma}_{S^*}^{[m]}$, $m = 1, 2$.

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

$$\hat{\theta}_{n,S}^{\mathrm{JK}} := 2\hat{\theta}_{n,S} - \frac{1}{2}\left(\hat{\theta}_{n,S^*}^{[1]} + \hat{\theta}_{n,S^*}^{[2]}\right), \tag{18}$$

[14] See Hahn and Newey (2004) and Dhaene and Jochmans (2010) 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

$$\begin{aligned} E\left[\hat{\theta}^{\text{JK}}_{n,S} - \hat{\theta}_n\right] &= 2E\left[\hat{\theta}_{n,S} - \hat{\theta}_n\right] - \frac{1}{2}\left(E\left[\hat{\theta}^{[1]}_{n,S^*} - \hat{\theta}_n\right] + E\left[\hat{\theta}^{[2]}_{n,S^*} - \hat{\theta}_n\right]\right) \\ &\simeq b_1\left[2S^{-\beta} - (S^*)^{-\beta}\right] + b_2\left[2S^{-\alpha} - (S^*)^{-\alpha_2}\right], \end{aligned}$$

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

$$S^* = \frac{S}{2^{1/\beta}} \text{ or } S^* = \frac{S}{2^{1/\alpha_2}}, \tag{19}$$

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

$$\hat{\theta}^{\text{JK}}_{n,S} = M\hat{\theta}_{n,S} - \sum_{m=1}^{M} p_m \hat{\theta}^{[m]}_{n,S_m}. \tag{20}$$

Dhaene and Jochmans (2010, Section 3.1) demonstrate in a panel data context that the optimal procedure is to choose $M = 2$ and $p_m = 1/2$ if the goal is to remove the leading bias term. We expect that a similar result extends to parametric simulation-based estimators in our setting. On the other hand, the generalized adjustment as given in eq. (20) 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 (2002): For each $m = 1, 2$, compute

$$\hat{\theta}^{[m,k+1]}_{n,S^*} = \hat{\theta}^{[m,k]}_{n,S^*} - \left[\frac{\partial G_n(\hat{\theta}^{[m,k]}_{n,S^*}, \hat{\gamma}^{[m]}_{S^*})}{\partial \theta}\right]^{-1} G(\hat{\theta}^{[m,k]}_{n,S^*}, \hat{\gamma}^{[m]}_{S^*}), \quad k = 1, 2, 3, ... \tag{21}$$

with starting value $\hat{\theta}^{[m,1]}_{n,S^*} = \hat{\theta}_{n,S}$, and compute $\hat{\theta}^{\text{JK}}_{n,S}$ 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, as we did in section 2. Define

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

It is easy to see that the estimator defined by $G_n^*(\tilde{\theta}_{n,S}^{\mathrm{JK}}, \hat{\gamma}_S) = 0$ is equivalent to $\hat{\theta}_{n,S}^{\mathrm{JK}}$ given in eq. (18) 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. Often the exact estimator suffers from biases of order n^{-1} *relative to the true value* in finite samples:

$$E\left[\hat{\theta}_{n,S} - \theta_0\right] \simeq b_1 S^{-\beta} + b_2 S^{-\alpha_2} + b_3 n^{-1},$$

where again higher-order terms are suppressed. Note that we now consider $E[\hat{\theta}_{n,S} - \theta_0]$ instead of $E[\hat{\theta}_{n,S} - \hat{\theta}_n]$. Then, by the same arguments as before, it is easily seen that $\hat{\theta}_{n,S}^{\mathrm{JK}}$ also removes the third term, $b_3 n^{-1}$, 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 the initial, unadjusted estimator, $\hat{\theta}_{n,S}$, one or more Newton-Raphson iterations based on the approximate objective function with a finer approximation $S^* > S$ produce an estimator that has the presumably higher precision of $\hat{\theta}_{n,S^*}$. The resulting estimator based on k iterations, $\hat{\theta}_{n,S}^{(k+1)}$, is defined in eq. (4).

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

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

Combining this with Robinson (1988, Theorem 2), we obtain the following theorem:

Theorem 3 *Assume that A.1-A.4, A.5(3) and A.7(6) hold. Let the initial estimate $\bar{\theta}_{n,S}$ be chosen as either $\hat{\theta}_{n,S}$, $\hat{\theta}_{n,S}^{\mathrm{AB}}$, or $\hat{\theta}_{n,S}^{\mathrm{JK}}$. Then the NR-estimator $\hat{\theta}_{n,S}^{(k+1)}$ defined in eq. (4) satisfies:*

$$||\hat{\theta}_{n,S}^{(k+1)} - \hat{\theta}_n|| = O_P\left(||\bar{\theta}_{n,S} - \hat{\theta}_n||^{2^k}\right) + O_P\left(||\bar{\theta}_{n,S^*} - \hat{\theta}_n||\right) \tag{22}$$

as n, S and S^* go to infinity with $S^* > S$.

The above result formalizes the intuition that a (large enough) number of NR-steps with the score and Hessian evaluated at γ_{S^*} yields an estimator that is equivalent to the extremum estimator obtained from full optimization of the objective function based on γ_{S^*}. This holds irrespective of the convergence rate of the initial estimator. However, the number of NR iterations, k, needed to obtain this result does depend on the precision of the initial estimator. For unadjusted parametric simulation-based estimators in the EIA scheme for instance, we know from Theorem 1 that $\|\bar{\theta}_{n,S} - \hat{\theta}_n\| = O_P(1/S)$. Then the first term on the right-hand side of the inequality in Theorem 3 is asymptotically dominated by the second term if $S^* = o(S^{2^k})$. Taking $k = 1$ and having S^*/S converge to some positive number would be enough in this case.

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 3 has to be adjusted, cf. Robinson (1988, Theorem 5). In particular, more iterations are required to obtain a given level of precision.

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.

We take the true model to have parameters $a = 1, s = 1, b = 0$. In this specification, the mean probability of $y = 1$ is close to 0.5. For $\tau = 1$ (resp. $\tau = 2$) the generalized R^2 is 0.11 (resp. 0.21); in the corresponding simple logit model, which has $s = 0$, the R^2 would be 0.17 (resp. 0.39.)

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. In empirical IO, the simulated method of moments is 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. Instead, we focus here on SML, which is perhaps the most popular method outside of empirical IO.

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

$$\Pr(y=1|x) = \int \frac{\phi(u)}{1+\exp(-(b+(a+su)x)}du. \tag{23}$$

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.

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

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-diameter 0.2 for the mean slope a. With $\tau = 1$ it would take about $(7.2 * 1.96/0.2)^2 \simeq 5,200$ observations; and still about $4,500$ for $\tau = 2$, even though the generalized R^2 almost doubles. With such sample sizes, the estimate of the size of the heterogeneity s would still be very noisy: its 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 $\hat{\theta}_{nS} - \hat{\theta}_n$ in our theorems. We will be using SML under the EIA scheme (independent draws across observations). 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 $S = 100$ simulations and $\tau = 1$ for instance, the bias on a is -0.09, and the bias on s is -0.23. For sample sizes of a few 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, on which we will focus here.

We ran experiments for several sets of parameter values, sample sizes n, explanatory power (through τ), and numbers of draws S. Since the results are similar, we only present here those we obtained for a sample of $25,000$ observations when the true model has $a = 1, s = 1, b = 0$, the covariate has a standard error $\tau = 2$.

We present below the results for $S = 50, 100, 20$, and 500 simulations. We ran 5,000 simulations in each case, starting from initial values of the parameters drawn randomly from uniform distributions: $a \sim U[0.5, 1.5]$, $b \sim U[-0.5, 0.5]$, and $s \sim U[0.5, 1.5]$. Our exact ML estimator relies on Gaussian quadrature as in equation (23), with 64 points. For each simulated sample, we estimated the model using both uncorrected SML, and SML with analytic bias adjustment (ABA), resampling, and Newton-Raphson. The ABA was done on the objective function, while the resampling was done on the estimator itself. For the Newton-Raphson correction, we use only $k = 1$ step, with $S^* = 10 \times S$ draws. Finally, our resampling correction uses $S^* = S/2$ draws.

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 1.4% to 1.6% of the 5,000 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. “ABA” refers to our analytical bias adjustment.

Tables 2, 3 report our results for the mean error of our various SML methods, relative to the ML estimator. Each row corresponds to a value of the number of simulations S. All numbers in the last five columns 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 5,000 samples (minus the small number that were eliminated due to numerical issues). The standard error of these averages is about 0.001, so that many of the biases from the corrected estimates are close to insignificant.

S	SML	SML+Newton	SML+resampling	SML+ABA	SML+ABA+Newton
50	-0.136	-0.029	-0.034	0.011	0.003
100	-0.072	-0.009	-0.008	0.005	0.003
200	-0.036	-0.002	0.000	0.004	0.003
500	-0.013	0.002	0.003	0.003	0.003

Table 2: Mean error on a

The ”SML” columns in the tables report the biases of the uncorrected SML estimator. It follows from Table 1 that the theoretical values of the leading term of the bias are -0.165 for

S	SML	SML+Newton	SML+resampling	SML+ABA	SML+ABA+Newton
50	−0.226	−0.027	−0.054	0.015	0.002
100	−0.121	−0.012	−0.016	0.005	0.002
200	−0.062	−0.005	−0.003	0.002	0.002
500	−0.025	−0.001	0.001	0.002	0.002

Table 3: Mean error on s

a and -0.270 for s when using $S = 50$ simulations; since the leading term is in $1/S$, it should be ten times smaller for $S = 500$. The leading term appears to be a good approximation to the actual size of the bias in these simulations, and the measured bias is close to proportional to $1/S$. This suggests that the two methods that focus on correcting for the leading term of the bias, our analytical bias adjustment and the resampling method, should work very well. ABA in fact does eliminate most of the bias; resampling also works quite well, at least for $S \geq 100$. The Newton step with ten times more simulations reduces the bias, as expected; but it does not do it quite as effectively as our analytical bias adjustment. In addition, applying a Newton step after ABA does not reduce the bias further as it is already very small. We note in passing that for the estimation of s in particular, the Newtonized SML estimator for $S = 50$ works about as well as the SML estimator for $S = 500$, as theory suggests.

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. Figure 1 plots the estimated densities of the error terms $\hat{\theta}_{n,S} - \hat{\theta}_n$. 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 ten times more simulations, it reduces the efficiency loss relative to the MLE. On the other hand, resampling corrects the S-simulations estimator by using an average of estimators with $S^* = S/2$, and so it introduces more noise.

S	SML	SML+Newton	SML+resampling	SML+ABA	SML+ABA+Newton
50	0.137	0.032	0.047	0.025	0.012
100	0.074	0.015	0.028	0.018	0.011
200	0.038	0.011	0.022	0.015	0.011
500	0.017	0.010	0.017	0.012	0.011

Table 4: RMSE on a

These trade-offs are reflected in the RMSEs of the error terms, as collected in tables 4 and 5. Correcting the error using analytical bias adjustment or a Newton step reduces the

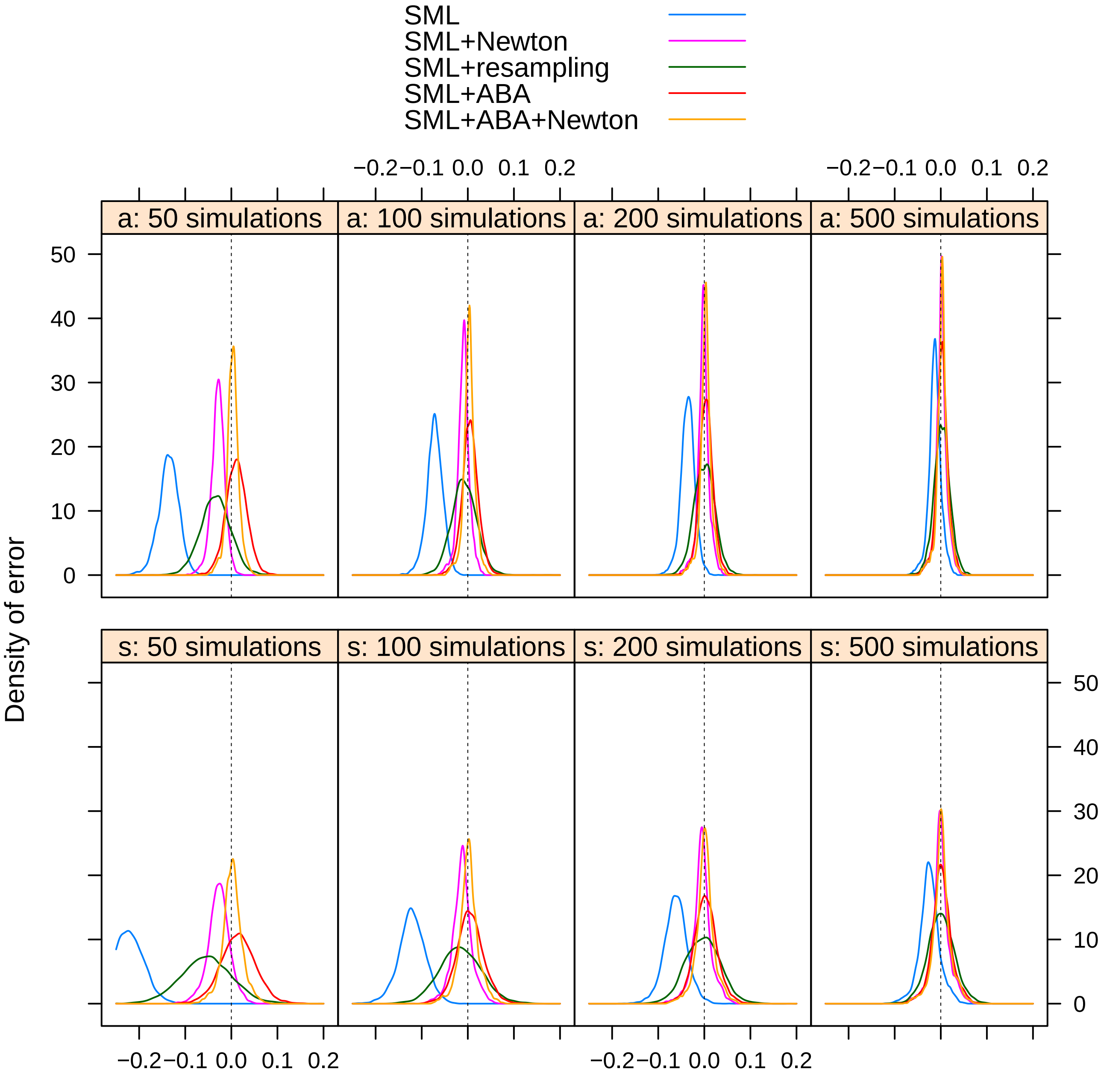

Figure 1: Estimation errors due to simulation

S	SML	SML+Newton	SML+resampling	SML+ABA	SML+ABA+Newton
50	0.229	0.035	0.077	0.040	0.020
100	0.124	0.023	0.048	0.029	0.018
200	0.066	0.018	0.038	0.024	0.017
500	0.031	0.017	0.028	0.020	0.017

Table 5: RMSE on s

RMSE by similar amounts; but doing both combines the bias-reducing effect of the ABA and the dispersion-reducing effect of the Newton step to yield a spectacular reduction in the RMSE. 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 models with more than a few parameters, as it requires a reasonably accurate evaluation of the matrix of second derivatives. In our experiment, we relied on the fact that the minimization algorithm itself proceeds by Newton-Raphson steps; after multiplying by ten the number of simulations, we let the algorithm do exactly one iteration of its line search. This appears to work very well, and is very easy to implement.

- *Computer time:* We report in Table 6 the average time each of our five methods took to produce an estimator (in seconds per sample.) The analytical bias adjustment wins this comparison hands down. For SML for instance, the evaluation of the corrected objective function requires the variance of the simulated choice probabilities in addition to their mean—a very small computational cost. Resampling, as implemented in this study, roughly doubles the cost of the uncorrected estimator. Newton can be more costly still, depending on the structure of the model and the care needed to approximate the Hessian.

S	SML	SML+Newton	SML+resampling	SML+ABA	SML+ABA+Newton
50	1.8	4.9	4.1	1.9	6.0
100	3.4	9.4	7.2	3.6	10.8
200	6.7	18.6	13.7	6.9	20.6
500	16.4	46.2	33.3	16.9	50.3

Table 6: Computing times (in seconds)

Like all Monte Carlo study, 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 approximate 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. Finally, combining analytical bias correction with a Newton step spectacularly reduces the RMSE of the error.

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 where objective function and approximator (as functions of θ) were both smooth. In principle, one could import the arguments of Chen et al (2003) to handle non-smooth cases. 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.

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 (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

Ackerberg, D., J. Geweke and J. Hahn (2009) Comments on "Convergence Properties of the Likelihood of Computed Dynamic Models". *Econometrica* 77, 2009–2017.

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

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

Brownlees, C.T., D. Kristensen and Y. Shin (2011) Smooth Filtering and Likelihood Inference in Dynamic Latent Variables 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 (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.

Dubé, J.P., J. Fox, and C-L. Su (2009) Improving the Numerical Performance of BLP Static and Dynamic Discrete Choice Random Coefficients Demand Estimation, mimeo, Chicago Booth.

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

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

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

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

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

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 (2010) Constrained Optimization Approaches to Estimation of Structural models. Working paper, CMS-EMS.

Kristensen, D. and B. Salanié (2010) Higher Order Improvements for Approximate Estimators. CAM Working Paper 2010-04, University of Copenhagen.

Kristensen, D. and B. Schjerning (2011) Implementation and Estimation of Discrete Markov Decision Models by Sieve Approximations. Manuscript, University of Copenhagen.

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.

Mariano, R., 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. 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. and R. Smith (2004) Higher-order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica* 72, 219–255.

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

Norets, A. (2011) Estimation of Dynamic Discrete Choice Models Using Artificial Neural Network Approximations. Forthcoming in *Econometric Reviews.*

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

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.

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.

Rothenberg, T.J. (1984) Approximating the Distributions of Econometric Estimators and Test Statistics. In *Handbook of Econometrics*, vol. 2, eds. K. Arrow and M. Intriligator. North Holland.

Rust, J. (1997) Using Randomization to Break the Curse of Dimensionality. *Econometrica* 65, 487-516.

Tauchen, G. and R. Hussey(1991) Quadrature-Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models. *Econometrica* 59, 371-396.

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

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

A Proofs

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

$$\sup_{\theta\in\Theta}\|G_n(\theta,\gamma_0)-G(\theta,\gamma_0)\|\to^P 0, \tag{24}$$

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 $\lambda \leq 2$,

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

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 $\lambda \leq 2$,

$$\begin{aligned} E\left[\sup_{\theta\in\Theta}\left\|G_n(\theta, \hat{\gamma}_S) - G_n(\theta, \gamma_0)\right\|\right] &\leq \frac{1}{n}\sum_{i=1}^{n} E\left[\sup_{\theta\in\Theta}\left\|g(z_i; \theta, \hat{\gamma}_S) - g(z_i; \theta, \gamma_0)\right\|\right] \\ &\leq \bar{G}_0 \frac{1}{n}\sum_{i=1}^{n} E\left[\left\|\hat{\gamma}_{i,S} - \gamma_0\right\|^{\lambda}\right] \\ &= o_P(1) \end{aligned} \qquad (25)$$

Combining this result with eq. (24), we obtain $\sup_{\theta\in\Theta}\|G_n(\theta, \hat{\gamma}_S) - G(\theta, \gamma_0)\| \to^P 0$. 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:

$$o_P\left(n^{-1/2}\right) = G_n(\hat{\theta}_{n,S}, \hat{\gamma}_S) = G_n(\theta_0, \hat{\gamma}_S) + H_n(\bar{\theta}_{n,S}, \hat{\gamma}_S)(\hat{\theta}_{n,S} - \theta_0), \qquad (26)$$

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. (24)-(25), Assumption A.4 then ensures that,

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

Going back to eq. (26), we have now shown that

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

Subtracting gives

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

We now use the expansion given in eq. (9) with $m=2$ and $\theta=\theta_0$, to get

$$\left\|\hat{\theta}_{n,S}-\hat{\theta}_n\right\|=O_P\left(\left\|\nabla G_n(\theta_0)\left[\Delta\hat{\gamma}_S\right]+\frac{1}{2}\nabla^2 G_n(\theta_0)\left[\Delta\hat{\gamma}_S,\Delta\hat{\gamma}_S+R_{n,S}\right]\right\|\right)+o_P\left(1/\sqrt{n}\right), \tag{27}$$

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

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

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:

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

The rates of the first and second order functional differentials of $G_n(\theta_0,\gamma)$ are given in Lemmas 6 and 7 depending on whether the ECA approximator of (10) or the EIA approximator of eq. (11) is used. By plugging those into eq. (27) 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 $\hat{\theta}_{n,S}^{\mathrm{AB}}$ is consistent.

Next, we make a Taylor expansion of eq. (16),

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

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

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

Suppressing any dependence on θ_0, use eq. (9) to write

$$\begin{aligned} G_n\left(\hat{\gamma}_S\right) - \dot{\Delta}_{n,S} - G_n\left(\gamma\right) &= \left\{\frac{1}{2}\nabla^2 G_n[\psi_{n,S}, \psi_{n,S}] - \dot{\Delta}_{n,S}\right\} + \nabla G_n[\hat{\gamma}_S - \gamma] \\ &\quad + \frac{1}{2}\left\{\nabla^2 G_n[\hat{\gamma}_S - \gamma, \hat{\gamma}_S - \gamma] - \nabla^2 G_n[\psi_{n,S}, \psi_{n,S}]\right\} + R_{n,S}. \end{aligned} \tag{28}$$

The rates of the second and third terms of eq. (28) are derived in Lemma 7 while Lemma 8 delivers a refinement of $R_{n,S}$ relative to the rate obtained in the proof of Theorem 1 that ensures it is negiglible. The crucial term is the first term of eq. (28). 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. (28) 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 4 with $r = 1$ to $W_{i,S} := S^{-2} \sum_{s \neq t} \nabla^2 g_i[e_{is}, e_{it}]$, getting

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

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.7), we can employ U-statistic results for absolutely regular sequences: Yoshihara (1976, Lemma 3) states that $E\left[\|W_{i,S}\|^4 |z_i\right] = O\left(S^{-4}\right)$. By inspection of the proof of Yoshihara (1976, Lemma 3), it is easily checked that in fact, for some constant $C > 0$ we have $E\left[\|W_{i,S}\|^4 |z_i\right] \leq CS^{-4} M_S(z_i)$, where

$$M_S(z_i) := \sup_{s<t} E\left[\left\|\nabla^2 g(z_i)[e_{is}, e_{it}]\right\|^{4+\epsilon} |z_i\right]^{4/(4+\epsilon)}, \text{ for some } \epsilon > 0.$$

Thus, with $\delta = 2$ and using the Lipschitz condition on $\nabla^2 g$, we obtain

$$\begin{aligned}
E\left[\|W_{i,S}\|^4\right] &\leq CS^{-4} E\left[M_S(z_i)\right] \\
&\leq CS^{-4} E\left[\sup_{s<t} E\left[\left\|\nabla^2 g(z_i)[e_{is}, e_{it}]\right\|^{4+\epsilon} |z_i\right]^{4/(4+\epsilon)}\right] \\
&\leq CS^{-4} E\left[b^4(z_i) \sup_{s<t} E\left[\|e_{is}(z)]\|^{4+\epsilon} \|e_{it}(z)]\|^{4+\epsilon} |z_i\right]^{4/(4+\epsilon)}\right] \\
&\leq CS^{-4} E\left[b^4(z_i) E\left[\|e_{is}(z)]\|^{8+\epsilon} |z_i\right]^{4/(8+\epsilon)}\right] \\
&\leq CS^{-4} \sqrt{E\left[b^8(z_i)\right]} E\left[\|e_{is}\|^{8+2\epsilon}\right]^{4/(8+2\epsilon)} \\
&= O\left(S^{-4+\mu_8/2}\right).
\end{aligned}$$

It follows that:

$$\frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s\neq t}\nabla^2 g_i[e_{is}, e_{it}] = O_P(n^{-1/2}S^{-1+\mu_8/4}).$$

As for the second term, by definition $\hat{\gamma}_i - \bar{w}_i = \psi_{S,i}$; and it follows from Lemma 5 that $E\left[\nabla^2 g_i[\psi_{S,i}, \psi_{S,i}]\right] = O\left(S^{-\alpha_2}\right)$ and

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

Summing up, $\tilde{B}_2 = H_0^{-1} E\left[\nabla^2 G_n[\psi_{n,S}, \psi_{n,S}]/2 - \dot{\Delta}_{n,S}\right] = O\left(S^{-2+\mu_2}\right)$ while

$$\operatorname{Var}\left(\nabla^2 G_n[\psi_{n,S}, \psi_{n,S}]/2 - \dot{\Delta}_{n,S}\right) = O(n^{-1}S^{-2+\mu_8/2}) + O\left(n^{-1}S^{-2+\alpha_4}\right).$$

This completes the proof. ■

Proof of Theorem 3. We wish to apply the general result in Robinson (1988, Theorem 2), and so need to check that his conditions A.1 and A.3 are satisfied in our application. His condition A.1 is satisfied by our Assumptions A.1-A.7 since these implies consistency of the initial estimator for suitable choice of S. Robinson's condition A.3 is satisfied by the smoothness conditions imposed on $G_T(\theta, \hat{\gamma}_S)$ in our Assumption A.1 ■

B Lemmas

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

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

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

Proof. From Rio (1994), we obtain that

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^n W_i\right\|^{2r}\right] \leq C_r\left[n^{-r}M_{2,\alpha,n}^r + n^{1-2r}M_{2r,\alpha,n}\right], \tag{29}$$

where $M_{p,\alpha,n}$, $p \geq 2$, is defined in Rio (1994) and $n^{1-2r} = o\left(n^{-r}\right)$ 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,

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

In particular,

$$M^r_{2,\alpha,n} \leq C\left(r, A, a\right) E\left[\left\|W_i\right\|^{2+\delta}\right]^{2r/(2+\delta)}, \quad M_{2r,\alpha} \leq C\left(r, A, a\right) E\left[\left\|W_i\right\|^{2r+\delta}\right]^{2r/(2r+\delta)}. \quad (30)$$

The claimed result now follows by combining eqs. (29) and (30). ■

Lemma 5 *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 $m\left(z; d\gamma\right)$ be a functional satisfying:*

$$E\left[\left\|m\left(z; d\gamma\right)\right\|^{2r+\delta}\right] < \infty, \quad E\left[\left\|m\left(z; d\gamma\right)\right\|^{2+\delta}\right] \leq \bar{M}\left\|d\gamma\right\|^{k(2+\delta)}, \quad (31)$$

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

Then, with b_S and ψ_S given in A.5, $M_S\left(\psi\right) = E\left[m\left(z; \psi_S\right)\right]$, and $M_S\left(b\right) = E\left[m\left(z; b_S\right)\right]$ the following hold:

(i) For EIA's,

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

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

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

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

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

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

(iii) The means satisfy:

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

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. (31) implies that $E\left[\left\|W_{i,S}\right\|^{2r+\delta}\right] < \infty$.

We can therefore apply Lemma 4

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

where $C=C\left(r,a,A\right)$ only depends on r and the mixing coefficients of $\{z_i\}$ and $\left\{\psi_{i,S}\right\}$. By eq. (31),

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

and

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

It is easily seen that the above inequalities 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 $W_{S,i}:=m\left(z_i;\psi_S\right)-\bar{m}\left(\psi_S\right)$. Conditional on ψ_S, it is easily seen that $W_{S,i}$ satisfies the conditions of Lemma 4 such that

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

where $C=C\left(r,a,A\right)$; in particular, it does not depend on ψ_S. Next, observe that

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

and we conclude that

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

Finally,

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

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

Lemma 6 *Under A.1-A.4, A.5(2) and A.6(4), the first and second order differentials of G_n for the ECA yield the rates given in Theorem 1.*

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 of eq. (11), the functional differentials are given by

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

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

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

Now,

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

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

By Lemma 5(i) with $m\left(z; d\gamma\right) = \nabla g\left(z\right)\left[d\gamma\right]$, $k = 1$ and $r = 1$,

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

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

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

Moreover, define $\nabla\bar{g}\left[\gamma\right] = E[\nabla g_i[\gamma]]$ (where expectations are taken w.r.t. the observation z_i); then

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

Recalling the definition of $\nabla\bar{g}\left[\psi_S\right]$, it follows from Lemma 5(ii) with $m\left(z; d\gamma\right) = \nabla g\left(z\right)\left[d\gamma\right]$ 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

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

moreover,

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

and, applying Lemma 5(ii) with $m\left(z; d\gamma\right) = \nabla^2 g\left(z\right)\left[d\gamma, d\gamma\right]$, $k = 2$ and $r = 1$,

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

To bound the variance component, define $\nabla^2\bar{g}\left[\gamma,\gamma\right]=E\left[\nabla^2 g_i\left[\gamma,\gamma\right]\right]$, and write

$$\nabla^2 G_n[\psi_S,\psi_S]=\nabla^2\bar{g}\left[\psi_S,\psi_S\right]+\frac{1}{n}\sum_{i=1}^{n}\left(\nabla^2 g_i\left[\psi_S,\psi_S\right]-\nabla^2\bar{g}\left[\psi_S,\psi_S\right]\right).$$

Applying Lemma 5(ii) with $m\left(z;d\gamma\right)=\nabla^2 g\left(z\right)\left[d\gamma,d\gamma\right]$ and $r=1,k=2$, we obtain that $E\left\|\nabla^2 G_n\left[\psi_S,\psi_S\right]\right\|=O_P\left(S^{-2\alpha_2}\right)$.

Finally, by the same arguments as before, $E\left[\nabla^2 G_n[\psi_S,b_S]\right]=0$ while $\text{Var}\left(\nabla^2 G_n[\psi_S,b_S]\right)=O(n^{-1}S^{-\alpha_4})$ and $\text{Var}\left(\nabla^2 G_n[\psi_S,b_S]\right)=O(n^{-1}S^{-\alpha_2-2\beta})$. ■

Lemma 7 *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 (10) yield the rates given in Theorem 1.*

Proof. Again, we suppress dependence on θ_0. For the EIA, the first and second order differentials are $\nabla G_n\left[d\gamma\right]=\sum_{i=1}^{n}\nabla g_i\left[d\gamma_i\right]/n$ and $\nabla^2 G_n)\left[d\gamma,d\gamma'\right]=\sum_{i=1}^{n}\nabla^2 g_i\left[d\gamma_i,d\gamma_i'\right]/n$, 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 6, 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,

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

while its variance satisfies, using Lemma 5(i),

$$\text{Var}\left(\nabla G_n[\psi_S]\right)\leq\frac{1}{n}CE\left[\left\|\psi_S\right\|^{2+}\right]=O\left(n^{-1}S^{-\alpha_2}\right).$$

Applying Lemma 5(i) and (iii) with $m\left(z;d\gamma\right)=\nabla^2 g\left(z\right)\left[d\gamma,d\gamma\right]$ and $k=2$, the mean and the variance of the second order differential satisfy

$$E\left[\nabla^2 G_n[\psi_S,\psi_S]\right]=E\left[\nabla^2 g_i\left[\psi_{i,S},\psi_{i,S}\right]\right]\leq CE\left[\left\|\psi_{i,S}\right\|^2\right]=O\left(S^{-\alpha_2}\right),$$

and $\text{Var}\left[\nabla^2 G_n[\psi_S,\psi_S]\right]=O(n^{-1}S^{-\alpha_4})$. The cross term satisfies $E\left[\nabla^2 G_n[\psi_S,b_S]\right]=0$ while $\text{Var}\left(\nabla^2 G_n[\psi_S,b_S]\right)=O(n^{-1}S^{-\alpha_2}S^{-2\beta})$, and so we can ignore this term since it is of lower order. ■

Lemma 8 *Assume that A.1-A.4, A.5*(3) *and A.7*(6) *hold. Then the rate of the remainder term $R_{n,S}$ can be sharpened to:*

$$R_{n,S}=O_P\left(S^{-3\beta}\right)+O_P\left(S^{-(2-\mu_4)}\right)+O\left(S^{-(2-\mu_3)}\right)+O\left(n^{-1/2}S^{-(3-\mu_6)/2}\right).$$

Proof. Since the third-order differential exists, the remainder term in eq. (9) can be further expanded to obtain $R_{n,S} = \nabla^3 G_n \left[\Delta\hat{\gamma}_S, \Delta\hat{\gamma}_S, \Delta\hat{\gamma}_S\right]/6 + \bar{R}_{n,S}$ where, by A.4(3) and the same arguments used in the proof of Theorem 1, $E\left[\left\|\bar{R}_{n,S}\right\|\right] \leq \bar{G}_0 E\left[\left\|\Delta\hat{\gamma}_{i,S}\right\|^4\right] = O\left(S^{-4\beta}\right) + O\left(S^{-(2-\mu_4)}\right)$. 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 6, so what remains is the variance component.

In the case of EIA, the variance component can be written as $\nabla^3 G_n\left[\psi_S, \psi_S, \psi_S\right] = \sum_{i=1}^n \nabla^3 g_i\left[\psi_S, \psi_S, \psi_S\right]/n$. By Lemma 5, we obtain:

$$\nabla^3 G_n\left[\psi_S, \psi_S, \psi_S\right] - E\left[\nabla^3 G_n\left[\psi_S, \psi_S, \psi_S\right]\right] = O\left(n^{-1/2}S^{-(3-\mu_6)/2}\right),$$

while, due to the independence between simulations,

$$\begin{aligned}\left|E\left[\nabla^3 G_n\left[\psi_S, \psi_S, \psi_S\right]\right]\right| &\leq \frac{1}{S^3}\sum_{s,t,u=1}^{S}\left|E\left[\nabla^3 g_i\left[e_{i,s}, e_{i,t}, e_{i,u}\right]\right]\right| \\ &= \frac{\left|E\left[\nabla^3 g_i\left[e_{i,s}, e_{i,s}, e_{i,s}\right]\right]\right|}{S^2} \\ &\leq \frac{C}{S^2}E\left[e_{i,s}^3\right] = O(S^{-(2-\mu_3)}).\end{aligned}$$

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

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

Applying Lemma 5(ii) with $m\left(z; d\gamma\right) = \nabla^3 g\left(z\right)\left[d\gamma, d\gamma, d\gamma\right]$, the two terms are $O_P\left(S^{-(3/2-\mu_3)}\right)$ and $O_P(n^{-1/2}S^{-(3-\mu_6)/2})$ respectively. ■