The ABC of Simulation Estimation with Auxiliary Statistics

Jean-Jacques Forneron* Serena Ng†

August 2016

Abstract

The frequentist method of simulated minimum distance (SMD) is widely used in economics to estimate complex models with an intractable likelihood. In other disciplines, a Bayesian approach known as Approximate Bayesian Computation (ABC) is far more popular. This paper connects these two seemingly related approaches to likelihood-free estimation by means of a Reverse Sampler that uses both optimization and importance weighting to target the posterior distribution. Its hybrid features enable us to analyze an ABC estimate from the perspective of SMD. We show that an ideal ABC estimate can be obtained as a weighted average of a sequence of SMD modes, each being the minimizer of the deviations between the data and the model. This contrasts with the SMD, which is the mode of the average deviations. Using stochastic expansions, we provide a general characterization of frequentist estimators and those based on Bayesian computations including Laplace-type estimators. Their differences are illustrated using analytical examples and a simulation study of the dynamic panel model.

JEL Classification: C22, C23.

Keywords: Indirect Inference, Simulated Method of Moments, Efficient Method of Moments, Laplace Type Estimator.

*Department of Economics, Columbia University. Email: jmf2209@columbia.edu

†Department of Economics, Columbia University, and NBER. Email Serena.Ng at Columbia.edu.
Correspondence Address: 420 W. 118 St. Room 1117, New York, NY 10025.
Financial support is provided by the National Science Foundation (SES-0962431 and SES-1558623). We thank Richard Davis for discussions that initiated this research, Neil Shephard, Christopher Drovandi, two anonymous referees, and the editors for many helpful suggestions. Comments from seminar participants at Columbia, Harvard/MIT, UPenn, and Wisconsin are greatly appreciated. All errors are our own.

1 Introduction

As knowledge accumulates, scientists and social scientists incorporate more and more features into their models to have a better representation of the data. The increased model complexity comes at a cost; the conventional approach of estimating a model by writing down its likelihood function is often not possible. Different disciplines have developed different ways of handling models with an intractable likelihood. An approach popular amongst evolutionary biologists, geneticists, ecologists, psychologists and statisticians is Approximate Bayesian Computation (ABC). This work is largely unknown to economists who mostly estimate complex models using frequentist methods that we generically refer to as the method of Simulated Minimum Distance (SMD), and which include such estimators as Simulated Method of Moments, Indirect Inference, or Efficient Methods of Moments.[1]

The ABC and SMD share the same goal of estimating parameters θ using auxiliary statistics $\widehat{\psi}$ that are informative about the data. An SMD estimator minimizes the L_2 distance between $\widehat{\psi}$ and an average of the auxiliary statistics simulated under θ, and this distance can be made as close to zero as machine precision permits. An ABC estimator evaluates the distance between $\widehat{\psi}$ and the auxiliary statistics simulated for each θ drawn from a proposal distribution. The posterior mean is then a weighted average of the draws that satisfy a distance threshold of $\delta > 0$. There are many ABC algorithms, each differing according to the choice of the distance metric, the weights, and sampling scheme. But the algorithms can only approximate the desired posterior distribution because δ cannot be zero, or even too close to zero, in practice.

While both SMD and ABC use simulations to match $\psi(\theta)$ to $\widehat{\psi}$ (hence likelihood-free), the relation between them is not well understood beyond the fact that they are asymptotically equivalent under some high level conditions. To make progress, we focus on the MCMC-ABC algorithm due to Marjoram et al. (2003). The algorithm applies uniform weights to those θ satisfying $\|\widehat{\psi} - \psi(\theta)\| \leq \delta$ and zero otherwise. Our main insight is that this δ can be made very close to zero if we combine optimization with Bayesian computations. In particular, the desired ABC posterior distribution can be targeted using a 'Reverse Sampler' (or RS for short) that applies importance weights to a sequence of SMD solutions. Hence, seen from the perspective of the RS, the ideal MCMC-ABC estimate with $\delta = 0$ is a weighted average of SMD modes. This offers a useful contrast with the SMD estimate, which is the mode of the average deviations between the model and the data. We then use stochastic expansions to study sources of variations in the two estimators in the case of exact identification. The differences are illustrated using simple analytical examples as well as simulations of the dynamic panel model.

Optimization of models with a non-smooth objective function is challenging, even when the

[1] Indirect Inference is due to Gouriéroux et al. (1993), the Simulated Method of moments is due to Duffie and Singleton (1993), and the Efficient Method of Moments is due to Gallant and Tauchen (1996).

model is not complex. The Quasi-Bayes (LT) approach due to Chernozhukov and Hong (2003) use Bayesian computations to approximate the mode of a likelihood-free objective function. Its validity rests on the Laplace (asymptotic normal) approximation of the posterior distribution with the goal of valid asymptotic frequentist inference. The simulation analog of the LT (which we call SLT) further uses simulations to approximate the intractable relation between the model and the data. We show that both the LT and SLT can also be represented as a weighted average of modes with appropriately defined importance weights.

A central theme of our analysis is that the mean computed from many likelihood-free posterior distributions can be seen as a weighted average of solutions to frequentist objective functions. Optimization permits us to turn the focus from computational to analytical aspects of the posterior mean, and to provide a bridge between the seemingly related approaches. Although our optimization-based samplers are not intended to compete with the many ABC algorithms that are available, they can be useful in situations when numerical optimization of the auxiliary model is fast. This aspect is studied in our companion paper Forneron and Ng (2016) in which implementation of the RS in the overidentified case is also considered. The RS is independently proposed in Meeds and Welling (2015) with emphasis on efficient and parallel implementations. Our focus on the analytical properties complements their analysis.

The paper proceeds as follows. After laying out the preliminaries in Section 2, Section 3 presents the general idea behind ABC and introduces an optimization view of the ideal MCMC-ABC. Section 4 considers Quasi-Bayes estimators and interprets them from an optimization perspective. Section 5 uses stochastic expansions to study the properties of the estimators. Section 6 uses analytical examples and simulations to illustrate their differences. Throughout, we focus the discussion on features that distinguish the SMD from ABC which are lesser known to economists.[2]

2 Preliminaries

As a matter of notation, we use $L(\cdot)$ to denote the likelihood, $p(\cdot)$ to denote posterior densities, $q(\cdot)$ for proposal densities, and $\pi(\cdot)$ to denote prior densities. A 'hat' denotes estimators that correspond to the mode and a 'bar' is used for estimators that correspond to the posterior mean. We use (s, S) and (b, B) to denote the (specific, total number of) draws in frequentist and Bayesian type analyses respectively. A superscript s denotes a specific draw and a subscript S denotes the average over S draws. For a function $f(\theta)$, we use $f_\theta(\theta_0)$ to denote $\frac{\partial}{\partial\theta} f(\theta)$ evaluated at θ_0, $f_{\theta\theta_j}(\theta_0)$ to denote

[2]The class of SMD estimators considered are well known in the macro and finance literature and with apologies, many references are omitted. We also do not consider discrete choice models; though the idea is conceptually similar, the implementation requires different analytical tools. Smith (2008) provides a concise overview of these methods. The finite sample properties of the estimators are studied in Michaelides and Ng (2000). Readers are referred to the original paper concerning the assumptions used.

$\frac{\partial}{\partial\theta_j} f_\theta(\theta)$ evaluated at θ_0 and $f_{\theta,\theta_j,\theta_k}(\theta_0)$ to denote $\frac{\partial^2}{\partial\theta_j\theta_k} f_\theta(\theta)$ evaluated at θ_0.

Throughout, we assume that the data $\mathbf{y} = (y_1, \ldots, y_T)'$ are covariance stationary and can be represented by a parametric model with probability measure $\mathcal{P}_\theta$ where $\theta \in \Theta \subset \mathbb{R}^K$. The true value of θ is denoted by θ_0. Unless otherwise stated, we write $\mathbb{E}[\cdot]$ for expectations taken under P_{θ_0} instead of $\mathbb{E}_{\mathcal{P}_{\theta_0}}[\cdot]$. If the likelihood $L(\theta) = L(\theta|\mathbf{y})$ is tractable, maximizing the log-likelihood $\ell(\theta) = \log L(\theta)$ with respect to θ gives

$$\widehat{\theta}_{ML} = \text{argmax}_\theta \ell(\theta).$$

Bayesian estimation combines the likelihood with a prior $\pi(\theta)$ to yield the posterior density

$$p(\theta|\mathbf{y}) = \frac{L(\theta) \cdot \pi(\theta)}{\int_\Theta L(\theta)\pi(\theta) d\theta}. \tag{1}$$

For any prior $\pi(\theta)$, it is known that $\widehat{\theta}_{ML}$ solves $\text{argmax}_\theta \ell(\theta) = \lim_{\lambda\to\infty} \frac{\int_\Theta \theta \exp(\lambda\ell(\theta))\pi(\theta)d\theta}{\int_\Theta \exp(\lambda\ell(\theta))\pi(\theta)d\theta}$. That is, the maximum likelihood estimator is a limit of the Bayes estimator using $\lambda \to \infty$ replications of the data $\mathbf{y}$.[3] The parameter λ is the cooling temperature in simulated annealing, a stochastic optimizer due to Kirkpatrick et al. (1983) for handling problems with multiple modes.

In the case of conjugate problems, the posterior distribution has a parametric form which makes it easy to compute the posterior mean and other quantities of interest. For non-conjugate problems, the method of Monte-Carlo Markov Chain (MCMC) allows sampling from a Markov Chain whose ergodic distribution is the target posterior distribution $p(\theta|\mathbf{y})$, and without the need to compute the normalizing constant. We use the Metropolis-Hastings (MH) algorithm in subsequent discussion. In classical Bayesian estimation with proposal density $q(\cdot)$, the acceptance ratio is

$$\rho_{BC}(\theta^b, \theta^{b+1}) = \min\Big(\frac{L(\theta^{b+1})\pi(\theta^{b+1})q(\theta^b|\theta^{b+1})}{L(\theta^b)\pi(\theta^b)q(\theta^{b+1}|\theta^b)}, 1\Big).$$

When the posterior mode $\widehat{\theta}_{BC} = \text{argmax}_\theta p(\theta|y)$ is difficult to obtain, the posterior mean

$$\overline{\theta}_{BC} \quad = \quad \frac{1}{B}\sum_{b=1}^{B} \theta^b \approx \int_\Theta \theta p(\theta|y) d\theta$$

is often the reported estimate, where θ^b are draws from the Markov Chain upon convergence. Under quadratic loss, the posterior mean minimizes the posterior risk $Q(a) = \int_\Theta |\theta - a|^2 p(\theta|\mathbf{y}) d\theta$.

2.1 Minimum Distance Estimators

The method of generalized method of moments (GMM) is a likelihood-free frequentist estimator developed in Hansen (1982); Hansen and Singleton (1982). For example, it allows for the estimation

[3] See Robert and Casella (2004, Corollary 5.11), Jacquier et al. (2007).

of K parameters in a dynamic model without explicitly solving the full model. It is based on a vector of $L \geq K$ moment conditions $g_t(\theta)$ whose expected value is zero at $\theta = \theta_0$, i.e. $\mathbb{E}[g_t(\theta_0)] = 0$. Let $\overline{g}(\theta) = \frac{1}{T}\sum_{t=1}^T g_t(\theta)$ be the sample analog of $\mathbb{E}[g_t(\theta)]$. The estimator is

$$\widehat{\theta}_{GMM} \quad = \quad \text{argmin}_\theta J(\theta), \qquad J(\theta) = \frac{T}{2} \cdot \overline{g}(\theta)' W \overline{g}(\theta) \tag{2}$$

where W is a $L \times L$ positive-definite weighting matrix. Most estimators can be put in the GMM framework with suitable choice of g_t. For example, when g_t is the score of the likelihood, the maximum likelihood estimator is obtained.

Let $\widehat{\psi} \equiv \widehat{\psi}(\mathbf{y}(\theta_0))$ be L auxiliary statistics with the property that $\sqrt{T}(\widehat{\psi} - \psi(\theta_0)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. It is assumed that the mapping $\psi(\theta) = \lim_{T\to\infty} \mathbb{E}[\widehat{\psi}(\theta)]$ is continuously differentiable in θ and locally injective at θ_0. Gouriéroux et al. (1993) refer to $\psi(\theta)$ as the *binding function* while Jiang and Turnbull (2004) use the term *bridge function*. The minimum distance estimator is a GMM estimator which specifies

$$\overline{g}(\theta) = \widehat{\psi} - \psi(\theta),$$

with efficient weighting matrix $W = \Sigma^{-1}$. Classical MD estimation assumes that the binding function $\psi(\theta)$ has a closed form expression so that in the exactly identified case, one can solve for θ by inverting $\overline{g}(\theta)$.

2.2 SMD Estimators

Simulation estimation is useful when the asymptotic binding function $\psi(\theta_0)$ is not analytically tractable but can be easily evaluated on simulated data. The first use of this approach in economics appears to be due to Smith (1993). The simulated analog of MD, which we will call SMD, minimizes the weighted difference between the auxiliary statistics evaluated at the observed and simulated data:

$$\widehat{\theta}_{SMD} \quad = \quad \text{argmin}_\theta J_S(\theta) = \text{argmin}_\theta \overline{g}'_S(\theta) W \overline{g}_S(\theta).$$

where

$$\overline{g}_S(\theta) = \widehat{\psi} - \frac{1}{S}\sum_{s=1}^S \widehat{\psi}^s(\mathbf{y}^s(\theta)),$$

$\mathbf{y}^s(\theta) \equiv \mathbf{y}^s(\varepsilon^s, \theta)$ are data simulated under θ with errors ε^s drawn from an assumed distribution F_ε, and $\widehat{\psi}^s(\theta) \equiv \widehat{\psi}^s(\mathbf{y}^s(\varepsilon^s, \theta))$ are the auxiliary statistics computed using $\mathbf{y}^s(\theta)$. Of course, $\overline{g}_S(\theta)$ is also the average over S deviations between $\widehat{\psi}$ and $\widehat{\psi}^s(\mathbf{y}^s(\theta))$. To simplify notation, we will write $\mathbf{y}^s$ and $\widehat{\psi}^s(\theta)$ when the context is clear. As in MD estimation, the auxiliary statistics $\psi(\theta)$ should 'smoothly embed' the properties of the data in the terminology of Gallant and Tauchen (1996). But SMD estimators replace the asymptotic binding function $\psi(\theta_0) = \lim_{T\to\infty} \mathbb{E}[\widehat{\psi}(\theta_0)]$ by a finite sample analog using Monte-Carlo simulations. While the SMD is motivated with the estimation of complex models in mind, Gouriéroux et al. (1999) show that simulation estimation has a bias reduction effect like the bootstrap. Hence in the econometrics literature, SMD estimators are used even when the likelihood is tractable, as in Gouriéroux et al. (2010).

The steps for implementing the SMD are as follows:

0 For $s = 1, \ldots, S$, draw $\varepsilon^s = (\varepsilon_1^s, \ldots, \varepsilon_T^s)'$ from F_ε. These are innovations to the structural model that will be held fixed during iterations.

1 Given θ, repeat for $s = 1, \ldots S$:

   a Use (ε^s, θ) and the model to simulate data $\mathbf{y}^s = (y_1^s, \ldots, y_T^s)'$.

   b Compute the auxiliary statistics $\widehat{\psi}^s(\theta)$ using simulated data $\mathbf{y}^s$.

2 Compute: $\overline{g}_S(\theta) = \widehat{\psi}(\mathbf{y}) - \frac{1}{S}\sum_{s=1}^S \widehat{\psi}^s(\theta)$. Minimize $J_S(\theta) = \overline{g}_S(\theta)' W \overline{g}_S(\theta)$.

The SMD is the θ that makes $J_S(\theta)$ smaller than the tolerance specified for the numerical optimizer. In the exactly identified case, the tolerance can be made as small as machine precision permits. When $\widehat{\psi}$ is a vector of unconditional moments, the SMM estimator of Duffie and Singleton (1993) is obtained. When $\widehat{\psi}$ are parameters of an auxiliary model, we have the 'indirect inference' estimator of Gouriéroux et al. (1993). These are Wald-test based SMD estimators in the terminology of Smith (2008). When $\widehat{\psi}$ is the score function associated with the likelihood of the auxiliary model, we have the EMM estimator of Gallant and Tauchen (1996), which can also be thought of as an LM-test based SMD. If $\widehat{\psi}$ is the likelihood of the auxiliary model, $J_S(\theta)$ can be interpreted as a likelihood ratio and we have a LR-test based SMD. Gouriéroux and Monfort (1996) provide a framework that unifies these three approaches to SMD estimation. Nickl and Potscher (2010) show that an SMD based on non-parametrically estimated auxiliary statistics can have asymptotic variance equal to the Cramer-Rao bound if the tuning parameters are optimally chosen.

The Wald, LM, and LR based SMD estimators minimize a weighted L_2 distance between the data and the model as summarized by auxiliary statistics. Creel and Kristensen (2013) consider a class of estimators that minimize the Kullback-Leibler distance between the model and the data.[4] Within this class, their MIL estimator maximizes an 'indirect likelihood', defined as the likelihood of the auxiliary statistics. Their BIL estimator uses Bayesian computations to approximate the mode of the indirect likelihood. In practice, the indirect likelihood is unknown. Estimating it by kernel smoothing of the simulated statistics, the SBIL estimator combines Bayesian computations with non-parametric estimation. Gao and Hong (2014) show that using local linear regressions

[4] In the sequel, we take the more conventional L_2 definition of SMD as given above.

instead of kernel estimation can reduce the variance and the bias. Using non-parametric estimation in ABC has previously been considered in Beaumont et al. (2002). Creel et al. (2016) show that not only can such an ABC implementation bypass MCMC altogether, it can provide asymptotically valid frequentist inference. Bounds for the number of simulations that achieve the parametric rate of convergence and asymptotic normality are derived.

3 Approximate Bayesian Computations

The ABC literature often credits Donald Rubin to be the first to consider the possibility of estimating the posterior distribution when the likelihood is intractable. Diggle and Gratton (1984) propose to approximate the likelihood by simulating the model at each point on a parameter grid and appear to be the first implementation of simulation estimation for models with intractable likelihoods. Subsequent developments adapted the idea to conduct posterior inference, giving the prior an explicit role. The first ABC algorithm was implemented by Tavare et al. (1997) and Pritchard et al. (1996) to study population genetics. Their Accept/Reject algorithm is as follows: (i) draw θ^b from the prior distribution $\pi(\theta)$, (ii) simulate data using the model under θ^b (iii) accept θ^b if the auxiliary statistics computed using the simulated data are close to $\widehat{\psi}$. As in the SMD literature, the auxiliary statistics can be parameters of a regression or unconditional sample moments. Heggland and Frigessi (2004), Drovandi et al. (2011, 2015) use simulated auxiliary statistics.

Since simulating from a non-informative prior distribution is inefficient, subsequent work suggests to replace the rejection sampler by one that takes into account the features of the posterior distribution. The likelihood of the full dataset $L(y|\theta)$ is intractable, as is the likelihood of the finite dimensional statistic $L(\widehat{\psi}|\theta)$. However, the latter can be consistently estimated using simulations. The general idea is to set as a target the intractable posterior density

$$p^*_{ABC}(\theta|\widehat{\psi}) \propto \pi(\theta)L(\widehat{\psi}|\theta)$$

and approximate it using Monte-Carlo methods. Some algorithms are motivated from the perspective of non-parametric density estimation, while others aim to improve properties of the Markov chain.[5] The main idea is, however, using data augmentation to consider the joint density $p_{ABC}(\theta, x|\widehat{\psi}) \propto L(\widehat{\psi}|x, \theta)L(x|\theta)\pi(\theta)$, putting more weight on the draws with x close to $\widehat{\psi}$. When $x = \widehat{\psi}$, $L(\widehat{\psi}|\widehat{\psi}, \theta)$ is a constant, $p_{ABC}(\theta, \widehat{\psi}|\widehat{\psi}) \propto L(\widehat{\psi}|\theta)\pi(\theta)$, and the target posterior is recovered. If $\widehat{\psi}$ are sufficient statistics, one recovers the posterior distribution associated with the intractable likelihood $L(\theta|y)$, not just an approximation.

[5] Recent surveys on ABC can be found in Marin et al. (2012), Blum et al. (2013) among others. See Drovandi et al. (2015, 2011) for differences amongst ABC estimators.

To better understand the ABC idea and its implementation, we will write $\mathbf{y}^b$ instead of $\mathbf{y}^b(\varepsilon^b, \theta^b)$ and $\widehat{\psi}^b$ instead of $\widehat{\psi}^b(\mathbf{y}^b(\varepsilon^b, \theta^b))$ to simplify notation. Let $\mathbb{K}_\delta(\widehat{\psi}^b, \widehat{\psi}|\theta) \geq 0$ be a kernel function that weighs deviations between $\widehat{\psi}$ and $\widehat{\psi}^b$ over a window of width δ. Suppose we keep only the draws that satisfy $\widehat{\psi}^b = \widehat{\psi}$ and hence $\delta = 0$. Note that $\mathbb{K}_0(\widehat{\psi}^b, \widehat{\psi}|\theta) = 1$ if $\widehat{\psi} = \widehat{\psi}^b$ for any choice of the kernel function. Once the likelihood of interest

$$L(\widehat{\psi}|\theta) = \int L(x|\theta)\mathbb{K}_0(x, \widehat{\psi}|\theta)dx$$

is available, moments and quantiles can be computed. In particular, for any measurable function φ whose expectation exists, we have:

$$\mathbb{E}\left[\varphi(\theta)|\widehat{\psi} = \widehat{\psi}^b\right] = \frac{\int_\Theta \varphi(\theta^b)\pi(\theta)L(\widehat{\psi}|\theta^b)d\theta^b}{\int_\Theta \pi(\theta^b)L(\widehat{\psi}|\theta^b)d\theta^b} = \frac{\int_\Theta \int \varphi(\theta^b)\pi(\theta^b)L(x|\theta^b)\mathbb{K}_0(x, \widehat{\psi}|\theta^b)dxd\theta^b}{\int_\Theta \int \pi(\theta^b)L(x|\theta^b)\mathbb{K}_0(x, \widehat{\psi}|\theta^b)dxd\theta^b}.$$

Since $\widehat{\psi}^b|\theta^b \sim L(\cdot|\theta^b)$, the expectation can be approximated by averaging over draws from $L(\cdot|\widehat{\theta}^b)$. More generally, draws can be taken from an importance density $q(\cdot)$. In particular,

$$\widehat{\mathbb{E}}\left[\varphi(\theta)|\widehat{\psi} = \widehat{\psi}^b\right] = \frac{\sum_{b=1}^B \varphi(\theta^b)\mathbb{K}_0(\widehat{\psi}^b, \widehat{\psi}|\theta^b)\frac{\pi(\theta^b)}{q(\theta^b)}}{\sum_{b=1}^B \mathbb{K}_0(\widehat{\psi}^b, \widehat{\psi}|\theta^b)\frac{\pi(\theta^b)}{q(\theta^b)}}.$$

The importance weights are then

$$w_0^b \propto \mathbb{K}_0(\widehat{\psi}^b, \widehat{\psi}|\theta^b)\frac{\pi(\theta^b)}{q(\theta^b)}.$$

By a law of large numbers, $\widehat{\mathbb{E}}\left[\varphi(\theta)|\widehat{\psi}\right] \to \mathbb{E}\left[\varphi(\theta)|\widehat{\psi}\right]$ as $B \to \infty$.

There is, however, a caveat. When $\widehat{\psi}$ has continuous support, $\widehat{\psi}^b = \widehat{\psi}$ is an event of measure zero. Replacing $\mathbb{K}_0$ with $\mathbb{K}_\delta$ where δ is close to zero yields the approximation:

$$\mathbb{E}\left[\varphi(\theta)|\widehat{\psi} = \widehat{\psi}^b\right] \approx \frac{\int_\Theta \int \varphi(\theta^b)\pi(\theta^b)L(x|\theta^b)\mathbb{K}_\delta(x, \widehat{\psi}|\theta^b)dxd\theta^b}{\int_\Theta \int \pi(\theta^b)L(x|\theta^b)\mathbb{K}_\delta(x, \widehat{\psi}|\theta^b)dxd\theta^b}.$$

Since $\mathbb{K}_\delta(\cdot)$ is a kernel function, consistency of the non-parametric estimator for the conditional expectation of $\varphi(\theta)$ follows from, for example, Pagan and Ullah (1999). This is the approach considered in Beaumont et al. (2002), Creel and Kristensen (2013) and Gao and Hong (2014). The case of a rectangular kernel $\mathbb{K}_\delta(\widehat{\psi}, \widehat{\psi}^b) = \mathbb{I}_{\|\widehat{\psi}-\widehat{\psi}^b\| \leq \delta}$ corresponds to the ABC algorithm proposed in Marjoram et al. (2003). This is the first ABC algorithm that exploits MCMC sampling. Hence we refer to it as MCMC-ABC. Our analysis to follow is based on this algorithm. Accordingly, we now explore it in more detail.

Algorithm MCMC-ABC Let $q(\cdot)$ be the proposal distribution. For $b = 1, \dots, B$ with θ^0 given,

1 Generate $\theta^{b+1} \sim q(\theta^{b+1}|\theta^b)$.

2 Draw ε^{b+1} from F_ε and simulate data $\mathbf{y}^{b+1}$. Compute $\widehat{\psi}^{b+1}$.

3 Accept θ^{b+1} with probability $\rho_{\text{ABC}}(\theta^b, \theta^{b+1})$ and set it equal to θ^b with probability $1 - \rho_{\text{ABC}}(\theta^b, \theta^{b+1})$ where

$$\rho_{\text{ABC}}(\theta^b, \theta^{b+1}) = \min\left(\mathbb{I}_{\|\widehat{\psi}-\widehat{\psi}^{b+1}\|\le\delta} \frac{\pi(\theta^{b+1})q(\theta^b|\theta^{b+1})}{\pi(\theta^b)q(\theta^{b+1}|\theta^b)}, 1\right). \tag{3}$$

As with all ABC algorithms, the success of the MCMC-ABC lies in augmenting the posterior with simulated data $\widehat{\psi}^b$, i.e. $p^*_{ABC}(\theta^b, \widehat{\psi}^b|\widehat{\psi}) \propto L(\widehat{\psi}|\theta^b, \widehat{\psi}^b)L(\widehat{\psi}^b|\theta^b)\pi(\theta^b)$. The joint posterior distribution that the MCMC-ABC would like to target is

$$p^0_{\text{ABC}}\left(\theta^b, \widehat{\psi}^b|\widehat{\psi}\right) \propto \pi(\theta^b)L(\widehat{\psi}^b|\theta^b)\mathbb{I}_{\|\widehat{\psi}^b-\widehat{\psi}\|=0}$$

since integrating out ε^b would yield $p^*_{ABC}(\theta|\widehat{\psi})$. But it would not be possible to generate draws such that $\|\widehat{\psi}^b - \widehat{\psi}\|$ equals zero exactly. Hence as a compromise, the MCMC-ABC algorithm allows $\delta > 0$ and targets

$$p^\delta_{\text{ABC}}\left(\theta^b, \widehat{\psi}^b|\widehat{\psi}\right) \propto \pi(\theta^b)L(\widehat{\psi}^b|\theta^b)\mathbb{I}_{\|\widehat{\psi}^b-\widehat{\psi}\|\le\delta}.$$

The adequacy of p^δ_{ABC} as an approximation of p^0_{ABC} is a function of the tuning parameter δ.

To understand why this algorithm works, we follow the argument in Sisson and Fan (2011). If the initial draw θ^1 satisfies $\|\widehat{\psi} - \widehat{\psi}^1\| \le \delta$, then all subsequent $b > 1$ draws are such that $\mathbb{I}_{\|\widehat{\psi}^b-\widehat{\psi}\|\le\delta} = 1$ by construction. Furthermore, since we draw θ^{b+1} and then independently simulate data $\widehat{\psi}^{b+1}$, the proposal distribution becomes $q(\theta^{b+1}, \widehat{\psi}^{b+1}|\theta^b) = q(\theta^{b+1}|\theta^b)L(\widehat{\psi}^{b+1}|\theta^{b+1})$. The two observations together imply that

$$\begin{aligned}
\mathbb{I}_{\|\widehat{\psi}-\widehat{\psi}^{b+1}\|\le\delta} \frac{\pi(\theta^{b+1})q(\theta^b|\theta^{b+1})}{\pi(\theta^b)q(\theta^{b+1}|\theta^b)} &= \frac{\mathbb{I}_{\|\widehat{\psi}-\widehat{\psi}^{b+1}\|\le\delta}}{\mathbb{I}_{\|\widehat{\psi}-\widehat{\psi}^{b}\|\le\delta}} \frac{\pi(\theta^{b+1})q(\theta^b|\theta^{b+1})}{\pi(\theta^b)q(\theta^{b+1}|\theta^b)} \frac{L(\widehat{\psi}^{b+1}|\theta^{b+1})}{L(\widehat{\psi}^b|\theta^b)} \frac{L(\widehat{\psi}^b|\theta^b)}{L(\widehat{\psi}^{b+1}|\theta^{b+1})} \\
&= \frac{\mathbb{I}_{\|\widehat{\psi}-\widehat{\psi}^{b+1}\|\le\delta}}{\mathbb{I}_{\|\widehat{\psi}-\widehat{\psi}^{b}\|\le\delta}} \frac{\pi(\theta^{b+1})L(\widehat{\psi}^{b+1}|\theta^{b+1})}{\pi(\theta^b)L(\widehat{\psi}^b|\theta^b)} \frac{q(\theta^b|\theta^{b+1})L(\widehat{\psi}^b|\theta^b)}{q(\theta^{b+1}|\theta^b)L(\widehat{\psi}^{b+1}|\theta^{b+1})} \\
&= \frac{p^\delta_{\text{ABC}}\left(\theta^{b+1}, \widehat{\psi}^{b+1}|\widehat{\psi}\right)}{p^\delta_{\text{ABC}}\left(\theta^b, \widehat{\psi}^b|\widehat{\psi}\right)} \frac{q(\theta^b, \widehat{\psi}^b|\theta^{b+1})}{q(\theta^{b+1}, \widehat{\psi}^{b+1}|\theta^b)}.
\end{aligned}$$

The last equality shows that the acceptance ratio is in fact the ratio of two ABC posteriors times the ratio of the proposal distribution. Hence the MCMC-ABC effectively targets the joint posterior distribution p^δ_{ABC}.

3.1 The Reverse Sampler

Thus far, we have seen that the SMD estimator is the θ that makes $\|\widehat{\psi} - \frac{1}{S}\sum_{s=1}^{S}\widehat{\psi}^s(\theta)\|$ no larger than the tolerance of the numerical optimizer. We have also seen that the feasible MCMC-ABC accepts draws θ^b satisfying $\|\widehat{\psi} - \widehat{\psi}^b(\theta^b)\| \leq \delta$ with $\delta > 0$. To view the MCMC-ABC from a different perspective, suppose that setting $\delta = 0$ was possible. Then each accepted draw θ^b would satisfy:

$$\widehat{\psi}^b(\theta^b) = \widehat{\psi}.$$

For fixed ε^b and assuming that the mapping $\widehat{\psi}^b : \theta \to \widehat{\psi}^b(\theta)$ is continuously differentiable and one-to-one, the above statement is equivalent to:

$$\theta^b = \text{argmin}_\theta \left(\widehat{\psi}^b(\theta) - \widehat{\psi}\right)' \left(\widehat{\psi}^b(\theta) - \widehat{\psi}\right).$$

Hence each accepted θ^b is the solution to a SMD problem with $S = 1$. Next, suppose that instead of drawing θ^b from a proposal distribution, we draw ε^b and solve for θ^b as above. Since the mapping $\widehat{\psi}^b$ is invertible by assumption, a change of variable yields the relation between the distribution of $\widehat{\psi}^b$ and θ^b. In particular, the joint density, say $h(\theta^b, \varepsilon^b)$, is related to the joint density $L(\widehat{\psi}^b(\theta^b), \varepsilon^b)$ via the determinant of the Jacobian $|\widehat{\psi}^b_\theta(\theta^b)|$ as follows:

$$h(\theta^b, \varepsilon^b|\widehat{\psi}) = |\widehat{\psi}^b_\theta(\theta^b)| L(\widehat{\psi}^b(\theta^b), \varepsilon^b|\widehat{\psi}).$$

Multiplying the quantity on the right-hand-side by $w^b(\theta^b) = \pi(\theta^b)|\widehat{\psi}^b_\theta(\theta^b)|^{-1}$ yields $\pi(\theta^b)L(\widehat{\psi}, \varepsilon^b|\theta^b)$ since $\widehat{\psi}^b(\theta^b) = \widehat{\psi}$ and the mapping from θ^b to $\psi^b(\theta^b)$ is one-to-one. This suggests that if we solve the SMD problem B times each with $S = 1$, re-weighting each of the B solutions by $w^b(\theta^b)$ would give the target the joint posterior $p^*_{ABC}(\theta|\widehat{\psi})$ after integrating out ε^b.

Algorithm RS

1 For $b = 1, \ldots, B$ and a given θ,

- i Draw ε^b from F_ε and simulate data $\mathbf{y}^b$ using θ. Compute $\widehat{\psi}^b(\theta)$ from $\mathbf{y}^b$.
- ii Let $\theta^b = \text{argmin}_\theta J_1^b(\theta)$, $J_1^b(\theta) = (\widehat{\psi} - \widehat{\psi}^b(\theta))' W (\widehat{\psi} - \widehat{\psi}^b(\theta))$.
- iii Compute the Jacobian $\widehat{\psi}^b_\theta(\theta^b)$ and its determinant $|\widehat{\psi}^b_\theta(\theta^b)|$. Let $w^b(\theta^b) = \pi(\theta^b)|\widehat{\psi}^b_\theta(\theta^b)|^{-1}$.

2 Compute the posterior mean $\overline{\theta}_{RS} = \sum_{b=1}^{B} \overline{w}^b(\theta^b)\theta^b$ where $\overline{w}^b(\theta^b) = \frac{w^b(\theta^b)}{\sum_{c=1}^{B} w^c(\theta^c)}$.

The RS has the optimization aspect of SMD as well as the sampling aspect of the MCMC-ABC. We call the RS the reverse sampler for two reasons. First, typical Bayesian estimation starts with an evaluation of the prior probabilities. The RS terminates with the evaluation of the prior. Furthermore, we use the SMD estimates to reverse engineer the posterior distribution.

Consistency of each RS solution (i.e. θ^b) is built on the fact that the SMD is consistent even with $S = 1$. The RS estimate is thus an average of a sequence of SMD modes. In contrast, the SMD is the mode of an objective function defined from a weighted average of the simulated auxiliary statistics. Optimization effectively allows δ to be as close to zero as machine precision permits. This puts the joint posterior distribution as close to the infeasible target as possible, but has the consequence of shifting the distribution from $(\mathbf{y}^b, \widehat{\psi}^b)$ to $(\mathbf{y}^b, \theta^b)$. Hence a change of variable is required. The importance weight depends on the Jacobian matrix, making the RS an optimization based importance sampler.

Lemma 1 *Suppose that $\psi : \theta \to \widehat{\psi}^b(\theta)$ is one-to-one and $\psi^b_\theta(\theta)$ has full column rank. The posterior distribution produced by the reverse sampler converges to the infeasible posterior distribution $p^*_{ABC}(\theta|\widehat{\psi})$ as $B \to \infty$.*

The proof is given in Forneron and Ng (2016). By convergence, we mean that for any measurable function $\varphi(\theta)$ such that the expectation exists, a law of large numbers implies that $\sum_{b=1}^B \overline{w}^b(\theta^b)\varphi(\theta^b) \xrightarrow{a.s.} \mathbb{E}_{p^*(\theta|\widehat{\psi})}(\varphi(\theta))$. In general, $\overline{w}^b(\theta^b) \neq \frac{1}{B}$. The RS draws and moments can be interpreted as if they were taken from p^*_{ABC}, the posterior distribution had the likelihood $p(\widehat{\psi}|\theta)$ been available.

That the draws of the MCMC-ABC at $\delta = 0$ can be seen from an optimization perspective allows us to subsequently use the RS as a conceptual framework to understand the differences between the ideal MCMC-ABC and SMD. It should be noted that the RS is not the same as the MCMC-ABC or any ABC estimator implemented with $\delta > 0$ as they necessarily have an acceptance rate strictly less than one. Indeed, a challenge of many ABC implementations is the low acceptance rate. The RS draws are always accepted and can be useful in situations when numerical optimization of the auxiliary model is easy. Properties of the RS are further analyzed in Forneron and Ng (2016). Meeds and Welling (2015) independently propose an ABC sampling algorithm similar to the RS. Their focus is on ways to implement it efficiently using embarrassingly parallel methods.

4 Quasi-Bayes Estimators

The GMM objective function $J(\theta)$ defined in (2) is not a proper density. Noting that $\exp(-J(\theta))$ is the kernel of the Gaussian density, Jiang and Turnbull (2004) define an *indirect likelihood* (distinct from the one defined in Creel and Kristensen (2013)) as

$$L_{IND}(\theta|\widehat{\psi}) \equiv \frac{1}{\sqrt{2\pi}}|\Sigma|^{-1}\exp(-J(\theta)).$$

Associated with the indirect likelihood is the indirect score, indirect Hessian, and a generalized information matrix equality, just like a conventional likelihood. Though the indirect likelihood is not a proper density, its maximizer has properties analogous to the maximum likelihood estimator provided by $\mathbb{E}[g_t(\theta_0)] = 0$.

In Chernozhukov and Hong (2003), the authors observe that extremum estimators can be difficult to compute if the objective function is highly non-convex, especially when the dimension of the parameter space is large. These difficulties can be alleviated by using Bayesian computational tools, but this is not possible when the objective function is not a likelihood. Chernozhukov and Hong (2003) take an exponential of $-J(\theta)$, as in Jiang and Turnbull (2004), but then combine $\exp(-J(\theta))$ with a prior density $\pi(\theta)$ to produce a quasi-posterior density. Chernozhukov and Hong initially termed their estimator 'Quasi-Bayes' because $\exp(-J(\theta))$ is not a standard likelihood. They settled on the term 'Laplace-type estimator' (LT), so-called because Laplace suggested to approximate a smooth pdf with a well defined peak by a normal density, see Tierney and Kadane (1986). If $\pi(\theta)$ is strictly positive and continuous over a compact parameter space Θ, the 'quasi-posterior' LT distribution

$$p_{LT}(\theta|\mathbf{y}) = \frac{\exp(-J(\theta))\pi(\theta)}{\int_\Theta \exp(-J(\theta)\pi(\theta))d\theta} \propto \exp(-J(\theta))\pi(\theta) \tag{4}$$

is proper. The LT posterior mean is thus well-defined even when the prior may not be proper. As discussed in Chernozhukov and Hong (2003), one can think of the LT under a flat prior as using simulated annealing to maximize $\exp(-J(\theta))$ and setting the cooling parameter τ to 1. Frequentist inference is asymptotically valid because as the sample size increases, the prior is dominated by the pseudo likelihood which, by the Laplace approximation, is asymptotically normal.[6]

In practice, the LT posterior distribution is targeted using MCMC methods. Upon replacing the likelihood $L(\theta)$ by $\exp(-J(\theta))$, the MH acceptance probability is

$$\rho_{LT}(\theta^b, \vartheta) = \min\Big(\frac{\exp(-J(\vartheta))\pi(\vartheta)q(\theta^b|\vartheta)}{\exp(-J(\theta^b))\pi(\theta^b)q(\vartheta|\theta^b)}, 1\Big).$$

The quasi-posterior mean is $\overline{\theta}_{LT} = \frac{1}{B}\sum_{b=1}^B \theta^b$ where each θ^b is a draw from $p_{LT}(\theta|\mathbf{y})$. Chernozhukov and Hong suggest to exploit the fact that the quasi-posterior mean is much easier to compute than the mode and that, under regularity conditions, the two are first order equivalent. In practice, the weighting matrix can be based on some preliminary estimate of θ, or estimated simultaneously with θ. In exactly identified models, it is well known that the MD estimates do not depend on the choice of W. This continues to be the case for the LT posterior mode $\widehat{\theta}_{LT}$. However, the posterior mean is affected by the choice of the weighting matrix even in the just-identified case.[7]

The LT estimator is built on the validity of the asymptotic normal approximation in the second order expansion of the objective function. Nekipelov and Kormilitsina (2015) show that in small

[6]For loss function $d(\cdot)$, the LT estimator is $\widehat{\theta}_{LT}(\vartheta) = \text{argmin}_\theta \int_\Theta d(\theta - \vartheta)p_{LT}(\theta|\mathbf{y})d\theta$. If $d(\cdot)$ is quadratic, the posterior mean minimizes quasi-posterior risk.

[7]Kormiltsina and Nekipelov (2014) suggests to scale the objective function to improve coverage of the confidence intervals.

samples, this approximation can be poor so that the LT posterior mean may differ significantly from the extremum estimate that it is meant to approximate. To see the problem in a different light, we again take an optimization view. Specifically, the asymptotic distribution $\sqrt{T}(\widehat{\psi}(\theta_0) - \psi(\theta_0)) \xrightarrow{d} \mathcal{N}(0, \Sigma(\theta_0)) \equiv \mathbb{A}_\infty(\theta_0)$ suggests to use

$$\widehat{\psi}^b(\theta) \approx \psi(\theta) + \frac{\mathbb{A}^b_\infty(\theta_0)}{\sqrt{T}}$$

where $\mathbb{A}^b_\infty(\theta_0) \sim \mathcal{N}(0, \widehat{\Sigma}(\theta))$. Given a draw of $\mathbb{A}^b_\infty$, there will exist a θ^b such that $(\widehat{\psi}^b(\theta) - \widehat{\psi})' W (\widehat{\psi}^b(\theta) - \widehat{\psi})$ is minimized. In the exactly identified case, this discrepancy can be driven to zero up to machine precision. Hence we can define

$$\theta^b = \text{argmin}_\theta \|\widehat{\psi}^b(\theta) - \widehat{\psi}\|.$$

Arguments analogous to the RS suggest the following will produce draws of θ from $p_{LT}(\theta|\mathbf{y})$.

1 For $b = 1, \ldots$B:

  i Draw $\mathbb{A}^b_\infty(\theta_0)$ and define $\widehat{\psi}^b(\theta) = \psi(\theta) + \frac{\mathbb{A}^b_\infty(\theta)}{\sqrt{T}}$.

  ii Solve for θ^b such that $\widehat{\psi}^b(\theta^b) = \widehat{\psi}$ (up to machine precision).

  iii Compute $w^b(\theta^b) = |\widehat{\psi}^b_\theta(\theta^b)|^{-1}\pi(\theta^b)$.

2 Compute $\overline{\theta}_{LT} = \sum \overline{w}^b(\theta^b)\theta^b$, where $\overline{w}^b = \frac{w^b(\theta^b)}{\sum_{c=1}^B w^c(\theta^c)}$.

Seen from an optimization perspective, the LT is a weighted average of MD modes with the determinant of the Jacobian matrix as importance weight, similar to the RS. It differs from the RS in that the Jacobian here is computed from the asymptotic binding function $\psi(\theta)$, and the draws are based on the asymptotic normality of $\widehat{\psi}$. As such, simulation of the structural model is not required.

4.1 The SLT

When $\psi(\theta)$ is not analytically tractable, a natural modification is to approximate it by simulations as in the SMD. This is the approach taken in Lise et al. (2015). We refer to this estimator as the Simulated Laplace-type estimator, or SLT. The steps are as follows:

0 Draw structural innovations $\varepsilon^s = (\varepsilon^s_1, \ldots, \varepsilon^s_T)'$ from F_ε. These are held fixed across iterations.

1 For $b = 1, \ldots, B$, draw ϑ from $q(\vartheta|\theta^b)$.

  i. For $s = 1, \ldots S$: use $(\vartheta, \varepsilon^s)$ and the model to simulate data $\mathbf{y}^s = (\mathbf{y}^s_1, \ldots, \mathbf{y}^s_T)'$. Compute $\widehat{\psi}^s(\vartheta)$ using $\mathbf{y}^s$.

ii. Form $J_S(\vartheta) = \overline{g}_S(\vartheta)' W \overline{g}_S(\vartheta)$, where $\overline{g}_S(\vartheta) = \widehat{\psi}(\mathbf{y}) - \frac{1}{S}\sum_{s=1}^{S} \widehat{\psi}^s(\vartheta)$.

iii. Set $\theta^{b+1} = \vartheta$ with probability $\rho_{SLT}(\theta^b, \vartheta)$, else reset ϑ to θ^b with probability $1 - \rho_{SLT}$ where the acceptance probability is:

$$\rho_{SLT}(\theta^b, \vartheta) = \min\Big(\frac{\exp(-J_S(\vartheta))\pi(\vartheta)q(\theta^b|\vartheta)}{\exp(-J_S(\theta^b))\pi(\theta^b)q(\vartheta|\theta^b)}, 1\Big).$$

2 Compute $\overline{\theta}^b_{SLT} = \frac{1}{B}\sum_{b=1}^{B} \theta^b$.

The SLT algorithm has two loops, one using S simulations for each b to approximate the asymptotic binding function, and one using B draws to approximate the 'quasi-posterior' SLT distribution

$$p_{SLT}(\theta|\mathbf{y}, \varepsilon^1, \ldots, \varepsilon^S) \quad = \quad \frac{\exp(-J_S(\theta))\pi(\theta)}{\int_\Theta \exp(-J_S(\theta))\pi(\theta)d\theta} \propto \exp(-J_S(\theta))\pi(\theta) \tag{5}$$

The above SLT algorithm has features of SMD, ABC, and LT, it also requires simulations of the full model. As a referee pointed out, though the SLT resembles the ABC algorithm when used with a Gaussian kernel, $\exp(-J_S(\theta))$ is not a proper density, and $p_{SLT}(\theta|\mathbf{y}, \varepsilon^1, \ldots, \varepsilon^S)$ is not a conventional likelihood-based posterior distribution. While the SLT targets the pseudo likelihood, ABC algorithms target the proper but intractable likelihood. Furthermore, the asymptotic distribution of $\widehat{\psi}$ is known from a frequentist perspective. In ABC estimation, lack of knowledge of the likelihood of $\widehat{\psi}$ motivates the Bayesian computation.

The optimization implementation of SLT presents a clear contrast with the ABC.

1 Given $\varepsilon^s = (\varepsilon^s_1, \ldots, \varepsilon^s_T)'$ for $s = 1, \ldots S$, repeat for $b = 1, \ldots B$:

i Draw $\widehat{\psi}^b(\theta) = \frac{1}{S}\sum_{s=1}^{S} \widehat{\psi}^s(\theta) + \frac{\mathbb{A}^b_\infty(\theta)}{\sqrt{T}}$.

ii Solve for θ^b such that $\widehat{\psi}^b(\theta^b) = \widehat{\psi}$ (up to machine precision).

iii Compute $w^b(\theta^b) = |\widehat{\psi}^b_\theta(\theta^b)|^{-1}\pi(\theta^b)$.

2. Compute $\overline{\theta}_{SLT} = \sum \overline{w}^b(\theta^b)\theta^b$, where $\overline{w}^b = \frac{w^b(\theta^b)}{\sum_{c=1}^{B} w^c(\theta^c)}$.

While the SLT is a weighted average of SMD modes, the draws of $\widehat{\psi}^b(\theta)$ are taken from the (frequentist) asymptotic distribution of $\widehat{\psi}$ instead of solving the model at each b. Gao and Hong (2014) use a similar idea to make draws of what we refer to as $\overline{g}(\theta)$ in their extension of the BIL estimator of Creel and Kristensen (2013) to non-separable models.

The SMD, RS, ABC, and SLT all require specification and simulation of the full model. At a practical level, the innovations $\varepsilon^1, \ldots, \varepsilon^s$ used in SMD and SLT are only drawn from F_ε once and held fixed across iterations. Equivalently, the seed of the random number generator is fixed so that the only difference in successive iterations is due to change in the parameters to be estimated. In contrast, ABC draws new innovations from F_ε each time a θ^{b+1} is proposed. We need to simulate B sets of innovations of length T, not counting those used in draws that are rejected, and B is generally much bigger than S. The SLT takes B draws from an asymptotic distribution of $\widehat{\psi}$. Hence even though some aspects of the algorithms considered seem similar, there are subtle differences.

5 Properties of the Estimators

This section studies the finite sample properties of the various estimators. Our goal is to compare the SMD with the RS, and by implication, the infeasible MCMC-ABC. Note that our RS is different from the original kernel based ABC methods. To do so in a tractable way, we only consider the expansion up to order $\frac{1}{T}$. As a point of reference, we first note that under assumptions in Rilstone et al. (1996); Bao and Ullah (2007), $\widehat{\theta}_{ML}$ admits a second order expansion

$$\widehat{\theta}_{ML} = \theta_0 + \frac{A_{ML}(\theta_0)}{\sqrt{T}} + \frac{C_{ML}(\theta_0)}{T} + o_p(\frac{1}{T}).$$

where $A_{ML}(\theta_0)$ is a mean-zero asymptotically normal random vector and $C_{ML}(\theta_0)$ depends on the curvature of the likelihood. These terms are defined as

$$A_{ML}(\theta_0) = \mathbb{E}[\ell_{\theta\theta}(\theta_0)]^{-1} Z_S(\theta_0) \tag{6a}$$

$$C_{ML}(\theta_0) = \mathbb{E}[-\ell_{\theta\theta}(\theta_0)]^{-1} \left[Z_H(\theta_0) Z_S(\theta_0) - \frac{1}{2} \sum_{j=1}^{K} (-\ell_{\theta\theta\theta_j}(\theta_0)) Z_S(\theta_0) Z_{S,j}(\theta_0) \right] \tag{6b}$$

where the normalized score $\frac{1}{\sqrt{T}}\ell_\theta(\theta_0)$ and centered Hessian $\frac{1}{\sqrt{T}}(\ell_{\theta\theta}(\theta_0) - \mathbb{E}[\ell_{\theta\theta}(\theta_0)])$ converge in distribution to the normal vectors Z_S and Z_H respectively. The order $\frac{1}{T}$ bias is large when Fisher information is low.

Classical Bayesian estimators are likelihood based. Hence the posterior mode $\widehat{\theta}_{BC}$ exhibits a bias similar to that of $\widehat{\theta}_{ML}$. However, the prior $\pi(\theta)$ can be thought of as a constraint, or penalty since the posterior mode maximizes $\log p(\theta|\mathbf{y}) = \log L(\theta|\mathbf{y}) + \log \pi(\theta)$. Furthermore, Kass et al. (1990) show that the posterior mean deviates from the posterior mode by a term that depends on the second derivatives of the log-likelihood. Accordingly, there are three sources of bias in the posterior mean $\overline{\theta}_{BC}$: a likelihood component, a prior component, and a component from approximating the mode by the mean. Hence

$$\widehat{\theta}_{BC} = \theta_0 + \frac{A_{ML}(\theta_0)}{\sqrt{T}} + \frac{1}{T}\left[C_{BC}(\theta_0) + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)} C^P_{BC}(\theta_0) + C^M_{BC}(\theta_0) \right] + o_p(\frac{1}{T}).$$

Note that the prior component is under the control of the researcher.

In what follows, we will show that posterior means based on auxiliary statistics $\widehat{\psi}$ generically have the above representation, but the composition of the terms differ.

5.1 Properties of $\widehat{\theta}_{SMD}$

Minimum distance estimators depend on auxiliary statistics $\widehat{\psi}$. Its properties have been analyzed in Newey and Smith (2004, Section 4.2) within an empirical-likelihood framework. To facilitate subsequent analysis, we follow Gouriéroux and Monfort (1996, Ch.4.4) and directly expand $\widehat{\psi}$ around $\psi(\theta_0)$, under the assumption that it admits a second-order expansion. In particular, since $\widehat{\psi}$ is $\sqrt{T}$ consistent for $\psi(\theta_0)$, $\widehat{\psi}$ has expansion

$$\widehat{\psi} = \psi(\theta_0) + \frac{\mathbb{A}(\theta_0)}{\sqrt{T}} + \frac{\mathbb{C}(\theta_0)}{T} + o_p(\frac{1}{T}). \tag{7}$$

It is then straightforward to show that the minimum distance estimator $\widehat{\theta}_{MD}$ has expansion

$$A_{MD}(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1}\mathbb{A}(\theta_0) \tag{8a}$$

$$C_{MD}(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1}\Big[\mathbb{C}(\theta_0) - \frac{1}{2}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A_{MD}(\theta_0)A_{MD,j}(\theta_0)\Big]. \tag{8b}$$

The bias in $\widehat{\theta}_{MD}$ depends on the curvature of the binding function and the bias in the auxiliary statistic $\widehat{\psi}$, $\mathbb{C}(\theta_0)$. Then following Gouriéroux et al. (1999), we can analyze the SMD as follows. In view of (7), we have, for each s:

$$\widehat{\psi}^s(\theta) = \psi(\theta) + \frac{\mathbb{A}^s(\theta)}{\sqrt{T}} + \frac{\mathbb{C}^s(\theta)}{T} + o_p(\frac{1}{T}).$$

The estimator $\widehat{\theta}_{SMD}$ satisfies $\widehat{\psi} = \frac{1}{S}\sum_{s=1}^S \widehat{\psi}^s(\widehat{\theta}_{SMD})$ and has expansion $\widehat{\theta}_{SMD} = \theta_0 + \frac{A_{SMD}(\theta_0)}{\sqrt{T}} + \frac{C_{SMD}(\theta_0)}{T} + o_p(\frac{1}{T})$. Plugging it in the Edgeworth expansions gives:

$$\psi(\theta_0) + \frac{\mathbb{A}(\theta_0)}{\sqrt{T}} + \frac{\mathbb{C}(\theta_0)}{T} + O_p(\frac{1}{T}) = \frac{1}{S}\sum_{s=1}^{S}\Big[\psi(\widehat{\theta}_{SMD}) + \frac{\mathbb{A}^s(\widehat{\theta}_{SMD})}{\sqrt{T}} + \frac{\mathbb{C}^s(\widehat{\theta}_{SMD})}{T} + o_p(\frac{1}{T})\Big].$$

Expanding $\psi(\widehat{\theta}_{SMD})$ and $\mathbb{A}^s(\widehat{\theta}_{SMD})$ around θ_0 and equating terms in the expansion of $\widehat{\theta}_{SMD}$,

$$A_{SMD}(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1}\Big(\mathbb{A}(\theta_0) - \frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s(\theta_0)\Big) \tag{9a}$$

$$C_{SMD}(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1}\Big(\mathbb{C}(\theta_0) - \frac{1}{S}\sum_{s=1}^{S}\mathbb{C}^s(\theta_0) - (\frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s_\theta(\theta_0))A_{SMD}(\theta_0)\Big) \tag{9b}$$

$$-\frac{1}{2}\Big[\psi_\theta(\theta_0)\Big]^{-1}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A_{SMD}(\theta_0)A_{SMD,j}(\theta_0).$$

The first order term can be written as $A_{SMD} = A_{MD} + \frac{1}{B}[\psi_\theta(\theta_0)]^{-1}\sum_{b=1}^B \mathbb{A}^b(\theta_0)$, the last term has variance of order $1/B$ which accounts for simulation noise. Note also that $\mathbb{E}\left(\frac{1}{S}\sum_{s=1}^S \mathbb{C}^s(\theta_0)\right) =$

$\mathbb{E}[\mathbb{C}(\theta_0)]$. Hence, unlike the MD, $\mathbb{E}[C_{SMD}(\theta_0)]$ does not depend on the bias $\mathbb{C}(\theta_0)$ in the auxiliary statistic. In the special case when $\widehat{\psi}$ is a consistent estimator of θ_0, $\psi_\theta(\theta_0)$ is the identity map and the term involving $\psi_{\theta\theta_j}(\theta_0)$ drops out. Consequently, the SMD has no bias of order $\frac{1}{T}$ when $S \to \infty$ and $\psi(\theta) = \theta$. In general, the bias of $\widehat{\theta}_{SMD}$ depends on the curvature of the binding function as

$$\mathbb{E}[C_{SMD}(\theta_0)] \overset{S\to\infty}{\to} -\frac{1}{2}\Big[\psi_\theta(\theta_0)\Big]^{-1}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)\mathbb{E}\Big[A_{MD}(\theta_0)A_{MD,j}(\theta_0)\Big]. \tag{10}$$

This is an improvement over $\widehat{\theta}_{MD}$ because as seen from (8b),

$$\mathbb{E}[C_{MD}(\theta_0)] = \Big[\psi_\theta(\theta_0)\Big]^{-1}\mathbb{C}(\theta_0) - \frac{1}{2}\Big[\psi_\theta(\theta_0)\Big]^{-1}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)\mathbb{E}\Big[A_{MD}(\theta_0)A_{MD,j}(\theta_0)\Big]. \tag{11}$$

The bias in $\widehat{\theta}_{MD}$ has an additional term in $\mathbb{C}(\theta_0)$.

5.2 Properties of $\overline{\theta}_{RS}$

The convergence properties of the ABC algorithms have been well analyzed but the theoretical properties of the estimates are less understood. Dean et al. (2011) establish consistency of the ABC in the case of hidden Markov models. The analysis considers a scheme so that maximum likelihood estimation based on the ABC algorithm is equivalent to exact inference under the perturbed hidden Markov scheme. The authors find that the asymptotic bias depends on the ABC tolerance δ. Calvet and Czellar (2015) provide an upper bound for the mean-squared error of their ABC filter and study how the choice of the bandwidth affects properties of the filter. Under high level conditions and adopting the empirical likelihood framework of Newey and Smith (2004), Creel and Kristensen (2013) show that the infeasible BIL is second order equivalent to the MIL after bias adjustments, while MIL is in turn first order equivalent to the continuously updated GMM. The feasible SBIL (which is also an ABC estimator) has additional errors compared to the BIL due to simulation noise and kernel smoothing, but these errors vanish as $S \to \infty$ for an appropriately chosen bandwidth. Gao and Hong (2014) show that local-regressions have better variance properties compared to kernel estimations of the indirect likelihood. Creel et al. (2016) show that the number of simulations can affect the parametric convergence rate and asymptotic normality of the estimator, which is important for frequentist inference.

ABC algorithms are traditionally implemented using kernel smoothing, the first implementation being Beaumont et al. (2002). The bias due to kernel smoothing is rigorously studied in Creel et al. (2016) under the assumption that the draws are taken directly from the prior. Our RS is an importance sampler that does not use kernel smoothing. Instead it uses optimization to set δ equal to zero. This offers different insight as we look at the bias in the ideal case where δ is exactly zero.

As shown above, $\overline{\theta}_{RS}$ is the weighted average of a sequence of SMD modes. Analysis of the weights $w^b(\theta^b)$ requires an expansion of $\widehat{\psi}^b_\theta(\theta^b)$ around $\psi_\theta(\theta_0)$. From such an analysis, shown in the Appendix, we find that

$$\overline{\theta}_{RS} = \sum_{b=1}^{B} \overline{w}^b(\theta^b)\theta^b = \theta_0 + \frac{A_{RS}(\theta_0)}{\sqrt{T}} + \frac{C_{RS}(\theta_0)}{T} + o_p(\frac{1}{T})$$

where

$$A_{RS}(\theta_0) = \frac{1}{B}\sum_{b=1}^{B} A^b_{RS}(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1}\Big(\mathbb{A}(\theta_0) - \frac{1}{B}\sum_{b=1}^{B}\mathbb{A}^b(\theta_0)\Big) \quad (12a)$$

$$C_{RS}(\theta_0) = \frac{1}{B}\sum_{b=1}^{B} C^b_{RS}(\theta_0) + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\Big[\frac{1}{B}\sum_{b=1}^{B}(A^b_{RS}(\theta_0) - \overline{A}_{RS}(\theta_0))A^b_{RS}(\theta_0)\Big] + C^M_{RS}(\theta_0). \quad (12b)$$

Proposition 1 *Let $\widehat{\psi}(\theta)$ be the auxiliary statistic that admits the expansion as in (7) and suppose that the prior $\pi(\theta)$ is positive and continuously differentiable around θ_0 when $dim(\widehat{\psi}) = dim(\theta)$. Then $\mathbb{E}[A_{RS}(\theta_0)] = 0$ but $\mathbb{E}[C_{RS}(\theta_0)] \neq 0$ for an arbitrary choice of prior.*

The SMD and RS are first order equivalent, but $\overline{\theta}_{RS}$ has an order $\frac{1}{T}$ bias. The bias, given by $C_{RS}(\theta_0)$, has three components. The $C^M_{RS}(\theta_0)$ term (defined in Appendix A) can be traced directly to the weights, or to the interaction of the weights with the prior, and is a function of $A_{RS}(\theta_0)$. Some but not all the terms vanish as $B \to \infty$. The second term will be zero if a uniform prior is chosen since $\pi_\theta = 0$. A similar result is obtained in Creel and Kristensen (2013). The first term is

$$\frac{1}{B}\sum_{b=1}^{B} C^b_{RS}(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1}\frac{1}{B}\sum_{b=1}^{B}\Big(\mathbb{C}(\theta_0) - \mathbb{C}^b(\theta_0) - \frac{1}{2}\sum_{j=1}^{K}\psi_{\theta\theta_j}(\theta_0)A^b_{RS}(\theta_0)A^b_{RS,j}(\theta_0) - \mathbb{A}^b_\theta(\theta_0)A^b_{RS}(\theta_0)\Big).$$

The term $\mathbb{C}(\theta_0) - \frac{1}{B}\sum_{b=1}^{B}\mathbb{C}^b(\theta_0)$ is exactly the same as in $C_{SMD}(\theta_0)$. The middle term involves $\psi_{\theta\theta_j}(\theta_0)$ and is zero if $\psi(\theta) = \theta$. But because the summation is over θ^b instead of $\widehat{\psi}^s$,

$$\frac{1}{B}\sum_{b=1}^{B}\mathbb{A}^b_\theta(\theta_0)A^b_{RS}(\theta_0) \overset{B\to\infty}{\to} \mathbb{E}[\mathbb{A}^b_\theta(\theta_0)A^b_{RS}(\theta_0)] \neq 0.$$

As a consequence $\mathbb{E}[C_{RS}(\theta_0)] \neq 0$ even when $\psi(\theta) = \theta$. In contrast, $\mathbb{E}[C_{SMD}(\theta_0)] = 0$ when $\psi(\theta) = \theta$ as seen from (10). The reason is that the comparable term in $C_{SMD}(\theta_0)$ is

$$\Big(\frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s_\theta(\theta_0)\Big)A_{SMD}(\theta_0) \overset{S\to\infty}{\to} \mathbb{E}[\mathbb{A}^s_\theta(\theta_0)]A_{SMD}(\theta_0) = 0.$$

The difference boils down to the fact that the SMD is the mode of the average over simulated auxiliary statistics, while the RS is a weighted average over the modes. As will be seen below, this difference is also present in the LT and SLT and comes from averaging over θ^b. The result is based on fixing δ at zero and holds for any B. Proposition 1 implies that the ideal MCMC-ABC with $\delta = 0$ also has a non-negligible second-order bias. Note that Proposition 1 is stated for the exactly identified case. When $dim(\widehat{\psi}) > dim(\theta)$, the analysis is more complicated. Essentially, when the model is overidentified, weighting is needed since all moments cannot be made equal to zero simultaneously in general. This introduces additional biases. A result analogous to Proposition 1 is given in Forneron and Ng (2016) for the overidentified case.

In theory, the order $\frac{1}{T}$ bias can be removed if $\pi(\theta)$ can be found to put the right hand side of $C^{RS}(\theta_0)$ defined in (12b) to zero. Then $\overline{\theta}_{RS}$ will be second order equivalent to SMD when $\psi(\theta) = \theta$ and may have a smaller bias than SMD when $\psi(\theta) \neq \theta$ since SMD has a non-removable second order bias in that case. That the choice of prior will have bias implications for likelihood-free estimation echoes the findings in the parametric likelihood setting. Arellano and Bonhomme (2009) show in the context of non-linear panel data models that the first-order bias in Bayesian estimators can be eliminated with a particular prior on the individual effects. Bester and Hansen (2006) also show that in the estimation of parametric likelihood models, the order $\frac{1}{T}$ bias in the posterior mode and mean can be removed using objective Bayesian priors. They suggest to replace the population quantities in a differential equation with sample estimates. Finding the bias-reducing prior for the RS involves solving the differential equation:

$$0 = \mathbb{E}[C^b_{RS}(\theta_0)] + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\mathbb{E}[(A^b_{RS}(\theta_0) - \overline{A}_{RS}(\theta_0))A^b_{RS}(\theta_0)] + \mathbb{E}[C^M_{RS}(\theta_0), \pi(\theta_0)]$$

which has the additional dependence on π in $C^M_{RS}(\theta_0, \pi(\theta_0))$ that is not present in Bester and Hansen (2006). A closed-form solution is available only for simple examples as we will see Section 6.1 below. For realistic problems, how to find and implement the bias-reducing prior is not a trivial problem. A natural starting point is the plug-in procedure of Bester and Hansen (2006) but little is known about its finite sample properties even in the likelihood setting for which it was developed.

This section has studied the RS, which is the best that the MCMC-ABC can achieve in terms of δ. This enables us to make a comparison with the SMD holding the same L_2 distance between $\widehat{\psi}$ and $\psi(\theta)$ at zero by machine precision. However, the MCMC-ABC algorithm with $\delta > 0$ will not produce draws with the same distribution as the RS. To see the problem, suppose that the RS draws are obtained by stopping the optimizer before $\|\widehat{\psi} - \psi(\theta^b)\|$ reaches the tolerance guided by machine precision. This is analogous to equating $\psi(\theta^b)$ to the pseudo estimate $\widehat{\psi} + \delta$. Inverting the binding function will yield an estimate of θ that depends on the random δ in an intractable way. The RS estimate will thus have an additional bias from $\delta \neq 0$. By implication, the MCMC-ABC with $\delta > 0$ will be second order equivalent to the SMD only after a bias adjustment even when $\psi(\theta) = \theta$.

5.3 The Properties of LT and SLT

The mode of $\exp(-J(\theta))\pi(\theta)$ will inherit the properties of a MD estimator. However, the quasi-posterior mean has two additional sources of bias, one arising from the prior, and another one from approximating the mode by the mean. The optimization view of $\overline{\theta}_{LT}$ facilitates an understanding of these effects. As shown in Appendix B, each draw θ^b_{LT} has expansion terms

$$\begin{aligned} A^b_{LT}(\theta_0) &= \Big[\psi_\theta(\theta_0)\Big]^{-1}\Big(\mathbb{A}(\theta_0)-\mathbb{A}^b_\infty(\theta_0)\Big) \\ C^b_{LT}(\theta_0) &= \Big[\psi_\theta(\theta_0)\Big]^{-1}\left(\mathbb{C}(\theta_0)-\frac{1}{2}\sum_{j=1}^K \psi_{\theta,\theta_j}(\theta_0)(A^b_{LT}(\theta_0)A^b_{LT,j}(\theta_0)-\mathbb{A}^b_{\infty,\theta}(\theta_0)A^b_{LT}(\theta_0)\right). \end{aligned}$$

Even though the LT has the same objective function as MD, simulation noise enters both $A^b_{LT}(\theta_0)$ and $C^b_{LT}(\theta_0)$. Compared to the extremum estimate $\widehat{\theta}_{MD}$, we see that $A_{LT}=\frac{1}{B}\sum_{b=1}^B A^b_{LT}(\theta_0)\neq A_{MD}(\theta_0)$ and $C_{LT}(\theta_0)\neq C_{MD}(\theta_0)$. Although $C_{LT}(\theta_0)$ has the same terms as $C_{RS}(\theta_0)$, they are different because the LT uses the asymptotic binding function, and hence $A^b_{LT}(\theta_0)\neq A^b_{RS}(\theta_0)$.

A similar stochastic expansion of each θ^b_{SLT} gives:

$$\begin{aligned} A^b_{SLT}(\theta_0) &= \Big[\psi_\theta(\theta_0)\Big]^{-1}\left(\mathbb{A}(\theta_0)-\frac{1}{S}\sum_{s=1}^S\mathbb{A}^s(\theta_0)-\mathbb{A}^b_\infty(\theta_0)\right) \\ C^b_{SLT}(\theta_0) &= \Big[\psi_\theta(\theta_0)\Big]^{-1}\left(\mathbb{C}(\theta_0)-\frac{1}{S}\sum_{s=1}^S\mathbb{C}^s(\theta_0)-\frac{1}{2}\sum_{j=1}^K\psi_{\theta,\theta_j}(\theta_0)A^b_{SLT}A^b_{SLT,j}]\right) \\ &\quad -\Big[\psi_\theta(\theta_0)\Big]^{-1}\left(\frac{1}{S}\sum_{s=1}^S(\mathbb{A}^s_\theta(\theta_0)+\mathbb{A}^b_{\infty,\theta}(\theta_0))A^b_{SLT}(\theta_0)\right) \end{aligned}$$

Following the same argument as in the RS, an optimally chosen prior can reduce bias, at least in theory, but finding this prior will not be a trivial task. Overall, the SLT has features of the RS (bias does not depend on $\mathbb{C}(\theta_0)$) and the LT (dependence on $\mathbb{A}^b_\infty$) but is different from both. Because the SLT uses simulations to approximate the binding function $\psi(\theta)$, $\mathbb{E}[\mathbb{C}(\theta_0)-\frac{1}{S}\sum_{s=1}^S\mathbb{C}^s(\theta_0)]=0$. The improvement over the LT is analogous to the improvement of SMD over MD. However, the $A^b_{SLT}(\theta_0)$ is affected by estimation of the binding function (the term with superscript s) and of the quasi-posterior density (the terms with superscript b). This results in simulation noise with variance of order $1/S$ plus another of order $1/B$. Note also that the SLT bias has an additional term

$$\frac{1}{B}\sum_{b=1}^B\left(\frac{1}{S}\sum_{s=1}^S(\mathbb{A}^s_\theta(\theta_0)+\mathbb{A}^b_{\infty,\theta}(\theta_0))A^b_{SLT}(\theta_0)\right)\overset{S\to\infty}{\to}\frac{1}{B}\sum_{b=1}^B\mathbb{A}^b_{\infty,\theta}(\theta_0)A^b_{LT}(\theta_0).$$

The main difference with the RS is that $\mathbb{A}^b$ is replaced with $\mathbb{A}^b_\infty$. For $S=\infty$ this term matches that of the LT.

5.4 Overview

We started this section by noting that the Bayesian posterior mean has two components in its bias, one arising from the prior which acts like a penalty on the objective function, and another due to approximating the mean with the mode. We are now in a position to use the results in the foregoing subsections to show that for d=(MD, SMD, RS, LT) and SLT and D = (RS,LT,SLT) these estimators can be represented as

$$\widehat{\theta}_d = \theta_0 + \frac{A_d(\theta_0)}{\sqrt{T}} + \frac{C_d(\theta_0)}{T} + \frac{\mathbb{1}_{d\in D}}{T}\left[\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}C_d^P(\theta_0)) + C_d^M(\theta_0)\right] + o_p(\frac{1}{T}) \tag{13}$$

where with $A_d^b(\theta_0) = [\psi_\theta(\theta_0)]^{-1}\Big(\mathbb{A}(\theta_0) - \mathbb{A}_d^b(\theta_0)\Big)$,

$$\begin{aligned}
A_d(\theta_0) &= [\psi_\theta(\theta_0)]^{-1}\Big(\mathbb{A}(\theta_0) - \frac{1}{B}\sum_{b=1}^{B}\mathbb{A}_d^b(\theta_0)\Big) \\
C_d(\theta_0) &= [\psi_\theta(\theta_0)]^{-1}\Big(\mathbb{C}(\theta_0) - \mathbb{C}_d(\theta_0) - \frac{1}{2}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A_d^b(\theta_0)A_{d,j}^b(\theta_0) - \mathbb{A}_{d,\theta}^b A_d^b(\theta_0)\Big) \\
C_d^P(\theta_0) &= \frac{1}{B}\sum_{b=1}^{B}(A_d^b(\theta_0) - A_d(\theta_0))A_d^b(\theta_0),
\end{aligned}$$

The term $C_d^P(\theta_0)$ is a bias directly due to the prior. The term $C_d^M(\theta_0)$, defined in the Appendix, depends on $A_d(\theta_0)$, the curvature of the binding function, and their interaction with the prior. Hence at a general level, the estimators can be distinguished by whether or not Bayesian computation tools are used, as the indicator function is null only for the two frequentist estimators (MD and SMD). More fundamentally, the estimators differ because of $A_d(\theta_0)$ and $C_d(\theta_0)$, which in turn depend on $\mathbb{A}_d^b(\theta_0)$ and $\mathbb{C}_d(\theta_0)$. We compactly summarize the differences as follows:

d	$\mathbb{A}_d^b(\theta_0)$	$\mathbb{C}_d(\theta_0)$	$\text{var}(\mathbb{A}_d(\theta_0))$	$\mathbb{E}[\mathbb{C}(\theta_0) - \mathbb{C}_d(\theta_0)]$
MD	0	0	0	$\mathbb{E}[\mathbb{C}(\theta_0)]$
LT	$\mathbb{A}_\infty^b(\theta_0)$	0	$\frac{1}{B}\text{var}[\mathbb{A}_\infty^b(\theta_0)]$	$\mathbb{E}[\mathbb{C}(\theta_0)]$
RS	$\mathbb{A}^b(\theta_0)$	$\frac{1}{B}\sum_{b=1}^{B}\mathbb{C}^b(\theta_0)$	$\frac{1}{B}\text{var}[\mathbb{A}^b(\theta_0)]$	0
SMD	$\frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s(\theta_0)$	$\frac{1}{S}\sum_{s=1}^{S}\mathbb{C}^s(\theta_0)$	$\frac{1}{S}\text{var}[\mathbb{A}^s(\theta_0)]$	0
SLT	$\mathbb{A}_{SMD}(\theta_0) + \mathbb{A}_{LT}^b(\theta_0)$	$\frac{1}{S}\sum_{s=1}^{S}\mathbb{C}^s(\theta_0)$	$\text{var}[\mathbb{A}_{\text{SMD}}(\theta_0)] + \text{var}[\mathbb{A}_{\text{LT}}(\theta_0)]$	0

The MD is the only estimator that is optimization based and does not involve simulations. Hence it does not depend on b or s and has no simulation noise. The SMD does not depend on b because the optimization problem is solved only once. The LT simulates from the asymptotic binding function. Hence its errors are associated with parameters of the asymptotic distribution.

The MD and LT have a bias due to asymptotic approximation of the binding function. In such cases, Cabrera and Fernholz (1999) suggest to adjust an initial estimate $\widetilde{\theta}$ such that if the new estimate $\widehat{\theta}$ were the true value of θ, the mean of the original estimator equals the observed value $\widetilde{\theta}$. Their *target estimator* is the θ such that $\mathbb{E}_{\mathcal{P}_\theta}[\widehat{\theta}] = \widetilde{\theta}$. While the bootstrap directly estimates the bias, a target estimator corrects for the bias implicitly. Cabrera and Hu (2001) show that the bootstrap estimator corresponds to the first step of a target estimator. The latter improves upon the bootstrap estimator by providing more iterations.

An auxiliary statistic based target estimator is the θ that solves $\mathbb{E}_{\mathcal{P}_\theta}[\widehat{\psi}(\mathbf{y}(\theta))] = \widehat{\psi}(\mathbf{y}(\theta_0))$. It replaces the asymptotic binding function $\lim_{T\to\infty} \mathbb{E}[\widehat{\psi}(\mathbf{y}(\theta_0))]$ by $\mathbb{E}_{\mathcal{P}_\theta}[\widehat{\psi}(\mathbf{y}(\theta))]$ and approximates the expectation under $\mathcal{P}_\theta$ by stochastic expansions. The SMD and SLT can be seen as target estimators that approximate the expectation by simulations. Thus, they improve upon the MD estimator even when the binding function is tractable and is especially appealing when it is not. However, the improvement in the SLT is partially offset by having to approximate the mode by the mean.

6 Two Examples

The preceding section can be summarized as follows. A posterior mean computed through auxiliary statistics generically has a component due to the prior, and a component due to the approximation of the mode by the mean. The binding function is better approximated by simulations than asymptotic analysis. It is possible for simulation estimation to perform better than $\widehat{\psi}_{MD}$ even if $\psi(\theta)$ were analytically and computationally tractable.

In this section, we first illustrate the above findings using a simple analytical example. We then evaluate the properties of the estimators using the dynamic panel model with fixed effects.

6.1 An Analytical Example

We consider the simple DGP $y_i \sim N(m, \sigma^2)$. The parameters of the model are $\theta = (m, \sigma^2)'$. We focus on σ^2 since the estimators have more interesting properties.

The MLE of θ is

$$\widehat{m} = \frac{1}{T}\sum_{t=1}^{T} y_t, \qquad \widehat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \overline{y})^2.$$

While the posterior distribution is dominated by the likelihood in large samples, the effect of the prior is not negligible in small samples. We therefore begin with a analysis of the effect of the prior on the posterior mean and mode in Bayesian analysis. Details of the calculations are provided in Appendix D.1.

We consider the prior $\pi(m,\sigma^2) = (\sigma^2)^{-\alpha}\mathbb{I}_{\sigma^2>0}$, $\alpha > 0$ so that the log posterior distribution is

$$\log p(\theta|y) = \log p(\theta|\widehat{m},\widehat{\sigma}^2) \propto \frac{-T}{2}\left[\log(2\pi\sigma^2) - \alpha\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - m)^2\right]\mathbb{I}_{\sigma^2>0}.$$

The posterior mode and mean of σ^2 are $\sigma^2_{mode} = \frac{T\widehat{\sigma}^2}{T+2\alpha}$ and $\sigma^2_{mean} = \frac{T\widehat{\sigma}^2}{T+2\alpha-5}$. respectively. Using the fact that $E[\widehat{\sigma}^2] = \frac{(T-1)}{T}\sigma^2$, we can evaluate σ^2_{mode}, σ^2_{mean} and their expected values for different α. Two features are of note. For a given prior (here indexed by α), the mean does not coincide with

Table 1: Mean $\overline{\theta}_{BC}$ vs. Mode $\widehat{\theta}_{BC}$

α	$\overline{\theta}_{BC}$	$\widehat{\theta}_{BC}$	$\mathbb{E}[\overline{\theta}_{BC}]$	$\mathbb{E}[\widehat{\theta}_{BC}]$
0	$\widehat{\sigma}^2\frac{T}{T-5}$	$\widehat{\sigma}^2$	$\sigma^2\frac{T-1}{T-5}$	$\sigma^2\frac{T-1}{T}$
1	$\widehat{\sigma}^2\frac{T}{T-3}$	$\widehat{\sigma}^2\frac{T}{T+2}$	$\sigma^2\frac{T-1}{T-3}$	$\sigma^2\frac{T-1}{T+2}$
2	$\widehat{\sigma}^2\frac{T}{T-1}$	$\widehat{\sigma}^2\frac{T}{T+4}$	σ^2	$\sigma^2\frac{T-1}{T+4}$
3	$\widehat{\sigma}^2\frac{T}{T+1}$	$\widehat{\sigma}^2\frac{T}{T+6}$	$\sigma^2\frac{T-1}{T+1}$	$\sigma^2\frac{T-1}{T+6}$

the mode. Second, the statistic (be it mean or mode) varies with α. The Jeffrey's prior corresponds to $\alpha = 1$, but the bias-reducing prior is $\alpha = 2$. In the Appendix, we show that the bias reducing prior for this model is $\pi^R(\theta) \propto \frac{1}{\sigma^4}$.

Next, we consider estimators based on auxiliary statistics:

$$\widehat{\psi}(\mathbf{y})' = \begin{pmatrix}\widehat{m} & \widehat{\sigma}^2\end{pmatrix}.$$

As these are sufficient statistics, we can also consider (exact) likelihood-based Bayesian inference. For SMD estimation, we let $(\widehat{m}_S, \widehat{\sigma}^2_S) = (\frac{1}{S}\sum_{s=1}^{S}\widehat{m}^s, \frac{1}{S}\sum_{s=1}^{S}\widehat{\sigma}^{2,s})$. The LT quasi-likelihood using the variance of preliminary estimates of m and σ^2 as weights is:

$$\exp(-J(m,\sigma^2)) = \exp\Bigg(-\frac{T}{2}\left[\frac{(\widehat{m}-m)^2}{\widehat{\sigma}^2} + \frac{(\widehat{\sigma}^2-\sigma^2)^2}{2\widehat{\sigma}^4}\right]\Bigg).$$

The LT posterior distribution is $p(m,\sigma^2|\widehat{m},\widehat{\sigma}^2) \propto \pi(m,\sigma^2)\exp(-J(m,\sigma^2))$. Integrating out m gives $p(\sigma^2|\widehat{m},\widehat{\sigma}^2)$. We consider a flat prior $\pi^U(\theta) \propto \mathbb{I}_{\sigma^2\geq 0}$ and the bias-reducing prior $\pi^R(\theta) \propto 1/\sigma^4\mathbb{I}_{\sigma^2\geq 0}$.

The RS is the same as the SMD under a bias-reducing prior. Thus,

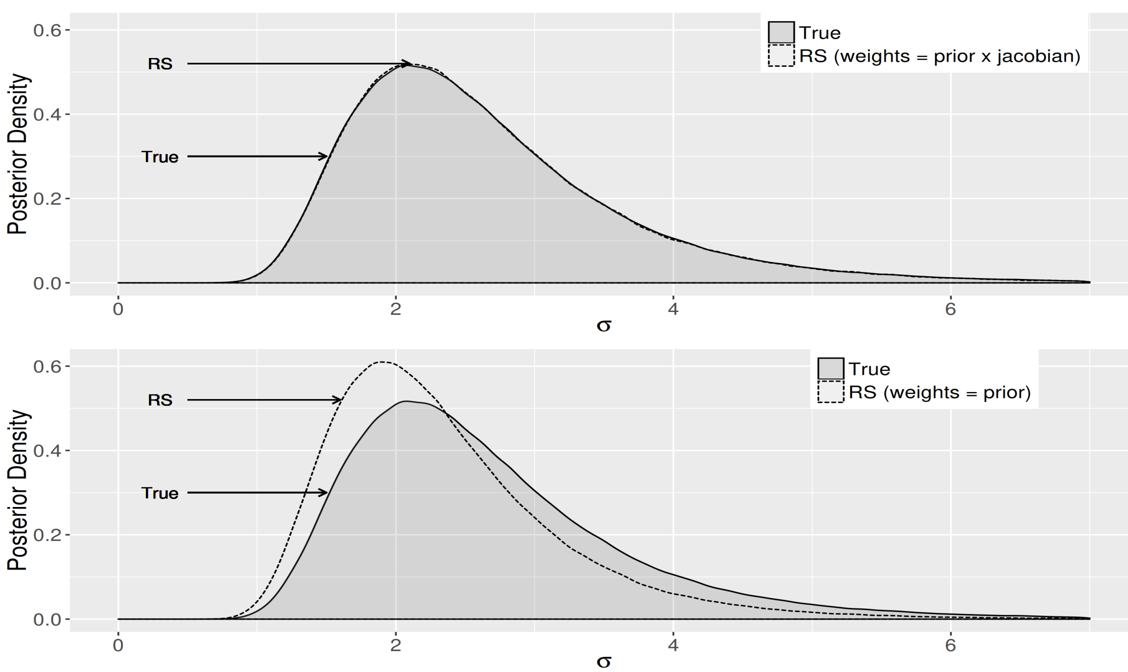
$$\begin{aligned}
\widehat{\sigma}^2_{SMD} &= \frac{\widehat{\sigma}^2}{\frac{1}{ST}\sum_{s=1}^S\sum_{t=1}^T(e_t^s-\overline{e}^s)^2}\\
\widehat{\sigma}^{2,R}_{RS} &= \frac{\widehat{\sigma}^2}{\frac{1}{BT}\sum_{b=1}^B\sum_{t=1}^T(e_t^b-\overline{e}^b)^2}\\
\widehat{\sigma}^{2,U}_{RS} &= \sum_{b=1}^B \frac{\frac{\widehat{\sigma}^2}{[\sum_{t=1}^T(e_t^b-\overline{e}^b)^2/T]^2}}{\sum_{b'=1}^B\frac{1}{\sum_{t=1}^T(e_t^{b'}-\overline{e}^{b'})^2/T}}.
\end{aligned}$$

For completeness, the parametric Bootstrap bias corrected estimator $\widehat{\sigma}^2_{\text{Bootstrap}}=2\widehat{\sigma}^2-\mathbb{E}_{\text{Bootstrap}}(\widehat{\sigma}^2)$ is also considered:

$$\widehat{\sigma}^2_{\text{Bootstrap}} = 2\widehat{\sigma}^2-\widehat{\sigma}^2\frac{T-1}{T}=\widehat{\sigma}^2(1+\frac{1}{T}).$$

$\mathbb{E}_{\text{Bootstrap}}(\widehat{\sigma}^2)$ computes the expected value of the estimator replacing the true value σ^2 with $\widehat{\sigma}^2$, the plug-in estimate. In this example the bias can be computed analytically since $\mathbb{E}(\widehat{\sigma}^2(1+\frac{1}{T}))=\sigma^2(1-\frac{1}{T})(1+\frac{1}{T})=\sigma^2(1-\frac{1}{T^2})$. While the bootstrap does not involve inverting the binding function, this computational simplicity comes at the cost of adding a higher order bias term (in $1/T^2$).

Figure 1: ABC vs. RS Posterior Density



A main finding of this paper is that the reverse sampler can replicate draws from $p^*_{ABC}(\theta_0)$, which in turn equals the Bayesian posterior distribution if $\widehat{\psi}$ are sufficient statistics. The weight for each SMD estimate is the prior times the Jacobian. To illustrate the importance of the Jacobian transformation, the top panel of Figure 1 plots the Bayesian/ABC posterior distribution and the one obtained from the reverse sampler. They are indistinguishable. The bottom panel shows an incorrectly constructed reverse sampler that does not apply the Jacobian transformation. Notably, the two distributions are not the same.

Table 2: Properties of the Estimators

Estimator	Prior	$\mathbf{E}[\widehat{\theta}]$	Bias	Variance
$\widehat{\theta}_{ML}$	-	$\sigma^2 \frac{T-1}{T}$	$-\frac{\sigma^2}{T}$	$2\sigma^4 \frac{T-1}{T^2}$
$\overline{\theta}_{BC}$	1	$\sigma^2 \frac{T-1}{T-5}$	$\frac{2\sigma^2}{T-5}$	$2\sigma^4 \frac{T-1}{(T-5)^2}$
$\overline{\theta}_{BC}^R$	$1/\sigma^4$	σ^2	0	$2\sigma^4 \frac{1}{T-1}$
$\overline{\theta}_{RS}^U$	1	$\sigma^2 \frac{T-1}{T-5}$	$\frac{2\sigma^2}{T-5}$	$2\sigma^4 \frac{T-1}{(T-5)^2}$
$\overline{\theta}_{RS}^R$	$\frac{1}{\sigma^4}$	$\sigma^2 \frac{B(T-1)}{B(T-1)-2}$	$\frac{2\sigma^2}{B(T-1)-2}$	$2\sigma^4 \frac{\kappa_1}{T-1}$
$\widehat{\theta}_{SMD}$	-	$\sigma^2 \frac{S(T-1)}{S(T-1)-2}$	$\frac{2\sigma^2}{S(T-1)-2}$	$2\sigma^4 \frac{\kappa_1}{T-1}$
$\overline{\theta}_{LT}^U$	1	$\sigma^2 \frac{T-1}{T}(1+\kappa_{LT})$	$\sigma^2 \frac{T-1}{T}\kappa_{LT} - \frac{\sigma^2}{T}$	$2\sigma^4 \frac{T-1}{T^2}(1+\kappa_{LT})^2$
$\widehat{\theta}_{SLT}^U$	1	$\sigma^2 \frac{S(T-1)}{S(T-1)-2} + \kappa_{SLT}$	$\frac{\sigma^2}{S(T-1)-2} + \sigma^2 \frac{T-1}{T}\mathbb{E}[\kappa_{SLT}]$	$2\sigma^4 \frac{\kappa_{LT}}{T-1} + \Delta_{SLT}$
$\widehat{\theta}_{\text{Bootstrap}}$	-	$\sigma^2(1-\frac{1}{T^2})$	$\frac{-\sigma^2}{T^2}$	$2\sigma^4 \frac{T-1}{T^2}(1+\frac{1}{T})^2$

Notes to Table 2: Let $M(x) = \frac{\phi(x)}{1-\Phi(x)}$ be the Mills ratio.

i $\kappa_1(S,T) = \frac{(S(T-1))^2(T-1+S(T-1)-2)}{(S(T-1)-2)^2(S(T-1)-4)} > 1$, κ_1 tends to one as B, S tend to infinity.

ii $\kappa_{LT} = c_{LT}^{-1}M(-c_{LT})$, $c_{LT}^2 = \frac{T}{2}$, $\kappa_{LT} \to 0$ as $T \to \infty$.

iii $\kappa_{SLT} = \kappa_{LT} \cdot S \cdot T \cdot \text{Inv}\chi^2_{S(T-1)}$, $\Delta_{,SLT} = 2\sigma^4 \text{var}(\kappa_{SLT}) + 4\sigma^4 \frac{T-1}{T^2}\text{cov}(\kappa_{SLT}, S \cdot T\text{Inv}\chi^2_{S(T-1))})$.

The properties of the estimators are summarized in Table 2. It should be reminded that increasing S improves the approximation of the binding function in SMD estimation while increasing B improves the approximation to the target distribution in Bayesian type estimation. For fixed T, only the Bayesian estimator with the bias reducing prior is unbiased. The SMD and RS (with bias reducing prior) have the same bias and mean-squared error in agreement with the analysis in the previous section. These two estimators have smaller errors than the RS estimator with a uniform prior. The SLT posterior mean differs from that of the SMD by κ_{SLT} that is not mean-zero. This term, which is a function of the Mills-ratio, arises as a consequence of the fact that the σ^2 in SLT are drawn from the normal distribution and then truncated to ensure positivity.

6.2 The Dynamic Panel Model with Fixed Effects

The dynamic panel model $y_{it} = \alpha_i + \rho y_{it-1} + \sigma e_{it}$ is known to be severely biased when T is small because the unobserved heterogeneity α_i is imprecisely estimated. Various approaches have been suggested to improve the precision of the least squares dummy variable (LSDV) estimator $\widehat{\beta}$.[8] An interesting approach, due to Gouriéroux et al. (2010), is to exploit the bias reduction properties of the indirect inference estimator. Using the dynamic panel model as auxiliary equation, i.e. $\psi(\theta) = \theta$, the authors reported estimates of β that are sharply more accurate than the LSDV, even when an exogenous regressor and a linear trend is added to the model. Their simulation experiments hold σ^2 fixed. We reconsider their exercise but also estimate σ^2.

With $\theta = (\rho, \beta, \sigma^2)'$, we simulate data from the model:

$$y_{it} = \alpha_i + \rho y_{it-1} + \beta x_{it} + \sigma \varepsilon_{it}.$$

Let $A = I_T - 1_T 1_T'/T$ $\underline{A} = A \otimes I_T$, $\underline{y} = \underline{A}\ vec(y), \underline{y}_{-1} = \underline{A}\ vec(y_{-1}), \underline{x} = \underline{A}\ vec(x)$, where y_{-1} are the lagged y. For this model, Bayesian inference is possible since the likelihood in de-meaned data is

$$L(\underline{\mathbf{y}}, \underline{\mathbf{x}}|\theta) = \frac{1}{\sqrt{2\pi|\sigma^2\Omega|}^N} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=2}^{N} (\underline{y}_i - \rho \underline{y}_{i,-1} - \beta \underline{x}_i)' \Omega^{-1} (\underline{y}_i - \rho \underline{y}_{i,-1} - \beta \underline{x}_i) \right)$$

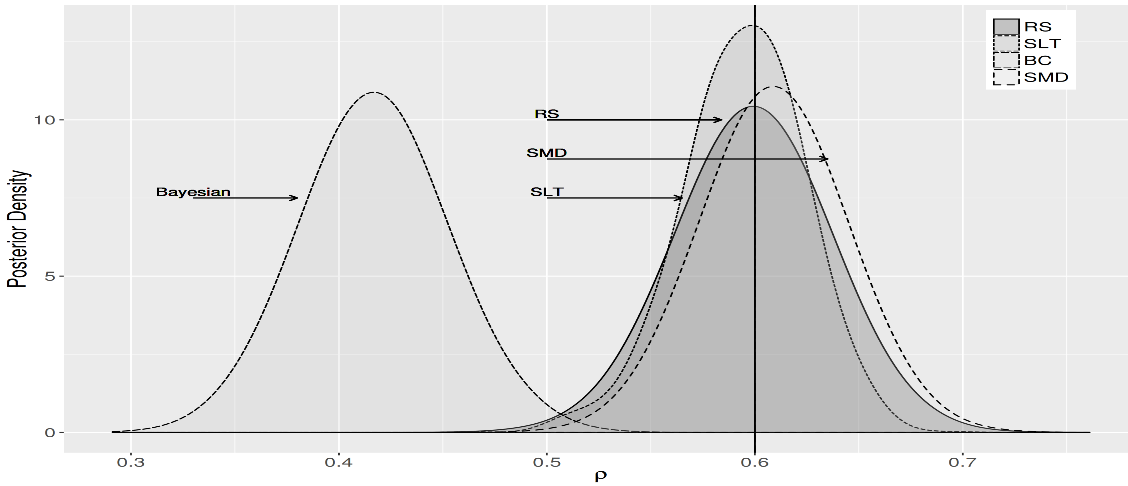
where $\Omega = I_{T-1} - 1_{T-1} 1_{T-1}'/T$. We use the following moment conditions for MD estimation:

$$\overline{g}(\rho, \beta, \sigma^2) = \begin{pmatrix} \underline{y}_{-1}(\underline{y} - \rho \underline{y}_{-1} - \beta \underline{x}) \\ \underline{x}(\underline{y} - \rho \underline{y}_{-1} - \beta \underline{x}) \\ (\underline{y} - \rho \underline{y}_{-1} - \beta \underline{x})^2 - \sigma^2(1 - 1/T) \end{pmatrix}.$$

with $\overline{g}(\widehat{\rho}, \widehat{\beta}, \widehat{\sigma}^2) = 0$. The simulated quantity $\overline{g}_S(\theta)$ for SMD and $\overline{g}^b(\theta)$ for ABC are defined analogously. The MD estimator in this case is also the LSDV. The auxiliary estimates for the ABC, RS, SLT and SMD are the LSDV estimates. Recall that while the weighting matrix W is irrelevant to finding the mode in exactly identified models, W affects computation of the posterior mean. We use $W = (\frac{1}{NT}\sum_{i,t} g_{it}' g_{it} - \overline{g}'\overline{g})^{-1}$ for LT, MCMC-ABC, and SMD. The prior is $\pi(\theta) = \mathbb{I}_{\sigma^2 \geq 0, \rho \in [-1,1], \beta \in \mathbb{R}}$. Since the demeaned data are used in LSDV estimation, the estimates are invariant to the specification of the fixed effects. Accordingly, we set them to zero both in the assumed DGP and the auxiliary model. The innovations ε^s used to simulate the auxiliary model and to construct $\widehat{\psi}^s$ are drawn from the standard normal distribution once and held fixed.

[8] See Hsiao (2003) for a detailed account of this incidental parameter problem.

Figure 2: Frequentist, Bayesian, and Approximate Bayesian Inference for ρ



$p_{BC}(\rho|\widehat{\psi})$ is the likelihood based Bayesian posterior distribution,
$p_{SLT}(\rho|\widehat{\psi})$ is the Simulated Laplace type quasi-posterior distribution.
$p_{RS}(\rho|\widehat{\psi})$ is the approximate posterior distribution based on the RS .
The frequentist distribution of $\widehat{\theta}_{SMD}$ is estimated by $\mathcal{N}(\widehat{\theta}_{SMD}, \widehat{\text{var}}(\widehat{\theta}_{SMD}))$.

Table 3 reports results from 5000 replications for $T = 6$ time periods and $N = 100$ cross-section units, as in Gouriéroux et al. (2010). Both $\widehat{\rho}$ and $\widehat{\sigma}^2$ are significantly biased. The LT is the same as the MD except that it is computed using Bayesian tools. Hence its properties are similar to the MD. The simulation estimators have much improved properties. The properties of $\overline{\theta}_{RS}$ are similar to those of the SMD. Figure 2 illustrates for one simulated dataset how the posteriors for RS /SLT are shifted towards the true value compared to the one based on the direct likelihood.

The MCMC-ABC results in Table 3 are for $\delta = 0.10$ which has an acceptance rate of 0.58. These estimates are clearly more precise than MLE but more biased than SMD or RS. The dependence of MCMC-ABC on δ is investigated in further detail in Forneron and Ng (2016). In brief, when we set $\delta = 0.25$, we achieve an acceptance ratio of 0.72 but the estimates are severely biased, as shown in Figure 3. Bias similar to SMD and RS can be obtained if we set δ to 0.025. But the corresponding acceptance rate is 0.28, meaning that the MCMC-ABC needs at least three times more draws than the RS for a comparable level of bias. The choice of δ is more important for the properties of MCMC-ABC than the RS which associates δ with the tolerance of optimization.

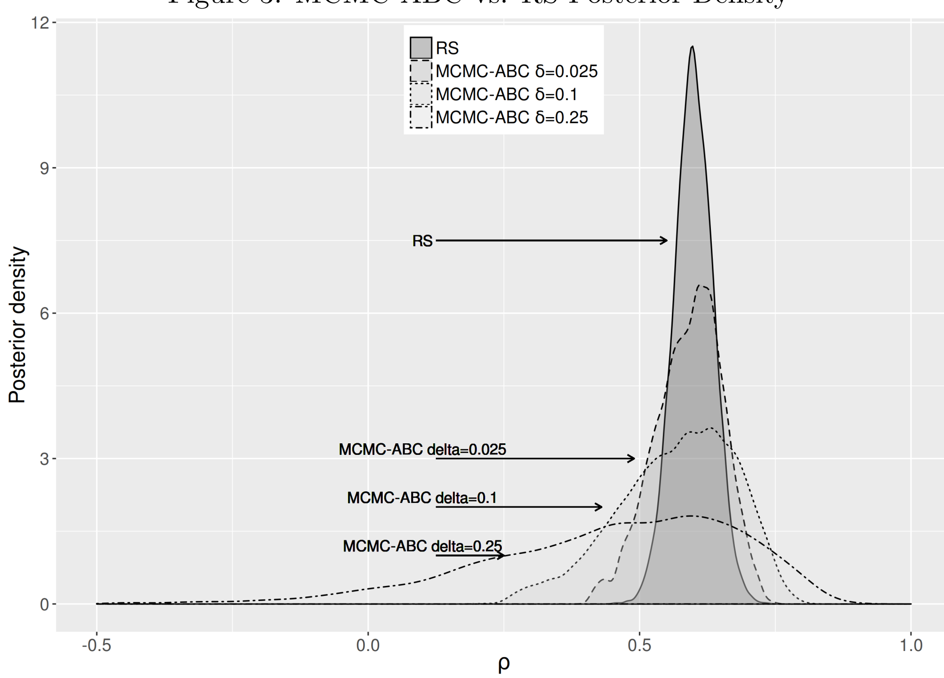
Table 3: Dynamic Panel $\rho = 0.6, \beta = 1, \sigma^2 = 2$

Mean over 1000 replications

		MLE	LT	SLT	SMD	$\frac{\text{MCMC}}{\text{ABC}}$	RS	Boot
	Mean	0.419	0.419	0.593	0.598	0.544	0.599	0.419
$\widehat{\rho}$:	SD	0.037	0.037	0.038	0.035	0.036	0.035	0.074
	Bias	-0.181	-0.181	-0.007	-0.002	-0.056	-0.001	-0.181
	Mean	0.940	0.940	0.997	1.000	0.974	1.000	0.940
$\widehat{\beta}$:	SD	0.070	0.071	0.073	0.073	0.075	0.073	0.139
	Bias	-0.060	-0.060	-0.003	0.000	-0.026	0.000	-0.060
	Mean	1.869	1.878	1.973	1.989	1.921	2.099	1.869
$\widehat{\sigma}^2$:	SD	0.133	0.146	0.144	0.144	0.149	0.152	0.267
	Bias	-0.131	-0.122	-0.027	-0.011	-0.079	0.099	-0.131
	S	–	–	500	500	1	1	–
	B	–	500	500	–	500	500	500

Note: MLE=MD. The MCMC-ABC uses $\delta_{\text{ABC}} = 0.10$.

Figure 3: MCMC-ABC vs. RS Posterior Density



7 Conclusion

Different disciplines have developed different estimators to overcome the limitations posed by an intractable likelihood. These estimators share many similarities: they rely on auxiliary statistics and use simulations to approximate quantities that have no closed form expression. We suggest an optimization framework that helps understand the estimators from the perspective of classical minimum distance estimation. All estimators are first-order equivalent as $S \to \infty$ and $T \to \infty$ for any choice of $\pi(\theta)$. Nonetheless, up to order $1/T$, the estimators are distinguished by biases due to the prior and approximation of the mode by the mean, the very two features that distinguish Bayesian and frequentist estimation.

We have only considered regular problems when θ_0 is in the interior of Θ and the objective function is differentiable. When these conditions fail, the posterior is no longer asymptotically normal around the MLE with variance equal to the inverse of the Fisher Information Matrix. Understanding the properties of these estimators under non-standard conditions is the subject for future research.

Appendix

The terms $\mathbb{A}(\theta)$ and $\mathbb{C}(\theta)$ in $\widehat{\theta}_{MD}$ are derived for the just identified case as follows. Recall that $\widehat{\psi}$ has a second order expansion:

$$\widehat{\psi} = \psi(\theta_0) + \frac{\mathbb{A}(\theta_0)}{\sqrt{T}} + \frac{\mathbb{C}(\theta_0)}{T} + o_p(\frac{1}{T}). \tag{A.1}$$

Now $\widehat{\theta} = \theta_0 + \frac{A(\theta_0)}{\sqrt{T}} + \frac{C(\theta_0)}{T} + o_p(\frac{1}{T})$. Thus expanding $\psi(\widehat{\theta})$ around $\widehat{\theta} = \theta_0$:

$$\begin{aligned}\psi(\widehat{\theta}) &= \psi\left(\theta_0 + \frac{A(\theta_0)}{\sqrt{T}} + \frac{C(\theta_0)}{T} + o_p(\frac{1}{T})\right)\\ &= \psi(\theta_0) + \psi_\theta(\theta_0)\left(\frac{A(\theta_0)}{\sqrt{T}} + \frac{C(\theta_0)}{T} + o_p(\frac{1}{T})\right) + \frac{1}{2T}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A(\theta_0)A_j(\theta_0) + o_p(\frac{1}{T}).\end{aligned}$$

Equating with $\psi(\theta_0) + \frac{\mathbb{A}(\theta_0)}{\sqrt{T}} + \frac{\mathbb{C}(\theta_0)}{T} + o_p(\frac{1}{T})$ and solving for A, C we get:

$$\begin{aligned}A(\theta_0) &= \Big[\psi_\theta(\theta_0)\Big]^{-1}\mathbb{A}(\theta_0)\\ C(\theta_0) &= \Big[\psi_\theta(\theta_0)\Big]^{-1}\left(\mathbb{C}(\theta_0) - \frac{1}{2}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A(\theta_0)A_j(\theta_0)\right).\end{aligned}$$

For estimator specific A_d^b and a_d^b, define $a_d^b = \text{trace}([\psi_\theta(\theta_0)]^{-1}[\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A_{d,j}^b(\theta_0) + \mathbb{A}_{d,\theta}^b(\theta_0)])$,

$$\begin{aligned}C_d^M(\theta_0) \quad &= \quad 2\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\overline{A}_d(\theta_0)\overline{a}_d(\theta_0)\theta_0 - \overline{a}_d(\theta_0)^2\theta_0 - \left[\frac{\pi_\theta(\theta_0)\pi_\theta(\theta_0)'}{\pi(\theta_0)^2}\right]\overline{A}_d(\theta_0)'\overline{A}_d(\theta_0)\theta_0\\ &\quad -\frac{1}{B}\sum_{b=1}^{B}(a_d^b(\theta_0) - \overline{a}_d(\theta_0))A_d^b(\theta_0).\end{aligned} \tag{A.2}$$

Where $\overline{a}_d = \frac{1}{B}\sum_{b=1}^{B}a_d^b$, $\overline{A}_d$ is defined analogously. Note that $\overline{a}(\theta_0) \to 0$ as $B \to \infty$ if $\psi(\theta) = \theta$ and the first two terms drop out.

A.1 Proof of Proposition 1, RS

To prove Proposition 1, we need an expansion for $\widehat{\psi}^b(\theta^b)$ and the weights using

$$\theta^b \quad = \theta_0 + \tfrac{A^b(\theta_0)}{\sqrt{T}} + \tfrac{C^b(\theta_0)}{T} + o_p(\tfrac{1}{T}). \tag{A.3}$$

i. Expansion of $\widehat{\psi}^b(\theta_0)$ and $\widehat{\psi}^b_\theta(\theta_0)$:

$$
\begin{aligned}
\widehat{\psi}^b(\theta^b) &= \psi(\theta^b) + \frac{\mathbb{A}^b(\theta^b)}{\sqrt{T}} + \frac{\mathbb{C}^b(\theta^b)}{T} + o_p(\frac{1}{T}) \\
&= \psi(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})) + \frac{\mathbb{A}^b(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T}))}{\sqrt{T}} \\
&+ \frac{\mathbb{C}^b(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T}))}{T} + o_p(\frac{1}{T}) \\
&= \psi(\theta_0) + \frac{\mathbb{A}^b(\theta_0)}{\sqrt{T}} + \frac{\psi_\theta(\theta_0)A^b(\theta_0)}{\sqrt{T}} + \frac{\mathbb{C}^b(\theta_0)}{T} + \frac{\mathbb{A}^b_\theta(\theta_0)A^b(\theta_0)}{T} \\
&+ \frac{1}{2}\sum_{j=1}^{K} \frac{\psi_{\theta,\theta_j}(\theta_0)A^b(\theta_0)A^b_j(\theta_0)}{T} + o_p(\frac{1}{T}).
\end{aligned}
$$

Since $\widehat{\psi}^b(\theta^b)$ equals $\widehat{\psi}$ for all b,

$$
\begin{aligned}
A^b(\theta_0) &= \Big[\psi_\theta(\theta_0)\Big]^{-1} \big(\mathbb{A}(\theta_0) - \mathbb{A}^b(\theta_0)\big) && \text{(A.4)} \\
C^b(\theta_0) &= \Big[\psi_\theta(\theta_0)\Big]^{-1} \left(\mathbb{C}(\theta_0) - \mathbb{C}^b(\theta_0) - \frac{1}{2}\sum_{j=1}^{K} \psi_{\theta,\theta_j}(\theta_0)A^b(\theta_0)A^b_j(\theta_0) - \mathbb{A}^b_\theta(\theta_0)A^b(\theta_0)\right), && \text{(A.5)}
\end{aligned}
$$

it follows that

$$
\begin{aligned}
\widehat{\psi}^b_\theta(\theta^b) &= \widehat{\psi}^b_\theta\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right) \\
&= \psi_\theta\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right) + \frac{\mathbb{A}^b_\theta\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)}{\sqrt{T}} \\
&+ \frac{\mathbb{C}^b_\theta\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)}{T} + o_p(\frac{1}{T}) \\
&= \psi_\theta(\theta_0) + \sum_{j=1}^{K} \frac{\psi_{\theta,\theta_j}(\theta_0)A^b_j(\theta_0)}{\sqrt{T}} + \frac{\mathbb{A}^b_\theta(\theta_0)}{\sqrt{T}} + \frac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K} \frac{\psi_{\theta,\theta_j,\theta_k}(\theta_0)A^b_j(\theta_0)A^b_k(\theta_0)}{T} \\
&+ \sum_{j=1}^{K} \frac{\psi_{\theta,\theta_j}(\theta_0)C^b_j(\theta_0)}{T} + \sum_{j=1}^{K} \frac{\mathbb{A}^b_{\theta,\theta_j}(\theta_0)A^b_j(\theta_0)}{T} + \frac{\mathbb{C}^b(\theta_0)}{T} + o_p(\frac{1}{T}).
\end{aligned}
$$

To obtain the determinant of $\widehat{\psi}^b_\theta(\theta^b)$, let $a^b(\theta_0) = \text{trace}(\mathcal{A}^b(\theta_0))$, $a^b_2(\theta_0) = \text{trace}(\mathcal{A}^b(\theta_0)^2)$, $c^b(\theta_0) = \text{trace}(\mathcal{C}^b(\theta_0))$, where

$$
\mathcal{A}^b(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1} \left(\sum_{j=1}^{K} \psi_{\theta,\theta_j}(\theta_0)A^b_j(\theta_0) + \mathbb{A}^b_\theta(\theta_0)\right)
$$

$$
\mathcal{C}^b(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1} \left(\frac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K} \frac{\psi_{\theta,\theta_j,\theta_k}(\theta_0)A^b_j(\theta_0)A^b_k(\theta_0)}{T} + \sum_{j=1}^{K} \frac{\psi_{\theta,\theta_j}(\theta_0)C^b_j(\theta_0)}{T} + \sum_{j=1}^{K} \mathbb{A}^b_{\theta,\theta_j}(\theta_0)A^b_j(\theta_0) + \mathbb{C}^b(\theta_0)\right).
$$

Now for any matrix X with all eigenvalues smaller than 1 we have: $\log(I_K + X) = X - \frac{1}{2}X^2 + o(X)$. Furthermore, for any matrix M the determinant $|M| = \exp(\text{trace}(\log M)))$. Together, these imply that for arbitrary X_1, X_2:

$$\left|I + \frac{X_1}{\sqrt{T}} + \frac{X_2}{T} + o_p(\frac{1}{T})\right| = \exp\left(\text{trace}\left(\frac{X_1}{\sqrt{T}} + \frac{X_2}{T} + \frac{X_1^2}{T} + o_p(\frac{1}{T})\right)\right)$$
$$= 1 + \frac{\text{trace}\,(X_1)}{\sqrt{T}} + \frac{\text{trace}\,(X_2)}{T} + \frac{\text{trace}\,\left(X_1^2\right)}{T} + o_p(\frac{1}{T}).$$

Hence the required determinant is

$$\left|\widehat{\psi}_\theta^b(\theta^b)\right| = \left|\widehat{\psi}_\theta(\theta_0)\right|\left|I + \frac{\mathcal{A}^b(\theta_0)}{\sqrt{T}} + \frac{\mathcal{C}^b(\theta_0)}{T} + o_p(\frac{1}{T})\right| = \left|\widehat{\psi}_\theta(\theta_0)\right|\left(1 + \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{a_2^b(\theta_0)}{T} + \frac{c^b(\theta_0)}{T} + o_p(\frac{1}{T})\right).$$

ii. Expansion of $w^b(\theta^b) = |\widehat{\psi}_\theta(\theta^b)|^{-1}\pi(\theta^b)$:

$$\left|\widehat{\psi}_\theta^b(\theta^b)\right|^{-1}\pi(\theta^b) = \left|\widehat{\psi}_\theta(\theta_0)\right|^{-1}\left(1 + \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{a_2^b(\theta_0)}{T} + \frac{c^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)^{-1}\pi(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T}))$$
$$= \left|\widehat{\psi}_\theta(\theta_0)\right|^{-1}\left(1 - \frac{a^b(\theta_0)}{\sqrt{T}} - \frac{a_2^b(\theta_0)}{T} - \frac{c^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)$$
$$\times\left(\pi(\theta_0) + \pi_\theta(\theta_0)\frac{A^b(\theta_0)}{\sqrt{T}} + \pi_\theta(\theta_0)\frac{C^b(\theta_0)}{T} + \frac{1}{2}\sum_{j=1}^{K}\frac{\pi_{\theta,\theta_j}(\theta_0)A^b(\theta_0)A_j^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)$$
$$= \left|\widehat{\psi}_\theta(\theta_0)\right|^{-1}\pi(\theta_0)\left(1 - \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)}{\sqrt{T}} - \frac{a_2^b(\theta_0)}{T} - \frac{c^b(\theta_0)}{T}\right.$$
$$\left. - \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{a^b(\theta_0)A^b(\theta_0)}{T} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{C^b(\theta_0)}{T} + \frac{1}{2}\frac{A^b(\theta_0)\pi_{\theta,\theta'}(\theta_0)A^{b\prime}(\theta_0)}{T} + o_p(\frac{1}{T})\right).$$

Now $\overline{A}(\theta_0) = \frac{1}{B}\sum_{b=1}^{B}A^b(\theta_0)$. Similarly define $\overline{C}(\theta_0) = \frac{1}{B}C^b(\theta_0)$. Also, denote the term in $1/T$ by:

$$e^b(\theta_0) = -a_2^b(\theta_0) - c^b(\theta_0) - \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}a^b(\theta_0)A^b(\theta_0) + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}C^b(\theta_0) + \frac{1}{2}A^b(\theta_0)\pi_{\theta,\theta'}(\theta_0)A^{b\prime}(\theta_0).$$

The normalized weight for draw b is:

$$\overline{w}^b(\theta^b) = \frac{\left|\widehat{\psi}_\theta^b(\theta^b)\right|^{-1}\pi(\theta^b)}{\sum_{c=1}^{B}\left|\widehat{\psi}_\theta^c(\theta^c)\right|^{-1}\pi(\theta^c)} = \frac{1}{B}\left(\frac{1 - \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{e^b(\theta_0)}{T} + o_p(\frac{1}{T})}{1 + \frac{1}{B}\sum_{c=1}^{B} - \frac{a^c(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^c(\theta_0)}{\sqrt{T}} + \frac{e^c(\theta_0)}{T} + o_p(\frac{1}{T})}\right)$$
$$= \frac{1}{B}\left(\frac{1 - \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{e^b(\theta_0)}{T} + o_p(\frac{1}{T})}{1 - \frac{\overline{a}(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{\overline{A}(\theta_0)}{\sqrt{T}} + \frac{\overline{e}(\theta_0)}{T} + o_p(\frac{1}{T})}\right)$$
$$= \frac{1}{B}\left(1 - \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{e^b(\theta_0)}{T} + o_p(\frac{1}{T})\right) \times \left(1 + \frac{\overline{a}(\theta_0)}{\sqrt{T}} - \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{\overline{A}(\theta_0)}{\sqrt{T}} - \frac{\overline{e}(\theta_0)}{T} + o_p(\frac{1}{T})\right)$$
$$= \frac{1}{B}\left(1 - \frac{a^b(\theta_0) - \overline{a}(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0) - \overline{A}(\theta_0)}{\sqrt{T}} + \frac{e^b(\theta_0) - \overline{e}(\theta_0)}{T} - \frac{a^b(\theta_0)\overline{a}(\theta_0)}{T} - \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)\overline{a}(\theta_0)}{T}\right.$$
$$\left. - \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{\overline{A}(\theta_0)a^b(\theta_0)}{T} - \left[\frac{\pi_\theta(\theta_0)\pi_\theta(\theta_0)'}{\pi(\theta_0)^2}\right]\frac{A^b(\theta_0)'\overline{A}(\theta_0)}{T} + o_p(\frac{1}{T})\right).$$

The posterior mean is $\overline{\theta}_{RS} = \sum_{b=1}^{B}\overline{w}^b(\theta^b)\theta^b$. Using θ^b defined in (A.3), A and C defined in (A.4) and (A.5):

$$\overline{\theta}_{RS} = \theta_0 + \frac{1}{B}\sum_{b=1}^{B}\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{1}{B}\sum_{b=1}^{B}\frac{C^b(\theta_0)}{T} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{1}{B}\sum_{b=1}^{B}\frac{(A^b(\theta_0) - \overline{A}(\theta_0))A^b(\theta_0)}{T} + C^M(\theta_0) + o_p(\frac{1}{T}).$$

B.1 Proof of Results for LT

From

$$\theta^b = \theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T}),$$

we have, given that $\widehat{\psi}_b$ is drawn from the asymptotic distribution of $\widehat{\psi}$

$$\begin{aligned}\widehat{\psi}^b(\theta^b) &= \psi(\theta^b) + \frac{\mathbb{A}^b_\infty(\theta^b)}{\sqrt{T}} \\ &= \psi\bigg(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\bigg) + \frac{\mathbb{A}^b_\infty(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T}))}{\sqrt{T}} \\ &= \psi(\theta_0) + \frac{\mathbb{A}^b_\infty(\theta_0)}{\sqrt{T}} + \frac{\psi_\theta(\theta_0)A^b(\theta_0)}{\sqrt{T}} \frac{\mathbb{A}^b_{\infty,\theta}(\theta_0)A^b(\theta_0)}{T} + \frac{1}{2}\sum_{j=1}^{K} \frac{\psi_{\theta,\theta_j}(\theta_0)A^b(\theta_0)A^b_j(\theta_0)}{T} + o_p(\frac{1}{T})\end{aligned}$$

which is equal to $\widehat{\psi}$ for all b. Hence

$$A^b(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1}\left(\mathbb{A}(\theta_0) - \mathbb{A}^b_\infty(\theta_0)\right) \tag{B.1}$$

$$C^b(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1}\left(\mathbb{C}(\theta_0) - \frac{1}{2}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A^b(\theta_0)A^b_j(\theta_0) - \mathbb{A}^b_{\infty,\theta}(\theta_0)A^b(\theta_0)\right). \tag{B.2}$$

Note that the bias term C^b depends on the bias term $\mathbb{C}$. For the weights, we need to consider

$$\begin{aligned}\widehat{\psi}^b_\theta(\theta^b) &= \psi_\theta\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right) + \frac{\mathbb{A}^b_{\infty,\theta}\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)}{\sqrt{T}} \\ &= \psi_\theta(\theta_0) + \sum_{j=1}^{K}\frac{\psi_{\theta,\theta_j}(\theta_0)A^b_j(\theta_0)}{\sqrt{T}} + \frac{\mathbb{A}^b_{\infty,\theta}(\theta_0)}{\sqrt{T}} + \sum_{j=1}^{k}\frac{\psi_{\theta,\theta_j}(\theta_0)C^b_j(\theta_0)}{T} + \sum_{j=1}^{K}\frac{\mathbb{A}^b_{\infty,\theta,\theta_j}A^b_j(\theta_0)}{T} \\ &\quad + \frac{1}{2}\sum_{j,k=1}^{K}\frac{\psi_{\theta,\theta_j,\theta_k}(\theta_0)A^b_j(\theta_0)A^b_k(\theta_0)}{T} + o_p(\frac{1}{T}).\end{aligned}$$

Let

$$\mathcal{A}^b(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1}\left(\mathbb{A}^b_{\infty,\theta}(\theta_0) + \sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A^b_j(\theta_0)\right)$$

$$\mathcal{C}^b(\theta_0) = \Big[\psi_\theta(\theta_0)\Big]^{-1}\left(\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)C^b_j(\theta_0) + \sum_{j=1}^{K}\mathbb{A}^b_{\infty,\theta,\theta_j}(\theta_0)A^b_j(\theta_0) + \frac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K}\psi_{\theta,\theta_j,\theta_k}(\theta_0)A^b_j(\theta_0)A^b_k(\theta_0)\right)$$

$$a^b(\theta_0) = \text{trace}(\mathcal{A}^b(\theta_0)), \quad a^b_2(\theta_0) = \text{trace}(\mathcal{A}^b(\theta_0)^2), \quad c^b(\theta_0) = \text{trace}(\mathcal{C}^b(\theta_0)).$$

The determinant is

$$\begin{aligned}\Big|\widehat{\psi}^b_\theta(\theta_0)\Big|^{-1} &= \Big|\psi_\theta(\theta_0)\Big|^{-1}\Big|I + \frac{\mathcal{A}^b(\theta_0)}{\sqrt{T}} + \frac{\mathcal{C}^b(\theta_0)}{T} + o_p(\frac{1}{T})\Big|^{-1} = \Big|\psi_\theta(\theta_0)\Big|^{-1}\left(1 + \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{a^b_2(\theta_0)}{T} + \frac{c^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)^{-1} \\ &= \Big|\psi_\theta(\theta_0)\Big|^{-1}\left(1 - \frac{a^b(\theta_0)}{\sqrt{T}} - \frac{a^b_2(\theta_0)}{T} - \frac{c^b(\theta_0)}{T} + o_p(\frac{1}{T})\right).\end{aligned}$$

The prior is

$$\pi(\theta^b) = \pi\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)$$
$$= \pi(\theta_0) + \pi_\theta(\theta_0)\frac{A^b(\theta_0)}{\sqrt{T}} + \pi_\theta(\theta_0)\frac{C^b(\theta_0)}{T} + \frac{1}{2}\frac{A^b(\theta_0)\pi_{\theta,\theta'}A^{b\prime}(\theta_0)}{T} + o_p(\frac{1}{T}).$$

Let: $e^b(\theta_0) = -c^b(\theta_0) - a_2^b(\theta_0) + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}C^b(\theta_0) + A^b(\theta_0)\frac{\pi_{\theta,\theta'}}{\pi}(\theta_0)A^{b\prime}(\theta_0)$. After some simplification, the product is

$$\left|\widehat{\psi}_\theta^b(\theta_0)\right|^{-1}\pi(\theta^b) = \left|\psi_\theta(\theta_0)\right|^{-1}\pi(\theta_0)\Big(1 - \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{e^b(\theta_0)}{T} + o_p(\frac{1}{T})\Big).$$

Hence, the normalized weight for draw b is

$$\overline{w}^b(\theta^b) = \frac{\left|\widehat{\psi}_\theta^b(\theta_0)\right|^{-1}\pi(\theta^b)}{\sum_{c=1}^B\left|\widehat{\psi}_\theta^c(\theta_0)\right|^{-1}\pi(\theta^c)} = \frac{1}{B}\frac{1 - \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{e^b(\theta_0)}{T} + o_p(\frac{1}{T})}{1 - \frac{\overline{a}(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{\overline{A}(\theta_0)}{\sqrt{T}} + \frac{\overline{e}(\theta_0)}{T} + o_p(\frac{1}{T})}$$
$$= \frac{1}{B}\left(1 - \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{e^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)\left(1 + \frac{\overline{a}(\theta_0)}{\sqrt{T}} - \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{\overline{A}(\theta_0)}{\sqrt{T}} - \frac{\overline{e}(\theta_0)}{T} + o_p(\frac{1}{T})\right)$$
$$= \frac{1}{B}\Big(1 - \frac{a^b(\theta_0) - \overline{a}(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0) - \overline{A}(\theta_0)}{\sqrt{T}} + \frac{e^b(\theta_0) - \overline{e}(\theta_0)}{T} - \frac{a^b(\theta_0)\overline{a}(\theta_0)}{T} - \frac{\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}A^b(\theta_0)\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\overline{A}(\theta_0)}{T}$$
$$+ \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{a^b(\theta_0)\overline{A}(\theta_0)}{T} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{\overline{a}(\theta_0)A^b(\theta_0)}{T} + o_p(\frac{1}{T})\Big).$$

Hence the posterior mean is $\overline{\theta}_{\mathrm{LT}} = \sum_{b=1}^B \overline{w}^b(\theta^b)\theta^b$ and $\theta^b = \Big(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\Big)$. After simplification, we have

$$\overline{\theta}_{\mathrm{LT}} = \theta_0 + \frac{\overline{A}(\theta_0)}{\sqrt{T}} + \frac{\overline{C}(\theta_0)}{T} - \frac{1}{B}\sum_{b=1}^B\frac{(a^b(\theta_0) - \overline{a}(\theta_0))A^b(\theta_0)}{T} - \frac{[\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\overline{A}(\theta_0)]^2\theta_0}{T} + \frac{1}{B}\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\sum_{b=1}^B\frac{(A^b(\theta_0) - \overline{A}(\theta_0))A^b(\theta_0)}{T}$$
$$- \frac{\overline{a}(\theta_0)^2\theta_0}{T} + 2\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{\overline{a}(\theta_0)\overline{A}(\theta_0)\theta_0}{T} + o_p(\frac{1}{T})$$
$$= \theta_0 + \frac{\overline{A}(\theta_0)}{\sqrt{T}} + \frac{\overline{C}(\theta_0)}{T} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{1}{B}\sum_{b=1}^B\frac{(A^b(\theta_0) - \overline{A}(\theta_0))A^b(\theta_0)}{T} + C^M(\theta_0) + o_p(\frac{1}{T}),$$

where all terms are based on $A^b(\theta_0)$ defined in (B.1) and $C^b(\theta_0)$ in (B.2).

C.1 Results for SLT:

From

$$\widehat{\psi}^b(\theta) = \frac{1}{S}\sum_{s=1}^S\widehat{\psi}^s(\theta) + \frac{\mathbb{A}_\infty^b(\theta)}{\sqrt{T}}$$
$$\widehat{\psi}^s(\theta) = \psi(\theta) + \frac{\mathbb{A}^s(\theta)}{\sqrt{T}} + \frac{\mathbb{C}^s(\theta)}{T} + o_p(\frac{1}{T})$$
$$\theta^b = \theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T}),$$

we have

$$\widehat{\psi}^s(\theta^b) = \frac{1}{S}\sum_{s=1}^{S}\widehat{\psi}^s\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right) + \frac{\mathbb{A}^b_\infty(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T}))}{\sqrt{T}}$$

$$= \psi(\theta_0) + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{A}^s(\theta_0)}{\sqrt{T}} + \frac{\mathbb{A}^b_\infty(\theta_0)}{\sqrt{T}} + \psi_\theta(\theta_0)\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{A}^s_\theta(\theta_0)A^b(\theta_0)}{T} + \frac{\mathbb{A}^b_{\infty,\theta}(\theta_0)A^b(\theta_0)}{T}$$

$$+ \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{C}^s(\theta_0)}{T} + \frac{1}{2}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)\frac{A^b(\theta_0)A^b_j(\theta_0)}{T} + \psi_\theta(\theta_0)\frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T}).$$

Thus,

$$A^b(\theta_0) = \left[\psi_\theta(\theta_0)\right]^{-1}\left(\mathbb{A}(\theta_0) - \frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s(\theta_0) - \mathbb{A}^b_\infty(\theta_0)\right) \tag{C.1}$$

$$C^b(\theta_0) = \left[\psi_\theta(\theta_0)\right]^{-1}\left(\mathbb{C}(\theta_0) - \frac{1}{S}\sum_{s=1}^{S}\mathbb{C}^s(\theta_0) - \frac{1}{2}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A^b(\theta_0)A^b_j(\theta_0)\right)$$
$$-\left[\psi_\theta(\theta_0)\right]^{-1}\left[\frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s_\theta(\theta_0) + \mathbb{A}^b_{\infty,\theta}(\theta_0)\right]A^b(\theta_0). \tag{C.2}$$

Note that we have $\mathbb{A}^b_\infty \sim \mathcal{N}$ while $\mathbb{A}^s \xrightarrow{d} \mathcal{N}$. To compute the weight for draw b, consider

$$\widehat{\psi}^b(\theta^b) = \psi_\theta\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right) + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{A}^s\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)}{\sqrt{T}}$$

$$+ \frac{\mathbb{A}^b_\infty\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)}{\sqrt{T}} + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{C}^s\left(\theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)}{T} + o_p(\frac{1}{T})$$

$$= \psi_\theta(\theta_0) + \sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)\frac{A^b_j(\theta_0)}{\sqrt{T}} + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{A}^s_\theta(\theta_0)}{\sqrt{T}} + \frac{\mathbb{A}^b_{\infty,\theta}(\theta_0)}{\sqrt{T}} + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{C}^s(\theta_0)}{T} + \sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)\frac{C^b_j(\theta_0)}{T}$$

$$+ \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{K}\frac{\mathbb{A}^s_{\theta,\theta_j}(\theta_0)A^b_j(\theta_0)}{T} + \sum_{j=1}^{K}\frac{\mathbb{A}^b_{\infty,\theta,\theta_j}(\theta_0)A^b_j(\theta_0)}{T} + \frac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K}\psi_{\theta,\theta_j,\theta_k}(\theta_0)\frac{A^b_k(\theta_0)A^b_j(\theta_0)}{T} + o_p(\frac{1}{T}).$$

Let:

$$\mathcal{A}^b(\theta_0) = \left[\psi_\theta(\theta_0)\right]^{-1}\left(\frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s_\theta(\theta_0) + \mathbb{A}^b_{\infty,\theta}(\theta_0) + \sum_{j=1}^{K}\psi_{\theta,\theta_j}A^b_j(\theta_0)\right)$$

$$\mathcal{C}^b(\theta_0) = \left[\psi_\theta(\theta_0)\right]^{-1}\left(\frac{1}{S}\sum_{s=1}^{S}\mathbb{C}^s(\theta_0) + \sum_{j=1}^{K}\left[\psi_{\theta,\theta_j}(\theta_0)C^b_j(\theta_0) + \frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s_{\theta,\theta_j}(\theta_0)A^b_j(\theta_0) + \mathbb{A}^b_{\infty,\theta,\theta_j}(\theta_0)A^b_j(\theta_0)\right]\right)$$
$$+ \left[\psi_\theta(\theta_0)\right]^{-1}\left(\frac{1}{2}\sum_{j,k=1}^{K}\psi_{\theta,\theta_j,\theta_k}(\theta_0)A^b_k(\theta_0)A^b_j(\theta_0)\right)$$

$$a^b(\theta_0) = \text{trace}(\mathcal{A}^b(\theta_0)), \quad a^b_2(\theta_0) = \text{trace}(\mathcal{A}^b(\theta_0)^2), \quad c^b(\theta_0) = \text{trace}(\mathcal{C}^b(\theta_0)).$$

The determinant is

$$\left|\widehat{\psi}^b(\theta^b)\right|^{-1} = \left|\psi_\theta(\theta_0)\right|^{-1}\left(1 - \frac{a^b(\theta_0)}{\sqrt{T}} - \frac{a_2^b(\theta_0)}{T} - \frac{c^b(\theta_0)}{T} + o_p(\frac{1}{T})\right).$$

Hence

$$\begin{aligned}
\left|\widehat{\psi}^b(\theta^b)\right|^{-1}\pi(\theta^b) &= \left|\psi_\theta(\theta_0)\right|^{-1}\pi(\theta_0)\left(1 - \frac{a^b(\theta_0)}{\sqrt{T}} - \frac{a_2^b(\theta_0)}{T} - \frac{c^b(\theta_0)}{T} + o_p(\frac{1}{T})\right) \\
&\times\left(1 + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{C^b(\theta_0)}{T} + \frac{1}{2}\sum_{j=1}^{K}\frac{\pi_{\theta,\theta_j}(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)A_j^b(\theta_0)}{T} + o_p(\frac{1}{T})\right) \\
&= \left|\psi_\theta(\theta_0)\right|^{-1}\pi(\theta_0)\left(1 - \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{e^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)
\end{aligned}$$

where $e^b(\theta_0) = -a^b(\theta_0)\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}A^b(\theta_0) - a_2^b(\theta_0) - c^b(\theta_0) + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}C^b(\theta_0) + \frac{1}{2}\sum_{j=1}^{K}\frac{\pi_{\theta,\theta_j}(\theta_0)}{\pi(\theta_0)}A^b(\theta_0)A_j^b(\theta_0)$. The normalized weights are

$$\begin{aligned}
\overline{w}^b(\theta^b) &= \frac{\left|\widehat{\psi}^b(\theta^b)\right|^{-1}\pi(\theta^b)}{\sum_{c=1}^{B}\left|\widehat{\psi}^c(\theta^c)\right|^{-1}\pi(\theta^c)} \\
&= \frac{1}{B}\left(1 - \frac{a^b(\theta_0)}{\sqrt{T}} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{A^b(\theta_0)}{\sqrt{T}} + \frac{e^b(\theta_0)}{T} + o_p(\frac{1}{T})\right)\left(1 + \frac{\overline{a}(\theta_0)}{\sqrt{T}} - \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{\overline{A}(\theta_0)}{\sqrt{T}} - \frac{\overline{e}(\theta_0)}{T} + o_p(\frac{1}{T})\right).
\end{aligned}$$

The posterior mean $\overline{\theta}_{\text{SLT}} = \sum_{b=1}^{B}\overline{w}^b(\theta^b)\theta^b$ with $\theta^b = \theta_0 + \frac{A^b(\theta_0)}{\sqrt{T}} + \frac{C^b(\theta_0)}{T} + o_p(\frac{1}{T})$. After some simplification,

$$\begin{aligned}
\overline{\theta}_{SLT} &= \theta_0 + \frac{\overline{A}(\theta_0)}{\sqrt{T}} + \frac{\overline{C}(\theta_0)}{T} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{1}{B}\sum_{B=1}^{B}\frac{(A^b(\theta_0) - \overline{A}(\theta_0))A^b(\theta_0)}{T} - \frac{1}{B}\sum_{b=1}^{B}\frac{(a^b(\theta_0) - \overline{a}(\theta_0))A^b(\theta_0)}{T} \\
&\quad + 2\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{\overline{a}(\theta_0)\overline{A}(\theta_0)\theta_0}{T} - \frac{\overline{a}^2(\theta_0)\theta_0}{T} - [\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\overline{A}(\theta_0)]^2\frac{\theta_0}{T} + o_p(\frac{1}{T}) \\
&= \theta_0 + \frac{\overline{A}(\theta_0)}{\sqrt{T}} + \frac{\overline{C}(\theta_0)}{T} + \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\frac{1}{B}\sum_{B=1}^{B}\frac{(A^b(\theta_0) - \overline{A}(\theta_0))A^b(\theta_0)}{T} + C^M(\theta_0) + o_p(\frac{1}{T})
\end{aligned}$$

where terms in A and C are defined from (C.1) and (C.2).

D.1 Results For The Example in Section 6.1

The data generating process is $y_t = m_0 + \sigma_0 e_t$, $e_t \sim iid\, \mathcal{N}(0,1)$. As a matter of notation, a hat is used to denote the mode, a bar denotes the mean, superscript s denotes a specific draw and a subscript S to denote average over S draws. For example, $\overline{e}_S = \frac{1}{ST}\sum_{s=1}^S \sum_{t=1}^T e_t^s = \frac{1}{S}\sum_{s=1}^S \overline{e}^s$.

MLE: Define $\overline{e} = \frac{1}{T}\sum_{t=1}^T e_t$. Then the mean estimator is $\widehat{m} = m_0 + \sigma_0 \overline{e} \sim N(0, \sigma_0^2/T)$. For the variance estimator, $\widehat{e} = y - \widehat{m} = \sigma_0(e - \overline{e}) = \sigma_0 M e$, $M = I_T - 1(1'1)^{-1}1'$ is an idempotent matrix with $T-1$ degrees of freedom. Hence $\widehat{\sigma}^2_{ML} = \widehat{e}'\widehat{e}/T \sim \sigma_0^2 \chi^2_{T-1}$.

BC: Expressed in terms of sufficient statistics $(\widehat{m}, \widehat{\sigma}^2)$, the joint density of $\mathbf{y}$ is

$$p(\mathbf{y}; m, \sigma^2) = (\frac{1}{2\pi\sigma^2})^{T/2} \exp\left(- \frac{\sum_{t=1}^T (m - \widehat{m})^2}{2\sigma^2} \times \frac{-T\widehat{\sigma}^2}{2\sigma^2} \right).$$

The flat prior is $\pi(m, \sigma^2) \propto 1$. The marginal posterior distribution for σ^2 is $p(\sigma^2|\mathbf{y}) = \int_{-\infty}^{\infty} p(\mathbf{y}|m, \sigma^2) dm$. Using the result that $\int_{-\infty}^{\infty} \exp(-\frac{T}{2\sigma^2}(m - \widehat{m})^2) dm = \sqrt{2\pi\sigma^2}$, we have

$$p(\sigma^2|\mathbf{y}) \quad \propto \quad (2\pi\sigma^2)^{-(T-1)/2} \exp(-T\widehat{\sigma}^2/2\sigma^2) \sim \text{inv}\Gamma\left(\frac{T-3}{2}, \frac{T\widehat{\sigma}^2}{2}\right).$$

The mean of an $\text{inv}\Gamma(\alpha, \beta)$ is $\frac{\beta}{\alpha - 1}$. Hence the BC posterior is $\overline{\sigma}^2_{BC} = E(\sigma^2|\mathbf{y}) = \widehat{\sigma}^2 \frac{T}{T-5}$.

SMD: The estimator equates the auxiliary statistics computed from the sample with the average of the statistics over simulations. Given σ, the mean estimator $\widehat{m}_S$ solves $\widehat{m} = \widehat{m}_S + \sigma \frac{1}{S}\sum_{s=1}^S \overline{e}^s$. Since we use sufficient statistics, $\widehat{m}$ is the ML estimator. Thus, $\widehat{m}_S \sim \mathcal{N}(m, \frac{\sigma_0^2}{T} + \frac{\sigma^2}{ST})$. Since $y_t^s - \overline{y}_t^s = \sigma(e_t^s - \overline{e}^s)$, the variance estimator $\widehat{\sigma}^2_S$ is the σ^2 that solves $\widehat{\sigma}^2 = \sigma^2(\frac{1}{ST}\sum_{s=1}^S \sum_{t=1}^T (e_t^s - \overline{e}^s)^2)$ Hence

$$\widehat{\sigma}^2_S = \frac{\widehat{\sigma}^2}{\frac{1}{ST}\sum_s \sum_t (\widehat{e}_t^s - \overline{e}^s)^2} = \sigma^2 \frac{\chi^2_{T-1}/T}{\chi^2_{S(T-1)}/(ST)} = \sigma^2 F_{T-1, S(T-1)}.$$

The mean of a F_{d_1, d_2} random variable is $\frac{d_2}{d_2 - 2}$. Hence $E(\widehat{\sigma}^2_{SMD}) = \sigma^2 \frac{(T-1)}{S(T-1)-2}$.

LT: The LT is defined as

$$p_{\text{LT}}(\sigma^2|\widehat{\sigma}^2) \propto \mathbb{1}_{\sigma^2 \geq 0} \exp\left(-\frac{T}{2} \frac{\left(\widehat{\sigma}^2 - \sigma^2\right)^2}{2\widehat{\sigma}^4} \right)$$

which implies

$$\sigma^2|\widehat{\sigma}^2 \sim_{\text{LT}} \mathcal{N}\left(\widehat{\sigma}^2, \frac{2\widehat{\sigma}^4}{T}\right) \text{ truncated to } [0, +\infty[.$$

For $X \sim \mathcal{N}(\mu, \sigma^2)$ we have $\mathbb{E}(X|X > a) = \mu + \frac{\phi(\frac{a-\mu}{\sigma})}{1 - \Phi(\frac{a-\mu}{\sigma})}\sigma$ (Mills-Ratio). Hence:

$$\mathbb{E}_{\text{LT}}(\sigma^2|\widehat{\sigma}^2) = \widehat{\sigma}^2 + \frac{\phi(\frac{0 - \widehat{\sigma}^2}{\sqrt{2/T}\widehat{\sigma}^2})}{1 - \Phi(\frac{0 - \widehat{\sigma}^2}{\sqrt{2/T}\widehat{\sigma}^2})}\sqrt{2/T}\widehat{\sigma}^2 = \widehat{\sigma}^2\left(1 + \sqrt{\frac{2}{T}}\frac{\phi(-\sqrt{T/2})}{1 - \Phi(-\sqrt{T/2})}\right).$$

Let $\kappa_{\text{LT}} = \sqrt{\frac{2}{T}}\frac{\phi(-\sqrt{T/2})}{1 - \Phi(-\sqrt{T/2})}$. We have $\mathbb{E}_{\text{LT}}(\sigma^2|\widehat{\sigma}^2) = \widehat{\sigma}^2(1 + \kappa_{\text{LT}})$. The expectation of the estimator is

$$\mathbb{E}\left(\mathbb{E}_{\text{LT}}(\sigma^2|\widehat{\sigma}^2)\right) = \sigma^2 \frac{T-1}{T}(1 + \kappa_{\text{LT}})$$

from which we deduce the bias of the estimator

$$\mathbb{E}\left(\mathbb{E}_{\mathrm{LT}}(\sigma^2|\widehat{\sigma}^2)\right) - \sigma^2 = \sigma^2\left(\frac{T-1}{T}\kappa_{\mathrm{LT}} - \frac{1}{T}\right).$$

The variance of the estimator is $2\sigma^4\frac{T-1}{T^2}\left(1+\kappa_{\mathrm{LT}}\right)^2$ and the Mean-Squared Error (MSE)

$$\sigma^4\left(2\frac{T-1}{T^2}\left(1+\kappa_{\mathrm{LT}}\right)^2 + \left(\frac{T-1}{T}\kappa_{\mathrm{LT}} - \frac{1}{T}\right)^2\right)$$

which is the squared bias of MLE plus terms that involve the Mills-Ratio (due to the truncation).

SLT: The SLT is defined as

$$p_{\mathrm{SLT}}(\sigma^2|\widehat{\sigma}^2) \propto \mathbb{1}_{\sigma^2\geq 0}\exp\left(-\frac{T}{2}\frac{\left(\widehat{\sigma}^2 - \sigma^2\frac{\chi^2_{S(T-1)}}{ST}\right)^2}{2\widehat{\sigma}^4}\right) = \mathbb{1}_{\sigma^2\geq 0}\exp\left(-\frac{T[\frac{\chi^2_{S(T-1)}}{ST}]^2}{2}\frac{\left(\widehat{\sigma}^2/\frac{\chi^2_{S(T-1)}}{ST} - \sigma^2\right)^2}{2\widehat{\sigma}^4}\right)$$

where

$$\widehat{\sigma}^2_S = \sigma^2\frac{1}{S}\sum_{s=1}^{2}\frac{1}{T}\sum_{t=1}^{T}(e^s_t - \overline{e}^s)^2 = \sigma^2\frac{\chi^2_{S(T-1)}}{ST}.$$

This yields the slightly more complicated formula

$$\sigma^2|\widehat{\sigma}^2, (e^s)_{s=1,\ldots,S} \sim \mathcal{N}\left(\widehat{\sigma}^2/\frac{\chi^2_{S(T-1)}}{ST}, \frac{2\widehat{\sigma}^4}{T}[\frac{ST}{\chi^2_{S(T-1)}}]^2\right)$$

and the posterior mean becomes

$$\begin{aligned}
\mathbb{E}_{\mathrm{SLT}}(\sigma^2|\widehat{\sigma}^2) &= \widehat{\sigma}^2\frac{ST}{\chi^2_{S(T-1)}} + \frac{\phi\left(-\frac{\widehat{\sigma}^2 ST/\chi^2_{S(T-1)}}{\sqrt{\frac{2\widehat{\sigma}^4}{T}(\frac{ST}{\chi^2_{S(T-1)}})^2}}\right)}{1-\Phi\left(-\frac{\widehat{\sigma}^2 ST/\chi^2_{S(T-1)}}{\sqrt{\frac{2\widehat{\sigma}^4}{T}(\frac{ST}{\chi^2_{S(T-1)}})^2}}\right)}\sqrt{2/T}\frac{ST}{\chi^2_{S(T-1)}}\widehat{\sigma}^2 \\
&= \widehat{\sigma}^2\frac{ST}{\chi^2_{S(T-1)}} + \frac{\phi\left(-\sqrt{T/2}\right)}{1-\Phi\left(-\sqrt{T/2}\right)}\sqrt{2/T}\frac{ST}{\chi^2_{S(T-1)}}\widehat{\sigma}^2.
\end{aligned}$$

Let $\kappa_{\mathrm{SLT}} = \frac{\phi(-\sqrt{T/2})}{1-\Phi(-\sqrt{T/2})}\sqrt{2/T}\frac{ST}{\chi^2_{S(T-1)}} = \kappa_{\mathrm{LT}}\frac{ST}{\chi^2_{S(T-1)}}$ (random). We can compute

$$\mathbb{E}\left(\mathbb{E}_{\mathrm{SLT}}(\sigma^2|\widehat{\sigma}^2)\right) = \sigma^2\frac{S(T-1)}{S(T-1)-2} + \sigma^2\frac{T-1}{T}\mathbb{E}(\kappa_{\mathrm{SLT}})$$

and the bias

$$\mathbb{E}\left(\mathbb{E}_{\mathrm{SLT}}(\sigma^2|\widehat{\sigma}^2)\right) - \sigma^2 = \sigma^2\frac{2}{S(T-1)-2} + \sigma^2\frac{T-1}{T}\mathbb{E}(\kappa_{\mathrm{SLT}})$$

which is the bias of SMD and the Mills-Ratio term that comes from taking the mean of the truncated normal rather than the mode. The variance is similar to the LT and the SMD

$$2\sigma^4\kappa_1\frac{1}{T-1} + 2\sigma^4\mathbb{V}(\kappa_{\mathrm{SLT}}) + 4\sigma^4\frac{T-1}{T^2}\mathrm{Cov}(\kappa_{\mathrm{SLT}}, \frac{S}{\chi^2_{S(T-1)}}).$$

The extra term is due to κ_{SLT} being random. We could simplify further noting that $\kappa_{\text{SLT}} = \kappa_{\text{LT}} \frac{ST}{\chi^2_{S(T-1)}}$, $\mathbb{E}(\kappa_{\text{SLT}}) = \kappa_{\text{LT}} \frac{ST}{S(T-1)-2}$, $\mathbb{V}(\kappa_{\text{SLT}}) = \kappa^2_{\text{LT}} \frac{S^2T^2}{(S(T-1)-2)^2(S(T-1)-4)}$ and $\text{Cov}(\kappa_{\text{SLT}}, \frac{S}{\chi^2_{S(T-1)}}) = \kappa_{\text{LT}} S^2 T \mathbb{V}(1/\chi^2_{S(T-1)}) = \kappa_{\text{LT}} \frac{S^2T}{(S(T-1)-2)^2(S(T-1)-4)}$.

The MSE is

$$\sigma^4 \left[\frac{2}{S(T-1)-2} + \frac{T-1}{T} \mathbb{E}(\kappa_{\text{SLT}}) \right]^2 + 2\sigma^4 \kappa_1 \frac{1}{T-1} + 2\sigma^4 \mathbb{V}(\kappa_{\text{SLT}}) + 4\sigma^4 \frac{T-1}{T^2} \text{Cov}(\kappa_{\text{SLT}}, \frac{S}{\chi^2_{S(T-1)}})$$

$$= \underbrace{2\sigma^4 \left[\frac{2}{[S(T-1)-2]^2} + \kappa_1 \frac{1}{T-1} \right]}_{\text{MSE of SMD}} + \frac{(T-1)^2}{T^2} \mathbb{E}(\kappa^2_{\text{SLT}} + \frac{4\sigma^4}{S(T-1)-2} \frac{T-1}{T} \mathbb{E}(\kappa_{\text{SLT}})$$

$$+ 2\sigma^4 \mathbb{V}(\kappa_{\text{SLT}}) + 4\sigma^4 \frac{T-1}{T^2} \text{Cov}(\kappa_{\text{SLT}}, \frac{S}{\chi^2_{S(T-1)}}).$$

RS: The auxiliary statistic for each draw of simulated data is matched to the sample auxiliary statistic. Thus, $\widehat{m} = m^b + \sigma^b \overline{e}^b$. Thus conditional on $\widehat{m}$ and $\sigma^{2,b}$, $m^b = \widehat{m} - \sigma^b \overline{e}^b \sim \mathcal{N}(0, \sigma^{2,b}/T)$. For the variance, $\widehat{\sigma}^{2,b} = \sigma^{2,b} \sum_t (e^b_t - \overline{e}^b)^2/T$. Hence

$$\sigma^{2,b} = \frac{\widehat{\sigma}^2}{\sum_t (e^b_t - \overline{e}^b)^2/T} = \sigma^2 \frac{\sum_t (e_t - \overline{e})^2/T}{\sum_t (e^b_t - \overline{e}^b)^2/T} \sim \text{inv}\Gamma\left(\frac{T-1}{2}, \frac{T\widehat{\sigma}^2}{2} \right)$$

Note that $p_{BC}(\sigma^2|\widehat{\sigma}^2) \sim \text{inv}\Gamma\left(\frac{T-3}{2}, \frac{T\widehat{\sigma}^2}{2} \right)$ under a flat prior, the Jacobian adjusts to the posterior to match the true posterior. To compute the posterior mean, we need to compute the Jacobian of the transformation: $|\psi_\theta|^{-1} = \frac{\partial \sigma^{2,s}}{\partial \widehat{\sigma}^2}$[9]. Since $\sigma^{2,b} = \frac{T\widehat{\sigma}^2}{\sum_t (e^b_t - \overline{e}^b)^2}$, $|\psi_\theta|^{-1} = \frac{T}{\sum_t (e^b_t - \overline{e}^b)^2}$.

Under the prior $p(\sigma^{2,s}) \propto 1$, the posterior mean without the Jacobian transformation is

$$\overline{\sigma}^2 = \sigma^2 \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_t (e_t - \overline{e})^2/T}{\sum_t (e^b_t - \overline{e}^b)^2/T} \stackrel{B\to\infty}{\longrightarrow} \widehat{\sigma}^2 \frac{T}{T-3}$$

The posterior mean after adjusting for the Jacobian transformation is

$$\overline{\sigma}^2_{RS} \quad = \quad \frac{\sum_{b=1}^B \sigma^{2,b} \cdot \frac{T}{\sum_t (e^b_t - \overline{e}^b)^2}}{\sum_{b=1}^B 1/\sigma^{2,b}} = \widehat{\sigma}^2 \frac{\sum_b (\frac{T}{\sum_t (e^b_t - \overline{e}^b)^2})^2}{\sum_{b=1} \sum_t (e^b_t - \overline{e}^b)^2/T} = T\widehat{\sigma}^2 \frac{\frac{1}{B} \sum_b (z^b)^2}{\frac{1}{B} \sum_b z^b}$$

where $1/z^b = \sum_t (e^b_t - \overline{e}^b)^2$. As $B \to \infty$, $\frac{1}{B} \sum_b (z^b)^2 \stackrel{p}{\longrightarrow} E[(z^b)^2]$ and $\frac{1}{S} \sum_b z^b \stackrel{p}{\longrightarrow} E[z^b]$. Now $z^b \sim \text{inv}\chi^2_{T-1}$ with mean $\frac{1}{T-3}$ and variance $\frac{2}{(T-3)^2(T-5)}$ giving $E[(z^b)^2] = \frac{1}{(T-3)(T-5)}$. Hence as $B \to \infty$, $\overline{\sigma}^2_{RS,R} = \widehat{\sigma}^2 \frac{T}{T-5} = \overline{\sigma}^2_{BC}$.

Derivation of the Bias Reducing Prior The bias of the MLE estimator has $\mathbb{E}(\widehat{\sigma}) = \sigma^2 - \frac{1}{T}\sigma^2$ and variance $V(\widehat{\sigma}^2) = 2\sigma^4(\frac{1}{T} - \frac{1}{T^2})$. Since the auxiliary parameters coincide with the parameters of interest, $\nabla_\theta \psi(\theta)$ and $\nabla_{\theta\theta'} \psi(\theta) = 0$. For $Z \sim \mathcal{N}(0,1)$, $A(v; \sigma^2) = \sqrt{2}\sigma^2(1 - \frac{1}{T})Z$, Thus $\partial_{\sigma^2} A(v; \sigma^2) = \sqrt{2}(1 - \frac{1}{T})Z, a^s =$

[9]This holds because $\widehat{\sigma}^{2,b}(\sigma^{2,b}) = \widehat{\sigma}^2$ so that $|d\widehat{\sigma}^{2,b}/d\sigma^{2,b}|^{-1} = |d\sigma^{2,b}/d\widehat{\sigma}^2|$.

$\sqrt{2}\sigma^2(1-\frac{1}{T})(Z-Z^s)$. The terms in the asymptotic expansion are therefore

$$\begin{aligned}
\partial_{\sigma^2} A(v^s;\sigma^2)a^s &= 2\sigma^2(1-\frac{1}{T})^2 Z^s(Z-Z^s) \Rightarrow \mathbb{E}(\partial_{\sigma^2} A(v^s;\sigma^2)a^s) = -\sigma^2 2(1-\frac{1}{T})^2 \\
V(a^s) &= 4\sigma^4(1-\frac{1}{T})^2 \\
cov(a^s,a^{s'}) &= 2(1-\frac{1}{T})^2\sigma^4 \\
(1-\frac{1}{S})V(a^s)+\frac{S-1}{S}cov(a^s,a^{s'}) &= \sigma^4(1-\frac{1}{T})^2\Big(4(1-\frac{1}{S})+2\frac{S-1}{S}\Big) = \frac{\sigma^2 S}{3(S-1)}
\end{aligned}$$

Noting that $|\partial_{\widehat{\sigma}^2}\sigma^{2,b}| \propto \sigma^{2,b}$, it is analytically simpler in this example to solve for the weights directly, i.e. $w(\sigma^2)=\pi(\sigma^2)|\partial_{\widehat{\sigma}^2}\sigma^{2,b}|$ rather than the bias reducing prior π itself. Thus the bias reducing prior satisfies

$$\partial_{\sigma^2} w(\sigma^2) = \frac{-2\sigma^2(1-\frac{1}{T})^2}{\sigma^4(1-\frac{1}{T})^2\Big(4(1-\frac{1}{S})+2\frac{S-1}{S}\Big)} = -\frac{1}{\sigma^2}\frac{2}{4(1-\frac{1}{S})+2\frac{S-1}{S}}.$$

Taking the integral on both sides we get:

$$\log(w(\sigma^2)) \propto -\log(\sigma^2) \Rightarrow w(\sigma^2) \propto \frac{1}{\sigma^2} \Rightarrow \pi(\sigma^2) \propto \frac{1}{\sigma^4}$$

which is the Jeffreys prior if there is no re-weighting and the square of the Jeffreys prior when we use the Jacobian to re-weight. Since the estimator for the mean was unbiased, $\pi(m) \propto 1$ is the prior for m.

The posterior mean under the Bias Reducing Prior $\pi(\sigma^{2,s}) = 1/\sigma^{4,s}$ is the same as the posterior without weights but using the Jeffreys prior $\pi(\sigma^{2,s}) = 1/\sigma^{2,s}$:

$$\overline{\sigma}^2_{RS} = \frac{\sum_{s=1}^S \sigma^{2,s}(1/\sigma^{2,s})}{\sum_{s=1}^S 1/\sigma^{2,s}} = \frac{S}{\sum_{s=1}^S 1/\sigma^{2,s}} = \sigma^2\frac{\sum_{t=1}^T (e_t-\overline{e})^2/T}{\sum_{s=1}^S\sum_{t=1}^T (e_t^s-\overline{e}^s)^2/(ST)} \equiv \widehat{\sigma}^2_{SMD}.$$

D.2 Further Results for Dynamic Panel Model with Fixed Effects

Table 4: Dynamic Panel $\rho = 0.9, \beta = 1, \sigma^2 = 2$

Mean over 1000 replications

		MLE	LT	SLT	SMD	ABC	RS	Bootstrap
	Mean	0.751	0.751	0.895	0.898	0.889	0.899	0.751
$\widehat{\rho}$:	SD	0.030	0.030	0.026	0.025	0.025	0.025	0.059
	Bias	-0.149	-0.149	-0.005	-0.002	-0.011	-0.001	-0.149
	Mean	0.934	0.934	0.998	1.000	0.996	1.000	0.935
$\widehat{\beta}$:	SD	0.070	0.071	0.074	0.073	0.073	0.073	0.139
	Bias	-0.066	-0.066	-0.002	0.000	-0.004	0.000	-0.065
	Mean	1.857	1.865	1.972	1.989	2.054	2.097	1.858
$\widehat{\sigma}^2$:	SD	0.135	0.141	0.145	0.145	0.151	0.153	0.269
	Bias	-0.143	-0.135	-0.028	-0.011	0.054	0.097	-0.142
S		–	–	500	500	1	1	500
B		–	500	500	–	500	500	–

See note to Table 3.

References

Arellano, M. and Bonhomme, S. 2009, Robust Priors in Nonlinear Panel Data Models, *Econometrica* **77**(2), 489–536.

Bao, Y. and Ullah, A. 2007, The Second-Order Bias and Mean-Squared Error of Estimators in Time Series Models, *Journal of Econometrics* **140**(2), 650–669.

Beaumont, M., Zhang, W. and Balding, D. 2002, Approximate Bayesian Computation in Population Genetics, *Genetics* **162**, 2025–2035.

Bester, A. and Hansen, C. 2006, Bias Reduction for Bayesian and Frequentist Estimators, Mimeo, University of Chicago.

Blum, M., Nunes, M., Prangle, D. and Sisson, A. 2013, A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation, *Statistical Science* **28**(2), 189–208.

Cabrera, J. and Fernholz, L. 1999, Target Estimation for Bias and Mean Square Error Reduction, *Annals of Statistics* **27**, 1080–1104.

Cabrera, J. and Hu, I. 2001, Algorithms for Target Estimation Using Stochastic Approximation, *InterStat* **2**(4), 1–18.

Calvet, L. and Czellar, V. 2015, Accurate Methods for Approximate Bayesian Computation Filtering, *Journal of Financial Econometrics* **13**(4), 798–838.

Chernozhukov, V. and Hong, H. 2003, An MCMC Approach to Classical Estimation, *Journal of Econometrics* **115:2**, 293–346.

Creel, M. and Kristensen, D. 2013, Indirect Likelihood Inference, *mimeo, UCL.*

Creel, M., Gao, J., Hong, H. and Kristensen, D. 2016, Bayesian Indirect Inference and the ABC of GMM, unpublished manuscript.

Dean, T., Singh, S., Jasra, A. and Peters, G. 2011, Parameter Estimation for Hidden Markov Models with Intractable Likelihoods, arXiv:1103.5399.

Diggle, P. and Gratton, J. 1984, Monte Carlop Methods of Inference for Implicit Statistical Methods, *Journal of the Royal Statistical Association Series B* **46**, 193–227.

Drovandi, C., Pettitt, A. and Faaddy, M. 2011, Approximate Bayesian Computation using Indirect Inference, *Journal of the Royal Statistical Society, Series C* **60**(3), 503–524.

Drovandi, C., Pettitt, A. and Lee, A. 2015, Bayesian Indirect Inference Using a Parametric Auxiliary Model, *Statistical Science* **30**(1), 72–95.

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

Forneron, J. J. and Ng, S. 2016, A Likelihood Free Reverse Sampler of the Posterior Distribution, *in* G. Gonzalez-Rivera, R. C. Hill and T.-H. Lee (eds), *Advances in Econometrics, Essays in Honor of Aman Ullah*, Vol. 36, Emerald Group Publishing, pp. 389–415.

Gallant, R. and Tauchen, G. 1996, Which Moments to Match, *Econometric Theory* **12**, 657–681.

Gao, J. and Hong, H. 2014, A Computational Implementation of GMM, SSRN Working Paper 2503199.

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

Gouriéroux, C., Monfort, A. and Renault, E. 1993, Indirect Inference, *Journal of Applied Econometrics* **85**, 85–118.

Gouriéroux, C., Renault, E. and Touzi, N. 1999, Calibration by Simulation for Small Sample Bias Correction,, *in* R. Mariano, T. Schuermann and M. Weeks (eds), *Simulation-based Inference in Econometrics: Methods and Applications*, Cambridge University Press.

Gouriéroux, G., Phillips, P. and Yu, J. 2010, Indirect Inference of Dynamic Panel Models, *Journal of Econometrics* **157**(1), 68–77.

Hansen, L. P. 1982, Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica* **50**, 1029–1054.

Hansen, L. P. and Singleton, K. J. 1982, Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models, *Econometrica* **50**, 1269–1296.

Heggland, K. and Frigessi, A. 2004, Estimating Functions in Indirect Inference, *Journal of the Royal Statistical Association Series B* **66**, 447–462.

Hsiao, C. 2003, *Analysis of Panel Data*, Cambridge University Press.

Jacquier, E., Johannes, M. and Polson, N. 2007, MCMC Maximum Likelihood for Latent State Models, *Journal of Econometrics* **137**(2), 615–640.

Jiang, W. and Turnbull, B. 2004, The Indirect Method: Inference Based on Intermediate Statistics-A Synthesis and Examples, *Statistical Science* **19**(2), 239–263.

Kass, R., Tierney, L. and Kadane, J. 1990, The Validity of Posterior Expansion Based on Lapalce's Method, *in* R. K. S. Gleisser and L. Wasserman (eds), *Bayesian and Likelihood Methods in Statistics and Econometrics*, Elsevier Science Publishers, North Holland.

Kirkpatrick, S., Gellatt, C. and Vecchi, M. 1983, Optimization by Simulated Annealing, *Science* **220**, 671–680.

Kormiltsina, A. and Nekipelov, D. 2014, Consistent Variance of the Laplace Type Estimators, SMU, mimeo.

Lise, J., Meghir, C. and Robin, J. M. 2015, Matching, Sorting, and Wages, *Review of Economic Dynamics*. Cowles Foundation Working Paper 1886.

Marin, J. M., Pudio, P., Robert, C. and Ryder, R. 2012, Approximate Bayesian Computation Methods, *Statistical Computations* **22**, 1167–1180.

Marjoram, P., Molitor, J., Plagnol, V. and Tavare, S. 2003, Markov Chain Monte Carlo Without Likelihoods, *Procedings of the National Academy of Science* **100**(26), 15324–15328.

Meeds, E. and Welling, M. 2015, Optimization Monte Carlo: Efficient and Embarrassingly Parallel Likelihood-Free Inference, arXiv:1506:03693v1.

Michaelides, A. and Ng, S. 2000, Estimating the Rational Expectations Model of Speculative Storage: A Monte Carlo Comparison of Three Simulation Estimators, *Journal of Econometrics* **96:2**, 231–266.

Nekipelov, D. and Kormilitsina, A. 2015, Approximation Properties of Laplace-Type Estimators, *in* N. Balke, F. Canova, F. Milani and M. Wynne (eds), *DSGE Models in Macroeconomics: Estimation, Evaluation, and New Developments*, Vol. 28, pp. 291–318.

Newey, W. and Smith, R. 2004, Higher Order Properties of GMM and Generalized Empirical Likeliood Estimators, *Econometrica* **71:1**, 219–255.

Nickl, R. and Potscher, B. 2010, Efficient Simulation-Based Minimum Distance Estimation and Indirect Inference, *Mathematical Methods of Statistics* **19**(4), 327–364.

Pagan, A. and Ullah, A. 1999, *Nonparametric Econometrics*, Vol. Themes in Modern Econometrics, Cambridge University Press.

Pritchard, J., Seielstad, M., Perez-Lezman, A. and Feldman, M. 1996, Population Growth of Human Y chromosomes: A Study of Y Chromosome MicroSatellites, *Molecular Biology and Evolution* **16**(12), 1791–1798.

Rilstone, P., Srivastara, K. and Ullah, A. 1996, The Second-Order Bias and Mean Squared Error of Nonlinear Estimators, *Journal of Econometrics* **75**, 369–385.

Robert, C. and Casella, G. 2004, *Monte Carlo Statistical Methods*, Textbooks in Statistics, second edn, Springer.

Sisson, S. and Fan, Y. 2011, Likelihood Free Markov Chain Monte Carlo, *in* S. Brooks, A. Celman, G. Jones and X.-L. Meng (eds), *Handbook of Markov Chain Monte Carlo*, Vol. Chapter 12, pp. 313–335. arXiv:10001.2058v1.

Smith, A. 1993, Estimating Nonlinear Time Series Models Using Simulated Vector Autoregressions, *Journal of Applied Econometrics* **8**, S63–S84.

Smith, A. 2008, Indirect Inference, *in* S. Durlauf and L. Blume (eds), *The New Palgrave Dictionary of Economics*, Vol. 2, Palgrave Mcmillian.

Tavare, S., Balding, J., Griffiths, C. and Donnelly, P. 1997, Inferring Coalescence Times From DNA Sequence Data, *Genetics* **145**, 505–518.

Tierney, L. and Kadane, J. 1986, Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association* **81**, 82–86.