Essays on Simulation-Based Estimation

Jean-Jacques Forneron

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

ESSAYS ON SIMULATION-BASED ESTIMATION

Jean-Jacques Forneron

Complex nonlinear dynamic models with an intractable likelihood or moments are increasingly common in economics. A popular approach to estimating these models is to match informative sample moments with simulated moments from a fully parameterized model using SMM or Indirect Inference. This dissertation consists of three chapters exploring different aspects of such simulation-based estimation methods. The following chapters are presented in the order in which they were written during my thesis.

Chapter 1, written with Serena Ng, provides an overview of existing frequentist and Bayesian simulation-based estimators. These estimators are seemingly computationally similar in the sense that they all make use of simulations from the model in order to do the estimation. To better understand the relationship between these estimators, this chapters introduces a Reverse Sampler which expresses the Bayesian posterior moments as a weighted average of frequentist estimates. As such, it highlights a deeper connection between the two class of estimators beyond the simulation aspect. This Reverse Sampler also allows us to compare the higher-order bias properties of these estimators. We find that while all estimators have an automatic bias correction property (as highlighted in Gouriéroux & Monfort, 1996) the Bayesian estimator introduces two additional biases. The first is due to computing a posterior mean rather than the mode. The second is due to the prior, which penalizes the estimates in a particular direction.

Chapter 2, also written with Serena Ng, proves that the Reverse Sampler described above targets the desired Approximate Bayesian Computation (ABC) posterior distribution. The idea relies on a change of variable argument: the frequentist optimization step implies a non-linear transformation. As a result, the unweighted draws follow a distribution that depends on the likelihood that comes from the simulations, and a Jacobian term that arises from the non-linear transformation. Hence, solving the frequentist estimation problem multiple times, with different numerical seeds, leads to an optimization-based importance sampler where the weights depend on the prior and the volume of the Jacobian of the non-linear transformation. In models where optimization is relatively fast, this Reverse Sampler is shown to compare favourably to existing ABC-MCMC or ABC-SMC sampling methods.

Chapter 3, relaxes the parametric assumptions on the distribution of the shocks in simulation-based estimation. It extends the existing SMM literature, where even though the choice of moments is flexible and potentially nonparametric, the model itself is assumed to be fully parametric. The large sample theory in this chapter allows for both time-series and short-panels which are the two most common data types found in empirical applications. Using a flexible sieve density reduces the sensitivity of estimates and counterfactuals to an *ad hoc* choice of distribution such as the Gaussian density. Compared to existing work on sieve estimation, the Sieve-SMM estimator involves dynamically generated data which implies non-standard bias and dependence properties. First, the dynamics imply an accumulation of the bias resulting in a larger nonparametric approximation error than in static models. To ensure that it does not accumulate too much, a set decay conditions on the data generating process are given and the resulting bias is derived. Second, by construction, the dependence properties of the simulated data vary with the parameter values so that standard empirical process results, which rely on a coupling argument, do not apply in this setting. This non-standard dependent empirical process is handled through an inequality built by adapting results from the existing literature. The results hold for bounded empirical processes under a geometric ergodicity condition. This is illustrated in the paper with Monte-Carlo simulations and two empirical applications.

Contents

List of Figures

List of Tables

Acknowledgements

> "I would like to conclude with two observations. First, writing a thesis should be fun. Second, writing a thesis is like cooking a pig: nothing goes to waste." (Umberto Eco, *How to Write a Thesis*)

During the six years in the making of this thesis, I have come to owe much to my advisor, Serena Ng, for her continuous support and guidance as well as the immense amount of time she has lent me. Long before I discovered the above quote, she taught me to pursue topics of interest to me and to not let any bit of effort go to waste.

I have benefited much from discussions with Jushan Bai, Sokbae Lee, José Luis Montiel Olea, Christoph Rothe and Bernard Salanié. Their comments and suggestions have helped shape this thesis into what it is today.

I extremely thankful to my wife, Kim Long-Forneron, for her patience and support throughout these six years as she read through numerous versions of the following chapters. Among other things, the writing quality of the following pages would be much degraded without her irreplaceable input.

I would also like to thank my classmates, office neighbours and former students for making these six years memorable. I am particularly thankful to Sakai Ando, Eugénie Dugoua, Andrew Kosenko, Charles Maurin, Anouch Missirian, Golvine de Rochambeau, Kerem Tuzcuoglu and Jason Wong.

Besides human interactions, I have benefited much from Columbia's computing facilities. In particular, I have abundantly abused the now defunct Yeti and Hotfoot clusters to produce the Monte-Carlo simulations presented in the following chapters.

To my beloved wife, Kim.

Chapter 1

The ABC of Simulation-Based Estimation with Auxiliary Statistics

JEAN-JACQUES FORNERON AND SERENA NG[†]

[†]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.

1.1 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

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

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 model is not complex. The Quasi-Bayes (LT) approach due to Chernozhukov & 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 & Ng (2016) in which implementation of the RS in the overidentified case is also considered. The RS is independently proposed in Meeds & 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] 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 & Ng

1.2 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 $\frac{\partial}{\partial\theta_j}f_\theta(\theta)$ 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 strictly 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

$$p(\theta|\mathbf{y}) = \frac{L(\theta)\cdot\pi(\theta)}{\int_\Theta L(\theta)\pi(\theta)d\theta}. \tag{1.1}$$

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 es-

(2000). Readers are referred to the original paper concerning the assumptions used.

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

timation 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

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

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 $Q(a) = \int_\Theta |\theta - a|^2 p(\theta|\mathbf{y})d\theta$.

Minimum Distance Estimators

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

$$\widehat{\theta}_{GMM} \;=\; \text{argmin}_\theta J(\theta), \qquad J(\theta) = \frac{T}{2}\cdot \overline{g}(\theta)' W \overline{g}(\theta) \tag{1.2}$$

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 & 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 $W = \widehat{\Sigma}^{-1}$. 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 $\overline{g}(\theta)$.

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:

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

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 & 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 an automatic bias reduction effect when $\widehat{\psi}$ is consistent for θ, which is comparable to bootstrap-based bias correction methods. 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 estimator 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 & 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 & 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 & Monfort (1996) provide a framework that unifies these three approaches to SMD estimation. Nonparametric estimation of the auxiliary statistics was considered in Gallant & Tauchen (1996), Fermanian & Salanié (2004), Carrasco et al. (2007a), among others. Nickl & Pötscher (2011) 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.[4].

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 & Kristensen (2013) consider a class of estimators that minimize the Kullback-Leibler distance between the model and the data.[5] 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 & Hong (2014) show that using local linear regressions 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. (2009). 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.

1.3 Approximate Bayesian Computation

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 & Gratton (1984) propose to approximate the likelihood by simulating the model at each point on

[4]Similar ideas in statistics include Mitrovic et al. (2016), Park et al. (2016), and Bernton et al. (2017).

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

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 formal 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 & 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.[6] 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.

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

[6] 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.

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 & Ullah (1999). This is the approach considered in Beaumont et al. (2009), Creel & Kristensen (2013) and Gao & Hong (2014). The case of a rectangular kernel $\mathbb{K}_\delta(\widehat{\psi},\widehat{\psi}^b)=\mathbb{I}_{\|\widehat{\psi}-\widehat{\psi}^b\|\le\delta}$ 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 $b=1,\ldots,B$ 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

$$\rho_{\text{ABC}}(\theta^b,\theta^{b+1})=\min\left(\mathbb{I}_{\|\widehat{\psi}-\widehat{\psi}^{b+1}\|\le\delta}\frac{\pi(\theta^{b+1})q(\theta^b|\theta^{b+1})}{\pi(\theta^b)q(\theta^{b+1}|\theta^b)},1\right). \tag{1.3}$$

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

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

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 & Fan (2011). If the initial draw θ^1 satisfies $\|\widehat{\psi} - \widehat{\psi}^1\| \leq \delta$, then all subsequent $b > 1$ draws are such that $\mathbb{I}_{\|\widehat{\psi}^b-\widehat{\psi}\|\leq\delta} = 1$ 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

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

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}.

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 & 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 & Ng (2016). Meeds & 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.

1.4 Quasi-Bayes Estimators

The GMM objective function $J(\theta)$ defined in (1.2) is not a proper density. Noting that $\exp(-J(\theta))$ is the kernel of the Gaussian density, Jiang & Turnbull (2004) define an *indirect likelihood* as

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

where $\widehat{\Sigma}$ is a consistent estimate of Σ. Note that $L_{IND}(\theta)$ is distinct from the indirect likelihood defined in Creel & Kristensen (2013), but analogous to the 'synthetic likelihood' defined in Wood (2010). Associated with the indirect likelihood is the indirect score, indirect Hessian, and a generalized information matrix equality, just like a conventional likeli-

hood. 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 & 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 & Hong (2003) take an exponential of $-J(\theta)$, as in Jiang & 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 probability density with a well defined peak by a normal density, see Tierney & Kadane (1986). If $\pi(\theta)$ is strictly positive and continuous over a compact parameter space Θ, the 'quasi-posterior' LT distribution

$$p_{LT}(\theta|\mathbf{y}) = \frac{\exp(-J(\theta))\pi(\theta)}{\int_\Theta \exp(-J(\theta)\pi(\theta))d\theta} \propto \exp(-J(\theta))\pi(\theta) \tag{1.4}$$

is proper. The LT posterior mean is thus well-defined even when the prior may not be proper. Wood (2010) considers similar idea, but replaces $J(\theta)$ with $L_{IND}(\theta)$. As discussed in Chernozhukov & 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.[7]

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

[7] 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.

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 and variance are affected by the choice of the weighting matrix even in the just-identified case.[8]

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

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

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 $\varepsilon^s = (\varepsilon_1^s, \ldots, \varepsilon_T^s)'$ 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 $\mathbf{y}^s = (\mathbf{y}_1^s, \ldots, \mathbf{y}_T^s)'$. 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 $\overline{\theta}_{SLT}^b = \frac{1}{B}\sum_{b=1}^B \theta^b$.

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

$$p_{SLT}(\theta|\mathbf{y}, \varepsilon^1, \ldots, \varepsilon^S) \;=\; \frac{\exp(-J_S(\theta))\pi(\theta)}{\int_\Theta \exp(-J_S(\theta))\pi(\theta)d\theta} \propto \exp(-J_S(\theta))\pi(\theta) \tag{1.5}$$

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 $\varepsilon^s = (\varepsilon_1^s, \ldots, \varepsilon_T^s)'$ for $s = 1, \ldots S$, repeat for $b = 1, \ldots B$:

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

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 & 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 & 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.

1.5 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 & 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

$$A_{ML}(\theta_0) = \mathbb{E}[\ell_{\theta\theta}(\theta_0)]^{-1} Z_S(\theta_0) \tag{1.6a}$$

$$C_{ML}(\theta_0) = \mathbb{E}[-\ell_{\theta\theta}(\theta_0)]^{-1}\left[Z_H(\theta_0)Z_S(\theta_0) - \frac{1}{2}\sum_{j=1}^K(-\ell_{\theta\theta\theta_j}(\theta_0))Z_S(\theta_0)Z_{S,j}(\theta_0)\right] \tag{1.6b}$$

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.

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

Minimum distance estimators depend on auxiliary statistics $\widehat{\psi}$. Its properties have been analyzed in Newey & Smith (2004, Section 4.2) within an empirical-likelihood framework. To facilitate subsequent analysis, we follow Gouriéroux & 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

$$\widehat{\psi} = \psi(\theta_0) + \frac{\mathbb{A}(\theta_0)}{\sqrt{T}} + \frac{\mathbb{C}(\theta_0)}{T} + o_p(\frac{1}{T}). \tag{1.7}$$

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

$$A_{MD}(\theta_0) = \left[\psi_\theta(\theta_0)\right]^{-1}\mathbb{A}(\theta_0) \tag{1.8a}$$

$$C_{MD}(\theta_0) = \left[\psi_\theta(\theta_0)\right]^{-1}\left[\mathbb{C}(\theta_0) - \frac{1}{2}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A_{MD}(\theta_0)A_{MD,j}(\theta_0)\right]. \tag{1.8b}$$

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 (1.7), 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 second-order expansions gives:

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

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

$$A_{SMD}(\theta_0) \;=\; \left[\psi_\theta(\theta_0)\right]^{-1}\left(\mathbb{A}(\theta_0) - \frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s(\theta_0)\right) \tag{1.9a}$$

$$\begin{aligned} C_{SMD}(\theta_0) \;=\; & \left[\psi_\theta(\theta_0)\right]^{-1}\left(\mathbb{C}(\theta_0) - \frac{1}{S}\sum_{s=1}^{S}\mathbb{C}^s(\theta_0) - \frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s_\theta(\theta_0)A_{SMD}(\theta_0)\right) \\ & -\frac{1}{2}\left[\psi_\theta(\theta_0)\right]^{-1}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)A_{SMD}(\theta_0)A_{SMD,j}(\theta_0). \end{aligned} \tag{1.9b}$$

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

$$\mathbb{E}[C_{SMD}(\theta_0)] \overset{S\to\infty}{\to} -\frac{1}{2}\left[\psi_\theta(\theta_0)\right]^{-1}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)\mathbb{E}\left[A_{MD}(\theta_0)A_{MD,j}(\theta_0)\right]. \tag{1.10}$$

This is an improvement over $\widehat{\theta}_{MD}$ because as seen from (1.8b),

$$\mathbb{E}[C_{MD}(\theta_0)] = \left[\psi_\theta(\theta_0)\right]^{-1}\mathbb{C}(\theta_0) - \frac{1}{2}\left[\psi_\theta(\theta_0)\right]^{-1}\sum_{j=1}^{K}\psi_{\theta,\theta_j}(\theta_0)\mathbb{E}\left[A_{MD}(\theta_0)A_{MD,j}(\theta_0)\right]. \tag{1.11}$$

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

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. ABC has also been applied to filter unobserved latent variables in intractable non-linear non-gaussian state-space models. Calvet & 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 & Smith (2004), Creel & 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 & 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. (2009). 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

$$A_{RS}(\theta_0) = \frac{1}{B}\sum_{b=1}^{B} A^b_{RS}(\theta_0) = \left[\psi_\theta(\theta_0)\right]^{-1}\left(\mathbb{A}(\theta_0) - \frac{1}{B}\sum_{b=1}^{B}\mathbb{A}^b(\theta_0)\right) \tag{1.12a}$$

$$\begin{aligned} C_{RS}(\theta_0) &= \frac{1}{B}\sum_{b=1}^{B} C^b_{RS}(\theta_0) \\ &+ \frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}\left[\frac{1}{B}\sum_{b=1}^{B}(A^b_{RS}(\theta_0) - \overline{A}_{RS}(\theta_0))A^b_{RS}(\theta_0)\right] + C^M_{RS}(\theta_0). \end{aligned} \tag{1.12b}$$

Proposition 1. *Let $\widehat{\psi}(\theta)$ be the auxiliary statistic that admits the expansion as in (1.7) 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 & Kristensen (2013). The first term is

$$\begin{gathered} \tfrac{1}{B}\textstyle\sum_{b=1}^{B} C^b_{RS}(\theta_0) = \\ \left[\psi_\theta(\theta_0)\right]^{-1} \tfrac{1}{B}\textstyle\sum_{b=1}^{B}\left(\mathbb{C}(\theta_0) - \mathbb{C}^b(\theta_0) - \tfrac{1}{2}\textstyle\sum_{j=1}^{K}\psi_{\theta\theta_j}(\theta_0)A^b_{RS}(\theta_0)A^b_{RS,j}(\theta_0) - \mathbb{A}^b_\theta(\theta_0)A^b_{RS}(\theta_0)\right). \end{gathered}$$

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 (1.10). The reason is that the comparable term in $C_{SMD}(\theta_0)$ is

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

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 & 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 (1.12b) 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 & 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 & 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 & 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 & 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 $\psi(\theta) = \theta$.

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:

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

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

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

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

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

$$\widehat{\theta}_d=\theta_0+\frac{A_d(\theta_0)}{\sqrt{T}}+\frac{C_d(\theta_0)}{T}+\frac{\mathbb{1}_{d\in D}}{T}\left[\frac{\pi_\theta(\theta_0)}{\pi(\theta_0)}C_d^P(\theta_0)+C_d^M(\theta_0)\right]+o_p(\frac{1}{T}) \tag{1.13}$$

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

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

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 & 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 & 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.

1.6 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 approxi-

mated 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.

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

Table 1.1: 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	$\widehat{\sigma}^2\frac{T}{T+1}$	$\widehat{\sigma}^2\frac{T}{T+6}$	$\sigma^2\frac{T-1}{T+1}$	$\sigma^2\frac{T-1}{T+6}$

does not coincide with 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:

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

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,

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

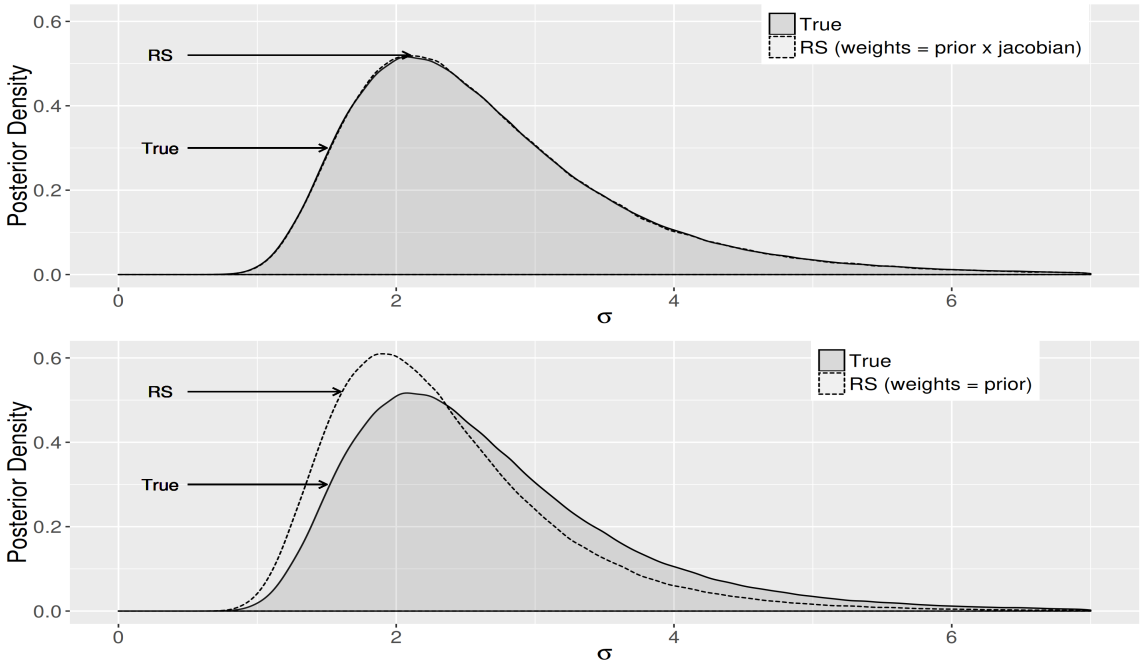
For completeness, the parametric Bootstrap bias corrected estimator $\widehat{\sigma}^2_{\text{Bootstrap}} = 2\widehat{\sigma}^2 - \mathbb{E}_{\text{Bootstrap}}(\widehat{\sigma}^2)$ is also considered:

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

$\mathbb{E}_{\text{Bootstrap}}(\widehat{\sigma}^2)$ 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 $1/T^2$).

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 2.2 plots

Figure 1.1: ABC vs. RS Posterior Density



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.

The properties of the estimators are summarized in Table 1.2. 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.

Table 1.2: 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	$\sigma^2\frac{T-1}{T-5}$	$\frac{2\sigma^2}{T-5}$	$2\sigma^4\frac{T-1}{(T-5)^2}$
$\overline{\theta}^R_{BC}$	$1/\sigma^4$	σ^2	0	$2\sigma^4\frac{1}{T-1}$
$\overline{\theta}^U_{RS}$	1	$\sigma^2\frac{T-1}{T-5}$	$\frac{2\sigma^2}{T-5}$	$2\sigma^4\frac{T-1}{(T-5)^2}$
$\overline{\theta}^R_{RS}$	$\frac{1}{\sigma^4}$	$\sigma^2\frac{B(T-1)}{B(T-1)-2}$	$\frac{2\sigma^2}{B(T-1)-2}$	$2\sigma^4\frac{\kappa_1}{T-1}$
$\widehat{\theta}_{SMD}$	-	$\sigma^2\frac{S(T-1)}{S(T-1)-2}$	$\frac{2\sigma^2}{S(T-1)-2}$	$2\sigma^4\frac{\kappa_1}{T-1}$
$\overline{\theta}^U_{LT}$	1	$\sigma^2\frac{T-1}{T}(1+\kappa_{LT})$	$\sigma^2\frac{T-1}{T}\kappa_{LT}-\frac{\sigma^2}{T}$	$2\sigma^4\frac{T-1}{T^2}(1+\kappa_{LT})^2$
$\widehat{\theta}^U_{SLT}$	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 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 $\widehat{\beta}$.[9] 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. Following their setting, we take $\alpha_i \overset{iid}{\sim} \mathcal{N}(0,1)$, $_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and $y_{i,0}|\alpha_i \overset{iid}{\sim} \mathcal{N}\left(\alpha_i/(1-\rho),(1-\rho^2)^{-1}\right)$.

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

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

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

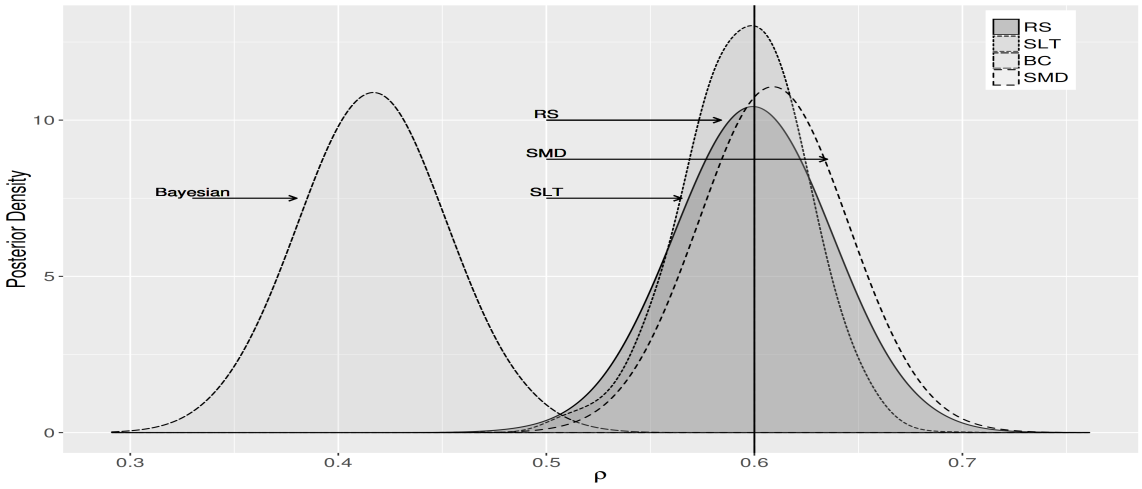
where $\Omega = I_{T-1} - 1_{T-1}1_{T-1}'/T$. We use the following moment conditions for MD estimation:

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

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.

Table 1.3 reports results from 5,000 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 1.2 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.

Figure 1.2: 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 $\mathcal{N}(\widehat{\theta}_{SMD}, \widehat{\text{var}}(\widehat{\theta}_{SMD}))$.

The MCMC-ABC results in Table 1.3 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 & 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 1.3. 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.

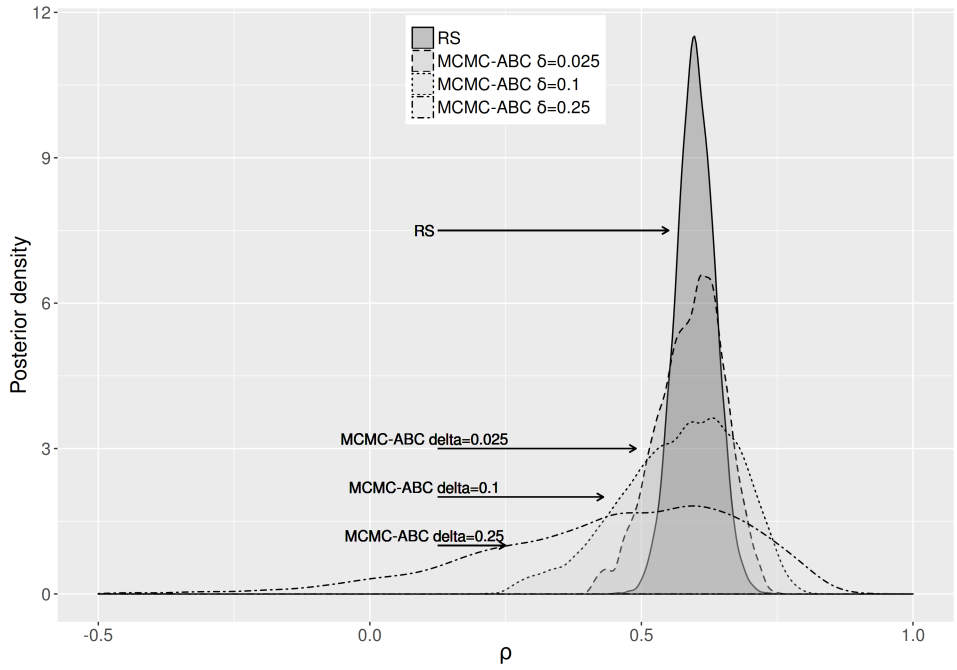
Table 1.3: Dynamic Panel $\rho = 0.6, \beta = 1, \sigma^2 = 2$

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 1.3: MCMC-ABC vs. RS Posterior Density



1.7 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.

Chapter 2

A Likelihood-Free Reverse Sample of the Posterior Distribution

JEAN-JACQUES FORNERON AND SERENA NG[†]

[†]Financial support is provided by the National Science Foundation, SES-0962431.
We thank Christopher Drovandi, Neil Shephard, and two anonymous referees for many helpful comments. The second author would like to thank Aman Ullah for his support and guidance.

2.1 Introduction

Maximum likelihood estimation rests on the ability of a researcher to express the joint density of the data, or the likelihood, as a function of K unknown parameters $\boldsymbol{\theta}$. Inference can be conducted using classical distributional theory once the mode of the likelihood function is determined by numerical optimization. Bayesian estimation combines the likelihood with a prior to form the posterior distribution from which the mean and other quantities of interest can be computed. Though the posterior distribution may not always be tractable, it can be approximated by Monte Carlo methods provided that the likelihood is available. When the likelihood is intractable but there exists $L \geq K$ auxiliary statistics $\widehat{\boldsymbol{\psi}}$ with model analog $\boldsymbol{\psi}(\boldsymbol{\theta})$ that is analytically tractable, one can still estimate $\boldsymbol{\theta}$ by minimizing the difference between $\widehat{\boldsymbol{\psi}}$ and $\boldsymbol{\psi}(\boldsymbol{\theta})$.

Increasingly, parametric models are so complex that neither the likelihood nor $\boldsymbol{\psi}(\boldsymbol{\theta})$ is tractable. But if the model is easy to simulate, the mapping $\boldsymbol{\psi}(\boldsymbol{\theta})$ can be approximated by simulations. Estimators that exploit this idea can broadly be classified into two types. One is simulated minimum distance estimator (SMD), a frequentist approach that is quite widely used in economic analysis. The other is the method of Approximate Bayesian Computation that is popular in other disciplines. This method, ABC for short, approximates the posterior distribution using auxiliary statistics $\widehat{\boldsymbol{\psi}}$ instead of the full dataset y. It takes draws of $\boldsymbol{\theta}$ from a prior distribution and keeps the draws that, when used to simulate the model, produces auxiliary statistics that are close to the sample estimates $\widehat{\boldsymbol{\psi}}$. Both the ABC and SMD can be regarded as likelihood free estimators in the sense that the likelihood that corresponds to the structural model of interest is not directly evaluated.

While both the SMD and ABC exploit auxiliary statistics to perform likelihood free estimation, there are important differences between them. The SMD solves for the $\boldsymbol{\theta}$ that makes $\widehat{\boldsymbol{\psi}}$ close to the average of $\boldsymbol{\psi}(\boldsymbol{\theta})$ over many simulated paths of the data. In contrast, the ABC evaluates $\boldsymbol{\psi}(\boldsymbol{\theta})$ for each draw from the prior and accepts the draw only if $\boldsymbol{\psi}(\boldsymbol{\theta})$ is close to $\widehat{\boldsymbol{\psi}}$. The ABC estimate is the average over the accepted draws, which is the posterior mean. In Forneron & Ng (2018), we focused on the case of exact identification and used a reverse sampler (RS) to better understand the difference between the two approaches. The RS approximates the posterior distribution by solving a sequence of SMD problems, each using only one simulated path of data. Using stochastic expansions as in Rilstone et al. (1996) and Bao & Ullah (2007), we reported that in the special case when $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ (i.e the auxiliary model is the assumed model), the SMD has an unambiguous bias advantage over the ABC. But in more general settings, the ABC can, by clever choice

of prior, eliminate biases that are inherent in the SMD.

In this paper, we extend the analysis to over-identified models and provide a deeper understanding of the reverse sampler. The RS is shown to be an optimization-based importance sampler that transforms the density from draws of $\boldsymbol{\psi}$ to draws of $\boldsymbol{\theta}$ so that when multiplied by the prior and properly weighted, the draws follow the desired posterior distribution. Section 2 considers the exactly identified case and shows that the importance ratio is the determinant of the Jacobian matrix. Section 3 considers the over-identified case when the dimension of $\boldsymbol{\psi}(\boldsymbol{\theta})$ exceeds that of $\boldsymbol{\theta}$. Because of the need to transform densities of different dimensions, the determinant of the Jacobian matrix is replaced by its volume. Using analytically tractable models, we show that the RS exactly reproduces the desired posterior distribution.

The RS was initially developed as a framework to better understand the different approaches to likelihood free estimation. While not intended to compete with existing implementations of ABC, the use of optimization in RS turns out to have a property that is of independent interest. Creating a long sequence of ABC draws such that the simulated statistic $\widehat{\boldsymbol{\psi}}^b$ and the data $\widehat{\boldsymbol{\psi}}$ deviate by no more than δ can take infinite time if δ is set to exactly zero as theory suggests. This has generated interests within the ABC community to control for δ. The RS by-passes this problem because SMD estimation makes $\widehat{\boldsymbol{\psi}}^b$ as close to $\widehat{\boldsymbol{\psi}}^b$ as machine precision permits. We elaborate on this feature in Section 4. Of course, the RS is useful only when the SMD objective function is well behaved and easy to optimize, which may not always be the case. But allowing optimization to play a role in ABC can be useful, as independent work by Meeds & Welling (2015) also found.

Preliminaries

In what follows, we use a 'hat' to denote estimators that correspond to the mode (or extremum estimators) and a 'bar' 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. These parameters S and B have different roles. The SMD uses S simulations to approximate the mapping $\boldsymbol{\psi}(\boldsymbol{\theta})$, while the ABC uses B simulations to approximate the posterior distribution of the infeasible likelihood.

We assume that the data $\mathbf{y} = (y_1, \ldots, y_T)'$ have finite fourth moments and can be represented by a parametric model with probability measure $\mathcal{P}_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^K$, $\boldsymbol{\theta}_0$ is the true value. The likelihood $L(\boldsymbol{\theta}|\mathbf{y})$ is intractable. Estimation of $\boldsymbol{\theta}$ is based on $L \geq K$

auxiliary statistics $\widehat{\psi}(\mathbf{y}(\boldsymbol{\theta}_0))$ which we simply denote by $\widehat{\psi}$ when the context is clear. The model implies statistics $\psi(\boldsymbol{\theta})$. The classical minimum distance estimator is

$$\widehat{\boldsymbol{\theta}}_{\text{CMD}} = \text{argmin}_{\boldsymbol{\theta}} J(\widehat{\psi}, \psi(\boldsymbol{\theta})) = \overline{g}(\boldsymbol{\theta})' \text{W} \overline{g}(\boldsymbol{\theta}), \quad \overline{g}(\boldsymbol{\theta}) = \widehat{\psi} - \psi(\boldsymbol{\theta}).$$

Assumption A :

- i There exists a unique interior point $\boldsymbol{\theta}_0 \in \Theta$ (compact) that minimizes the population objective function $(\psi(\boldsymbol{\theta}_0) - \psi(\boldsymbol{\theta}))' \text{W} (\psi(\boldsymbol{\theta}_0) - \psi(\boldsymbol{\theta}))$. The mapping $\boldsymbol{\theta} \to \psi(\boldsymbol{\theta}) = \lim_{T \to \infty} \mathbb{E}[\widehat{\psi}(\boldsymbol{\theta})]$ is continuously differentiable and injective. The $L \times K$ Jacobian matrix $\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ has full column rank, and the rank is constant in the neighborhood of $\boldsymbol{\theta}_0$.
- ii There is an estimator $\widehat{\psi}$ such that $\sqrt{T}(\widehat{\psi} - \psi(\boldsymbol{\theta}_0)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$.
- iii W is a $L \times L$ positive definite matrix and $\text{W} \psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)$ has rank K.

Assumption A ensures global identification and consistent estimation of $\boldsymbol{\theta}$, see Newey & McFadden (1994). In Gouriéroux et al. (1993), the mapping $\psi : \boldsymbol{\theta} \to \psi(\boldsymbol{\theta})$ is referred to as the binding function while in Jiang & Turnbull (2004), $\psi(\boldsymbol{\theta})$ is referred to as a bridge function. When $\psi(\boldsymbol{\theta})$ is analytically intractable, the simulated minimum distance estimator (SMD) is

$$\widehat{\boldsymbol{\theta}}_{\text{SMD}} = \text{argmin}_{\boldsymbol{\theta}} J_S(\widehat{\psi}, \widehat{\psi}_S(\boldsymbol{\theta})) = \text{argmin}_{\boldsymbol{\theta}} \overline{g}_S(\boldsymbol{\theta})' \text{W} \overline{g}_S(\boldsymbol{\theta}). \tag{2.1}$$

where $S \geq 1$ is the number of simulations,

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

Notably, the term $\mathbb{E}[\widehat{\psi}(\boldsymbol{\theta})]$ in CMD estimation is approximated by $\frac{1}{S} \sum_{s=1}^{S} \widehat{\psi}^s(\mathbf{y}^s(\boldsymbol{\theta}))$. The SMD was first used in Smith (1993). Different SMD estimators can be obtained by suitable choice of the moments $\overline{g}(\boldsymbol{\theta})$, including the indirect inference estimator of Gouriéroux et al. (1993), the simulated method of moments of Duffie & Singleton (1993), and the efficient method of moments of Gallant & Tauchen (1996).

The first ABC algorithm was implemented by Tavare et al. (1997) and Pritchard et al. (1996) to study population genetics. They draw $\boldsymbol{\theta}^b$ from the prior distribution $\pi(\boldsymbol{\theta})$, simulate the model under $\boldsymbol{\theta}^b$ to obtain data $\mathbf{y}^b$, and accept $\boldsymbol{\theta}^b$ if the vector of auxiliary statistics $\psi(\boldsymbol{\theta}^b)$ deviates from $\widehat{\psi}$ by no more than a tuning parameter δ. If $\widehat{\psi}$ are sufficient statistics and $\delta = 0$, the procedure produces samples from the true posterior distribution if $B \to \infty$.

The Accept-Reject ABC: For $b = 1, \ldots, B$

i Draw ϑ from $\pi(\boldsymbol{\theta})$ and ε^b from an assumed distribution F_ε
ii Generate $\mathbf{y}^b(\varepsilon^b, \vartheta)$ and $\widehat{\boldsymbol{\psi}}^b = \boldsymbol{\psi}(\mathbf{y}^b)$.
iii Accept $\boldsymbol{\theta}^b = \vartheta$ if $J_1^b = \left(\widehat{\boldsymbol{\psi}}^b - \widehat{\boldsymbol{\psi}}\right)' \mathrm{W} \left(\widehat{\boldsymbol{\psi}}^b - \widehat{\boldsymbol{\psi}}\right) \leq \delta$.

The accept-reject method (hereafter, AR-ABC) simply keeps those draws from the prior distribution $\pi(\boldsymbol{\theta})$ that produce auxiliary statistics which are close to the observed $\widehat{\boldsymbol{\psi}}$. As it is not easy to choose δ a priori, it is common in AR-ABC to fix a desired quantile q, repeat the steps $[B/q]$ times. Setting δ to the q-th quantile of the sequence of J_1^b that will produce exactly B draws is analogous to the idea of keeping $k-$nearest neighbors considered in Gao & Hong (2014).

Since simulating from a non-informative prior distribution is inefficient, the accept-reject sampler can be replaced by one that targets at features of the posterior distribution. There are many ways to target the posterior distribution. We consider the MCMC implementation of ABC proposed in Marjoram et al. (2003) (hereafter, MCMC-ABC).

The MCMC-ABC: For $b = 1, \ldots, B$ with $\boldsymbol{\theta}^0$ given and proposal density $q(\cdot|\boldsymbol{\theta}^b)$,

i Generate $\vartheta \sim q(\vartheta|\boldsymbol{\theta}^b)$
ii Draw errors ε^{b+1} from F_ε and simulate data $\mathbf{y}^{\mathbf{b+1}}(\varepsilon^{b+1}, \vartheta)$. Compute $\widehat{\boldsymbol{\psi}}^{b+1} = \boldsymbol{\psi}(\mathbf{y}^{b+1})$.
iii Set $\boldsymbol{\theta}^{b+1}$ to ϑ with probability $\rho_{ABC}(\boldsymbol{\theta}^b, \vartheta)$ and to $\boldsymbol{\theta}^{b+1}$ with probability $1 - \rho_{ABC}(\boldsymbol{\theta}^b, \vartheta)$ where

$$\rho_{ABC}(\boldsymbol{\theta}^b, \vartheta) = \min\left(\mathbb{I}_{\|\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\psi}}^{b+1}\| \leq \delta} \frac{\pi(\vartheta) q(\boldsymbol{\theta}^b|\vartheta)}{\pi(\boldsymbol{\theta}^b) q(\vartheta|\boldsymbol{\theta}^b)}, 1\right) \tag{2.2}$$

The AR and MCMC both produce an approximation to the posterior distribution of $\boldsymbol{\theta}$. It is common to use the posterior mean of the draws $\overline{\boldsymbol{\theta}} = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{\theta}^b$ as the ABC estimate. The MCMC-ABC uses a proposal distribution to account for features of the data so that it is less likely to have proposed values with low posterior probability. The tuning parameter δ affects the bias of the estimates. Too small a δ may require making many draws which can be computationally costly.

The ABC samples from the joint distribution of $(\boldsymbol{\theta}^b, \boldsymbol{\psi}^b(\varepsilon^b, \boldsymbol{\theta}^b))$ and then integrates out ε^b. The posterior distribution is thus

$$p(\boldsymbol{\theta}^b|\widehat{\boldsymbol{\psi}}) \propto \int p(\boldsymbol{\theta}^b, \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \varepsilon^b)|\widehat{\boldsymbol{\psi}}) \mathbb{I}_{\|\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^b)\| < \delta} d\varepsilon^b.$$

The indicator function (also the rectangular kernel) equals one if $\|\widehat{\psi} - \psi^b\|$ does not exceed δ. The ABC draws are dependent due to the Markov nature of the MCMC-ABC sampler.

Both the SMD and ABC assume that simulations provide an accurate approximation of $\psi(\theta)$ and that auxiliary statistics are chosen to permit identification of θ. Creel & Kristensen (2015) suggests a cross-validation method for selecting the auxiliary statistics. For the same choice of $\widehat{\psi}$, the SMD finds the θ that makes the average of the simulated auxiliary statistics close to $\widehat{\psi}$. The ABC takes the average of θ^b, drawn from the prior, with the property that each ψ^b is close to $\widehat{\psi}$. In an attempt to understand this difference, Forneron & Ng (2018), takes as starting point that each θ^b in the above ABC algorithm can be reformulated as an SMD problem with $S = 1$. We consider an algorithm that solves the SMD problem many times to obtain a distribution for θ^b, each time using one simulated path. The sampler terminates with an evaluation of the prior probability, in contrast to the ABC which starts with a draw from the prior distribution. Hence we call our algorithm a reverse sampler (hereafter, RS). The RS produces a sequence of θ^b that are independent optimizers and do not have a Markov structure.

In the next two sections, we explore additional features of the RS. As an overview, the distribution of draws that emerge from SMD estimation with $S = 1$ may not be from the desired posterior distribution. Hence the draws are re-weighted to target the posterior. In the exactly identified case, $\widehat{\psi}^b$ can be made exactly equal to $\widehat{\psi}$ by choosing the SMD estimate as θ^b. Thus the RS is simply an optimization based importance sampler using the determinant of Jacobian matrix as importance ratio. In the over-identified case, the volume of the (rectangular) Jacobian matrix is used in place of the determinant. Additional weighting is given to those $\widehat{\theta}^b$ that yields $\widehat{\psi}^b$ sufficiently close to $\widehat{\psi}$.

2.2 The Reverse Sampler: Case $K = L$

The algorithm for the case of exact identification is as follows. For $b = 1, \ldots, B$

i Generate ε^b from F_ε.
ii Find $\theta^b = \text{argmin}_\theta J_1^b(\widehat{\psi}^b(\theta, \varepsilon^b), \widehat{\psi})$ and let $\widehat{\psi}^b = \widehat{\psi}^b(\theta^b, \varepsilon^b)$.
iii Set $w(\theta^b, \varepsilon^b) = \pi(\theta^b)|\widehat{\psi}^b_\theta(\theta^b, \varepsilon^b)|^{-1}$.
iv Re-weigh the θ^b by $\frac{w(\theta^b)}{\sum_{b=1}^B w(\theta^b)}$.

Like the ABC, the draws $\boldsymbol{\theta}^b$ provides an estimate of the posterior distribution of $\boldsymbol{\theta}$ from which an estimate of the posterior mean:

$$\overline{\boldsymbol{\theta}}_{RS} = \sum_{b=1}^{B} \frac{w(\boldsymbol{\theta}^b)}{\sum_{b=1}^{B} w(\boldsymbol{\theta}^b)} \boldsymbol{\theta}^b$$

can be used as an estimate of $\boldsymbol{\theta}$. Each $\boldsymbol{\theta}^b$ is a function of the data $\widehat{\boldsymbol{\psi}}$ and the draws $\boldsymbol{\varepsilon}^b$ that minimizes $J_1^b(\boldsymbol{\psi}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b), \widehat{\boldsymbol{\psi}})$. The K first-order conditions are given by

$$\mathcal{F}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}) = \frac{\partial \overline{\boldsymbol{g}}_1(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\theta}}' \mathrm{W} \overline{\boldsymbol{g}}_1(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}) = 0 \tag{2.3}$$

where $\frac{\partial \overline{\boldsymbol{g}}_1(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\theta}}$ is the $L \times K$ matrix of derivatives with respect to $\boldsymbol{\theta}$ evaluated at the arguments. It is assumed that, for all b, this derivative matrix has full column rank K. For SMD estimation, $\frac{\partial \overline{\boldsymbol{g}}_1(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\theta}} = \widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})$. This Jacobian matrix plays an important role in the RS.

The importance density denoted $h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b | \widehat{\boldsymbol{\psi}})$ is obtained by drawing $\boldsymbol{\varepsilon}^b$ from the assumed distribution $F_{\boldsymbol{\varepsilon}}$ and finding $\boldsymbol{\theta}^b$ such that $J(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b), \widehat{\boldsymbol{\psi}})$ is smaller than a pre-specified tolerance. When $K = L$, this tolerance can be made arbitrarily small so that up to numerical precision, $\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) = \widehat{\boldsymbol{\psi}}$. This density $h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b | \widehat{\boldsymbol{\psi}})$ is related to $p_{\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b}(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)) \equiv p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b)$ by a change of variable:

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

Now $p(\boldsymbol{\theta}^b, \widehat{\boldsymbol{\psi}}^b | \widehat{\boldsymbol{\psi}}) \propto p(\widehat{\boldsymbol{\psi}} | \boldsymbol{\theta}^b, \widehat{\boldsymbol{\psi}}^b) p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b | \boldsymbol{\theta}^b) \pi(\boldsymbol{\theta}^b)$ and $p(\widehat{\boldsymbol{\psi}} | \boldsymbol{\theta}^b, \widehat{\boldsymbol{\psi}}^b)$ is constant since $\widehat{\boldsymbol{\psi}}^b = \widehat{\boldsymbol{\psi}}$. Hence

$$\begin{aligned} p(\boldsymbol{\theta}^b | \widehat{\boldsymbol{\psi}}) &\propto \int \pi(\boldsymbol{\theta}^b) p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b | \widehat{\boldsymbol{\psi}}) \mathbb{I}_{\|\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b\| = 0} d\boldsymbol{\varepsilon}^b \\ &= \int \pi(\boldsymbol{\theta}^b) |\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})|^{-1} h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b | \widehat{\boldsymbol{\psi}}) \mathbb{I}_{\|\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b\| = 0} d\boldsymbol{\varepsilon}^b \\ &= \int w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b | \widehat{\boldsymbol{\psi}}) d\boldsymbol{\varepsilon}^b \end{aligned}$$

where the weights are, assuming invertibility of the determinant:

$$w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) = \pi(\boldsymbol{\theta}^b) |\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})|^{-1}. \tag{2.4}$$

Note that in general, $\frac{w(\boldsymbol{\theta}^b)}{\sum_{b=1}^{B} w(\boldsymbol{\theta}^b)} \neq \frac{1}{B}$.

In the above, we have used the fact that $\mathbb{I}_{\|\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b\| = 0}$ is 1 with probability one when $K = L$. The Jacobian of the transformation appears in the weights because the draws $\boldsymbol{\theta}^b$

are related to the likelihood via a change of variable. Hence a crucial aspect of the RS is that it re-weighs the draws of $\boldsymbol{\theta}^b$ from $h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon})$. Put differently, the unweighted draws will not, in general, follow the target posterior distribution.

Consider a weighted sample $(\boldsymbol{\theta}^b, w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}))$ with $w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)$ defined in (2.4). The following proposition shows that as $B \to \infty$, RS produces the posterior distribution associated with the infeasible likelihood, which is also the ABC posterior distribution with $\delta = 0$.

Proposition 2. *Suppose that $\widehat{\boldsymbol{\psi}}^b : \boldsymbol{\theta} \to \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b)$ is one-to-one and the determinant $|\frac{\partial \boldsymbol{\psi}^b(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\theta}}| = |\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})|$ is bounded away from zero around $\boldsymbol{\theta}^b$. For any measurable function $\varphi(\boldsymbol{\theta})$ such that $\mathbb{E}_{p(\boldsymbol{\theta}|\widehat{\boldsymbol{\psi}})}(\varphi(\boldsymbol{\theta})) = \int \varphi(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\widehat{\boldsymbol{\psi}}) d\boldsymbol{\theta}$ exists, then*

$$\frac{\sum_b^B w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)\varphi(\boldsymbol{\theta}^b)}{\sum_b^B w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)} \xrightarrow{a.s.} \mathbb{E}_{p(\boldsymbol{\theta}|\widehat{\boldsymbol{\psi}})}(\varphi(\boldsymbol{\theta})).$$

Convergence to the target distribution follows from a strong law of large numbers. Fixing the event $\widehat{\boldsymbol{\psi}}^b = \widehat{\boldsymbol{\psi}}$ is crucial to this convergence result. To see why, consider first the numerator:

$$\begin{aligned}
\frac{1}{B}\sum_b w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)\varphi(\boldsymbol{\theta}^b) &\xrightarrow{a.s.} \iint \varphi(\boldsymbol{\theta})\, w(\boldsymbol{\theta}, \boldsymbol{\varepsilon})\, p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b|\boldsymbol{\theta})|\widehat{\boldsymbol{\psi}}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}, \widehat{\boldsymbol{\psi}})| d\boldsymbol{\varepsilon}^b d\boldsymbol{\theta} \\
&= \iint \varphi(\boldsymbol{\theta}) \left|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}, \widehat{\boldsymbol{\psi}})\right|^{-1} \pi(\boldsymbol{\theta}) p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b|\boldsymbol{\theta}) \left|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}, \widehat{\boldsymbol{\psi}})\right| d\boldsymbol{\varepsilon}^b d\boldsymbol{\theta} \\
&= \iint \varphi(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}) p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}|\boldsymbol{\theta}) d\boldsymbol{\varepsilon} d\boldsymbol{\theta} \\
&= \iint \varphi(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}) p(\widehat{\boldsymbol{\psi}}, \boldsymbol{\varepsilon}|\boldsymbol{\theta}) d\boldsymbol{\varepsilon} d\boldsymbol{\theta} \\
&= \int \varphi(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}) L(\widehat{\boldsymbol{\psi}}|\boldsymbol{\theta}) d\boldsymbol{\theta}.
\end{aligned}$$

Furthermore, the denominator converges to the integrating constant since $\frac{1}{B}\sum_b w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}) \xrightarrow{a.s.} \int \pi(\boldsymbol{\theta}) L(\widehat{\boldsymbol{\psi}}|\boldsymbol{\theta}) d\boldsymbol{\theta}$. Proposition 2 implies that the weighted average of $\boldsymbol{\theta}^b$ converges to the posterior mean. Furthermore, the posterior quantiles produced by the reverse sampler tends to those of the infeasible posterior distribution $p(\boldsymbol{\theta}|\widehat{\boldsymbol{\psi}})$ as $B \to \infty$. As discussed in Forneron & Ng (2018), the ABC can be presented as an importance sampler. Hence the accept-reject algorithm in Tavare et al. (1997) and Pritchard et al. (1996), as well as the Sequential Monte-Carlo approach to ABC in Sisson et al. (2007); Toni et al. (2009) and Beaumont et al. (2009) are all important samplers. The RS differs in that it is optimization based. It is also developed independently in Meeds & Welling (2015).

We now use examples to illustrate how the RS works in the exactly identified case.

Example 1: Suppose we have one observation $y \sim \mathcal{N}(\theta, 1)$ or $y = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$. The prior for θ is $\theta \sim \mathcal{N}(0,1)$. By drawing, $\theta^b, \varepsilon^b \sim \mathcal{N}(0,1)$, we obtain $y^b = \theta^b + \varepsilon^b \sim \mathcal{N}(0,2)$. The ABC keeps $\theta^b | y^b = y$. Since (θ^b, y^b) are jointly normal with covariance of 1, we deduce that $\theta^b | y^b = y \sim \mathcal{N}(y/2, 1/2)$. The exact posterior distribution for θ is $\mathcal{N}(y/2, 1/2)$.

The RS draws $\varepsilon^b \sim \mathcal{N}(0,1)$ and computes $\theta^b = y - \varepsilon^b$ which is $\mathcal{N}(y,1)$ conditional on y. The Jacobian of the transformation is 1. Re-weighting according to the prior, we have:

$$p_{\text{RS}}(\theta|y) \propto \phi(\theta)\phi(\theta - y) \propto \exp\left(-\tfrac{1}{2}\left(\theta^2 + (\theta - y)^2\right)\right) \propto \exp\left(-\frac{1}{2}\left(2\theta^2 - 2\theta y\right)\right)$$
$$\propto \exp\left(-\frac{2}{2}\left(\theta - y/2\right)^2\right).$$

This is the exact posterior distribution as derived above.

Example 2 Suppose $y = Q(u, \theta), \varepsilon \sim \mathcal{U}_{[0,1]}$ and Q is a quantile function that is invertible and differentiable in both arguments.[1] For a single draw, y is a sufficient statistic. The likelihood-based posterior is:

$$p(\theta|y) \propto \pi(\theta) f(y|\theta).$$

The RS simulates $y^b(\theta) = Q(\varepsilon^b|\theta)$ and sets $Q(\varepsilon^b|\theta^b) = y$. Or, in terms of the CDF:

$$\varepsilon^b = F(y|\theta^b)$$

Consider a small perturbation to y holding u^b fixed:

$$0 \;=\; dy\frac{dF(y|\theta^b)}{dy} + d\theta^b \frac{dF(y|\theta^b)}{d\theta^b} = dy F'_y(y|\theta^b) + d\theta^b F'_{\theta^b}(y|\theta^b).$$

In the above, $f \equiv F'_y(\cdot)$ is the density of y given θ. The Jacobian is:

$$\left|\frac{d\theta^b}{dy}\right| = \left|\frac{F'_y(y|\theta^b)}{F'_{\theta^b}(y|\theta^b)}\right| = \left|\frac{f(y|\theta^b)}{F'_{\theta^b}(y|\theta^b)}\right|.$$

To find the distribution of θ^b conditional on y, assume $F(y,.)$ is increasing in θ:

$$\mathbb{P}\left(\theta^b \leq t|y\right) = \mathbb{P}\left(F(y|\theta^b) \leq F(y|t)|y\right)$$
$$= \mathbb{P}\left(\varepsilon^b \leq F(y|t)|y\right)$$
$$= F(y|t).$$

[1] We thank Neil Shephard for suggesting the example.

By construction, $f(\theta|y) = F'_\theta(y|\theta)$.[2] Putting things together,[3]

$$p_{\text{RS}}(\theta|y) \propto \pi(\theta)|F'_\theta(y|\theta)|\left|\frac{f(y|\theta)}{F'_\theta(y|\theta)}\right| = \pi(\theta)f(y|\theta) \propto p(\theta|y).$$

Example 3: Normal Mean and Variance We now consider an example in which the estimators can be derived analytically, and given in Forneron & Ng (2018). We assume $y_t = \varepsilon_t \sim N(m, \sigma^2)$. The parameters of the model are $\boldsymbol{\theta} = (m, \sigma^2)'$. We consider the auxiliary statistics: $\widehat{\boldsymbol{\psi}}(\mathbf{y})' = \begin{pmatrix} \overline{y} & \widehat{\sigma}^2 \end{pmatrix}$. The parameters are exactly identified.

The MLE of $\boldsymbol{\theta}$ is

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

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

$$\log p(\boldsymbol{\theta}|\widehat{m}, \widehat{\sigma}^2) \propto \frac{-T}{2}\log(2\pi)\sigma^2 - \alpha \log \sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - m)^2.$$

Since $\widehat{\boldsymbol{\psi}}(\mathbf{y})$ are sufficient statistics, the RS coincides with the likelihood-based Bayesian estimator, denoted B below. This is also the infeasible ABC estimator. We focus discussion on estimators for σ^2 which have more interesting properties. Under a uniform prior, we obtain

$$\begin{aligned}
\overline{\sigma}^2_B &= \widehat{\sigma}^2\frac{T}{T-5} \\
\widehat{\sigma}^2_{\text{SMD}} &= \frac{\widehat{\sigma}^2}{\frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T}(\varepsilon_t^s - \overline{\varepsilon}^s)^2} \\
\widehat{\sigma}^2_{RS} &= \sum_{b=1}^{B}\frac{\widehat{\sigma}^2}{\frac{[\sum_{t=1}^{T}(\epsilon_t^b - \overline{\epsilon}^b)^2/T]^2}{\sum_{k=1}^{B}\frac{1}{\sum_{t=1}^{T}(\epsilon_t^k - \overline{\epsilon}^k)^2/T}}}
\end{aligned}$$

In this example, the RS is also the ABC estimator with $\delta = 0$. It is straightforward to show that the bias reducing prior is $\alpha = 1$ and coincides with the SMD. Table 2.1 shows that the estimators are asymptotically equivalent but can differ for fixed T.

To highlight the role of the Jacobian matrix in the RS, the top panel of Figure 2.2 plots the exact posterior distribution and the one obtained from the reverse sampler. They are

[2] If $F(y, \cdot)$ is decreasing in θ, we have $\mathbb{P}(\theta^b \leq t|y) = 1 - F(y, t)$.

[3] An alternative derivation is to note that $t = \mathbb{P}(u \leq t|y) = \mathbb{P}\left(u = F(y, \theta^b) \leq t|y\right) = \mathbb{P}\left(\theta^b \leq F^{-1}(y, t) = t'|y\right)$. Hence $f(\theta^b|y) = \frac{dt}{dt'} = \frac{1}{(F^{-1})'_\theta(y,t)} = F'_2(y, t)$ as above.

Table 2.1: Properties of the Estimators

Estimator	Prior	$\mathbf{E}[\widehat{\boldsymbol{\theta}}]$	Bias	Variance	MSE
$\widehat{\boldsymbol{\theta}}_{ML}$	-	$\sigma^2\frac{T-1}{T}$	$-\frac{\sigma^2}{T}$	$2\sigma^4\frac{T-1}{T^2}$	$2\sigma^4\frac{2T-1}{2T^2}$
$\overline{\boldsymbol{\theta}}_{B}$	1	$\sigma^2\frac{T-1}{T-5}$	$\frac{2\sigma^2}{T-5}$	$2\sigma^4\frac{T-1}{(T-5)^2}$	$2\sigma^4\frac{T+1}{(T-5)^2}$
$\overline{\boldsymbol{\theta}}_{RS}$	1	$\sigma^2\frac{T-1}{T-5}$	$\frac{2\sigma^2}{T-5}$	$2\sigma^4\frac{T-1}{(T-5)^2}$	$2\sigma^4\frac{T+1}{(T-5)^2}$
$\widehat{\boldsymbol{\theta}}_{\text{SMD}}$	-	$\sigma^2\frac{S(T-1)}{S(T-1)-2}$	$\frac{2\sigma^2}{S(T-1)-2}$	$2\sigma^4\kappa_1\frac{1}{T-1}$	$2\sigma^4\frac{\kappa_1}{T-1}+\frac{4\sigma^4}{(S(T-1)-2))^2}$

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

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. Re-weighting by the Jacobian matrix is crucial to targeting the desired posterior distribution.

Figure 2.1 presents the likelihood based posterior distribution, along with the likelihood free ones produced by ABC and the RS-JI (just identified) for one draw of the data. The ABC results are based on the accept-reject algorithm. The numerical results corroborate with the analytical ones: all the posterior distributions are very similar. The RS-JI posterior distribution is very close to the exact posterior distribution. Figure 2.1 also presents results for the over-identified case (denoted RS-OI) using two additional auxiliary statistics: $\widehat{\boldsymbol{\psi}}=(\overline{y},\widehat{\sigma}_y^2,\widehat{\mu}_3/\widehat{\sigma}_y^2,\widehat{\mu}_4/\widehat{\sigma}_y^4)$ where $\mu_k=\mathbb{E}(y^k)$. The weight matrix is $\text{diag}(1,1,1/2,1/2)$. The posterior distribution is very close to RS-JI obtained for exact identification. We now explain how the posterior distribution for the over-identified case is obtained.

2.3 The RS: Case $L \geq K$:

The idea behind the RS is the same when we go from the case of exact to overidentification. The precise implementation is as follows. Let $\mathbb{K}_\delta(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\psi}}^b)$ be a kernel function and δ be a tolerance level such that $\mathbb{K}_0(\widehat{\boldsymbol{\psi}},\boldsymbol{\psi}^b)=\mathbb{I}_{\|\widehat{\boldsymbol{\psi}}-\widehat{\boldsymbol{\psi}}^b\|=0}$.

For $b=1,\ldots,B$

i Generate $\boldsymbol{\varepsilon}^b$ from $F_{\boldsymbol{\varepsilon}}$.

ii Find $\boldsymbol{\theta}^b=\text{argmin}_{\boldsymbol{\theta}}J_1^b(\widehat{\boldsymbol{\psi}}^b,\widehat{\boldsymbol{\psi}})$ where $\widehat{\boldsymbol{\psi}}^b=\widehat{\boldsymbol{\psi}}(\boldsymbol{\theta},\boldsymbol{\varepsilon}^b)$;

iii Set $w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) = \pi(\boldsymbol{\theta}^b)\text{vol}\left(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})\right)^{-1} \mathbb{K}_\delta(J_1^b(\widehat{\boldsymbol{\psi}}^b, \widehat{\boldsymbol{\psi}}))$ where: $\text{vol}(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}) = \sqrt{\left|\widehat{\boldsymbol{\psi}}^{b\prime}_{\boldsymbol{\theta}}\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}\right|}$.

iv Re-weigh $\boldsymbol{\theta}^b$ by $\frac{w(\boldsymbol{\theta}^b)}{\sum_{b=1}^B w(\boldsymbol{\theta}^b)}$.

We now proceed to explain the two changes:- the use of volume in place of determinant in the importance ratio, and the need for $L - K$ dimensional kernel smoothing.

The usual change of variable formula evaluates the absolute value of the determinant of the Jacobian matrix when the matrix is square. The determinant then gives the infinitesimal dilatation of the volume element in passing from one set of variables to another. The main issue in the case of overidentification is that the determinant of a rectangular Jacobian matrix is not well defined. However, as shown in Ben-Israel (1999), the determinant can be replaced by the volume when transforming from sets of a higher dimension to a lower one.[4] For a $L \times K$ matrix A, its volume, denoted $\text{vol}(A)$, is the product of the (non-zero) singular values of A:

$$\text{vol}(A) = \begin{cases} \sqrt{|A'A|} & L \geq K, \ \text{rank}(A) = K \\ \sqrt{|AA'|} & L \leq K, \ \text{rank}(A) = L. \end{cases}$$

Furthermore, if $A = BC$, $\text{vol}(A) = \text{vol}(B)\text{vol}(C)$.

To verify that our target distribution is unaffected by whether we calculate the volume or the determinant of the Jacobian matrix when $K = L$, observe that

$$\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b(\widehat{\boldsymbol{\psi}}), \boldsymbol{\varepsilon}^b) = \frac{\partial \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})}{\partial \widehat{\boldsymbol{\psi}}} \frac{\partial \widehat{\boldsymbol{\psi}}}{\partial \boldsymbol{\theta}^b}. \tag{2.5}$$

The K first order conditions defined by (2.3) become:

$$\mathcal{F}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}) = \widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})'\text{W}\left(\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)\right) = 0. \tag{2.6}$$

Since $L = K$, W can be set to an identity matrix I_K. Furthermore, $\boldsymbol{\psi}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}) = \widehat{\boldsymbol{\psi}}$ since $J_1^b(\boldsymbol{\theta}^b) = 0$ under exact identification. As $\frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}$ is a square matrix when $K = L$, we can directly use the fact that $\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})d\boldsymbol{\theta} + \mathcal{F}_{\boldsymbol{\psi}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})d\widehat{\boldsymbol{\psi}} = 0$ to obtain the required determinant:

$$|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})|^{-1} = \text{I}_K \cdot |\frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}| = |-\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})^{-1}\mathcal{F}_{\widehat{\boldsymbol{\psi}}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})|. \tag{2.7}$$

[4]From Ben-Israel (2001), $\int_V f(v)dv = \int_U f(\phi(u))\text{vol}\left(\phi_u(u)\right)du$ for a real valued function f integrable on V. See also `http://www.encyclopediaofmath.org/index.php/Jacobian`.

Now to use the volume result, put A $= \mathrm{I}_K$, B $= \frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}$ and C $= \frac{\partial \widehat{\boldsymbol{\psi}}}{\partial \boldsymbol{\theta}}$. But A is just a K-dimensional identity matrix. Hence $\mathrm{vol}(\mathrm{I}_K) = \mathrm{vol}\left(\frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}\right)\mathrm{vol}\left(\frac{\partial \widehat{\boldsymbol{\psi}}}{\partial \widehat{\boldsymbol{\theta}}}\right)$ which evaluates to

$$\mathrm{vol}\left(\frac{\partial \widehat{\boldsymbol{\psi}}}{\partial \boldsymbol{\theta}}\right)^{-1} = \mathrm{vol}\left(\frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}\right), \qquad \text{or} \qquad \left|\frac{\partial \widehat{\boldsymbol{\psi}}}{\partial \boldsymbol{\theta}}\right|^{-1} = \left|\frac{\partial \boldsymbol{\theta}}{\partial \widehat{\boldsymbol{\psi}}}\right|$$

which is precisely $|\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\varepsilon})|^{-1}$ as given in (2.7)[5]. Hence in the exactly identified case, there is no difference whether one evaluates the determinant or the volume of the Jacobian matrix.

Next, we turn to the role of the kernel function $\mathbb{K}_\delta(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\psi}}^b)$. The joint density $h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)$ is related to $p_{\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b}(\widehat{\boldsymbol{\psi}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b)) = p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b)$ through a change a variable now expressed in terms of volume:

$$h(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^b | \widehat{\boldsymbol{\psi}}) = p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b | \widehat{\boldsymbol{\psi}}) \cdot \mathrm{vol}\left(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})\right)$$

When $L \geq K$, the objective function $\|\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b\|_{\mathrm{W}} = J_1^b \geq 0$ measures the extent to which $\widehat{\boldsymbol{\psi}}$ deviates from $\widehat{\boldsymbol{\psi}}^b$ when the objective function at its minimum. Consider the thought experiment that $J_1^b = 0$ with probability 1, such as enabled by a particular draw of $\boldsymbol{\varepsilon}^b$. Then the arguments above for $K = L$ would have applied. We would still have $p(\boldsymbol{\theta}^b | \widehat{\boldsymbol{\psi}}) = \int \pi(\boldsymbol{\theta}^b) p(\widehat{\boldsymbol{\psi}}^b, \boldsymbol{\varepsilon}^b | \widehat{\boldsymbol{\psi}}) \mathbb{I}_{\|\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b\| = 0} d\boldsymbol{\varepsilon}^b = \int w(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) h(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b | \widehat{\boldsymbol{\psi}}) d\boldsymbol{\varepsilon}^b$, except that the weights are now defined in terms of volume. Proposition 1 would then extend to the case with $L \geq K$.

But in general $J_1^b \neq 0$ almost surely. Nonetheless, we can use only those draws that yield $J_1^b(\boldsymbol{\theta}^b)$ that are sufficiently close to zero. The more draws we make, the tighter this criterion can be. Suppose there is a symmetric kernel $\mathbb{K}_\delta(\cdot)$ satisfying conditions in Pagan & Ullah (1999, p.96) for consistent estimation of conditional moments non-parametrically. Analogous to Proposition 2, the volume $\mathrm{vol}\big(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})\big)$ is assumed to be bounded away from zero. Then as the number of draws $B \to \infty$, the bandwidth $\delta(B) \to 0$ and $B\delta(B) \to \infty$ with

$$w_{\delta(B)}(\boldsymbol{\theta}^b, \widehat{\boldsymbol{\varepsilon}}^b) \;=\; \pi(\boldsymbol{\theta}^b) \mathrm{vol}\left(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})\right)^{-1} \mathbb{K}_{\delta(B)}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\psi}}^b), \tag{2.8}$$

[5]Using the implicit function theorem to compute the gradient gives the same result. Since $\widehat{\boldsymbol{\psi}}^b = \widehat{\boldsymbol{\psi}}$ we have: $\mathcal{F}_{\boldsymbol{\theta}} = -\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})' \mathrm{W} \widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}) + \sum_j \widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}, \boldsymbol{\theta}_j}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b) W \left(\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})\right) = -\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})' W \widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}})$. Then $\mathrm{vol}(\mathcal{F}_{\boldsymbol{\theta}}^{-1} \mathcal{F}_{\widehat{\boldsymbol{\psi}}}) = \mathrm{vol}(\mathcal{F}_{\boldsymbol{\theta}}^{-1}) \mathrm{vol}(\mathcal{F}_{\widehat{\boldsymbol{\psi}}}) = \mathrm{vol}(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}))^{-1} |W|^{-1} \mathrm{vol}(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}))^{-1} \mathrm{vol}(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}))^{-1} |W| = \mathrm{vol}(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b, \boldsymbol{\varepsilon}^b, \widehat{\boldsymbol{\psi}}))^{-1}$. Hence the weights are the same when we only consider the draws where $J_1^b = 0$ which are the draws we are interested in.

a result analogous to Proposition 1 can be obtained:

$$\begin{aligned}
&\frac{1}{B}\sum_b w_{\delta(B)}(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)\varphi(\boldsymbol{\theta}^b)\xrightarrow{p}\iint \varphi(\boldsymbol{\theta})w_0(\boldsymbol{\theta},\boldsymbol{\varepsilon})\text{vol}\left(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\varepsilon}^b;\widehat{\boldsymbol{\psi}})\right)p(\widehat{\boldsymbol{\psi}},\boldsymbol{\varepsilon}^b|\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\varepsilon}^b\\
&=\iint \varphi(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathbb{1}_{\|\widehat{\boldsymbol{\psi}}-\widehat{\boldsymbol{\psi}}^b\|=0}\text{vol}\left(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\varepsilon}^b;\widehat{\boldsymbol{\psi}})\right)^{-1}p(\widehat{\boldsymbol{\psi}}^b,\boldsymbol{\varepsilon}^b|\boldsymbol{\theta})\text{vol}\left(\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\varepsilon}^b;\widehat{\boldsymbol{\psi}})\right)d\boldsymbol{\theta}d\boldsymbol{\varepsilon}^b\\
&=\iint \varphi(\boldsymbol{\theta})\pi(\boldsymbol{\theta})1_{\|\widehat{\boldsymbol{\psi}}-\widehat{\boldsymbol{\psi}}^b\|=0}p(\widehat{\boldsymbol{\psi}},\boldsymbol{\varepsilon}^b|\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\varepsilon}^b\\
&=\int \varphi(\boldsymbol{\theta})\pi(\boldsymbol{\theta})L(\widehat{\boldsymbol{\psi}}|\boldsymbol{\theta})d\boldsymbol{\theta}.
\end{aligned}$$

Similarly, the integrating constant is consistent as $\frac{1}{B}\sum_b w_{\delta(B)}(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)\xrightarrow{p}\int \pi(\boldsymbol{\theta})L(\widehat{\boldsymbol{\psi}}|\boldsymbol{\theta})d\boldsymbol{\theta}$. Hence, the RS sampler still recovers the posterior distribution with the infeasible likelihood. Note that the kernel function was introduced for developing a result analogous to Proposition 1, but no kernel smoothing is required in practical implementation. What is needed for the RS in the over-identified case is B draws with sufficiently small $J_1(\boldsymbol{\theta}^b)$. Hence, we can borrow the idea used in the AR-ABC. Specifically, we fix a quantile q, repeat $[B/q]$ times until the desired number of draws is obtained. Discarding some draws seems necessary in many ABC implementations.

In summary, there are two changes in implementation of the RS in the over-identified case: the volume and the kernel function. Kernel smoothing has no role in the RS when $K = L$. It is interesting to note that while the ABC and RS both rely on the kernel $\mathbb{K}_\delta$ to keep draws close to $\widehat{\boldsymbol{\psi}}^b$ in the over-identified case, the non-parametric rate at which the sum converges to the integral are different. The RS uses the first order conditions $\widehat{\boldsymbol{\psi}}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)'\text{W}\left(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)-\widehat{\boldsymbol{\psi}}\right)=0$ to indicate which K combinations of $\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)-\widehat{\boldsymbol{\psi}}$ are set to zero, rendering the dimension of the smoothing problem $L-K$. To see this, note first that each draw $\boldsymbol{\theta}^b$ from the RS is consistent for $\boldsymbol{\theta}_0$ and asymptotically normal as shown in Forneron & Ng (2018). In consequence, the first order condition (FOC) can be re-written as: $\left(\frac{d\boldsymbol{\psi}(\boldsymbol{\theta})}{d\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}+O_p(\frac{1}{\sqrt{T}})\right)'\text{W}\left(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)-\widehat{\boldsymbol{\psi}}\right)=0$, or

$$\frac{d\boldsymbol{\psi}(\boldsymbol{\theta})}{d\boldsymbol{\theta}}|'_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\text{W}\left(\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)-\widehat{\boldsymbol{\psi}}\right)=o_p(\frac{1}{\sqrt{T}}).$$

Since $\frac{d\boldsymbol{\psi}(\boldsymbol{\theta})}{d\boldsymbol{\theta}}|'_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\text{W}$ is full rank, there exists a subspace of dimension K such that $\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b,\boldsymbol{\varepsilon}^b)-\widehat{\boldsymbol{\psi}}$ is zero asymptotically. Hence the kernel smoothing problem is effectively $L-K$ dimensional. The ABC does not use the FOC. Even in the exactly identified case, the kernel smoothing is a $L=K$ dimensional problem. In general, the convergence rate of the ABC is $L\geq K$, the dimension of $\widehat{\boldsymbol{\psi}}$.

The following two examples illustrate the properties of the ABC and RS posterior distributions. The first example uses sufficient statistics and the second example does not. Both the ABC and RS achieve the desired number of draws by setting the quantile, as discussed in Section 2.

Example 4: Exponential Distribution Let $y_1, \ldots, y_T \sim \mathcal{E}(\theta), T = 5, \theta_0 = 1/2$. Now $\widehat{\psi} = \overline{y}$ is a sufficient statistic for $y_1, \ldots, y_T$. For a flat prior $\pi(\theta) \propto 1_{\theta \geq 0}$ we have:

$$p(\theta|\overline{y}) \propto p(\theta|y_1, \ldots, y_T) = \theta^T \exp(-\theta^T \overline{y}) \sim \Gamma(T+1, T\overline{y})$$

In the just identified case, we let $u_t^b \sim \mathcal{U}_{[0,1]}$ and $y_t^b = -\log(1-u_t^b)/\theta^b$. This gives:

$$\widehat{\psi}^b = \frac{1}{T}\sum_{t=1}^{T} y_t^b = -\frac{1}{T}\sum_{t=1}^{T} \frac{\log(1-u_t^b)}{\theta^b}.$$

Since $\overline{y}^b = \overline{y}$, the Jacobian matrix is:

$$\widehat{\psi}_b(\theta^b) = \frac{d\widehat{\psi}^b(\theta)}{d\theta}\bigg|_{\theta^b} = \frac{1}{T}\sum_{t=1}^{T} \frac{\log(1-u_t^b)}{[\theta^b]^2} = -\frac{\overline{y}}{\theta^b}.$$

Hence for a given T, the weights are: $w(\theta^b, u^b) \propto \mathbb{I}_{\boldsymbol{\theta}^b \geq 0} \frac{\theta^b}{\overline{y}^b} = \frac{\theta^b}{\overline{y}}$. We verified that the numerical results agree with this analytical result.

In the over identified case, we consider two moments:

$$\widehat{\boldsymbol{\psi}}^b = \begin{pmatrix} \overline{y}^b \\ \widehat{\sigma}_y^{b,2} \end{pmatrix} = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} y_t^b \\ \frac{1}{T}\sum_{t=1}^{T}(y_t^b)^2 - (\frac{1}{T}\sum_{t=1}^{T} y_t^b)^2 \end{pmatrix}.$$

Since $\frac{dy_t^b}{d\theta} = \frac{\log(1-u_t^b)}{(\theta^b)^2} = -\frac{y_t^b}{\theta^b}$. If $\delta = 0$, the Jacobian matrix is

$$\widehat{\boldsymbol{\psi}}_\theta^b = -\begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} \frac{y_t^b}{\theta} \\ \frac{2}{\theta^b}\frac{1}{T}\sum_{t=1}^{T}(y_t^b)^2 - \frac{2}{\theta^b}\left[\frac{1}{T}\sum_{t=1}^{T} y_t^b\right]^2 \end{pmatrix} = -\begin{pmatrix} \frac{\overline{y}}{\theta^b} \\ \frac{2(\widehat{\sigma}_y)^2}{\theta^b} \end{pmatrix}.$$

The volume to be computed is $\text{vol}(\widehat{\boldsymbol{\psi}}_\theta^b) = \sqrt{|\widehat{\boldsymbol{\psi}}_\theta^{b\prime}\widehat{\boldsymbol{\psi}}_\theta^b|}$, as stated in the algorithm. Even if W = I, the volume is the determinant of $\widehat{\boldsymbol{\psi}}_\theta^b$ in the exactly identified case, plus a term relating to the variance of y^b. We computed $\widehat{\boldsymbol{\psi}}_\theta^b$ for draws with $J_1^b \approx 0$ using numerical differentiation[6] and verified that the values are very close to the ones computed analytically for this example.

[6]In practice, since the mapping $\theta \to \widehat{\psi}^b(\theta)$ is not known analytically, the derivatives are approximated using finite differences: $\partial_{\theta_j}\widehat{\psi}^b(\theta) \simeq \frac{\widehat{\psi}^b(\theta+e_j\varepsilon)-\widehat{\psi}^b(\theta-e_j\varepsilon)}{2\varepsilon}$ for $\varepsilon \simeq 0$.

Figure 2.3 depicts a particular draw of the ABC posterior distribution (which coincides with the likelihood-based posterior since the statistics are sufficient), along with two generated by the RS sampler. The first one uses the sample mean as auxiliary statistic and hence is exactly identified. The second uses two auxiliary statistics: the sample mean and the sample variance. For the AR-ABC, we draw from the prior ten million times and keep the ten thousand nearest draws. This corresponds to a value of $\delta = 0.0135$. For the RS, we draw one million times[7] and keep the ten thousand nearest draws which corresponds to a $\delta = 0.0001$. As for the weight matrix W, if we put $W_{11} > 0$ and zero elsewhere, we will recover the exactly identified distribution. Here, we intentionally put a positive weight on the variance (which is not a sufficient statistic) to check the effect on the posterior mean. With $W_{11} = 1/5$ and $W_{22} = 4/5$, the RS posterior means are 0.7452 and 0.7456 for the just and overidentifed cases. The corresponding values are are 0.7456 and .7474 for the exact posterior and the ABC-AR. They are very similar.

Example 5: ARMA(1,1): For $t = 1, \ldots, T = 200$ and $\boldsymbol{\theta}_0 = (\alpha_0, \theta_0, \sigma_0) = (0.5, 0.5, 1.0)$, the data are generated as

$$y_t = \alpha y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

Least squares estimation of the auxiliary model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + u_t$$

yields $L = 5 > K = 3$ auxiliary parameters

$$\widehat{\boldsymbol{\psi}} = (\widehat{\phi}_1, \widehat{\phi}_2, \widehat{\phi}_3, \widehat{\phi}_4, \widehat{\sigma}_u^2).$$

We let $\pi(\alpha, \theta, \sigma) = \mathbb{I}_{\alpha, \theta \in [-1,1], \sigma \geq 0}$ and $\mathrm{W} = \mathrm{I}_5$ which is inefficient. In this example, $\widehat{\boldsymbol{\psi}}$ are not sufficient statistics since y_t has an infinite order autoregressive representation.

We draw σ from a uniform distribution on $[0, 3]$ since $\mathcal{U}_{[0,\infty]}$ is not a proper density. The weights of the RS are obtained by numerical differentiation. The likelihood based posterior is computed by MCMC using the Kalman Filter with initial condition $\varepsilon_0 = 0$. As mentioned above, the desired number of draws is obtained by setting the quantile instead of setting the tolerance δ. For the RS, we keep the 1/10=10% closest draws corresponding

[7]This means that we solve the optimization problem one million times. Given that the optimization problem is one dimensional, the one dimensional R optimization routine *optimize* is used. It performs a combination of the golden section with parabolic interpolations. The optimum is found, up to a given tolerance level (the default is 10^{-4}), over the interval $[0, 10]$.

to a $\delta = 0.0007$. The Sequential Monte-Carlo implementation of ABC (SMC-ABC) is more efficient at targeting the posterior than the ABC-AR. Hence we also compare the RS with SMC-ABC as implemented in the Easy-ABC package of Lenormand et al. (2013).[8] The requirement for 10,000 posterior draws are as follows:

	AR-ABC	SMC-ABC	RS	Likelihood
Computation Time (hours)	63	25	5	0.1
Effective number of draws	100,000,000	36,805,000	10,153,108	
δ	0.0132	0.0283	0.0007	

The difference, both in terms of computation time and number of model simulations, is notable. As shown in figure 2.4 the quality of the approximation is also different, especially for α and σ. The difference can be traced to δ. The δ used for the SMC-ABC is effectively much larger than for the RS. A better approximation requires a smaller δ which implies longer computational time. Alternatively stated, the acceptance rate at a low value of δ is very low. The caveat is that the speed gain is possible only if the optimization problem can be solved in a few iterations and reasonably fast. In practice, there will be a trade-off between the number of draws and the number of iterations in the optimization step as we further explore below.

2.4 Acceptance Rate

The RS was initially developed in Forneron & Ng (2018) as a framework to help understand frequentist (SMD) and the Bayesian (ABC) way of likelihood-free estimation. But it turns out that the RS has one computation advantage that is worth highlighting. The issue pertains to the low acceptance rate of the ABC.

As noted above, the ABC exactly recovers the posterior distribution associated with the infeasible likelihood if $\widehat{\psi}$ are sufficient statistics and $\delta = 0$ as noted in Blum (2010). Of course, $\delta = 0$ is an event of measure zero, and the ABC has an approximation bias that depends on δ. In theory, a small δ is desired. The ABC needs a large number of draws to accurately approximate the posterior and can be computationally costly.

[8]We implemented the SMC-ABC in two ways. First, we use the procedure inVo et al. (2015) using code generously provided by Christopher Drovandi. We also use the Easy-ABC package in *R* of Lenormand et al. (2013). We thank an anonymous referee for this suggestion.

To illustrate this point, consider estimating the mean m in Example 3 with $\sigma^2 = 1$ assumed to be known, and $\pi(m) \propto 1$. All computations are based on the software package R. From a previous draw m^b, a random walk step gives $m^\star = m^b + \varepsilon, \varepsilon \sim \mathcal{N}(0,1)$. For small δ, we can assume $m^\star|\widehat{m} \sim \mathcal{N}(\widehat{m}, 1/T)$. From a simulated sample of T observations, we get an estimated mean $\widehat{m}^\star \sim \mathcal{N}(m^\star, 1/T)$. As is typical of MCMC chains, these draws are serially correlated. To see that the algorithm can be stuck for a long time if m^* is far from $\widehat{m}$, observe that the event $\widehat{m}^\star \in [\widehat{m} - \delta, \widehat{m} + \delta]$ occurs with probability

$$\mathbb{P}(\widehat{m}^\star \in [\widehat{m} - \delta, \widehat{m} + \delta]) = \Phi\left(\sqrt{T}(\widehat{m} + \delta - m^\star)\right) - \Phi\left(\sqrt{T}(\widehat{m} - \delta - m^\star)\right) \approx 2\sqrt{T}\delta\phi\left(\sqrt{T}(\widehat{m} - m^\star)\right).$$

The acceptance probability $\int_{m^*} \mathbb{P}(\widehat{m}^\star \in [\widehat{m} - \delta, \widehat{m} + \delta])dm^*$ is thus approximately linear in δ. To keep the number of accepted draws constant, we need to increase the number of draws as we decrease δ.

This result that the acceptance rate is linear in δ also applies in the general case. Assume that $\widehat{\psi}^\star(\boldsymbol{\theta}^\star) \sim \mathcal{N}(\psi(\boldsymbol{\theta}^\star), \Sigma/T)$. We keep the draw if $\|\widehat{\psi} - \widehat{\psi}^\star(\boldsymbol{\theta}^\star)\| \leq \delta$. The probability of this event can be bounded above by $\sum_{j=1}^K \mathbb{P}\left(|\widehat{\psi}_j - \widehat{\psi}_j^\star(\boldsymbol{\theta}^\star)| \leq \delta\right)$ i.e.:

$$\sum_{j=1}^K \Phi\left(\frac{\sqrt{T}}{\sigma_j}\left(\widehat{\psi}_j + \delta - \psi_j(\boldsymbol{\theta}^\star)\right)\right) - \Phi\left(\frac{\sqrt{T}}{\sigma_j}\left(\widehat{\psi}_j - \delta - \psi_j(\boldsymbol{\theta}^\star)\right)\right) \approx 2\sqrt{T}\delta\sum_{j=1}^K \frac{\phi}{\sigma_j}\left(\frac{\sqrt{T}}{\sigma_j}\left(\widehat{\psi}_j - \psi_j(\boldsymbol{\theta}^\star)\right)\right).$$

The acceptance probability is still at best linear in δ. In general we need to increase the number of draws at least as much as δ declines to keep the number of accepted draws fixed.

Table 2.2: Acceptance Probability as a function of δ

δ	10	1	0.1	0.01	0.001
$\mathbb{P}(\|\widehat{\psi} - \widehat{\psi}^b\|_W \leq \delta)$	0.72171	0.16876	0.00182	0.00002	<0.00001

Table 2.2 shows the acceptance rate for Example 3 for $\boldsymbol{\theta}_0 = (m_0, \sigma_0^2) = (0, 2)$, $T = 20$, and weighting matrix $W = \text{diag}(\widehat{\sigma}^2, 2\widehat{\sigma}^4)/T$, $\pi(m, \sigma^2) \propto \mathbb{I}_{\sigma^2 \geq 0}$. The results confirm that for small values of δ, the acceptance rate is approximately linear in δ. Even though in theory, the targeted ABC posterior should be closer to the true posterior when δ is small, this may not be true in practice because of the poor properties of the MCMC chain. At least for this example, the MCMC chain with moderate value of δ provides a better approximation to the true posterior density.

To overcome the low acceptance rate issue, Beaumont et al. (2009) suggests to use local regression techniques to approximate $\delta = 0$ without setting it equal to zero. The

convergence rate is then non-parametric. Gao & Hong (2014) analyzes the estimator of Creel & Kristensen (2013) and finds that to compensate for the large variance associated with the kernel smoothing, the number of simulations need to be larger than $T^{K/2}$ to achieve $\sqrt{T}$ convergence, where K is the number of regressors. Other methods that aim to increase the acceptance rate include the ABC-SMC algorithm of Sisson et al. (2007); Sisson & Fan (2011), as well as the adaptive weighting variant due to Bonassi & West (2015), referred to below as SMC-AW. These methods build a sequence of proposals to more efficiently target the posterior. The acceptance rate still declines rapidly with δ, however.

The RS circumvents this problem because each $\boldsymbol{\theta}^b$ is accepted by virtue of being the solution of an optimization problem, and hence $\widehat{\boldsymbol{\psi}} - \widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b)$ is the smallest possible. In fact, in the exactly identified case, $\delta = J_1^b = 0$. Furthermore, the sequence of optimizers are independent, and the sampler cannot be stuck. We use two more examples to highlight this feature.

Example 6: Mixture Distribution Consider the example in Sisson et al. (2007), also considered in Bonassi & West (2015). Let $\pi(\theta) \propto 1_{\theta \in [-10,10]}$ and

$$x|\theta \sim 1/2\mathcal{N}(\theta, 1) + 1/2\mathcal{N}(\theta, 1/100)$$

Suppose we observe one draw $x = 0$. Then the true posterior is $\theta|x \sim 1/2\mathcal{N}(0,1) + 1/2\mathcal{N}(0,1/100)$ truncated to $[-10,10]$. As in Sisson et al. (2007) and Bonassi & West (2015), we choose three tolerance levels: $(2, 0.5, 0.025)$ for AR-ABC. Figure 2.5 shows that the ABC posterior distributions computed using accept-reject sampling with $\delta = 0.025$ are similar to the ones using SMC with and without adaptive weighting. The RS posterior distribution is close to both ABC-SMC and ABC-SMC-AW, and all similar to Figure 3 reported in Bonassi & West (2015). However, they are quite different from the AR-ABC with $\delta = 2$ and 0.5 are 2, showing that the choice of δ is important in ABC.

While the SMC, RS, and ABC-AR sampling schemes can produce similar posterior distributions, Table 2.3 shows that their computational time differ dramatically. The two SMC algorithms need to sample from a multinomial distribution which are evidently more time consuming. When $\delta = 0.25$, the AR-ABC posterior distribution is close to the ones produced by the SMC samplers and the RS, but the computational cost is still high. The AR-ABC is computationally efficient when δ is large, but as seen from Figure 2.5, the posterior distribution is quite poorly approximated. The RS takes 0.0017 seconds to solve, which is amazingly fast because for this example, the solution is available analytically.

No optimization is involved, and there is no need to evaluate the Jacobian because the model is linear. Of course, in cases when the SMD problem is numerically challenging, numerical optimization can be time consuming as well. Our results nonetheless suggest a role for optimization in Bayesian computation; they need not be mutually exclusive. Combining the ideas is an interesting topic for future research.

Table 2.3: Computation Time (in seconds)

RS	ABC-AR			ABC-SMC	
	δ=2	δ=.5	δ=.025	Sisson et-al	Bonassi-West
.0017	0.4973	1.6353	33.8136	190.1510	199.1510

Example 7: Precautionary Savings The foregoing examples are simple and are serve illustrative purposes. We now consider an example that indeed has an infeasible likelihood. In Deaton (1991), agents maximize expected utility $\mathbb{E}_0\left(\sum_{t=0}^{\infty}\beta^t u(c_t)\right)$ subject to the constraint that assets $a_{t+1} = (1+r)(a_t + y_t - c_t)$ are bounded below by zero, where r is interest rate, y is income and c consumption. The desire for precautionary saving interacts with borrowing constraints to generate a policy function that is not everywhere concave, but is a piecewise linear when cash-on-hand is below an endogenous threshold. The policy function can only be solved numerically at assumed parameter values. SMD estimation thus consists of solving the model and simulating S auxiliary statistics at each guess $\boldsymbol{\theta}$. Michaelides & Ng (2000) evaluate the finite sample properties of several SMD estimators using a model with similar features. Since the likelihood for this model is not available analytically, Bayesian estimation of this model has not been implemented. Here, we use the RS to approximate the posterior distribution.

We generate $T = 400$ observations assuming that $U(c) = \frac{c^{1-\gamma}-1}{1-\gamma}$, $y_t \sim$ iid $\mathcal{N}(\mu, \sigma^2)$ with $r = 0.05$, $\beta = 10/11$, $\mu = 100, \sigma = 10, \gamma = 2$ as true values. We estimate $\boldsymbol{\theta} = (\gamma, \mu, \sigma)$ and assume (β, r) are known. We use 10 auxiliary statistics:

$$\widehat{\psi} = \begin{pmatrix} \overline{y} & \widehat{\Gamma}_{yy}(0) & \widehat{\Gamma}_{aa}(0) & \widehat{\Gamma}_{cc}(0) & \widehat{\Gamma}_{cc}(1) & \widehat{\Gamma}_{aa}(1) & \widehat{\Gamma}_{cc}(2) & \widehat{\Gamma}_{aa}(2) & \widehat{\Gamma}_{cy}(0) & \widehat{\Gamma}_{ay}(0) \end{pmatrix}'$$

where $\widehat{\Gamma}_{ab}(j) = \frac{1}{T}\sum_{t=1}^{T}(a_t - \overline{a})(b_{t-j} - \overline{b})$. We generate $B = 13,423$ draws and keep the $3,356$ (25%) nearest draws to $\widehat{\psi}$. After weighting using the volume of the Jacobian matrix we have an effective sample size of $1,421$ draws.[9] We use an identity weighting matrix

[9]The effective sample size is computed as $1/\sum_{b=1}^{B} w_b^2$ where the weights satisfy $\sum_{b=1}^{B} w^b = 1$.

so $J_{RS}(\boldsymbol{\theta}) = \overline{g}(\boldsymbol{\theta})'\overline{g}(\boldsymbol{\theta})$. The Jacobian is computed using finite differences for the RS. As benchmark, we also compute an SMD with $S = 100$, $J_S = \overline{g}_S(\boldsymbol{\theta})'\overline{g}_S(\boldsymbol{\theta})$. In this exercise, the SMD only needs to solve for the policy function once at each step of the optimization. Hence the binding function can be approximated using simulated data at a low cost. For this example, the programs are coded in PYTHON. The Nelder-Mead method is used for optimization.

Table 2.4: Deaton Model: RS, SMD with W = I

	Posterior Mean/Estimate			Posterior SD/SE		
	γ	μ	σ	γ	μ	σ
RS	1.86	99.92	10.48	0.19	0.84	0.37
SMD	1.76	99.38	10.31	0.12	0.60	0.34

Figure 2.6 shows the posterior distribution of the RS (blue) along with the SMD distribution (purple) as approximated by $\mathcal{N}(\widehat{\boldsymbol{\theta}}_{\text{SMD}}, \widehat{\text{V}}_{\text{SMD}}/T)$ according to asymptotic theory. Table 2.4 shows that the two sets of point estimates are similar. As explained in Forneron & Ng (2018), the SMD uses simulations to approximate the binding function while the RS (and by implication the ABC) uses simulations to approximate the infeasible posterior distribution. In this example, the difference in bias is quite small. We should note that the RS took well over a day to solve while the SMD took less than three hours to compute. Whether we use our own code for the ABC-MCMC or from available packages, the acceptance rate is too low for the exercise to be feasible.

2.5 Conclusion

This paper studies properties of the reverse sampler considered in Forneron & Ng (2018) for likelihood-free estimation. The sampler produce draws from the infeasible posterior distribution by solving a sequence of frequentist SMD problems. We showed that the reverse sampler uses the Jacobian matrix as importance ratio. In the over-identified case, the importance ratio can be computed using the volume of the Jacobian matrix. The reverse sampler does not suffer from the problem of low acceptance rate that makes the ABC computationally demanding.

Figure 2.1: Normally Distributed data

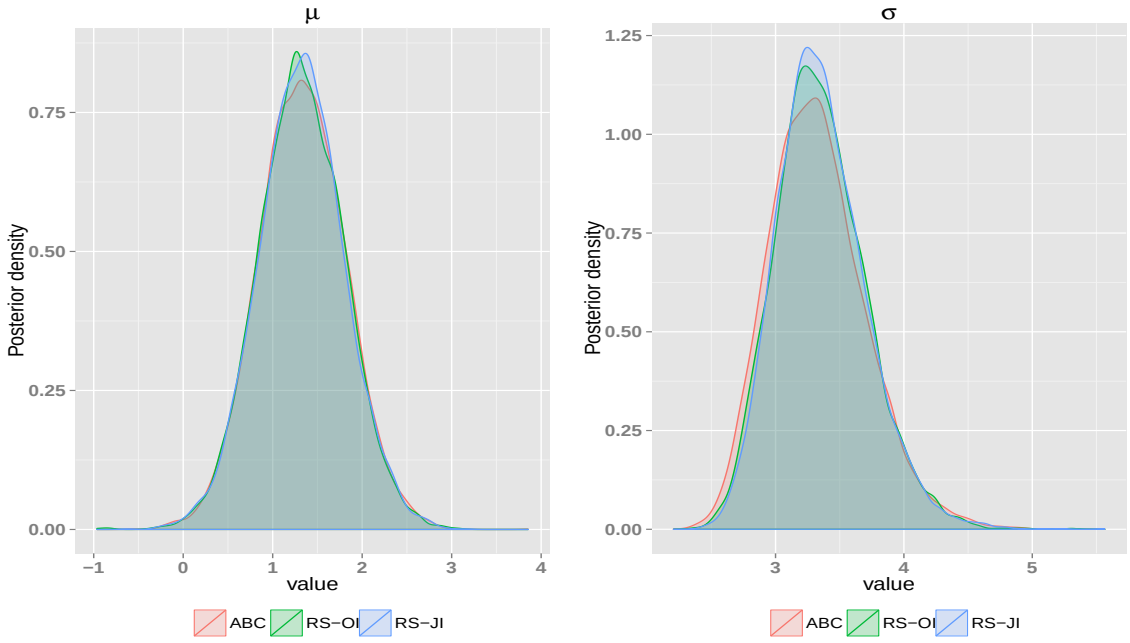


Figure 2.2: The Importance Weights in RS

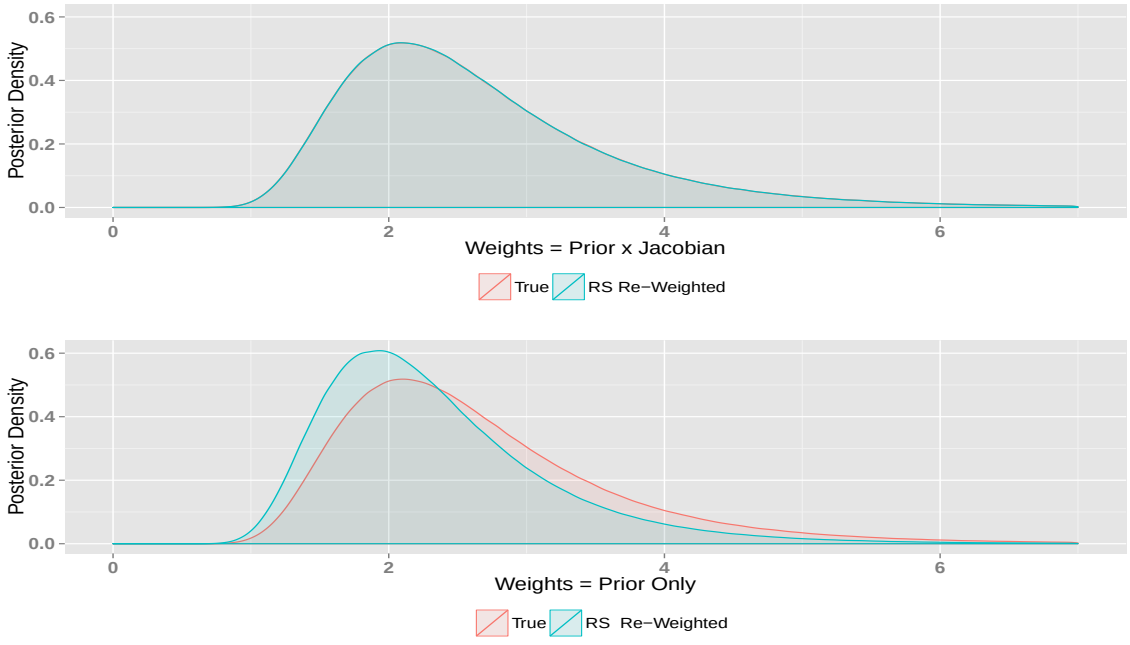


Figure 2.3: Exponential Distribution

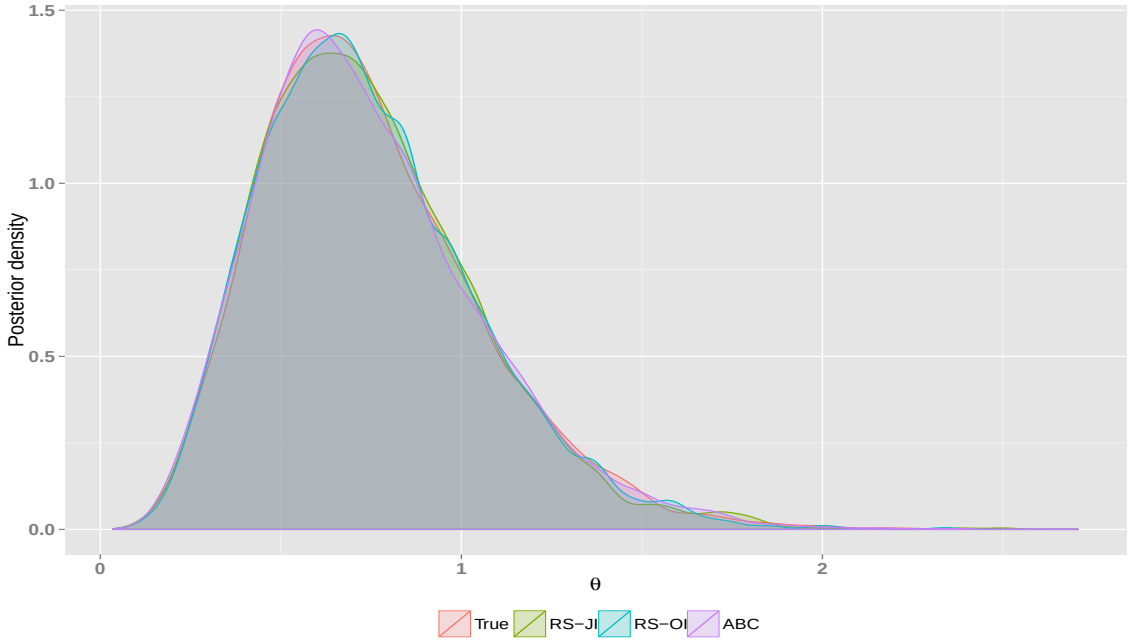


Figure 2.4: ARMA Model

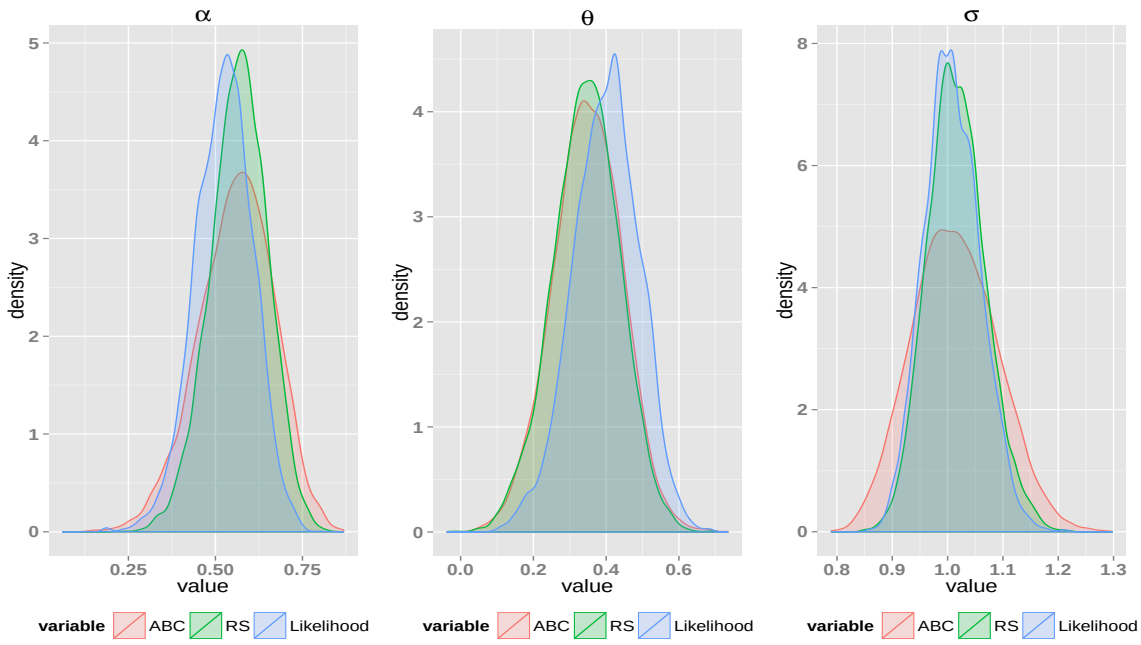


Figure 2.5: Mixture Distribution

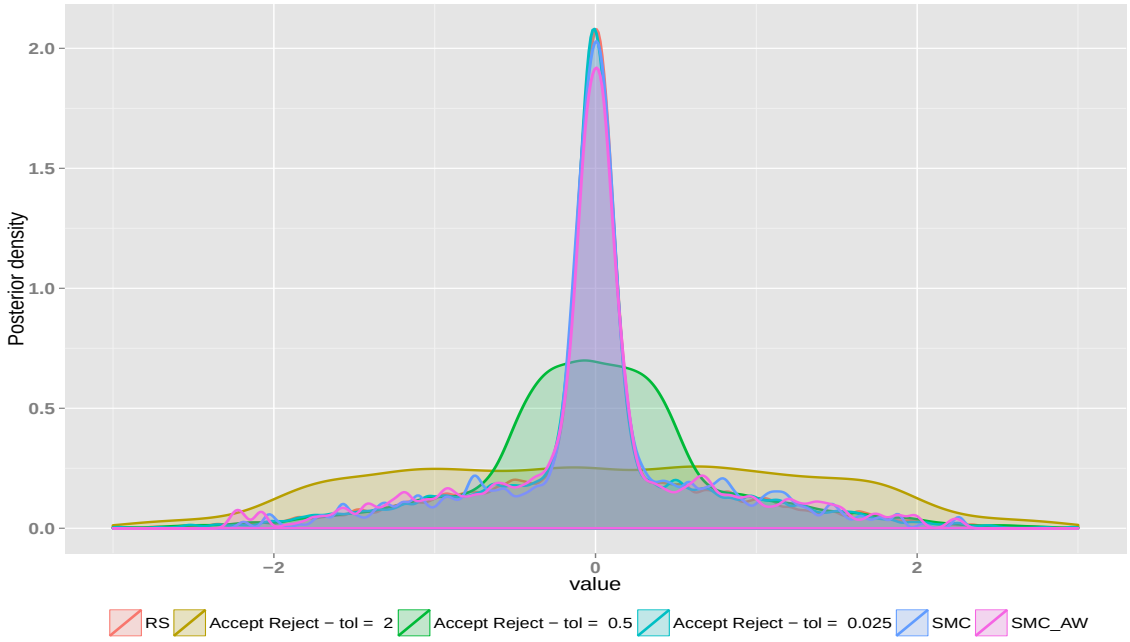
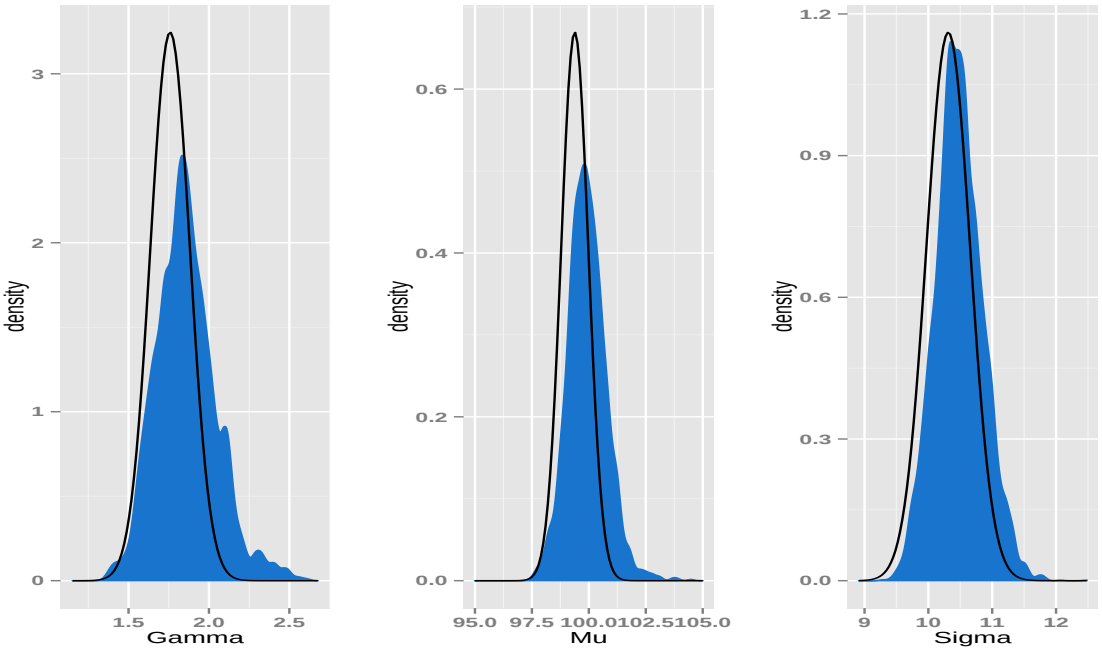


Figure 2.6: Deaton Model: RS and SMD



Note: Blue density: RS posterior, Black line: large sample approximation for the SMD estimator (identity weighting matrix).

Chapter 3

A Sieve-SMM Estimator for Dynamic Models

JEAN-JACQUES FORNERON[†]

[†]I am indebted to my advisor Serena Ng for her continuous guidance and support. I also greatly benefited from comments and discussions with Jushan Bai, Tim Christensen, Benjamin Connault, Gregory Cox, Ronald Gallant, Dennis Kristensen, Sokbae (Simon) Lee, Kim Long-Forneron, José Luis Montiel Olea, Christoph Rothe, Bernard Salanié and the participants of the Columbia Econometrics Colloquium.

3.1 Introduction

Complex nonlinear dynamic models with an intractable likelihood or moments are increasingly common in economics. A popular approach to estimating these models is to match informative sample moments with simulated moments from a fully parameterized model using SMM. However, economic models are rarely fully parametric since theory usually provides little guidance on the distribution of the shocks. The Gaussian distribution is often used in applications but in practice, different choices of distribution may have different economic implications; this is illustrated below. Yet to address this issue, results on semiparametric simulation-based estimation are few.

This paper proposes a Sieve Simulated Method of Moments (Sieve-SMM) estimator for both the structural parameters and the distribution of the shocks and explains how to implement it. The dynamic models considered here have the form:

$$y_t = g_{obs}(y_{t-1}, x_t, \boldsymbol{\theta}, f, u_t) \tag{3.1}$$

$$u_t = g_{latent}(u_{t-1}, \boldsymbol{\theta}, f, e_t), \quad e_t \sim f \tag{3.2}$$

The observed outcome variable is y_t, x_t are exogenous regressors and u_t is an unobserved latent process. The unknown parameters include $\boldsymbol{\theta}$, a finite dimensional vector, and the distribution f of the shocks e_t. The functions g_{obs}, g_{latent} are known, or can be computed numerically, up to $\boldsymbol{\theta}$ and f. The Sieve-SMM estimator extends the existing Sieve-GMM literature to more general dynamics with latent variables and the literature on sieve simulation-based estimation of some static models.

The estimator in this paper has two main building blocks: the first one is a sample moment function, such as the empirical characteristic function (CF) or the empirical CDF; infinite dimensional moments are needed to identify the infinite dimensional parameters. As in the finite dimensional case, the estimator simply matches the sample moment function with the simulated moment function. To handle this continuum of moment conditions, this paper adopts the objective function of Carrasco & Florens (2000); Carrasco et al. (2007a) in a semi-nonparametric setting.

The second building block is to nonparametrically approximate the distribution of the shocks using the method of sieves, as numerical optimization over an infinite dimension space is generally not feasible. Typical sieve bases include polynomials and splines which approximate smooth regression functions. Mixtures are particularly attractive to approximate densities for three reasons: they are computationally cheap to simulate from, they are known to have good approximation properties for smooth densities, and draws

from the mixture sieve are shown to satisfy the L^2-smoothness regularity conditions of the moments required for the asymptotic results. Restrictions on the number of mixture components, the tails and the smoothness of the true density ensure that the bias is small relative to the variance so that valid inferences can be made in large samples. To handle potentially fat tails, this paper introduces a Gaussian and tails mixture. The tail densities in the mixture are constructed to be easy to simulate from and also satisfy L^2-smoothness properties. The algorithm below summarizes the steps required to compute the estimator.

ALGORITHM: Computing the Sieve-SMM Estimator

Set a sieve dimension $k(n) \geq 1$ and a number of lags $L \geq 1$.
Compute $\widehat{\psi}_n$, the Characteristic Function (CF) of $(y_t, \ldots, y_{t-L}, x_t, \ldots, x_{t-L})$.
for $s = 1, \ldots, S$ **do**
Simulate the shocks e_t^s from $f_{\omega,\mu,\sigma}$: a $k(n)$ component Gaussian and tails mixture distribution with parameters (ω, μ, σ).
Simulate artificial samples $(y_1^s, \ldots, y_n^s)$ at $(\boldsymbol{\theta}, f_{\omega,\mu,\sigma})$ using e_t^s.
Compute $\widehat{\psi}_n^s(\boldsymbol{\theta}, f_{\omega,\mu,\sigma})$, the CF of the simulated data $(y_t^s, \ldots, y_{t-L}^s, x_t, \ldots, x_{t-L})$.
Compute the average simulated Characteristic Function $\widehat{\psi}_n^S(\boldsymbol{\theta}, f_{\omega,\mu,\sigma}) = \frac{1}{S}\sum_{s=1}^{S} \widehat{\psi}_n^s(\boldsymbol{\theta}, f_{\omega,\mu,\sigma})$.
Compute the objective function $\widehat{Q}_n^S(\boldsymbol{\theta}, f_{\omega,\mu,\sigma}) = \int \left|\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\boldsymbol{\theta}, f_{\omega,\mu,\sigma})\right|^2 \pi(\tau) d\tau$.
Find the parameters $(\widehat{\boldsymbol{\theta}}_n, \widehat{\omega}_n, \widehat{\mu}_n, \widehat{\sigma}_n)$ that minimize $\widehat{Q}_n^S$.

To illustrate the class of models considered and the usefulness of the mixture sieve for economic analysis, consider the first empirical application in section 3.6 where the growth rate of consumption $\Delta c_t = \log(C_t/C_{t-1})$ is assumed to follow the following process:

$$\Delta c_t = \mu_c + \rho_c \Delta c_{t-1} + \sigma_t e_{t,1}, \quad e_{t,1} \sim f \tag{3.3}$$

$$\sigma_t^2 = \mu_\sigma + \rho_\sigma \sigma_{t-1}^2 + \kappa_\sigma e_{t,2}, \quad e_{t,2} \sim \chi_1^2. \tag{3.4}$$

Compared to the general model (3.1)-(3.2), the Δc_t corresponds to the outcome y_t, the latent variable u_t is $(\sigma_t^2, e_{t,1})$ and the parameters are $\boldsymbol{\theta} = (\mu_c, \rho_c, \mu_\sigma, \rho_\sigma, \kappa_\sigma)$. This very simple model, with a flexible distribution f for the shocks $e_{t,1}$, can explain the low level of the risk-free rate with a simple power utility and recent monthly data. In comparison, the Long-Run Risks models relies on more complex dynamics and recursive utilities (Bansal & Yaron, 2004) and the Rare Disasters literature involves hard to quantify very large, low frequency shocks (Rietz, 1988; Barro, 2006b). Empirically, the Sieve-SMM estimates of distribution of f in the model (3.3)-(3.4) implies both a 25% larger higher welfare cost of business cycle fluctuations and an annualized risk-free rate that is up to 4 percentage

points lower than predicted by Gaussian shocks. Also, in this example the risk-free rate is tractable, up to a quadrature over σ_{t+1}, when using Gaussian mixtures:

$$r_t^{mixt} = -\log(\delta) + \gamma\mu_c + \gamma\rho_c\Delta c_t - \log\left(\sum_{j=1}^{k}\omega_j \mathbb{E}_t\left[e^{-\gamma\sigma_{t+1}\mu_j + \frac{\gamma^2}{2}\sigma_{t+1}^2[\sigma_j^2-1]}\right]\right).$$

In comparison, for a general distribution the risk-free rate depends on all moments but does not necessarily have closed form. The mixture thus combines flexible econometric estimation with convenient economic modelling.[1]

As in the usual sieve literature, this paper provides a consistency result and derives the rate of convergence of the structural and infinite dimensional parameters, as well as asymptotic normality results for finite dimensional functionals of these parameters. While the results apply to both static and dynamic models alike, two important differences arise in dynamic models compared to the existing literature on sieve estimation: proving uniform convergence of the objective function and controlling the dynamic accumulation of the nonparametric approximation bias.

The first challenge is to establish the rate of convergence of the objective function for dynamic models. To allow for the general dynamics (3.1)-(3.2) with latent variables, this paper adapts results from Andrews & Pollard (1994) and Ben Hariz (2005) to construct an inequality for uniformly bounded empirical processes which may be of independent interest. It allows the simulated data to be non-stationary when the initial (y_0, u_0) is not taken from the ergodic distribution. It requires a geometric ergodicity condition as in Duffie & Singleton (1993). The boundedness condition is satisfied by the CF and the CDF for instance. Also, the inequality implies a larger variance than typically found in the literature.[2]

The second challenge is that in the model (3.1)-(3.2) the nonparametric bias accumulates dynamically. At each time period the bias appears because draws are taken from a mixture approximation instead of the true f_0, this bias is also transmitted from one period to the next since (y_t^s, u_t^s) depends on (y_{t-1}^s, u_{t-1}^s). To ensure that this bias does not accumulate too much, a decay condition is imposed on the DGP. For the consumption process (3.3)-(3.4), this condition holds if both $|\rho_c|$ and $|\rho_\sigma|$ are strictly less than 1. The

[1]Gaussian mixtures are also convenient in more complicated settings where the model needs to be solved numerically. For instance, all the moments of a Gaussian mixture are tractable and quadrature is easy so that it can be applied to both the perturbation method and the projection method (see e.g. Judd, 1996, for a review of these methods) instead of the more commonly applied Gaussian distribution.

[2]See Chen (2007, 2011) for a summary of existing results with iid and dependent data.

resulting bias is generally larger than in static models and usual sieve estimation problems. Together, the increased variance and bias imply a slower rate of convergence for the Sieve-SMM estimates. Hence, in order to achieve the rate of convergence required for asymptotic normality, the Sieve-SMM requires additional smoothness of the true density f_0.

Monte-Carlo simulations illustrate the properties of the estimator and the effect of dynamics on the bias and the variance of the estimator. Two empirical applications highlight the importance of estimating the distribution of the shocks. The first is the example discussed above, and the second estimates a different stochastic volatility model on a long daily series of exchange rate data. The Sieve-SMM estimator suggests significant asymmetry and fat tails in the shocks, even after controlling for the time-varying volatility. As a result, commonly used parametric estimates for the persistence are significantly downward biased which has implications for forecasting; this effect is confirmed by the Monte-Carlo simulations.

Related Literature

The Sieve-SMM estimator presented in this paper combines two literatures: sieves and the Simulated Method of Moments (SMM). This section reviews the existing methods and results in each literature to introduce the new challenges arising from the combined Sieve-SMM setting.

A key aspect to simulation-based estimation is the choice of moments $\widehat{\psi}_n$. The Simulated Method of Moments (SMM) estimator of McFadden (1989) relies on unconditional moments, the Indirect Inference (IND) estimator of Gouriéroux et al. (1993) uses auxliary parameters from a simpler, tractable model and the Efficient Method of Moments (EMM) of Gallant & Tauchen (1996) uses the score of the auxiliary model. Simulation-based estimation has been applied to a wide array of economic settings: early empirical applications of these methods include the estimation of discrete choice models (Pakes, 1986; Rust, 1987), DSGE models (Smith, 1993) and models with occasionally binding constraints (Deaton & Laroque, 1992). More recent empirical applications include the estimation of earning dynamics (Altonji et al., 2013), of labor supply (Blundell et al., 2016) and the distribution of firm sizes (Gourio & Roys, 2014). Simulation-based estimation can also applied to models that are not fully specified as in Berry et al. (1995), these models are not considered in the Sieve-SMM estimation.

To achieve parametric efficiency a number of papers consider using nonparametric

moments but they assumed the distribution f is known.[3] To avoid dealing with the nonparametric rate of convergence of the moments Carrasco et al. (2007a) use the continuum of moments implied by the CF. This paper uses a similar approach in a semi-nonparametric setting. Bernton et al. (2017) use the Wasserstein, or Kantorovich distance, between the empirical and simulated distributions. This distance relies on unbounded moments and is thus excluded from the analysis in this paper.

General asymptotic results are given by Pakes & Pollard (1989) for SMM with iid data and Lee & Ingram (1991); Duffie & Singleton (1993) for time-series. Gouriéroux & Monfort (1996) provide an overview of existing results for a large number of simulation-based estimation methods.

While most of the literature discussed so far deals with fully parametric SMM models, there are a few papers concerned with sieve simulation-based estimation. Bierens & Song (2012) provide a consistency result for Sieve-SMM estimation of a static first-price auction model.[4] Newey (2001) uses a sieve simulated IV estimator for a measurement error model and proves consistency as both n and S go to infinity. These papers only consider specific static models and only provide limited asymptotic results. Furthermore, they consider sampling methods for the simulations that are very computationally costly (see section 3.2 for a discussion). Additionally, an incomplete working paper by Blasques (2011) uses the high-level conditions in Chen (2007) for a "Semi-NonParametric Indirect Inference" estimator. These conditions are very difficult to verify in practice and additional results are needed to handle the dynamics.[5]

An alternative to using sieves in SMM estimation involves using more general parametric families to model the first 3 or 4 moments flexibly. Ruge-Murcia (2012, 2017) considers the skew Normal and the Generalized Extreme Value distributions to model the first 3 moments of productivity and inflation shocks. Gospodinov & Ng (2015); Gospodinov et al. (2017) use the Generalized Lambda family to flexibly model the first 4 moments of the shocks in a non-invertible moving avergage and a measurement error

[3]See e.g. Gallant & Tauchen (1996); Fermanian & Salanié (2004); Kristensen & Shin (2012); Gach & Pötscher (2010); Nickl & Pötscher (2011).

[4]In order to do inference on f, they propose to invert a simulated version of Bierens (1990)'s ICM test statistic. A recent working paper by Bierens & Song (2017) introduces covariates in the same auction model and gives an asymptotic normality result for the coefficients $\widehat{\boldsymbol{\theta}}_n$ on the covariates.

[5]Also, to avoid using sieves and SMM in moment conditions models that are tractable up to a latent variable, Schennach (2014) proposes an Entropic Latent Variable Integration via Simulation (ELVIS) method to build estimating equations that only involve the observed variables. Dridi & Renault (2000) propose a Semi-Parametric Indirect Inference based on a partial encompassing principle.

model. However, in applications where the moments depend on the full distribution of the shocks, which is the case if the data y_t is non-separable in the shocks e_t, then the estimates $\widehat{\theta}_n$ will be sensitive to the choice of parametric family. Also, quantities of interest such as welfare estimates and asset prices that depend on the full distribution will also be sensitive to the choice of parametric family.

Another related literature is the sieve estimation of models defined by moment conditions. These models can be estimated using either Sieve-GMM, Sieve Empirical Likelihood or Sieve Minimum Distance (see Chen, 2007, for a review). Applications include nonparametric estimation of mean instrumental variables regressions[6], of quantile instrumental variables regressions,[7] and the semi-nonparametric estimation of asset pricing models,[8] for instance. Existing results cover the consistency and the rate of convergence of the estimator as well as asymptotic normality of functional of the parameters for both iid and dependent data. Recent general asymptotic results include Chen & Pouzo (2012, 2015) for iid data and Chen & Liao (2015) for dependent data.

In the empirical Sieve-GMM literature, an application closely related to the dynamics encountered in this paper appears in Chen et al. (2013). The authors show how to estimate an Euler equation with recursive preferences when the value function is approximated using sieves. Recursive preferences require a filtering step to recover the latent variable. This implies that the moments depend on the whole history of the data $(y_t, \ldots, y_1)$. However, general results based on coupling results (see e.g. Doukhan et al., 1995; Chen & Shen, 1998) do not apply to this class of moments. The authors use a Bootstrap for inference without formal asymptotic results.

Notation

The following notation and assumptions will be used throughout the paper: the parameter of interest is $\beta = (\theta, f) \in \Theta \times \mathcal{F} = \mathcal{B}$. The finite dimensional parameter space Θ is compact and the infinite dimensional set of densities $\mathcal{F}$ is possibly non-compact. The sets of mixtures satisfy $\mathcal{B}_k \subseteq \mathcal{B}_{k+1} \subseteq \mathcal{B}$, k is the data dependent dimension of the sieve set $\mathcal{B}_k$. The dimension k increases with the sample size: $k(n) \to \infty$ as $n \to \infty$. Using the notation of Chen (2007), $\Pi_{k(n)} f$ is the mixture approximation of the density f. The vector

[6] See e.g. Hall & Horowitz (2005); Carrasco et al. (2007b); Blundell et al. (2007); Darolles et al. (2011); Horowitz (2011).

[7] See e.g. Chernozhukov & Hansen (2005); Chernozhukov et al. (2007); Horowitz & Lee (2007).

[8] See e.g. Hansen & Richard (1987); Chen & Ludvigson (2009); Chen et al. (2013); Christensen (2017).

of shocks e has dimension $d_e \geq 1$ and density f. The total variation distance between two densities is $\|f_1 - f_2\|_{TV} = 1/2 \int |f_1(e) - f_2(e)|de$ and the supremum (or sup) norm is $\|f_1 - f_2\|_\infty = \sup_{e \in \mathbb{R}^{d_e}} |f_1(e) - f_2(e)|$. For simplification, the following convention will be used $\|\beta_1 - \beta_2\|_{TV} = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \|f_1 - f_2\|_{TV}$ and $\|\beta_1 - \beta_2\|_\infty = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \|f_1 - f_2\|_\infty$, where $\|\boldsymbol{\theta}\|$ and $\|e\|$ correspond the Euclidian norm of $\boldsymbol{\theta}$ and e respectively. $\|\beta_1\|_m$ is a norm on the mixture components: $\beta_1\|_m = \|\boldsymbol{\theta}\| + \|(\omega, \mu, \sigma)\|$ where $\|\cdot\|$ is the Euclidian norm and (ω, μ, σ) are the mixture parameters. For a functional ϕ, its pathwise, or Gâteau, derivative at β_1 in the direction β_2 is $\frac{d\phi(\beta_1)}{d\beta}[\beta_2] = \frac{d\phi(\beta_1 + \varepsilon\beta_2)}{d\varepsilon}\Big|_{\varepsilon=0}$, it will be assumed to be continuous in β_1 and linear in β_2. For two sequences a_n and b_n, the relation $a_n \asymp b_n$ implies that there exists $0 < c_1 \leq c_2 < \infty$ such that $c_1 a_n \leq b_n \leq c_2 a_n$ for all $n \geq 1$.

Structure of the Paper

The paper is organized as follows: Section 3.2 introduces the Sieve-SMM estimator, explains how to implement it in practice and provides important properties of the mixture sieve. Section 3.3 gives the main asymptotic results: under regularity conditions, the estimator is consistent. Its rate of convergence is derived, and under further conditions, finite dimensional functionals of the estimates are asymptotically normal. Section 3.4 provides two extensions, one to include auxiliary variables in the CF and another to allow for dynamic panels with small T. Section 3.5 provides Monte-Carlo simulations to illustrate the theoretical results. Section 3.6 gives empirical examples for the estimator. Section 3.7 concludes. Appendix 3.7 gives some information about the CF and details on how to compute the estimator in practice. Appendix 3.7 provides the proofs to the main results. Appendix 3.7 provides results for more general moment functions and sieve bases and Appendix 3.7 which provides the proofs for these results.

3.2 The Sieve-SMM Estimator

This section introduces the notation used in the remainder of the paper. It describes the class of DGPs considered in the paper and describes the DGP of the leading example in more details. It discusses the choice of mixture sieve, moments and objective function as well as some important properties of the mixture sieve. The running example used throughout the analysis is based on the empirical applications of section 3.6.

Example 1 (Stochastic Volatility Models)**.** *In both empirical applications, y_t follows an AR(1) process with log-normal stochastic volatility*

$$y_t = \mu_y + \rho_y y_{t-1} + \sigma_t e_{t,1}.$$

The first empirical application estimates a linear volatility process:

$$\sigma_t^2 = \mu_\sigma + \rho_\sigma \sigma_{t-1}^2 + \kappa_\sigma e_{t,2}$$

where $e_{t,2} \sim \chi_1^2$. The second empirical application estimates a log-normal stochastic volatility process:

$$\log(\sigma_t) = \mu_\sigma + \rho_\sigma \log(\sigma_{t-1}) + \kappa_\sigma e_{t,2}.$$

where $e_{t,2} \overset{iid}{\sim} \mathcal{N}(0,1)$. In both applications $e_{t,1} \overset{iid}{\sim} f$ with the restrictions $\mathbb{E}(e_{t,1}) = 0$ and $\mathbb{E}(e_{t,1}^2) = 1$. The first application approximates f with a mixture of Gaussian distributions, the second adds two tail components to model potential fat tails.

Stochastic volatility (SV) models in Example 1 are intractable because of the latent volatility. With log-normal volatility, the model becomes tractable after taking the transformation $\log([y_t - \mu_y - \rho_y y_{t-1}]^2)$ (see e.g. Kim et al., 1998) and the problem can be cast as a deconvolution problem (Comte, 2004). However, the transformation removes all the information about asymmetries in f, which turn out to empirically significant (see section 3.6). In the parametric case, alternatives to using the transformation involve Bayesian simulation-based estimators such as the Particle Filter and Gibbs sampling or EMM for frequentist estimation.

Sieve Basis - Gaussian and Tails Mixture

The following definition introduces the Gaussian and tails mixture sieve that will be used in the paper. It combines a simple Gaussian mixture with two tails densities which model asymmetric fat tails parametrically. Drawing from this mixture is computationally simple: draw uniforms and gaussian random variables, switch between the Gaussians and the tails depending on the uniform and the mixture weights ω. The tail draws are a simple function of uniform random variables.

Definition 1 (Gaussian and Tails Mixture)**.** *A random variable e follows a k component Gaussian and Tails mixture if its density has the form:*

$$f_{\omega,\mu,\sigma}(e) = \sum_{j=1}^{k} \frac{\omega_j}{\sigma_j}\phi(\frac{e-\mu_j}{\sigma_j}) + \frac{\omega_{k+1}}{\sigma_{k+1}}\mathbb{1}_{e\leq\mu_{k+1}} f_L\left(\frac{e-\mu_{k+1}}{\sigma_{k+1}}\right) + \frac{\omega_{k+2}}{\sigma_{k+2}}\mathbb{1}_{e\geq\mu_{k+2}} f_R\left(\frac{e-\mu_{k+2}}{\sigma_{k+2}}\right)$$

where ϕ is the standard Gaussian density and its left and right tail components are

$$f_L(e,\xi_L) = (2+\xi_L)\frac{|e|^{1+\xi_L}}{[1+|e|^{2+\xi_L}]^2} \quad \text{for } e \leq 0, \quad f_R(e,\xi_R) = (2+\xi_R)\frac{e^{1+\xi_R}}{[1+e^{2+\xi_R}]^2} \quad \text{for } e \geq 0$$

with $f_L(e,\xi_L) = 0$ for $e \geq 0$ and $f_R(e,\xi_L) = 0$ for $e \leq 0$. To simulate from the Gaussian and tails mixture, draw $Z_1,\ldots,Z_k \overset{iid}{\sim} \mathcal{N}(0,1)$, $\nu,\nu_L,\nu_R \overset{iid}{\sim} \mathcal{U}_{[0,1]}$ and compute $Z_{k+1} = -\left(\frac{1}{\nu_L}-1\right)^{\frac{1}{2+\xi_L}}$ and $Z_{k+2} = \left(\frac{1}{\nu_R}-1\right)^{\frac{1}{2+\xi_R}}$. Then, for $\omega_0 = 0$:

$$e = \sum_{j=1}^{k+2} \mathbb{1}_{\nu \in [\sum_{l=0}^{j-1}\omega_l, \sum_{l=0}^{j}\omega_l]} \left(\mu_j + \sigma_j Z_j\right)$$

follows the Gaussian and tails mixture $f_{\omega,\mu,\sigma}$.

For application where fat tails are deemed unlikely, as in the first empirical application, the weights $\omega_{k+1},\omega_{k+2}$ can be set to zero to use a Gaussian mixture. If $\frac{\omega_{k+1}}{\sigma_{k+1}} \neq 0$ and $\frac{\omega_{k+2}}{\sigma_{k+2}} \neq 0$ then the left and right tails satisfy:

$$f_L(e) \overset{e\to-\infty}{\sim} |e|^{-3-\xi_L}, \quad f_R(e) \overset{e\to+\infty}{\sim} e^{-3-\xi_R}.$$

If $\xi_L,\xi_R \geq 0$ then draws from the tail components have finite expectation, they also have finite variance if $\xi_L,\xi_R \geq 1$. More generally, for the j-th moment to be finite, $j \geq 1$, $\xi_L,\xi_R \geq j$ is necessary. Gallant & Nychka (1987) also add a parametric component to model fat tails by using a mixture of a Hermite polynomial with a Student density. However, neither the Hermite polynomial nor the Student t-distribution have closed-form quantiles, which is not practical for simulation. Here, the densities f_L, f_R are constructed to be easy to simulated from.

The indicator function $\mathbb{1}_{\nu_l^s \in [\sum_{l=0}^{j-1}\omega_l, \sum_{l=0}^{j}\omega_l]}$ introduces discontinuities in the parameter ω. Standard derivative-free optimization routines such as the Nelder-Mead algorithm (Nelder & Mead, 1965) as implemented in the NLopt library of Johnson (2014) can handle this estimation problem as illustrated in section 3.5.[9]

In the finite mixture literature, mixture components are known to be difficult to identify because of possible label switching and the likelihood is globally unbounded.[10] Using the characteristic function rather than the likelihood resolves the unbounded likelihood

[9]The NLopt library is available for C++, Fortran, Julia, Matlab, Python and R among others.

[10]See e.g. McLachlan & Peel (2000) for a review of estimation, identification and applications of finite mixtures. See also Chen et al. (2014b) for some recent results.

problem as discussed in Yu (1998). More importantly, the object of interest in this paper is the mixture density $f_{\omega,\mu,\sigma}$ itself rather than the mixture components. As a result, permutations of the mixture components are not a concern, since they do not affect the resulting mixture density $f_{\omega,\mu,\sigma}$.

Moments - Empirical Characteristic Function and Objective Function

As in the parametric case, the moments need to be informative enough to identify the parameters. In Sieve-SMM estimation, the parameter $\beta = (\boldsymbol{\theta}, f)$ is infinite dimensional so that no finite dimensional vector of moments could possibly identify β. As a result, this paper relies on moment functions which are themselves infinite dimensional.

The leading choice of moment function in this paper is the empirical characteristic function for the joint vector of lagged observations $(\mathbf{y}_t, \mathbf{x}_t) = (y_t, \ldots, y_{t-L}, x_t, \ldots, x_{t-L})$:

$$\widehat{\boldsymbol{\psi}}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} e^{i\tau'(\mathbf{y}_t,\mathbf{x}_t)}, \ \forall \tau \in \mathbb{R}^{d_\tau}$$

where i is the imaginary number such that $i^2 = -1$.[11] The CF is one-to-one with the joint distribution of $(\mathbf{y}_t, \mathbf{x}_t)$, so that the model is identified by $\widehat{\boldsymbol{\psi}}_n(\cdot)$ if and only if the distribution of $(\mathbf{y}_t, \mathbf{x}_t)$ identifies the true β_0. Using lagged variables allows to identify the dynamics in the data. Knight & Yu (2002) show how the characteristic function can identify parametric dynamic models. Some useful properties of the CF are given in Appendix 3.7.

Besides the CF, another choice of bounded moment function is the CDF. While the CF is a smooth transformation of the data, the empirical CDF has discontinuities at each point of support of the data $(\mathbf{y}_t, \mathbf{x}_t)$ which could make numerical optimization more challenging. Also, the CF around $\tau = 0$ summarizes the information about the tails of the distribution (see Ushakov, 1999, page 30). This information is thus easier to extract from the CF than the CDF. The main results of this paper can be extended to any bounded moment function satisfying a Lipschitz condition.[12]

Since the moments are infinite dimensional, this paper adopts the objective function of Carrasco & Florens (2000); Carrasco et al. (2007a) to handle the continuum of moment

[11] The moments can also be expressed in terms of sines and cosines since $e^{i\tau'(\mathbf{y}_t,\mathbf{x}_t)} = \cos(\tau'(\mathbf{y}_t, \mathbf{x}_t)) + i\sin(\tau'(\mathbf{y}_t, \mathbf{x}_t))$.

[12] Appendix 3.7 allows for more general non-Lipschitz moment functions and other sieve bases. However, the conditions required for these results are more difficult to check.

conditions:[13]

$$\widehat{Q}_n^S(\beta) = \int \left| \widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \beta) \right|^2 \pi(\tau) d\tau. \tag{3.5}$$

The objective function is a weighted average of the square norm between the empirical $\widehat{\psi}_n$ and the simulated $\widehat{\psi}_n^S$ moment functions. As discussed in Carrasco & Florens (2000) and Carrasco et al. (2007a), using the continuum of moments avoids the problem of constructing an increasing vector of moments. The weighting density π is chosen to be the multivariate normal density for the main results. Other choices for π are possible as long as it has full support and is such that $\int \sqrt{\pi(\tau)} d\tau < \infty$. As an example, the exponential distribution satisfies these two conditions, while the Cauchy distribution does not satisfy the second. In practice, choosing π to be the Gaussian density with same mean and variance as $(\mathbf{y_t}, \mathbf{x}_t)$ gave satisfying results in sections 3.5 and 3.6.[14] In the appendix, the results allow for a bounded linear operator B which plays the role of the weight matrix W in SMM and GMM as in Carrasco & Florens (2000). Carrasco & Florens (2000); Carrasco et al. (2007a) provide theoretical results for choosing and approximating the optimal operator B in the parametric setting. Similar work is left to future research in this semi-nonparametric setting.

Given the sieve basis, the moments and the objective function, the estimator $\widehat{\beta}_n = (\widehat{\boldsymbol{\theta}}_n, \widehat{f}_n)$ is defined as an approximate minimizer of $\widehat{Q}_n^S$:

$$\widehat{Q}_n^S(\widehat{\beta}_n) \leq \text{diag}_{\beta \in \mathcal{B}_{k(n)}} \widehat{Q}_n^S(\beta) + O_p(\eta_n) \tag{3.6}$$

where $\eta_n \geq 0$ and $\eta_n = o(1)$ corresponds to numerical optimization and integration errors. Indeed, since the integral in (3.5) needs to be evaluated numerically, some form of numerical integration is required. Quadrature and sparse quadrature were found to give satisfying results when $\dim(\tau)$ is not too large (less than 4). For larger dimensions, quasi-Monte-Carlo integration using either the Halton or Sobol sequence gave satisfying results.[15] All Monte-Carlo simulations and empirical results in this paper are based on quasi-Monte-Carlo integration. Additional details on the computation of the objective function are given in Appendix 3.7.

[13] Carrasco & Florens (2000) provide a general theory for GMM estimation with a continuum of moment conditions. They show how to efficiently weight the continuum of moments and propose a Tikhonov (ridge) regularization approach to invert the singular variance-covariance operator. Earlier results, without optimal weighting, include Koul (1986) for minimum distance estimation with a continuum of moments.

[14] Monte-Carlo experiments not reported in this paper showed similar results when using the exponential density for π instead of the Gaussian density.

[15] See e.g. Heiss & Winschel (2008); Holtz (2011) for an introduction to sparse quadrature in economics and finance, and Owen (2003) for quasi-Monte-Carlo sampling.

Approximation and L^2-Smoothness Properties of the Mixture Sieve

This subsection provides more details on the approximation and L^p-smoothness properties of the mixture sieve. It also provides the necessary restrictions on the true density f_0 to be estimated. Gaussian mixtures can approximate any smooth univariate density but the rate of this approximation depends on both the smoothness and the tails of the density (see e.g. Kruijer et al., 2010). The tail densities parametrically model asymmetric fat tails in the density. This is useful in the second empirical example since a thin tail assumption may not hold for exchange rate data. The following lemma extends the approximation results of Kruijer et al. (2010) to a multivariate density with independent components and potentially fat tails.

Lemma 2 (Approximation Properties of the Gaussian and Tails Mixture)**.** *Suppose that the shocks $e = (e_{t,1}, \ldots, e_{t,d_e})$ are independent with density $f = f_1 \times \cdots \times f_{d_e}$. Suppose that each marginal f_j can be decomposed into a smooth density $f_{j,S}$ and the two tails f_L, f_R of Definition 1:*

$$f_j = (1 - \omega_{j,1} - \omega_{j,2}) f_{j,S} + \omega_{j,1} f_L + \omega_{j,2} f_R.$$

Let each $f_{j,S}$ satisfy the assumptions of Kruijer et al. (2010):

i. Smoothness: $f_{j,S}$ is r-times continuously differentiable with bounded r-th derivative.

ii. Tails: $f_{j,S}$ has exponential tails, i.e. there exists $\bar{e}, M_f, a, b > 0$ such that:

$$f_{j,S}(e) \leq M_f e^{-a|e|^b}, \forall |e| \geq \bar{e}.$$

iii. Monotonicity in the Tails: $f_{j,S}$ is strictly positive and there exists $\underline{e} < \bar{e}$ such that $f_{j,S}$ is weakly decreasing on $(-\infty, \underline{e}]$ and weakly increasing on $[\bar{e}, \infty)$.

and $\|f_j\|_\infty \leq \overline{f}$ for all j. Then there exists a Gaussian and tails mixture $\Pi_k f = \Pi_k f_1 \times \cdots \times \Pi_k f_{d_e}$ satisfying the restrictions of Kruijer et al. (2010):

iv. Bandwidth: $\sigma_j \geq \underline{\sigma}_k = O(\frac{\log[k]^{2/b}}{k})$.

v. Location Parameter Bounds: $\mu_j \in [-\overline{\mu}_k, \overline{\mu}_k]$ with $\overline{\mu}_k = O\left(\log[k]^{1/b}\right)$

such that as $k \to \infty$:

$$\|f - \Pi_k f\|_{\mathcal{F}} = O\left(\frac{\log[k]^{2r/b}}{k^r}\right)$$

where $\|\cdot\|_{\mathcal{F}} = \|\cdot\|_{TV}$ or $\|\cdot\|_\infty$.

The space of true densities satisfying the assumptions will be denoted as $\mathcal{F}$ and $\mathcal{F}_k$ is the corresponding space of Gaussian and tails mixtures $\Pi_k f$.

Note that additional restrictions on f may be required for identification, such as mean zero, unit variance or symmetry. The assumption that the shocks are independent is not too strong for structural models where this, or a parametric factor structure is typically assumed. Note that under this assumption, there is no curse of dimensionality because the components f_j can be approximated separately. Also, the restriction $\|f_j\|_\infty \leq \overline{f}$ is only required for the approximation in supremum norm $\|\cdot\|_\infty$.

An important difficulty which arises in simulating from a nonparametric density is that draws are a very nonlinear transformation of the nonparametric density f. As a result, standard regularity conditions such as Hölder and L^p-smoothness are difficult to verify and may only hold under restrictive conditions. The following discusses these regularity conditions for the methods used in the previous literature and provides a L^p-smoothness result the mixture sieve (Lemma 3 below).

Bierens & Song (2012) use Inversion Sampling: they compute the CDF F_k from the nonparametric density and draw $F_k^{-1}(\nu_t^s), \nu_t^s \overset{iid}{\sim} \mathcal{U}_{[0,1]}$. Computing the CDF and its inverse to simulate is very computationally demanding. Also, while the CDF is linear in the density, its inverse is a highly non-linear transformation of the density. Hence, Hölder and L^p-smoothness results for the draws are much more challenging to prove without further restrictions.

Newey (2001) uses Importance Sampling for which Hölder conditions are easily verified but requires $S \to \infty$ for consistency alone. Furthermore, the choice of importance distribution is very important for the finite sample properties (the effective sample size) of the simulated moments. In practice, the importance distribution should give sufficient weight to regions for which the nonparametric density has more weight. Since the nonparametric density is unknown ex-ante, this is hard to achieve in practice.

Gallant & Tauchen (1993) use Accept/Reject (outside of an estimation setting): however, it is not practical for simulation-based estimation. Indeed, the required number of draws to generate an accepted draw depends on both the instrumental density and the target density $f_{\omega,\mu,\sigma}$. The latter varies with the parameters during the optimization. This also makes the L^p-smoothnes properties challenging to establish. In comparison, the following lemma shows that the required L^2-smoothness condition is satisfied by draws from a mixture sieve.

Lemma 3 (L^2-Smoothness of Simulated Mixture Sieves)**.** *Suppose that*

$$e_t^s = \sum_{j=1}^{k(n)} \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1}\omega_l, \sum_{l=0}^{j}\omega_l]} \left(\mu_j + \sigma_j Z_{t,j}^s\right), \quad \widetilde{e}_t^s = \sum_{j=1}^{k(n)} \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1}\widetilde{\omega}_l, \sum_{l=0}^{j}\widetilde{\omega}_l]} \left(\widetilde{\mu}_j + \widetilde{\sigma}_j Z_{t,j}^s\right)$$

with $|\mu_j|$ *and* $|\widetilde{\mu}_j| \leq \overline{\mu}_{k(n)}$, $|\sigma_j|$ *and* $|\widetilde{\sigma}_j| \leq \overline{\sigma}$. *If* $\mathbb{E}(|Z_{t,j}^s|^2) \leq C_Z^2 < \infty$ *then there exists a finite constant* C *which only depends on* C_Z *such that:*

$$\left[\mathbb{E}\left(sup_{\|f_{\omega,\mu,\sigma} - f_{\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma}}\|_m \leq \delta} \left|e_t^s - \widetilde{e}_t^s\right|^2\right)\right]^{1/2} \leq C\left(1 + \overline{\mu}_{k(n)} + \overline{\sigma} + k(n)\right)\delta^{1/2}.$$

Lemma 3 is key in proving the L^2-smoothness conditions of the moments $\widehat{\psi}_n^s$ required to establish the convergence rate of the objective function and stochastic equicontinuity results. Here, the L^p-smoothness constant depends on both the bound $\overline{\mu}_{k(n)}$ and the number of mixture components $k(n)$.[16] Kruijer et al. (2010) showed that both the total variation and supremum norms are bounded above by the pseudo-norm $\|\cdot\|_m$ on the mixture parameters (ω, μ, σ) up to a factor which depends on the bandwidth $\underline{\sigma}_{k(n)}$. As a result, the pseudo-norm $\|\cdot\|_m$ controls the distance between densities and the simulated draws as well. Furthermore, draws from the tail components are shown in the appendix to be L^2-smooth in their tail parameters ξ_L, ξ_R. Hence, draws from the Gaussian and tails mixture are L^2-smooth in both (ω, μ, σ) and ξ.

3.3 Asymptotic Properties of the Estimator

This section provides conditions under which the Sieve-SMM estimator in (3.6) is consistent. Its rate of convergence is derived and an asymptotic normality result for functionals of $\widehat{\beta}_n$ is given.

Consistency

Consistency results are given under low-level conditions on the DGP using the Gaussian and tails mixture sieve with the CF.[17] First, the population objective Q_n is:

$$Q_n(\beta) = \int \left|\mathbb{E}\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \beta)\right)\right|^2 \pi(\tau) d\tau. \tag{3.7}$$

[16] See e.g. Andrews (1994); Chen et al. (2003) for examples of L^p-smooth functions.

[17] Consistency results allowing for non-mixture sieves and other moments are given in Appendix 3.7.

The objective depends on n because (y_t^s, x_t) are not covariance stationary: the moments can depend on t. Under geometric ergodicity, it has a well-defined limit:[18]

$$Q_n(\beta) \overset{n\to\infty}{\to} Q(\beta) = \int \left| \lim_{n\to\infty} \mathbb{E}\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau,\beta)\right) \right|^2 \pi(\tau) d\tau.$$

In the definition of the objective Q_n and its limit Q, the expectation is taken over both the data $(\mathbf{y}_t, \mathbf{x}_t)$ and the simulated samples $(\mathbf{y}_t^s, \mathbf{x}_t)$. The following assumption, provide a set of sufficient conditions on the true density f_0, the sieve space and a first set of conditions on the model (identification and time-series properties) to prove consistency.

Assumption 1 (Sieve, Identification, Dependence)**.** *Suppose the following conditions hold:*

i. (Sieve Space) the true density f_0 and the mixture sieve space $\mathcal{F}_{k(n)}$ satisfy the assumptions of Lemma 2 with $k(n)^4 \log[k(n)]^4/n \to 0$ as $k(n)$ and $n \to \infty$. Θ is compact and $1 \leq \xi_L, \xi_R \leq \overline{\xi} < \infty$.

ii. (Identification) $\lim_{n\to\infty} \mathbb{E}\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^s(\tau,\beta)\right) = 0, \pi$ a.s. $\Leftrightarrow \|\beta - \beta_0\|_{\mathcal{B}} = 0$ where π is the Gaussian density. For any $n, k \geq 1$ and for all $\varepsilon > 0$, $diag_{\beta\in\mathcal{B}_k, \|\beta-\beta_0\|_{\mathcal{B}}\geq\varepsilon} Q_n(\beta)$ is strictly positive and weakly decreasing in both n and k.

iii. (Dependence) (y_t, x_t) is strictly stationary and α-mixing with exponential decay, the simulated $(y_t^s(\beta), x_t)$ are geometrically ergodic, uniformly in $\beta \in \mathcal{B}$.

Condition *i.* is stronger than the usual condition $k(n)/n \to 0$ in the sieve literature (see e.g Chen, 2007). The additional $\log[k(n)]$ term is due to the mixture being a non-linear sieve basis and the fourth power is due to the dependence. Indeed, the inequality in Lemma .0.4 implies that the variance is of order $k(n)^2 \log[k(n)]^2/\sqrt{n}$ instead of $\sqrt{k(n)\log[k(n)]/n}$ for iid data.

Condition *ii.* is the usual identification condition. It is assumed that the information from the joint distribution of $(\mathbf{y}_t, \mathbf{x}_t) = (y_t, \ldots, y_{t-L}, x_t, \ldots, x_{t-L})$ uniquely identifies $\beta = (\boldsymbol{\theta}, f)$. Proving general global identification results is quite challenging in this setting and is left to future research. Local identification in the sense of Chen et al. (2014a) is also challenging to prove here because the dynamics imply that the distribution of (y_t^s, x_t, u_t^s) is a convolution of f with the distribution of $(y_{t-1}^s, x_t, u_{t-1}^s)$. Since the stationary distributions of (y_t^s, x_t, u_t^s) and $(y_{t-1}^s, x_t, u_{t-1}^s)$ are the same, the resulting distribution is the fixed point of its convolution with f. This makes derivatives with respect to f difficult to compute in many dynamic models. Note that the identification assumption does not exclude

[18]Since the CF is bounded, the dominated convergence theorem can be used to prove the existence of the limit.

ill-posedness.[19] The space $\mathcal{F}$ is assumed to include the necessary restrictions (if any) for identification such as mean zero and unit variance. Global identification results for the stochastic volatility model in Example 1 are given in Appendix 3.7.

Condition *iii.* is common in SMM estimation with dependent data (see e.g. Duffie & Singleton, 1993). In this setting, it implies two important features: the simulated (y_t^s, x_t) are α-mixing (Liebscher, 2005), and the initial condition bias is negligible: $Q_n(\beta_0) = O(1/n^2)$.[20]

Assumption 2 (Data Generating Process)**.** *y_t^s is simulated according to the dynamic model (3.1)-(3.2) where g_{obs} and g_{latent} satisfy the following Hölder conditions for some $\gamma \in (0, 1]$:*

y(i). $\|g_{obs}(y_1, x, \beta, u) - g_{obs}(y_2, x, \beta, u)\| \leq C_1(x, u)\|y_1 - y_2\|$ *with* $\mathbb{E}\left(C_1(x_t, u_t^s)^2 | y_{t-1}^s\right) \leq \overline{C}_1 < 1$.

y(ii). $\|g_{obs}(y, x, \beta_1, u) - g_{obs}(y, x, \beta_2, u)\| \leq C_2(y, x, u)\|\beta_1 - \beta_2\|_{\mathcal{B}}^{\gamma}$ *with* $\mathbb{E}\left(C(y_t^s, x_t, u_t^s)^2\right) \leq \overline{C}_2 < \infty$.

y(iii). $\|g_{obs}(y, x, \beta, u_1) - g_{obs}(y, x, \beta, u_2)\| \leq C_3(y, x)\|u_1 - u_2\|^{\gamma}$ *with* $\mathbb{E}\left(C_3(y_t^s, x_t)^2 | u_t^s\right) \leq \overline{C}_3 < \infty$.

u(i). $\|g_{latent}(u_1, \beta, e) - g_{latent}(u_2, \beta, e)\| \leq C_4(e)\|u_1 - u_2\|$ *with* $\mathbb{E}\left(C_4(e_t^s)^2\right) \leq \overline{C}_4 < 1$.

u(ii). $\|g_{latent}(u, \beta_1, e) - g_{latent}(u, \beta_2, e)\| \leq C_5(u, e)\|\beta_1 - \beta_2\|_{\mathcal{B}}^{\gamma}$ *with* $\mathbb{E}\left(C_5(u_{t-1}^s, e_t^s)^2\right) \leq \overline{C}_5 < \infty$.

u(iii). $\|g_{latent}(u, \beta, e_1) - g_{latent}(u, \beta, e_2)\| \leq C_6(u)\|e_1 - e_2\|$ *with* $\mathbb{E}\left(C_6(u_{t-1}^s)^2\right) \leq \overline{C}_6 < \infty$.

for any $(\beta_1, \beta_2) \in \mathcal{B}$, $(y_1, y_2) \in \mathbb{R}^{dim(y)}$, $(u_1, u_2) \in \mathbb{R}^{dim(u)}$ and $(e_1, e_2) \in \mathbb{R}^{dim(e)}$. The norm $\|\cdot\|_{\mathcal{B}}$ is either the total variation or supremum norm.

Conditions *y(ii), u(ii)* correspond to the usual Hölder conditions in GMM and M-estimation but placed on the DGP itself rather than the moments. Since the cosine and sine functions are Lipschitz, it implies that the moments are Hölder continuous as well.[21]

The decay conditions *y(i), u(i)* together with condition *y(iii)* ensure that the differences due to $\|\beta_1 - \beta_2\|_{\mathcal{B}}$ do not accumulate too much with the dynamics. As a result, keeping the shocks fixed, the Hölder condition applies to (y_t^s, u_t^s) as a whole. It also implies that

[19] See e.g. Carrasco et al. (2007b) and Horowitz (2014) for a review of ill-posedness in economics.

[20] See Proposition .0.4 in the supplemental material for the second result.

[21] For any choice of moments that preserve identification and are Lipschitz, the main results will hold assuming $\|\tau\|_{\infty}\sqrt{\pi(\tau)}$ and $\int \sqrt{\pi(\tau)}d\tau$ are bounded. For both the Gaussian and the exponential density, these quantities turn out to be bounded. In general Lispchitz transformations preserve L^p-smoothness properties (see e.g. Andrews, 1994; van der Vaart & Wellner, 1996), here additional conditions on π are required to handle the continuum of moments with unbounded support.

the nonparametric approximation bias $\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}$ does not accumulate too much. These conditions are similar to the L^2-Unit Circle condition which Duffie & Singleton (1993) suggest as an stronger alternative to geometric ergodicity in a uniform LLN and a CLT. The decay conditions play a more important role here since they are needed to control the nonparametric bias of the estimator.

Condition *u(iii)* ensures that the DGP preserves the L^2-smoothness properties derived for mixture draws in Lemma 3. This condition does not appear in the usual sieve literature which does not simulate from a nonparametric density. In the SMM literature, a Lipschitz or Hölder condition is usually given on the moments directly. Note that a condition analogous to *u(iii)* would also be required for parametric SMM estimation of a parametric distribution.

Assumption 2 does not impose that the DGP be smooth. This allows for kinks in g_{obs} or g_{latent} as in the sample selection model or the models of Deaton (1991) and Deaton & Laroque (1992). Assumption 2′ in Appendix 3.7 extends Assumption 2 to allow for possible discontinuities in g_{obs}, g_{latent}. The following shows how to verify the conditions of Assumption 2 in Example 1 with χ^2_1 volatility shocks.[22]

Example 1 (Continued) (Stochastic Volatility)**.** *If $|\rho_y| < 1$ then assumption y(i) is satisfied. Also:*

$$|\mu_{y,1} + \rho_{y,1}y_{t-1} - \mu_{y,2} - \rho_{y,2}y_{t-1}| \leq (|\mu_{y,1} - \mu_{y,2}| + |\rho_{y,1} - \rho_{y,2}|)(1 + |y_{t-1}|)$$

and thus condition y(ii) is satisfied assuming $\mathbb{E}(y^2_{t-1})$ is bounded. Since f has mean zero and unit variance, $\mathbb{E}(y^2_{t-1})$ is bounded if $|\mu_\sigma| \leq \overline{\mu}_\sigma < \infty$, $|\rho_\sigma| \leq \overline{\rho}_\sigma < 1$ and $\kappa_\sigma \leq \overline{\kappa}_\sigma < \infty$ for some $\overline{\mu}_\sigma, \overline{\rho}_\sigma, \overline{\kappa}_\sigma$. For condition y(iii), take $u_t = (\sigma^2_t, e_{t,1})$ and $\widetilde{u}_t = (\widetilde{\sigma}^2_t, \widetilde{e}_{t,1})$:

$$|\sigma_t e_{t,1} - \widetilde{\sigma}_t e_{t,1}| \leq |e_{t,1}|\sqrt{|\sigma^2_t - \widetilde{\sigma}^2_t|}, \quad |\sigma_t e_{t,1} - \sigma_t \widetilde{e}_{t,1}| \leq \sigma_t |e_{t,1} - \widetilde{e}_{t,1}|.$$

The first inequality is due to the Hölder continuity of the square-root function.[23] *σ_t and $\widetilde{e}_{t,1}$ are independent, $\mathbb{E}(\sigma^2_t)$ is bounded above under the previous parameter bounds and $\mathbb{E}(e^2_{t,1}) = 1$ and so condition y(iii) holds term by term. If the volatility σ^2_t is bounded below by a strictly positive constant for all paramater values then the Hölder continuity y(iii) can be strengthened to a Lipschitz continuity result. Given that σ^2_t follows an AR(1) process, assumptions u(i), u(ii) and u(iii) are satisfied.*

[22]Some additional examples are given in Appendix 3.7. They are not tied to the use of mixtures, and as a result, impose stronger restrictions on the density f such as bounded support.

[23]For any two $x, y \geq 0$, $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x^2 - y^2|}$.

The Hölder coefficient in conditions *y(ii)*, *y(iii)* and *u(ii)* is assumed to be the same to simplify notation. If they were denoted γ_1, γ_2 and γ_3, in order of appearance, then the rate of convergence would depend on $\min(\gamma_1, \gamma_2\gamma_3)$ instead of γ^2. This could lead to sharper rates of convergence in section 3.3 and weaker condition for the stochastic equicontinuity result in section 3.3. As shown above, in Example 1 the Hölder coefficients are $\gamma_1 = \gamma_3 = 1$, $\gamma_2 = 1/2$ when σ_t does not have a strictly positive lower bound.

Lemma 4 (Assumption 2/2′ implies L^2-Smoothness of the Moments)**.** *Under either Assumption 2 or 2′, if the assumptions of Lemma 3 hold and π is the Gaussian density, then there exists $\overline{C} > 0$ such that for all $\delta > 0$, uniformly in $t \geq 1$, $(\beta_1, \beta_2) \in \mathcal{B}_{k(n)}$ and $\tau \in \mathbb{R}^{d_\tau}$:*

$$\mathbb{E}\left(sup_{\|\beta_1-\beta_2\|_m \leq \delta}\left|e^{i\tau'(\mathbf{y}_t^s(\beta_1),\mathbf{x}_t)} - e^{i\tau'(\mathbf{y}_t^s(\beta_2),\mathbf{x}_t)}\right|^2\sqrt{\pi(\tau)}\right) \leq \overline{C}\max\left(\frac{\delta^{\gamma^2}}{\underline{\sigma}_{k(n)}^{2\gamma^2}}, [k(n) + \overline{\mu}_{k(n)} + \overline{\sigma}]^\gamma \delta^{\gamma^2/2}\right)$$

where $\|\beta\|_m = \|\boldsymbol{\theta}\| + \|(\omega, \mu, \sigma)\|$ is the pseudo-norm on $\boldsymbol{\theta}$ and the mixture parameters (ω, μ, σ) from Lemma 3. Also, since π is the Gaussian density the integral $\int \sqrt{\pi(\tau)}d\tau$ is finite.

Lemma 4 gives the first implication of Assumption 2. It shows that the moments $\widehat{\psi}_t^s$ are L^2-smooth, uniformly in $t \geq 1$. The L^2-smoothness factor depends on the bounds of the sieve components. In the SMM and sieve literatures, the L^p-smoothness constant depends on neither k nor n by assumption. Here, drawing from the mixture distribution implies that the constant will increase with the sample size n. The rate at which it increases is implied by the assumptions of Lemma 2.[24] Furthermore, because the index τ has unbounded support, the L^2-smoothness result involves the weights via $\sqrt{\pi}$. Without π, the L^2-smoothness result may not hold uniformly in $\tau \in \mathbb{R}^{d_\tau}$.

Lemma 5 (Nonparametric Approximation Bias)**.** *Suppose Assumptions 1 and 2 (or 2′) hold. Furthermore suppose that $\mathbb{E}\left(\|y_t^s\|^2\right)$ and $\mathbb{E}\left(\|u_t^s\|^2\right)$ are bounded for $\beta = \beta_0$ and $\beta = \Pi_{k(n)}\beta_0$ for all $k(n) \geq 1$, $t \geq 1$ then:*

$$Q_n(\Pi_{k(n)}\beta_0) = O\left(\max\left[\frac{\log[k(n)]^{4r/b+2}}{k(n)^{2r}}, \frac{\log[k(n)]^{4\gamma^2 r/b}}{k(n)^{2\gamma^2 r}}, \frac{1}{n^2}\right]\right) = O\left(\frac{\log[k(n)]^{4r/b+2}}{k(n)^{2\gamma^2 r}}\right)$$

where $\Pi_{k(n)}\beta_0$ is the mixture sieve approximation of β_0, γ the Hölder coefficient in Assumption 2, b and r are the exponential tail index and the smoothness of the density f_S in Lemma 2.

[24] Under the assumption of Lemma 2: $\underline{\sigma}_{k(n)}^{-2\gamma^2} = O\left(k(n)^{2\gamma^2}/\log[(n)]^{4\gamma^2/b}\right)$ and $[k(n) + \overline{\mu}_{k(n)} + \overline{\sigma}]^\gamma = O\left(k(n)^\gamma\right)$. As a result, the maximum term is bounded above by $\max\left(k(n)^{2\gamma^2}, k(n)^\gamma\right)\delta^{\gamma^2/2}$ (up to a constant).

Lemma 5 gives the second implication of Assumption 2; it computes the value of the objective function Q_n at $\Pi_{k(n)}\beta_0$, which is directly related to the bias of the estimator $\widehat{\beta}_n$. Two terms are particularly important for the rate of convergence: the smoothness of the true density r and the roughness of the DGP as measured by the Hölder coefficient $\gamma \in (0,1]$. If r and γ are larger then the bias will be smaller. The rate in this lemma is different from the usual rate found in the sieve literature. Chen & Pouzo (2012) assume for instance that $Q_n(\Pi_{k(n)}\beta_0) \asymp \|\beta_0 - \Pi_{k(n)}\beta_0\|^2_{\mathcal{B}}$. In comparison, the rate derived here is:

$$Q_n(\Pi_{k(n)}\beta_0) \asymp \max\left(\|\beta_0 - \Pi_{k(n)}\beta_0\|^2_{\mathcal{B}} \log\left(\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}\right)^2, \|\beta_0 - \Pi_{k(n)}\beta_0\|^{2\gamma^2}_{\mathcal{B}}, 1/n^2\right)$$

with $\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}} = O(\log[k(n)]^{2r/b}/k(n)^r)$ as given in Lemma 2. The $1/n^2$ term corresponds to the bias due to the nonstationarity, its order is implied by the geometric ergodicity condition and the boundedness of the moments. The log-bias term $\log\left(\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}\right)$ is due to the dynamics: y_t^s depends on the full history $(e_t^s, \ldots, e_1^s)$ which are iid $\Pi_{k(n)}f_0$, so that the bias accumulates. The decay conditions *y(i), y(iii), u(i)* ensure that the resulting bias accumulation only inflates bias by a log term. The term $\|\beta_0 - \Pi_{k(n)}\beta_0\|^{2\gamma^2}_{\mathcal{B}}$ is due to the Hölder smoothness of the DGP. If the DGP is Lipschitz, i.e. $\gamma = 1$, and the model is static then the rate becomes $Q_n(\Pi_{k(n)}\beta_0) \asymp \|\beta_0 - \Pi_{k(n)}\beta_0\|^2_{\mathcal{B}}$, which is the rate assumed in Chen & Pouzo (2012).

Theorem 1 (Consistency)**.** *Suppose Assumptions 1 and 2 (or 2′) hold. Suppose that $\beta \to Q_n(\beta)$ is continuous on $(\mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}})$ and the numerical optimization and integration errors are such that $\eta_n = o(1/n)$. If for all $\varepsilon > 0$ the following holds:*

$$\max\left(\frac{log[k(n)]^{4r/b+2}}{k(n)^{2\gamma^2 r}}, \frac{k(n)^4 \log[k(n)]^4}{n}, \frac{1}{n^2}\right) = o\left(diag_{\beta\in\mathcal{B}_{k(n)}, \|\beta-\beta_0\|_{\mathcal{B}}\geq\varepsilon} Q_n(\beta)\right) \tag{3.8}$$

where r is the assumed smoothness of the smooth component f_S and b its exponential tail index. Then the Sieve-SMM estimator is consistent:

$$\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} = o_p(1).$$

Theorem 1 is a consequence of the general consistency lemma in Chen & Pouzo (2012) reproduced as Lemma .0.1 in the appendix. They provide high level conditions which Assumption 2 together with Lemmas 4 and 5 verify for simulation-based estimation of static and dynamic models. Condition (3.8) in Theorem 1 allows for ill-posedness but

requires the minimum to be well separated on the sieve space relative to the bias and the variance.

The variance term $k(n)^4 \log[k(n)]^4/n$ is derived using the inequality in Lemma .0.4 which is adapted from existing results of Andrews & Pollard (1994); Ben Hariz (2005). It is based on the moment inequalities for α-mixing sequences of Rio (2000) rather than coupling results (see e.g. Doukhan et al., 1995; Chen & Shen, 1998; Dedecker & Louhichi, 2002). This implies that the moments can be nonstationary, because of the initial condition, and depend on arbitrarily many lags as in Example 1 where y_t^s is a function of $e_t^s, \ldots, e_t^1$. It also allows for filtering procedures as in the first extension of the main results. The two main drawbacks of this inequality is that it requires uniformly bounded moments and implies a larger variance than, for instance, in the iid case. The boundedness restricts the class of moments used in Sieve-SMM and the larger variance implies a slower rate of convergence.

Rate of Convergence

Once the consistency of the estimator is established, the next step is to derive its rate of convergence. It is particularly important to derive rates that are as sharp as possible since a rate of a least $n^{-1/4}$ under the weak norm of Ai & Chen (2003) is required for the asymptotic normality results. This weak norm is introduced below for the continuum of complex valued moments. It is related to the objective function Q_n, and as such allows to derive the rate of convergence of $\widehat{\beta}_n$.[25] Ultimately, the norm of interest in the strong norm $\|\cdot\|_{\mathcal{B}}$ which is generally not equivalent to the weak norm since the space is infinite dimensional. The two are related by the local measure of ill-posedness of Blundell et al. (2007) which allows to derive the rate of convergence in the strong norm, that is in either the total variation or the supremum norm.

Assumption 3 (Weak Norm and Local Properties)**.** *Let* $\mathcal{B}_{osn} = \mathcal{B}_{k(n)} \cap \{\|\beta - \beta_0\|_{\mathcal{B}} \leq \varepsilon\}$ *for* $\varepsilon > 0$ *small and for* $(\beta_1, \beta_2) \in \mathcal{B}_{osn}$*:*

$$\|\beta_1 - \beta_2\|_{weak} = \left[\int \left| \frac{d\mathbb{E}\left(\widehat{\psi}_n^S(\tau, \beta_0)\right)}{d\beta} [\beta_1 - \beta_2] \right|^2 \pi(\tau) d\tau \right]^{1/2} \tag{3.9}$$

[25] For a discussion see Ai & Chen (2003) and Chen (2007).

is the weak norm of $\beta_1 - \beta_2$. Suppose that there exists $\underline{C}_w > 0$ such that for all $\beta \in \mathcal{B}_{osn}$:

$$\underline{C}_w \|\beta - \beta_0\|_{weak}^2 \leq \int \left| \mathbb{E}\left(\widehat{\psi}_n^S(\tau, \beta_0) - \widehat{\psi}_n^S(\tau, \beta) \right) \right|^2 \pi(\tau) d\tau. \tag{3.10}$$

Assumption 3 adapts the weak norm of Ai & Chen (2003) to an objective with a continuum of complex-valued moments. Note that $\int |\mathbb{E}\left(\widehat{\psi}_n^S(\tau, \beta_0) - \widehat{\psi}_n^S(\tau, \beta) \right)|^2 \pi(\tau) d\tau = Q_n(\beta_0) + O_p(1/n^2)$ under geometric ergodicity. As a result, Assumption 3 implies that the weak norm is Lipschitz continuous with respect to $\sqrt{Q_n}$. Additional assumptions on the norm and the objective are usually required such as: $Q_n(\beta) \asymp \|\beta - \beta_0\|_{weak}^2$ and $Q_n(\beta) \leq C_{\mathcal{B}} \|\beta - \beta_0\|_{\mathcal{B}}$ (see e.g. Chen & Pouzo, 2015, Assumption 3.4). Instead of these assumptions, the results in this paper rely on Lemma 5 to derive the bias of the estimator. The resulting bias is larger than in the usual sieve literature.

Theorem 2 (Rate of Convergence)**.** *Suppose that the assumptions for Theorem 1 hold and Assumption 3 also holds.The convergence rate in weak norm is:*

$$\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p\left(\max\left(\frac{\log[k(n)]^{r/b+1}}{k(n)^{\gamma^2 r}}, \frac{k(n)^2 \log[k(n)]^2}{\sqrt{n}} \right) \right). \tag{3.11}$$

The convergence rate in either the total variation or supremum norm $\|\cdot\|_{\mathcal{B}}$ is:

$$\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} = O_p\left(\frac{\log[k(n)]^{r/b}}{k(n)^r} + \tau_{\mathcal{B},n} \max\left(\frac{\log[k(n)]^{r/b+1}}{k(n)^{\gamma^2 r}}, \frac{k(n)^2 \log[k(n)]^2}{\sqrt{n}} \right) \right)$$

where $\tau_{\mathcal{B},n}$ is the local measure of ill-posedness of Blundell et al. (2007):

$$\tau_{\mathcal{B},n} = sup_{\beta \in \mathcal{B}_{osn}, \|\beta - \Pi_{k(n)}\beta_0\|_{weak} \neq 0} \frac{\|\beta - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}}{\|\beta - \Pi_{k(n)}\beta_0\|_{weak}}.$$

As usual in the (semi)-nonparametric estimation literature, the rate of convergence involves a bias/variance trade-off. As discussed before, the bias is larger than usual because of the dynamics and involves the Hölder smoothness γ of the DGP.

The variance term is of order $k(n)^2 \log[k(n)]^2 / \sqrt{n}$ instead of $\sqrt{k(n)} / \sqrt{n}$ in the iid case or strictly stationary case with fixed number of lags in the moments. This is because the inequality in Lemma .0.4 is more conservative than the inequalities found in Theorem 2.14.2 of van der Vaart & Wellner (1996) for iid observations or the inequalities based on a coupling argument in Doukhan et al. (1995); Chen & Shen (1998) for strictly stationary dependent data. However, in this simulation-based setting the dependence properties of y_t^s varies on $\boldsymbol{\theta}$ over the parameter space Θ so that a coupling approach may not apply unless it only depends on finitely many lags of e_t and x_t. Determining whether this inequality can be sharpened in subject to future research.

The increased bias and variance imply a slower rate of convergence than usual. The optimal rate of convergence equates the bias and variance terms in equation (3.11). This is achieved (up to a log term) by picking $k(n) = O(n^{\frac{1}{2(2+\gamma^2 r)}})$. To illustrate, for a Lipschitz DGP $\gamma = 1$ and f_0 twice continuously differentiable $r = 2$ and $k(n) \asymp n^{1/8}$, the rate of convergence becomes:

$$\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p(n^{-1/4} \log(n)^{\max(2/b+1,2)}).$$

In comparison, if (y_t^s, x_t) were iid, keeping $\gamma = 1$ and $r = 2$, the variance term would be $\sqrt{k(n) \log[k(n)]/n}$ and the optimal $k(n) \asymp n^{1/5}$. The rate of convergence becomes:

$$\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p\left(n^{-2/5} \log(n)^{\max(2/b+1,2)}\right).$$

To achieve a rate faster than $n^{-1/4}$, as required for asymptotic normality, the smoothness of the true density f_0 must satisfy $r \geq 3/\gamma^2$ where γ is the Hölder coefficient in Assumption 2. In the Lipschitz case, $\gamma = 1$, at 3 derivatives are needed compared to 12 derivatives when $\gamma = 1/2$. In comparison, in the iid case 2 and 8 derivatives are needed for $\gamma = 1$ and $\gamma = 1/2$ respectively.

The following corollary shows that the number of simulated samples S can significantly reduce the sieve variance. This changes the bias-variance trade-off and improves the rate of convergence in the weak norm.

Corollary 1 (Number of Simulated Samples S and the Rate of Convergence)**.** *If a long sample $(y_1^s, \ldots, y_{nS}^s)$ can be simulated then the variance term becomes:*

$$\min\left(\frac{k(n)^2 \log[n]^2}{\sqrt{n \times S}}, \frac{1}{\sqrt{n}}\right).$$

As a result, for $S(n) \asymp k(n)^4 \log[k(n)]^4$ the rate of convergence in weak norm is:

$$\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p\left(\max\left(\frac{\log[k(n)]^{r/b+1}}{k(n)^{\gamma^2 r}}, \frac{1}{\sqrt{n}}\right)\right).$$

And the rate of convergence in either the total variation or the supremum norm is:

$$\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} = O_p\left(\frac{\log[k(n)]^{r/b}}{k(n)^r} + \tau_{\mathcal{B},n} \max\left(\frac{\log[k(n)]^{r/b+1}}{k(n)^{\gamma^2 r}}, \frac{1}{\sqrt{n}}\right)\right)$$

where $\tau_{\mathcal{B},n}$ is the local measure of ill-posedness in Theorem 2.

The assumption that a long sample can be simulated is called the ECA assumption in Kristensen & Salanié (2017); it is more commonly found in dynamic models than cross-sectional or panel data models. In the parametric SMM and Indirect Inference literature, S has an effect on the asymptotic variance whereas in the Sieve-SMM setting, Corollary 1 shows that increasing S with the sample size n can also improve the rate of convergence in the weak norm. Assuming undersmoothing so that the rate in weak norm is of order $1/\sqrt{n}$, the rate of convergence in the stronger norm $\|\cdot\|_{\mathcal{B}}$ becomes $O_p(k(n)^{-r} + \tau_{\mathcal{B},n}/\sqrt{n})$, up to a log term. This is faster than the rates of convergence found in the literature.

In practice, the number of simulated sample $S(n)$ required to achieve the rate in Corollary 1 can be very large. For $n = 1,000$, $\gamma = 1$ and $r = 2$, the optimal $k(n) \simeq 5$ and $S(n) = k(n)^4 \simeq 625$. The total number of simulated y_t^s required is $n \times S(n) = 625,000$. For iid data, the required number of simulations is $n \times S(n) = 5,000$. As a result, improving the rate of convergence of the estimator can be computationally costly since it involves increasing both the number of samples to simulate and the number of parameters to be estimate.

Remark 1 (An Illustration of the Local Measure of Ill-Posedness)**.** *The sieve measure of ill-posedness is generally difficult to compute. To illustrate a source of ill-posedness and its order of magnitude, consider the following basic static model:*

$$y_t^s = e_t^s \overset{iid}{\sim} f.$$

The only parameter to be estimated is the density f which can also be approximated with kernel density estimates. For this model the characteristic function is linear in f and as a consequence the weak norm for $f_1 - f_2$ is the weighted difference of the CFs ψ_{f_1}, ψ_{f_1} for f_1, f_2:

$$\|f_1 - f_2\|_{weak} = \left[\int |\psi_{f_1}(\tau) - \psi_{f_2}(\tau)|^2 \pi(\tau) d\tau\right]^{1/2}.$$

The weak norm is bounded above by 2 for any two densities f_1, f_2. However, the total variation and supremum distances are not bounded above: as a result the ratio between the weak norm and these stronger norms is unbounded. To illustrate, simplify the problem further and assume there is only one mixture component:

$$f_{1,k(n)}(e) = \underline{\sigma}_{k(n)}^{-1} \phi\left(\frac{e}{\underline{\sigma}_{k(n)}}\right), \quad f_{2,k(n)}(e) = \underline{\sigma}_{k(n)}^{-1} \phi\left(\frac{e - \mu_{k(n)}}{\underline{\sigma}_{k(n)}}\right).$$

As the bandwidth $\sigma_{k(n)} \to 0$, the two densities approach Dirac masses. Unless $\mu_{k(n)} \to 0$, the total variation and supremum distances between the two densities go to infinity while the

distance in weak norm is bounded. The distance between f_1 and f_2 in weak, total-variation and supremum norm are given in Appendix 3.7. For a well chosen sequence $\mu_{k(n)}$, the total variation and supremum distances are bounded above and below while the weak norm goes to zero. The ratio provides the local measures of ill-posedness:

$$\tau_{TV,n} = O\left(\frac{k(n)}{\log[k(n)]^{2/b}}\right), \quad \tau_{\infty,n} = O\left(\frac{k(n)^2}{\log[k(n)]^{4/b}}\right).$$

Hence, this simple example suggests that Characteristic Function based Sieve-SMM estimation problems are at best mildly ill-posed.

Asymptotic Normality

This section derives asymptotic normality results for plug-in estimates $\phi(\widehat{\beta}_n)$ where ϕ are smooth functionals of the parameters. As in Chen & Pouzo (2015), the main result finds a normalizing sequence $r_n \to \infty$ such that:

$$r_n \times \left(\phi\left(\widehat{\beta}_n\right) - \phi\left(\beta_0\right)\right) \xrightarrow{d} \mathcal{N}(0,1)$$

where $r_n = \sqrt{n}/\sigma_n^*$, for some sequence of standard errors $(\sigma_n^*)_{n\geq 1}$ which can go to infinity. If $\sigma_n^* \to \infty$, the plug-in estimates will converge at a slower than $\sqrt{n}$-rate. In addition, sufficient conditions for $\widehat{\boldsymbol{\theta}}_n$ to be root-n asymptotically normal, that is $\lim_{n\to\infty} \sigma_n^* < \infty$, are given in Appendix 3.7 for the stochastic volatility model of Example 1.

To establish asymptotic normality results, stochastic equicontinuity results are required. However, the L^2-smoothness result only holds in the space of mixtures $\mathcal{B}_{k(n)}$ with the pseudo-norm $\|\cdot\|_m$ on the mixture parameters. This introduces two difficulties in deriving the results: a rate of convergence for the norm on the mixture components is required, and since $\beta_0 \notin \mathcal{B}_{k(n)}$ in general, the rate and the stochastic equicontinuity results need to be derived around a sequence of mixtures that are close enough to β_0 so that they extend to β_0. The following lemma provides the rate of convergence in the mixture norm.

Lemma 6 (Convergence Rate in Mixture Pseudo-Norm)**.** *Let $\delta_n = (k(n)\log[k(n)])^2/\sqrt{n}$ and $M_n = \log\log(n+1)$. Suppose the following undersmoothing assumptions hold:*

i. (Rate of Convergence) $\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p(\delta_n)$
ii. (Negligible Bias) $\|\Pi_{k(n)}\beta_0 - \beta_0\|_{weak} = o(\delta_n)$.

Furthermore, suppose that the population CF is smooth in β and satisfies:

iii. (Approximation Rate 1) Uniformly over $\beta \in \{\beta \in \mathcal{B}_{osn}, \|\beta - \beta_0\|_{weak} \leq M_n \delta_n\}$*:*

$$\int \left| \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \beta_0))}{d\beta}[\beta - \beta_0] - \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta}[\beta - \beta_0] \right|^2 \pi(\tau) d\tau = O(\delta_n^2).$$

iv. (Approximation Rate 2) The approximating mixture $\Pi_{k(n)}\beta_0$ *satisfies:*

$$\int \left| \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta}[\Pi_{k(n)}\beta_0 - \beta_0] \right|^2 \pi(\tau) d\tau = O(\delta_n^2).$$

Let $\underline{\lambda}_n$ *be the smallest eigenvalue of the matrix*

$$\int \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d(\boldsymbol{\theta}, \omega, \mu, \sigma)}' \overline{\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d(\boldsymbol{\theta}, \omega, \mu, \sigma)}} \pi(\tau) d\tau.$$

Suppose that $\underline{\lambda}_n > 0$ *and* $\delta_n \underline{\lambda}_n^{-1/2} = o(1)$ *then the convergence rate in the mixture pseudo-norm is:*

$$\|\widehat{\beta}_n - \Pi_{k(n)}\beta_0\|_m = O_p\left(\delta_n \underline{\lambda}_n^{-1/2}\right)$$

where $\|\beta\|_m = \|(\boldsymbol{\theta}, \omega, \mu, \sigma)\|$ *is the pseudo-norm on* $\boldsymbol{\theta}$ *and the mixture parameters* (ω, μ, σ)*.*

The rate of convergence in mixture norm $\|\cdot\|_m$ corresponds to the rate of convergence in the weak norm $\|\cdot\|_m$ times a measure of ill-posedness $\underline{\lambda}_n^{-1/2}$. Relations between the mixture norm and the strong norm $\|\cdot\|_{\mathcal{B}}$ imply that the local measure of ill-posedness in Theorem 2 can be computed using $\underline{\lambda}_n^{-1/2}$. Indeed, results in van der Vaart & Ghosal (2001); Kruijer et al. (2010) imply that $\|\beta - \Pi_{k(n)}\beta_0\|_{TV} \leq \underline{\sigma}_{k(n)}^{-1}\|\beta - \Pi_{k(n)}\beta_0\|_m$ and $\|\beta - \Pi_{k(n)}\beta_0\|_\infty \leq \underline{\sigma}_{k(n)}^{-2}\|\beta - \Pi_{k(n)}\beta_0\|_m$. These inequalities imply upper-bounds for ill-posedness in total variation and supremum norms:

$$\tau_{TV,n} \leq \underline{\lambda}_n^{-1/2}\underline{\sigma}_{k(n)}^{-1} \quad \text{and} \quad \tau_{\infty,n} \leq \underline{\lambda}_n^{-1/2}\underline{\sigma}_{k(n)}^{-2}.$$

The quantity $\underline{\lambda}_n^{-1/2}$ can be approximated numerically using sample estimates and $\underline{\sigma}_{k(n)}$ is the bandwidth in Lemma 2. As a result, even though the local measure of ill-posedness from Theorem 2 is generally not tractable, an upper bound can be computed using Lemma 6. Chen & Christensen (2017) shows how to achieve the optimal rate of convergence using plug-in estimates of the measure of ill-posedness in nonparametric instrumental variable regression, a similar approach should be applicable here using these bounds. This is left to future research.

Lemma 7 (Stochastic Equicontinuity Results)**.** *Let $\delta_{mn} = \delta_n \underline{\lambda}_n^{-1/2}$. Suppose that the assumptions of Lemma 6 hold and $(M_n\delta_{mn})^{\frac{\gamma^2}{2}} \max(\log[k(n)]^2, |\log[M_n\delta_{mn}]|^2)k(n)^2 = o(1)$, then a first stochastic equicontinuity result holds:*

$$sup_{\|\beta - \Pi_{k(n)}\beta_0\|_m \leq M_n\delta_{mn}} \int \left| [\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)] - \mathbb{E}[\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)] \right|^2 \pi(\tau)d\tau = o_p(1/n).$$

Also, suppose that $\beta \to \int \mathbb{E}\left|\widehat{\psi}_t^s(\tau,\beta_0) - \widehat{\psi}_t^s(\tau,\beta)\right|^2 \pi(\tau)d\tau$ is continuous with respect to $\|\cdot\|_{\mathcal{B}}$ at $\beta = \beta_0$, uniformly in $t \geq 1$, then a second stochastic equicontinuity result holds:

$$sup_{\|\beta - \Pi_{k(n)}\beta_0\|_m \leq M_n\delta_{mn}} \int \left| [\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\beta_0)] - \mathbb{E}[\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\beta_0)] \right|^2 \pi(\tau)d\tau = o_p(1/n).$$

Lemma 7 uses the rate of convergence in mixture norm to establish stochastic equicontinuity results. With these results, the moments $\widehat{\psi}_n^s(\tau,\beta) - \widehat{\psi}_n^s(\tau,\beta_0)$ can be substituted with a smoothed version under the integral of the objective function.

Remark 2 (Required Rate of Convergence)**.** *To achieve the rate of convergence required in Lemma 7, $k(n)$ must grow at a power of the sample size n, hence: $\log(n) \asymp \log[k(n)] \asymp |\log(\delta_{mn})|$. As a result, the condition on the rate of convergence in mixture norm*

$$(M_n\delta_{mn})^{\frac{\gamma^2}{2}} \max(\log[k(n)]^2, |\log[M_n\delta_{mn}]|^2)k(n)^2 = o(1)$$

in Lemma 7 can be simplified to:

$$M_n\delta_n = o\left(\frac{\sqrt{\underline{\lambda}_n}}{[k(n)\log(n)]^{4/\gamma^2}}\right).$$

The following definition adapts the tools used in the sieve literature to establish asymptotic normality of smooth functionals (see e.g. Wong & Severini, 1991; Ai & Chen, 2003; Chen & Pouzo, 2015; Chen & Liao, 2015) to a continuum of complex valued moments.

Definition 2 (Sieve Representer, Sieve Score, Sieve Variance)**.** *Let $\beta_{0,n}$ be such that $\|\beta_{0,n} - \beta_0\|_{weak} = diag_{\beta\in\mathcal{B}_{osn}}\|\beta - \beta_0\|_{weak}$, let $\overline{V}_{k(n)}$ be the closed span of $\mathcal{B}_{osn} - \{\beta_{0,n}\}$. The inner product $\langle\cdot,\cdot\rangle$ of $(v_1, v_2) \in \overline{V}_{k(n)}$ is defined as:*

$$\langle v_1, v_2\rangle = \frac{1}{2}\int \left[\psi_\beta(\tau, v_1)\overline{\psi_\beta(\tau, v_2)} + \overline{\psi_\beta(\tau, v_1)}\psi_\beta(\tau, v_2)\right]\pi(\tau)d\tau.$$

i. The Sieve Representer is the unique vector $v_n^* \in \overline{V}_{k(n)}$ such that $\forall v \in \overline{V}_{k(n)}$: $\langle v_n^*, v\rangle = \frac{d\phi(\beta_0)}{d\beta}[v]$.

ii. The Sieve Score S_n^* is:

$$\begin{aligned} S_n^* &= \frac{1}{2}\int \left[\psi_\beta(\tau, v_n^*)\overline{[\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n(\tau)]} + \overline{\psi_\beta(\tau, v_n^*)}[\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n(\tau)]\right]\pi(\tau)d\tau \\ &= \int Real\left(\psi_\beta(\tau, v_n^*)\overline{[\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n(\tau)]}\right)\pi(\tau)d\tau. \end{aligned}$$

iii. The Sieve Long Run Variance σ_n^* is:

$$\sigma_n^{*2} = n\mathbb{E}\left(S_n^{*2}\right) = n\mathbb{E}\left(\left[\int Real\left(\psi_\beta(\tau, v_n^*)\overline{[\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n(\tau)]}\right)\pi(\tau)d\tau\right]^2\right).$$

iv. The Scale Sieve Representer u_n^* is: $u_n^* = v_n^*/\sigma_n^*$.

Assumption 4 (Equivalence Condition)**.** *There exists $\underline{a} > 0$ such that for all $n \geq 1$: $\underline{a}\|v_n^*\|_{weak} \leq \sigma_n^*$. Furthermore, suppose that σ_n^* does not increase too fast: $\sigma_n^* = o(\sqrt{n})$.*

In Sieve-MD literature, Assumption 4 is implied by an eigenvalue condition on the conditional variance of the moments.[26] Because the moments are bounded and the data is geometrically ergodic, the long-run variance of the moments is bounded above uniformly in τ.[27] However, since τ has unbounded support, the eigenvalues of the variance may not have a strictly positive lower bound. Assumption 4 plays the role of the lower bound on the eigenvalues.[28]

Assumption 5 (Convergence Rate, Smoothness, Bias)**.** *$\mathcal{B}_{osn}$ is a convex neighborhood of β_0 where*

i. (Rate of Convergence) $M_n\delta_n = o(n^{-1/4})$ and $M_n\delta_n = o\left(\sqrt{\underline{\lambda}_n}/\left(k(n)\log(n)\right)^{4/\gamma^2}\right)$.

ii. (Smoothness) A linear expansion of ϕ is locally uniformly valid:

$$sup_{\|\beta-\beta_0\|\leq M_n\delta_n}\frac{\sqrt{n}}{\sigma_n^*}\left|\phi(\beta) - \phi(\beta_0) - \frac{d\phi(\beta_0)}{d\beta}[\beta-\beta_0]\right| = o(1).$$

A linear expansion of the moments is locally uniformly valid:

$$\begin{aligned} sup_{\|\beta-\beta_0\|_{weak}\leq M_n\delta_n}&\left(\int\left|\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta)) - \mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0)) - \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\beta-\beta_0]\right|^2\pi(\tau)d\tau\right)^{1/2} \\ &= O\left((M_n\delta_n)^2\right). \end{aligned}$$

[26] See e.g. assumption 3.1(iv) in Chen & Pouzo (2015).

[27] This is shown in Appendix 3.7.

[28] A discussion of this assumption is given in Appendix 3.7

The second derivative is bounded:

$$sup_{\|\beta-\beta_0\|_{weak}\leq M_n\delta_n}\left(\int\left|\frac{d^2\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta d\beta}[u_n^*,u_n^*]\right|^2\pi(\tau)d\tau\right)^{1/2}=O(1).$$

iii. (Bias) The approximation bias is negligible:

$$\frac{\sqrt{n}}{\sigma_n^*}\frac{d\phi(\beta_0)}{d\beta}[\beta_{0,n}-\beta_0]=o(1).$$

Note that if $\mathcal{B}_{osn}$ is a convex neighborhood of β_0 then $\boldsymbol{\theta}_0$ is in the interior of Θ. Assumption 5 is standard in the literature. The first rate condition ensure the nonparametric component converges fast enough so that the central limit theorem dominates the asymptotic distribution (Newey, 1994; Chen et al., 2003), the second rate condition is required in Lemma 7. The smoothness and bias conditions can also be found in Ai & Chen (2003) and Chen & Pouzo (2015). The bias condition implies undersmoothing so that the variance term dominates asymptotically.

Theorem 3 (Asymptotic Normality)**.** *Suppose the assumptions of Theorems 1, 2 and lemmas 6, 7 hold as well as Assumptions 4 and 5, then as n goes to infinity:*

$$r_n\times\left(\phi(\widehat{\beta}_n)-\phi(\beta_0)\right)\xrightarrow{d}\mathcal{N}(0,1)$$

where $r_n=\frac{\sqrt{n}}{\sigma_n^*}\to\infty$.

Theorem 3 shows that under the previous assumptions, inferences on $\phi(\beta_0)$ can be conducted using the confidence interval $[\phi(\widehat{\beta}_n)\pm 1.96\times\sigma_n^*/\sqrt{n}]$. The standard errors $\sigma_n^*>0$ adjust automatically so that $r_n=\sqrt{n}/\sigma_n^*$ gives the correct rate of convergence. If $\lim_{n\to\infty}\sigma_n^*<\infty$, then $\phi(\widehat{\beta}_n)$ is $\sqrt{n}-$convergent. A result for $\widehat{\boldsymbol{\theta}}_n$ is given in Proposition .0.1 in the Appendix for a smaller class of models that include the stochastic volatility model in Example 1.

As in Chen & Pouzo (2015) and Chen & Liao (2015), the sieve variance has a closed-form expression analogous to the parametric Delta-method formula. The notation is taken from Chen & Pouzo (2015), with sieve parameters $(\widehat{\omega}_n,\widehat{\mu}_n,\widehat{\sigma}_n)$ the sieve variance can be estimated using:

$$\widehat{\sigma}_n^{2*}=\frac{d\phi(\widehat{\boldsymbol{\theta}}_n,\widehat{\omega}_n,\widehat{\mu}_n,\widehat{\sigma}_n)}{d(\boldsymbol{\theta},\omega,\mu,\sigma)}'\widehat{D}_n\widehat{\mho}_n\widehat{D}_n\frac{d\phi(\widehat{\boldsymbol{\theta}}_n,\widehat{\omega}_n,\widehat{\mu}_n,\widehat{\sigma}_n)}{d(\boldsymbol{\theta},\omega,\mu,\sigma)}$$

where

$$\widehat{D}_n = \left(Real \left(\int \frac{d\widehat{\psi}_n^S(\tau, \widehat{\boldsymbol{\theta}}_n, \widehat{\omega}_n, \widehat{\mu}_n, \widehat{\sigma}_n)}{d(\boldsymbol{\theta}, \omega, \mu, \sigma)'} \overline{\frac{d\widehat{\psi}_n^S(\tau, \widehat{\boldsymbol{\theta}}_n, \widehat{\omega}_n, \widehat{\mu}_n, \widehat{\sigma}_n)}{d(\boldsymbol{\theta}, \omega, \mu, \sigma)}} \pi(\tau) d\tau \right) \right)^{-1}$$

$$\widehat{\mho}_n = \int \widehat{G}_n(\tau_1)' \widehat{\Sigma}_n(\tau_1, \tau_2) \widehat{G}_n(\tau_2) \pi(\tau_1) \pi(\tau_2) d\tau_1 d\tau_2.$$

$\widehat{G}_n$ stacks the real and imaginary components of the gradient:

$$\widehat{G}_n(\tau) = \begin{pmatrix} Real \left(\frac{d\widehat{\psi}_n^S(\tau, \widehat{\boldsymbol{\theta}}_n, \widehat{\omega}_n, \widehat{\mu}_n, \widehat{\sigma}_n)}{d(\boldsymbol{\theta}, \omega, \mu, \sigma)} \right) \\ Im \left(\frac{d\widehat{\psi}_n^S(\tau, \widehat{\boldsymbol{\theta}}_n, \widehat{\omega}_n, \widehat{\mu}_n, \widehat{\sigma}_n)}{d(\boldsymbol{\theta}, \omega, \mu, \sigma)} \right) \end{pmatrix}.$$

Let $Z_n^S(\tau, \beta) = \widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \beta)$ The covariance operator $\widehat{\Sigma}_n$ approximates the population long-run covariance operator Σ_n:

$$\Sigma_n(\tau_1, \tau_2) = n\mathbb{E} \begin{pmatrix} Real\left(Z_n^S(\tau_1, \beta_0)\right) Real\left(Z_n^S(\tau_2, \beta_0)\right) & Real\left(Z_n^S(\tau_1, \beta_0)\right) Im\left(Z_n^S(\tau_2, \beta_0)\right) \\ Im\left(Z_n^S(\tau_1, \beta_0)\right) Im\left(Z_n^S(\tau_2, \beta_0)\right) & Im\left(Z_n^S(\tau_1, \beta_0)\right) Real\left(Z_n^S(\tau_2, \beta_0)\right) \end{pmatrix}.$$

Carrasco et al. (2007a) gives results for the Newey-West estimator of Σ_n. In practice, applying the block Bootstrap to the quantity

$$Real \left(\frac{d\widehat{\psi}_n^S(\tau, \widehat{\boldsymbol{\theta}}_n, \widehat{\omega}_n, \widehat{\mu}_n, \widehat{\sigma}_n)}{d(\boldsymbol{\theta}, \omega, \mu, \sigma)} \overline{\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n(\tau, \widehat{\beta}_n) \right)} \right)$$

is more convenient than computing the large matrices $\widehat{G}_n, \widehat{\Sigma}_n$. $\widehat{\beta}_n$ is held fixed across Bootstrap iterations so that the model is only estimated once. The Gaussian and uniform draws $Z_{j,t}^s$ and ν_t^s are re-drawn at each Bootstrap iteration.

3.4 Extensions

This section considers two extensions to the main results: the first covers auxiliary variables in the CF and the seconds allows for panel datasets with small T.

Using Auxiliary Variables

The first extension involves adding transformations of the data, such as using simple functions of y_t or a filtered volatility from an auxiliary GARCH model, to the CF $\widehat{\psi}_n$. This

approach can be useful in cases where (y_t, u_t) is Markovian but y_t alone is not, in which case functions of the full history $(y_t, \dots, y_1)$ provide additional information about the unobserved u_t. It is used to estimate stochastic volatility models in sections 3.5 and 3.6. Other potential applications include filtering latent variables from an auxiliary linearized DSGE model to estimate a more complex, intractable non-linear DSGE model.

The auxiliary model consists of an auxiliary variable z_t^{aux} (the filtered GARCH volatility) and auxiliary parameters $\widehat{\eta}_n^{aux}$ (the estimated GARCH parameters). The estimates $\widehat{\eta}_n^{aux}$ are computed from the full sample $(y_1, \dots, y_n, x_1, \dots, x_n)$ and the auxiliary variables $z_t^{aux}, z_t^{s,aux}$ are computed using the full and simulated samples:[29]

$$z_t^{aux} = g_{t,aux}(y_t, \dots, y_1, x_t, \dots, x_1, \widehat{\eta}_n^{aux}), \quad z_t^{s,aux} = g_{t,aux}(y_t^s, \dots, y_1^s, x_t, \dots, x_1, \widehat{\eta}_n^{aux}).$$

The moment function $\widehat{\psi}_n$ is now the joint CF of the lagged data $(\mathbf{y}_t, \mathbf{x}_t)$ and the auxiliary $\mathbf{z}_t^{aux}$:

$$\widehat{\psi}_n(\tau, \widehat{\eta}_n^{aux}) = \sum_{t=1}^{n} e^{i\tau'(\mathbf{y}_t, \mathbf{x}_t, \mathbf{z}_t^{aux})}, \quad \widehat{\psi}_n^s(\tau, \widehat{\eta}_n^{aux}, \beta) = \sum_{t=1}^{n} e^{i\tau'(\mathbf{y}_t^s, \mathbf{x}_t, \mathbf{z}_t^{s,aux})}.$$

The following assumption provides sufficient conditions on the estimates $\widehat{\eta}_n^{aux}$ and the filtering process $g_{t,aux}$ for the asymptotic properties in section 3.3 to also hold with auxiliary variables.

Assumption 6 (Auxiliary Variables)**.** *The estimates $\widehat{\eta}_n^{aux}$ are such that:*

i. Compactness: with probability 1 $\widehat{\eta}_n^{aux} \in E$ finite dimensional, convex and compact.

ii. Convergence: there exists a $\eta^{aux} \in E$ such that:

$$\sqrt{n}\,(\widehat{\eta}_n^{aux} - \eta^{aux}) \xrightarrow{d} \mathcal{N}(0, V^{aux}).$$

iii. Lipschitz Continuity: for any two η_1^{aux}, η_2^{aux} and for both y_t^s and y_t:

$$\begin{aligned} &\|g_{t,aux}(y_t, \dots, y_1, x_t, \dots, x_1, \eta_1^{aux}) - z_{t,aux}(y_t, \dots, y_1, x_t, \dots, x_1, \eta_2^{aux})\| \\ &\leq C^{aux}(y_t, \dots, y_1, x_t, \dots, x_1) \times \|\eta_1^{aux} - \eta_2^{aux}\| \end{aligned}$$

with $\mathbb{E}(C^{aux}(y_t, \dots, y_1, x_t, \dots, x_1)^2) \leq \overline{C}^{aux} < \infty$ and $\mathbb{E}(C^{aux}(y_t^s, \dots, y_1^s, x_t, \dots, x_1)^2) \leq \overline{C}^{aux} < \infty$. The average of the Lipschitz constants $C_n^{aux} = \frac{1}{n}\sum_{t=1}^{n} C^{aux}(y_t, \dots, y_1, x_t, \dots, x_1)$ is uniformly stochastically bounded, it is $O_p(1)$, for both the data and the simulated data.

[29] Note that using the same estimates $\widehat{\eta}_n^{aux}$ for filtering the data and the simulated samples avoids the complication of proving uniform convergence of the auxiliary parameters over the sieve space.

iv. Dependence: for all $\eta^{aux} \in E$, (y_t, x_t, z_t^{aux}) is uniformly geometric ergodic.

v. Moments: for all $\eta^{aux} \in E$, $\beta = \beta_0$ and $\beta = \Pi_{k(n)}\beta_0$, the moments $\mathbb{E}(\|z_t^{aux}\|^2)$ and $\mathbb{E}(\|z_t^{s,aux}\|^2)$ exist and are bounded.

vi. Summability: for any $(y_t, \ldots, y_1)$, $(\widetilde{y}_t, \ldots, \widetilde{y}_1)$, any $\eta^{aux} \in E$ and for all $t \geq 1$:

$$\|g_{t,aux}(y_t, \ldots, y_1, x_t, \ldots, x_1, \eta^{aux}) - z_{t,aux}(\widetilde{y}_t, \ldots, \widetilde{y}_1, x_t, \ldots, x_1, \eta^{aux})\| \leq \sum_{j=1}^{t} \rho_j \|y_j - \widetilde{y}_j\|$$

with $\rho_j \geq 0$ for all $j \geq 1$ and $\sum_{j=1}^{\infty} \rho_j < \infty$.

vii. Central Limit Theorem for the Sieve Score:

$$\sqrt{n} Real\left(\int \psi_\beta(\tau, u_n^*, \eta^{aux}) \overline{\left(\widehat{\psi}_n(\tau, \widehat{\eta}_n^{aux}) - \widehat{\psi}_n^s(\tau, \widehat{\eta}_n^{aux}, \beta_0)\right)} \pi(\tau) d\tau\right) \xrightarrow{d} \mathcal{N}(0,1)$$

The summability condition *iv.* is key in preserving the Hölder continuity and bias accumulation results of section 3.3 when using auxiliary variables in the CF. For auxiliary variables generated using the Kalman Filter or a GARCH model, this corresponds to a stability condition in the Kalman Filter or the GARCH volatility equations.

Conditions *ii.* and *iii.* ensure that $\widehat{\eta}_n^{aux}$ is well behaved and does not affect the rate of convergence. Condition *iv* implies that the inequality for the supremum of the empirical process still applies. Condition *vii.* assumes a CLT applies to the leading term in the expansion of $\phi(\widehat{\beta}_n) - \phi(\beta_0)$. It could be shown by assuming an expansion of the form $\widehat{\eta}_n^{aux} = \frac{1}{n}\sum_{t=1}^{n} \eta^{aux}(y_t, x_t) + o_p(1/\sqrt{n})$ and expanding $\widehat{\psi}_n$, $\widehat{\psi}_n^s$ around the probability limit η^{aux}. The following illustrates the Lipschitz and summability conditions for the SV with GARCH filtered volatility.

Example 1 (Continued) (Stochastic Volatility and GARCH(1,1) Filtered Volatility)**.** *For simplicity, assume there are only volatility dynamics:*

$$y_t = \sigma_t e_{t,1}$$

For simplicity, consider the absolute value GARCH(1,1) auxiliary model:[30]

$$y_t = \sigma_t^{aux} e_{t,1}, \quad \sigma_t^{aux} = \eta_1^{aux} + \eta_2^{aux}|y_t| + \eta_3^{aux}\sigma_{t-1}^{aux}.$$

[30]The process is also known as the AVGARCH or TS-GARCH (see e.g. Bollerslev, 2010) and is a special case of the family GARCH model (see e.g. Hentschel, 1995). The method of proof is slightly more involved for a standard GARCH model, requiring for instance a lower bound on the volatility σ_t^{aux} together with finite and bounded fourth moments for y_t, y_t^s to prove the Lipschitz condition.

The focus here is on the Lipschitz and summability conditions in the GARCH auxiliary model. First, to prove the Lipschitz condition, consider a sequence (y_t) *and two sets of parameters* $\eta^{aux}, \widetilde{\eta}^{aux}$*, by recursion:*

$$|\sigma_t^{aux} - \widetilde{\sigma}_t^{aux}| = |\eta_1^{aux} - \widetilde{\eta}_1^{aux} + (\eta_2^{aux} - \widetilde{\eta}_2^{aux})|y_t| + (\eta_3^{aux} - \widetilde{\eta}_3^{aux})\sigma_{t-1}^{aux} + \widetilde{\eta}_3^{aux}(\sigma_{t-1}^{aux} - \widetilde{\sigma}_{t-1}^{aux})|$$
$$\leq \|\eta^{aux} - \widetilde{\eta}^{aux}\| \times (\frac{1+\sigma_0^{aux}}{1-\overline{\eta}_3^{aux}} + [1+\overline{\eta}_2^{aux}][|y_t| + \cdots + (\overline{\eta}_3^{aux})^{t-1}|y_1|])$$

$\overline{\eta}^{aux}$ *are upper-bounds on the parameters. If* $\mathbb{E}(|y_t|^2))$ *and* $\mathbb{E}(|y_t^s|^2))$ *are finite and bounded and* $0 \leq \overline{\eta}_3^{aux} < 1$ *then the Lispchitz condition holds with:*

$$\overline{C}^{aux} \leq \frac{1+\overline{\eta}_2^{aux}}{1-\overline{\eta}_3^{aux}}\left(1 + \sigma_0^{aux} + M_y\right)$$

where $\mathbb{E}(|y_t|^2)$ *and* $\mathbb{E}(|y_t^s|^2) \leq M_y$*, for all* $t \geq 1$ *and* $\beta \in \mathcal{B}$*. Next, the proof for the summability is very similar, consider two time-series* $y_t, \widetilde{y}_t$ *and a set of auxiliary parameters* η^{aux}*:*

$$|\sigma_t^{aux} - \widetilde{\sigma}_t^{aux}| \leq \overline{\eta}_2|y_t - \widetilde{y}_t| + \overline{\eta}_3^{aux}|\sigma_{t-1}^{aux} - \widetilde{\sigma}_{t-1}^{aux}|.$$

By a recursive argument, the inequality above becomes:

$$|\sigma_t^{aux} - \widetilde{\sigma}_t^{aux}| \leq$$
$$\overline{\eta}_2|y_t - \widetilde{y}_t| + \overline{\eta}_3^{aux}\overline{\eta}_2|y_{t-1} - \widetilde{y}_{t-1}| + \cdots + (\overline{\eta}_3^{aux})^{t-1}\overline{\eta}_2|y_1 - \widetilde{y}_1| + (\overline{\eta}_3^{aux})^{t-1}|\sigma_0^{aux} - \widetilde{\sigma}_0^{aux}|.$$

Suppose that σ_0^{aux} *only depends on* η^{aux} *or is fixed, for instance equal to* 0*. Then the summability condition holds, if the upper-bound* $\overline{\eta}_3^{aux} < 1$*, with:*

$$\rho_j = \overline{\eta}_2^{aux}(\overline{\eta}_3^{aux})^j, \quad \sum_{j=0}^{\infty}\rho_j = \frac{\overline{\eta}_2^{aux}}{1-\overline{\eta}_3^{aux}} < \infty.$$

The Lipschitz and summability conditions thus hold for the auxiliary GARCH model.

The following corollary shows that the results of section 3.3 also hold when addition auxiliary variables to the CF.

Corollary 2 (Asymptotic Properties using Auxiliary Variables)**.** *Suppose the assumptions for Theorems 1, 2 and 3 hold as well as Assumption 6, then the results of Theorems 1, 2 and 3 hold with auxiliary variables. The rate of convergence is unchanged.*

The proof of Corollary 2 is very similar to the proofs of the main results. Rather than repeating the full proofs, Appendix 3.7 shows where the differences with and without the auxiliary variables are and explains why the main results are unchanged.

To compute standard errors, a block Bootstrap is applied to compute the variance term for the difference $\widehat{\psi}_n(\cdot, \widehat{\eta}_n^{aux}) - \widehat{\psi}_n^S(\cdot, \beta_0, \widehat{\eta}_n^{aux})$ in the sandwich formula for the standard errors. The unknown β_0 is replaced by $\widehat{\beta}_n$ in practice.

Using Short Panels

The main theorems 1, 2 and 3 allow for either iid data or time-series. However, SMM estimation is also common in panel data settings where the time dimension T is small relative to the cross-sectional dimension n. The following provides a simple application of these results.

Example 2 (Dynamic Tobit Model)**.** *y_t follows a dynamic Tobit model:*

$$y_{j,t} = (x_{j,t}'\boldsymbol{\theta}_1 + u_{j,t})\mathbb{1}_{x_{j,t}'\boldsymbol{\theta}_1 + u_{j,t} \geq 0}$$
$$u_{j,t} = \rho u_{j,t-1} + e_{j,t}$$

where $|\rho| < 1$, $e_{j,t} \overset{iid}{\sim} f$, $\mathbb{E}(e_{j,t}) = 0$. The parameters to be estimated are $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \rho)$ and f.

An overview of the dynamic Tobit model is given in Arellano & Honoré (2001). Applications of the dynamic Tobit model include labor participation studies such as Li & Zheng (2008); Chang (2011). Li & Zheng (2008) find that estimates of ρ can be biased downwards under misspecification. This estimate matters for evaluating the probability of (re)-entering the labor market in the next period for instance.

Quantities of interest in the dynamic Tobit model includes the probability or re-entering the labor market $\mathbb{P}(y_{t+1} > 0 | x_{t+1}, \ldots, x_t, y_t = 0, y_{t-1}, \ldots, y_1)$ which depends on both the parameters $\boldsymbol{\theta}$ and the distribution f. Marginal effects such as $\partial_{x_{t+1}} \mathbb{P}(y_{t+1} > 0 | x_{t+1}, \ldots, x_t, y_t = 0, y_{t-1}, \ldots, y_1)$ also depend on the true distribution f. As a result these quantities are sensitive to a particular choice of distribution f, this motivates a semi-nonparametric estimation approach for this model.

Other applications of simulation-based estimation in panel data settings include Gourinchas & Parker (2010) and Guvenen & Smith (2014) who consider the problem of consumption choices with income uncertainty. For the simulation-based estimates, shocks to the income process are typically assumed to be Gaussian. Guvenen et al. (2015) use a very large and confidential panel data set from the U.S. Social Security Administration covering 1978 to 2013 to find that individual income shocks are display large negative skewness and excess kurtosis: the data strongly rejects Gaussian shocks.[31] They find that non-Gaussian income shocks help explain transitions between low and higher earnings

[31] Also, Geweke & Keane (2000) estimate the distribution of individual income shocks using Bayesian estimates of a finite Gaussian mixture. They also find evidence of non-Gaussianity in the shocks. Arellano et al. (2017) use non-linear panel data methods to study the relation between incomes shocks and consumption. They provide evidence of persitence in earnings and conditional skewness.

states. Hence, a Sieve-SMM approach should also be of interest in the estimation of precautionary savings behavior under income uncertainty.

Because of the fixed T dimension, the initial condition (y_0, u_0) cannot be systematically handled using a large time dimension and geometric ergodicity argument as in the time-series case. Some additional restrictions on the DGP are given in the assumption below.

Assumption 7 (Data Generating Process for Panel Data)**.** *The data $(y_{j,t}, x_{j,t})$ with $j = 1, \ldots, n$, $t = 1, \ldots, T$ is generated by a DGP with only one source of dynamics either:*

$$\begin{aligned} y_{j,t} &= g_{obs}(x_{j,t}, \beta, u_{j,t}) \\ u_{j,t} &= g_{latent}(u_{j,t-1}, \beta, e_{j,t}) \end{aligned} \tag{3.12}$$

or

$$y_{j,t} = g_{obs}(y_{j,t-1}, x_{j,t}, \beta, e_{j,t}) \tag{3.13}$$

where $e_{j,t} \overset{iid}{\sim} f$ in both models. The observations are iid over the cross-sectional dimension j.

In situations where the DGPs in Assumption 7 are too restrictive, an alternative approach would be to estimate the distribution of $u_{j,1}$ conditional on $(y_{j,1}, x_{j,1})$. The methodology of Norets (2010) would apply to this particular estimation problem, the dimension of $(y_{j,1}, x_{j,1})$ should not be too large to avoid a curse of dimensionality. This is left to future research.

For the DGP in equation (3.12), geometric ergodicity applies to $u^s_{j,t}$ when simulating a longer history $u^s_{j,-m}, \ldots, u^s_{j,0}, \ldots, u^s_{j,1}, \ldots, u^s_{j,T}$ and letting the history increase with n, the cross-sectional dimension: $m/n \to c > 0$ as $n \to \infty$. For the DGP in equation (3.13), fixing $y^s_{j,1} = y_{j,1}$ ensures that $(y^s_{j,1}, \ldots, y^s_{j,T}, x_{j,1}, \ldots, x_{j,T})$ and $(y_{j,1}, \ldots, y_{j,T}, x_{j,1}, \ldots, x_{j,T})$ have the same distribution when $\beta = \beta_0$ (the DGP is assumed to be correctly specified).

The moments $\widehat{\psi}_n, \widehat{\psi}^s_n$ are the empirical CF of $(\mathbf{y}_t, \mathbf{x}_t)$ and $(\mathbf{y}^s_t, \mathbf{x}_t)$ respectively where $\mathbf{y}_t = (y_t, \ldots, y_{t-L})$ for $1 \leq L \leq T-1$; $\mathbf{y}_t, \mathbf{x}_t, \mathbf{y}^s_t$ are defined similarly. The identification Assumption 1 is assumed to hold for the choice of L.

The following lemma derives the initial condition bias for dynamic panel models with fixed T.

Lemma 8 (Impact of the Initial Condition)**.** *Suppose that Assumption 7 holds. If the DGP is given by (3.12) and $(y^s_{j,t}, u^s_{j,t})$ with a long history for the latent variable $(u_{j,T}, \ldots, u_{j,0}, \ldots, u_{j,-m})$*

where $m/n \to c > 0$ as $n \to \infty$. Suppose that $\mathbf{u}_{j,t}^s$ is geometrically ergodic in t and the integrals

$$\int\int f(\mathbf{y}_{j,t}^s, \mathbf{x}_{j,t}|\mathbf{u}_{j,t}^s)^2 f(\mathbf{u}_{j,t}^s) d\mathbf{y}_{j,t}^s d\mathbf{x}_{j,t}^s d\mathbf{u}_{j,t}^s, \quad \int\int f(\mathbf{y}_{j,t}^s, \mathbf{x}_{j,t}|\mathbf{u}_{j,t}^s)^2 f^*(\mathbf{u}_{j,t}^s) d\mathbf{y}_{j,t}^s d\mathbf{x}_{j,t}^s d\mathbf{u}_{j,t}^s$$

are finite and bounded when $\beta = \beta_0$. Then, there exists a constant $\overline{\rho}_u \in (0,1)$ such that:

$$Q_n(\beta_0) = \int \left| \mathbb{E}\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \beta_0)\right) \right|^2 \pi(\tau) d\tau = O\left(\overline{\rho}_u^m\right).$$

The effect of the initial condition is exponentially decreasing in m for DGP (3.12). If the DGP is given by (3.13) and the data is simulated with $y_{j,1}^s = y_{j,1}$ fixed then there is no initial condition effect:

$$Q_n(\beta_0) = \int \left| \mathbb{E}\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \beta_0)\right) \right|^2 \pi(\tau) d\tau = 0$$

Simulating a long history $u_{j,T}^s, \ldots, u_{j,-m}^s$ implies that the impact of the initial condition $u_{j,m}^s = u_{-m}$ on the full simulated sample $y_{j,1}^s, \ldots, y_{j,T}^s$ delines exponentially fast in m. If m does not grow faster than n, that is $m/n \to c > 0$, than the dynamic bias accumulation is the same as in the time-series setting. In terms of bias, these m simulations play a similar role as the burn-in draws in MCMC estimation.

Corollary 3 (Asymptotic Properties for Short Panels)**.** *Suppose that Assumption 7 and Lemma 8 hold. For the DGP (3.12) in Assumption 7, assume that m is such that $\log[n]/m \to 0$ as $n \to \infty$. Suppose the assumptions for Theorems 1, 2 and 3 hold, then the resuls of Theorems 1, 2 and 3 hold. The rate of convergence in weak norm is the same as for iid data:*

$$\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p\left(\max\left(\frac{\log[k(n)]^{r/b+1}}{k(n)^{\gamma^2 r}}, \sqrt{\frac{k(n)\log[k(n)]}{n}}\right)\right).$$

The rate of convergence in total variance and supremum distance are:

$$\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} = O_p\left(\frac{\log[k(n)]^{r/b}}{k(n)^r} + \tau_{\mathcal{B},n} \max\left(\frac{\log[k(n)]^{r/b+1}}{k(n)^{\gamma^2 r}}, \sqrt{\frac{k(n)\log[k(n)]}{n}}\right)\right).$$

Remark 3. *For the DGP (3.13), the simulated history is finite and fixed so that the approximation bias is not inflated by the dynamics:*

$$\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p\left(\max\left(\frac{\log[k(n)]^{r/b}}{k(n)^{\gamma^2 r}}, \sqrt{\frac{k(n)\log[k(n)]}{n}}\right)\right).$$

As a result, the rate of convergence is the same as for static models.

The assumption that $\log[n]/m \to 0$ can be weakened to $m \to \infty$ and $\lim_{n\to\infty} \log[n]/m < -\log[\overline{\rho}_u]$. Heuristically, the requirement is $m \gg \log[n]$, for instance when $n = 1,000$ this implies $m \gg 7$: a short burn-in sample for $u_{j,t}$ is sufficient to reduce the impact of the initial condition. The following verifies some of the conditions in Assumption 2 for the Dynamic Tobit model.

Example 2 (Continued) (Dynamic Tobit)**.** *Since the function $x \to x\mathbb{1}_{x\geq 0}$ is Lipschitz the conditions y(i),y(ii) and y(iii) are satisfied as long as $\|\boldsymbol{\theta}_1\|$ is bounded, $\mathbb{E}(\|x_t\|_2^2)$ is finite and $\mathbb{E}(u_t^2)$ is finite and bounded. The last variance is bounded if $|\rho| \leq \overline{\rho} < 1$ and $\mathbb{E}(e_t^2)$ is bounded above. The last condition is a restriction on the density f. Since $|\rho| \leq \overline{\rho} < 1$, condition u(i) is automatically satisfied. Together, $\mathbb{E}(u_t^2)$ bounded and linearity in ρ imply u(ii). Finally, linearity in e_t implies u(iii).*

3.5 Monte-Carlo Illustrations

This section illustrates the finite sample properties of the Sieve-SMM estimator. First, two very simple examples illustrate the estimator in the static and dynamic case against tractable estimators. Then, Monte-Carlo simulations are conducted for the stochastic volatility model Example 1 and Dynamic Tobit Example 2 for panel data.

For all Monte-Carlo simulations, the initial value for the mixture is a Gaussian density in the optimization routine. In most examples the Nelder & Mead (1965) algorithm in the NLopt package of Johnson (2014) was sufficient for optimization. In more difficult problems, such as the SV model with tail mixture components, the DIRECT global search algorithm of Jones et al. (1993) was applied to initialize the Nelder-Mead algorithm. The Monte-Carlo simulations were conducted using R[32] for all examples except for the AR(1) for which Matlab was used.

The Generalized Extreme Value (GEV) distribution is used in all Monte-Carlo examples. For the chosen parametrization, it displays negative skewness (-0.9) and excess kurtosis (3.9). It was also chosen because the approximation bias is larger for both kernel and mixture sieve estimates, and is thus more visible than alternative designs with smoother densities not reported here. This is useful when illustrating the increased bias due to the dynamics.

[32]Some routines such as the computation of the CF and the simulation of mixtures were written in C++ and imported into R using Rcpp - see e.g. Eddelbuettel & Fran (2011a,b) for an introduction to Rcpp - and RcppArmadillo (Eddelbuettel & Sanderson, 2016) for linear algebra routines.

The Student t-distribution is also considered in the stochastic volatility design to illustrate the Sieve-SMM estimates with tail components. The density is smooth compared to the GEV. As a result, the bias is smaller and less visible.

Basic Examples

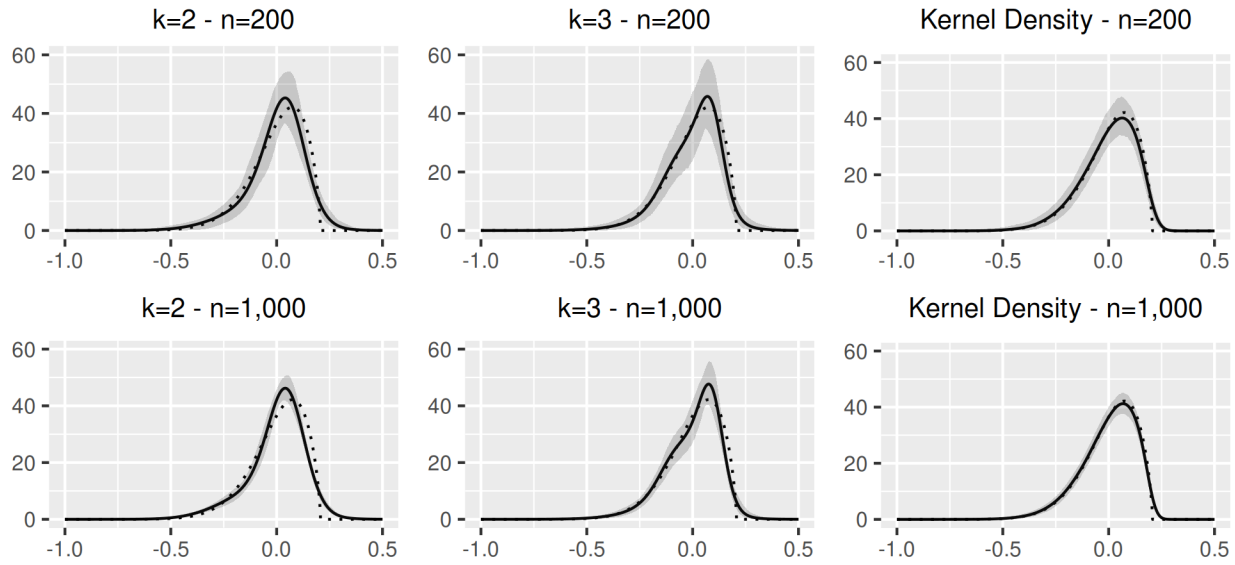
The following basic tractable examples are used as benchmarks to understand the basic properties of the Sieve-SMM estimator in terms of bias and dynamic bias accumulation as well as the impact of dependence on the variance. As a benchmark, the estimates are compared to feasible kernel density and OLS estimates.

A Static Model

To illustrate Remark 1, the first example uses the static DGP: $y_t = e_t \overset{iid}{\sim} f$, the only parameter to be estimated is f and kernel density estimation is feasible. The true distribution f is the Generalized Extreme Value (GEV) distribution. It is a 3 parameter distribution which allows for asymmetry and displays excess kurtosis.[33] In a recent application, Ruge-Murcia (2017) uses the GEV distribution to model the third moment in inflation and productivity shocks in a small asset pricing model. The Sieve-SMM estimates $\widehat{f_n}$ are compared to the feasible kernel density estimates $\widehat{f}_{n,kde}$.

[33]The GEV distribution was first introduced by McFadden (1978) to unify the Gumbel, Fréchet and Weibull families.

Figure 3.1: Static Model: Sieve-SMM vs. Kernel Density Estimates



Note: dotted line: true density, solid line: average estimate, bands: 95% pointwise interquantile range. Top panel $n = 200$ observation, bottom panel: $n = 1,000$ obervations. Left and middle: Sieve-SMM with $k = 2, 3$ Gaussian mixture components respectively and $S = 1$. Right: kernel density estimates.

Figure 3.1 plots the density estimates for $k = 2, 3$ with sample sizes $n = 200$ and $1,000$. The comparison between $k = 2$ and $k = 3$ illustrates the bias-variance trade-off: the bias is smaller for $k = 3$ but the variance of the estimates is larger compared to $k = 2$. Theorem 2 implies that when the sample size n increases, the number of mixture components k should increase as well to balance bias and variance. Here $k = 2$ appears to balance the bias and variance for $n = 200$ while $k \geq 3$ would be required for $n = 1,000$.

Autoregressive Dynamics

The second basic example considers an AR(1) model with an unknown distribution for the shocks:

$$y_t = \rho y_{t-1} + e_t,\ e_t \overset{iid}{\sim} (0, 1).$$

The shocks are drawn from a GEV density as in the previous example. The empirical CFs are computed using one lagged observation:

$$\widehat{\psi}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} e^{i\tau'(y_t, y_{t-1})}, \quad \widehat{\psi}^s_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} e^{i\tau'(y^s_t, y^s_{t-1})}.$$

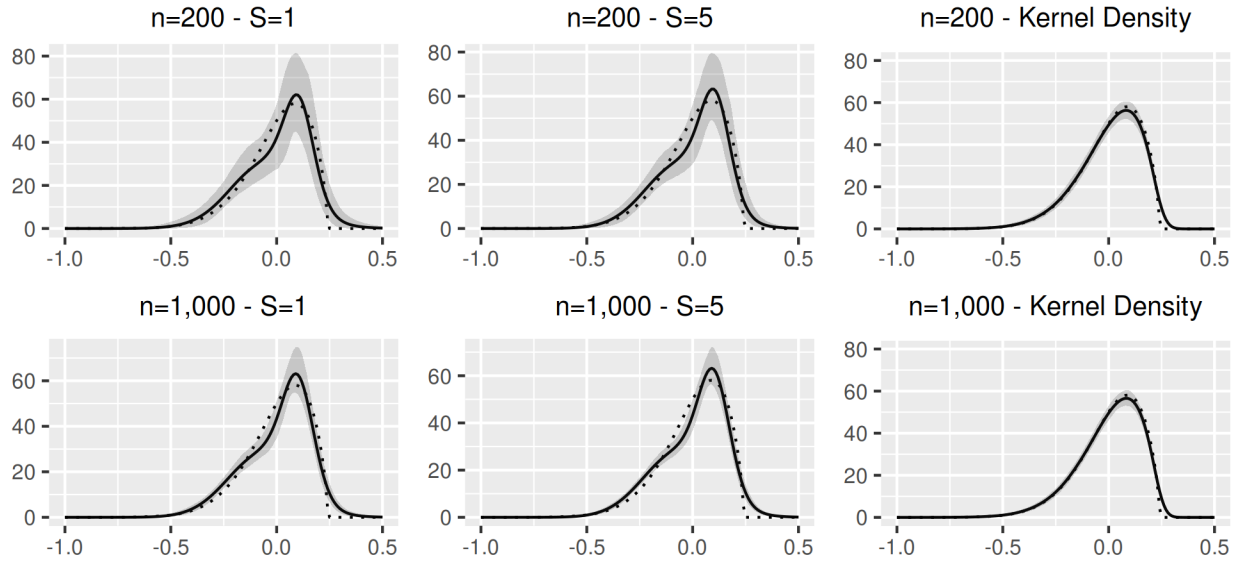
Knight & Yu (2002) note that additional lags do not improve the asymptotic properties of the estimator since y_t is Markovian of order 1.

This example illustrates Corollary 1 so the Monte-Carlo considers several choices of $S = 1, 5, 25$. Increasing S from 1 to 5 makes a notable difference on the variance of $\widehat{f}_n$. Further increasing S has a much smaller effect on the variance of the estimates. Table 3.1 compares the Sieve-SMM with OLS estimates for $\rho = 0.95$ for $n = 200$ and $n = 1,000$, $S = 1, 5, 25$. In all cases, $k = 2$ mixture components are used.

Table 3.1: Autoregressive Dynamics: Sieve-SMM vs. OLS Estimates

Parameter: ρ		Sieve-SMM			OLS	True
		$S = 1$	$S = 5$	$S = 25$		
$n = 200$	Mean Estimate	0.942	0.934	0.933	0.927	0.95
	$\sqrt{n}\times$ Std. Deviation	(0.54)	(0.45)	(0.44)	(0.46)	-
$n = 1,000$	Mean Estimate	0.949	0.947	0.947	0.946	0.95
	$\sqrt{n}\times$ Std. Deviation	(0.47)	(0.38)	(0.37)	(0.34)	-

Figure 3.2: Autoregressive Dynamics: Sieve-SMM vs. Kernel Density Estimates



Note: dotted line: true density, solid line: average estimate, bands: 95% pointwise interquantile range. Top panel: $n = 200$, bottom panel: $n = 1,000$. Left and middle: Sieve-SMM with $S = 1, 5$ repsectively and $k = 2$. Right: infeasible kernel density estimates.

Figure 3.2 compares the Sieve-SMM estimates with kernel density assuming the shocks e_t are observed - this is an infeasible estimator. The top panel shows results for $n = 200$ and the bottom panel illustrates the larger sample size $n = 1,000$.

There are several features to note. First, as discussed in section 3.3, the bias is more pronounced under AR(1) dynamics than in the static case. The variance is larger with AR(1) dynamics compared to the static model. Second, as shown in Corollary 1 the number of simulated samples S shifts the bias/variance trade-off so that $k(n)$ can be larger.

Example 1: Stochastic Volatility

The stochastic volatility model of Example 1, illustrates the properties of the Sieve-SMM estimator for an intractable, non-linear state-space model. As a simplification, there are no mean dynamics:

$$y_t = \sigma_t e_{t,1}, \quad \log(\sigma_t) = \mu_\sigma + \rho_\sigma \log(\sigma_{t-1}) + \kappa_\sigma e_{t,2}$$

where $e_{t,2} \overset{iid}{\sim} \mathcal{N}(0,1)$ and $e_{t,1} \overset{iid}{\sim} f$ with mean zero and unit variance. Using an extension of the main results, a GARCH(1,1) auxiliary model is introduced:

$$y_t^{aux} = \sigma_t^{aux} e_t^{aux}, \quad (\sigma_t^{aux})^2 = \mu^{aux} + \alpha_1^{aux}[e_{t-1}^{aux}]^2 + \alpha_2^{aux}(\sigma_{t-1}^{aux})^2.$$

Using the data y_t, the parameters $\widehat{\eta}_n^{aux} = (\mu_n^{aux}, \alpha_{1,n}^{aux}, \alpha_{2,n}^{aux})$ are estimated. The same $\widehat{\eta}_n^{aux}$ is used to compute both filtered volatilities $\widehat{\sigma}_t^{aux}, \widehat{\sigma}_t^{s,aux}$. The empirical CFs uses both y and $\widehat{\sigma}^{aux}$.[34]

$$\widehat{\psi}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} e^{\tau'(y_t, y_{t-1}, \widehat{\sigma}_t^{aux}, \log(\widehat{\sigma}_{t-1}^{aux}))}, \quad \widehat{\psi}_n^s(\tau,\beta) = \frac{1}{n}\sum_{t=1}^{n} e^{\tau'(y_t^s, y_{t-1}^s, \widehat{\sigma}_t^{s,aux}, \log(\widehat{\sigma}_{t-1}^{s,aux}))}.$$

The use of a GARCH model as an auxiliary model was suggested for indirect inference by Gouriéroux et al. (1993). Andersen et al. (1999) compare the EMM using ARCH, GARCH with the QML and GMM estimator using Monte-Carlo simulations. They find that EMM with GARCH(1,1) auxiliary model is more precise than GMM and QMLE in finite samples.

The parametrization is taken from Andersen et al. (1999): $\mu_\sigma = -0.736$, $\rho_\sigma = 0.90$, $\kappa_\sigma = 0.363$. Since Bayesian estimation is popular for SV models, the estimates are compared to a Gibbs sampling procedure, which assumes Gaussian shocks, using the R package *stochvol* of Kastner (2016). For Sieve-SMM estimation, the auxiliary GARCH filtered volatility estimates are computed using the R package *rugarch* of Ghalanos (2017).

[34] The simulation results are similar whether $\widehat{\sigma}^{aux}$ or $\log(\widehat{\sigma}^{aux})$ is used in the CF.

The Monte-Carlo consists of $1,000$ replications using $n = 1,000$ and $S = 2$. The distributions considered are the GEV and the Student t-distribution with 5 degrees of freedom. For the GEV density, $k = 4$ Gaussian mixture components are used and for the Student density, 4 Gaussian and 2 tail components are used.

Table 3.2: Stochastic Volatility: Sieve-SMM vs. Parametric Bayesian Estimates

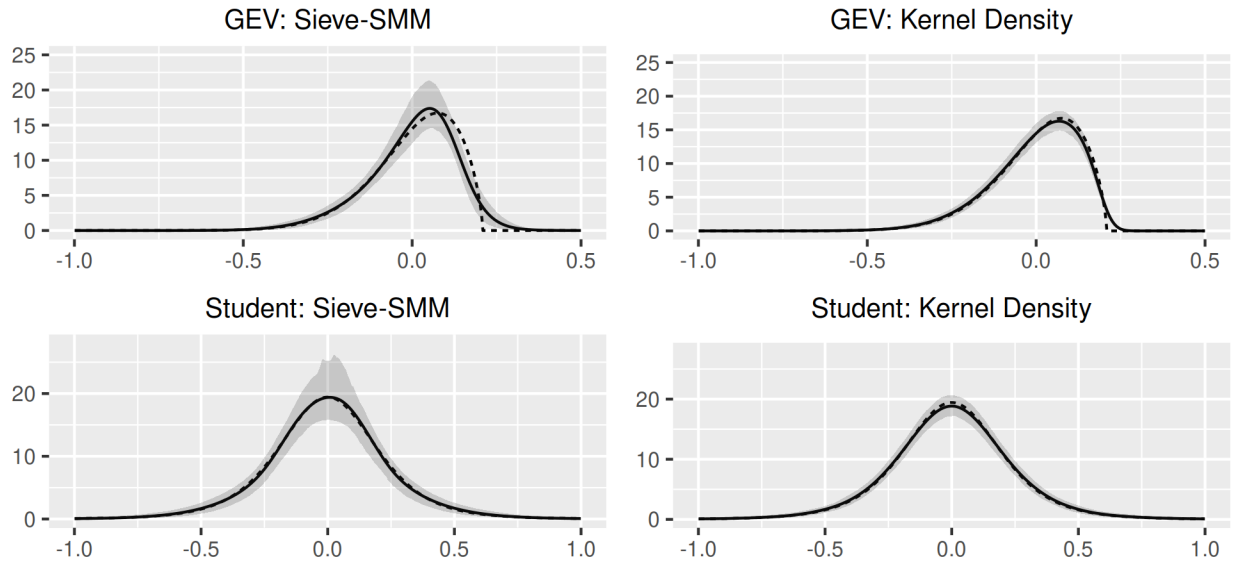
Parameter		True	GEV		Student	
			Sieve-SMM	Bayesian	Sieve-SMM	Bayesian
$\frac{\mu_\sigma}{1-\rho_\sigma}$	Mean Estimate	-7.36	-7.28	-7.37	-7.29	-7.63
	Std. Deviation	-	(0.16)	(0.13)	(0.15)	(0.13)
ρ_σ	Mean Estimate	0.90	0.90	0.88	0.92	0.71
	Std. Deviation	-	(0.03)	(0.04)	(0.08)	(0.10)
κ_σ	Mean Estimate	0.36	0.40	0.40	0.29	0.74
	Std. Deviation	-	(0.05)	(0.06)	(0.06)	(0.12)

The standard deviations are comparable to the EMM with GARCH(1,1) generator found in Andersen et al. (1999). Results based only on the CF of $\mathbf{y}_t = (y_t, \ldots, y_{t-2})$ (not reported here) were more comparable to the GMM estimates reported in Andersen et al. (1999) - both for SMM and Sieve-SMM. Applying some transformations such as $\log(y_t^2)$ provided somewhat better results but information about potential asymmetries in f is lost. This motivated the first extension of the main result in section 3.4 to allow for auxiliary variables. Also not reported here, the bias and standard deviations of parametric estimates with f_0 are comparable to the GEV results.

Table 3.2 shows that the parametric Bayesian estimates and the SMM estimator are well behaved when the true density is Gaussian. For the GEV distribution, both the Sieve-SMM and the misspecified parametric Bayesian estimates are well behaved. However, under heavier tails, the Student t-distribution implies a significant amount of bias for the misspecified Bayesian estimates. The Sieve-SMM estimates are only slightly biased compared with the Bayesian estimates.

Figure 3.3 compares the density estimates with the infeasible kernel density estimates based on $e_{t,1}$ directly. The top panel shows the results for the GEV density and the bottom panel for the Student t-distribution. The Sieve-SMM is less precise than the infeasible

Figure 3.3: Stochastic Volatility: Sieve-SMM vs. Kernel Density Estimates



Note: dotted line: true density, solid line: average estimate, bands: 95% pointwise interquantile range. Top panel: estimates of a GEV density, bottom panel: estimates of a Student t-distribution with 5 *degrees of freedom.*

estimator, as one would expect. As a comparison, the density is less precisely estimated than in the AR(1) case in figure 3.2. The two figures also illustrate bias reduction: the bias is larger for the AR(1) example which only uses $k = 2$ mixture components whereas the SV example uses $k = 4$.

The Monte-Carlo simulations for the stochastic volatility model highlight the lack of robustness of the parametric Bayesian estimates to the tail behavior of the shocks. This is particularly important for the second empirical application where Sieve-SMM and Bayesian estimates differ a lot and there is evidence of fat tails and asymmetry in the shocks.

Example 2: Dynamic Tobit Model

The dynamic Tobit model in Example 2 illustrates the properties of the estimator in a non-linear dynamic panel data setting:

$$y_{j,t} = (\boldsymbol{\theta}_1 + x'_{j,t}\boldsymbol{\theta}_2 + u_{j,t})\mathbb{1}_{\boldsymbol{\theta}_1 + x'_{j,t}\boldsymbol{\theta}_2 + u_{j,t} \geq 0}$$

$$u_{j,t} = \rho u_{j,t-1} + e_{j,t}$$

with $j = 1, \ldots, n$ and $t = 1, \ldots, T$. The Monte-Carlo simulations consider a sample with $n = 200$, $T = 5$ for a total of $1,000$ observations. The burn-in sample for the latent variable $u_{j,t}$, described in section 3.4, is $m = 10$ which is about twice the log of n. The regressors x_t follow an AR(1) with Gaussian shocks. The AR process is calibrated so that x has mean 2, autocorrelation 0.3 and variance 2. The other parameters are chosen to be: $(\rho, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (0.8, -1.25, 1)$ and f is the GEV distribution as in the other examples. As a result, about 40% of the sample is censored. The numbers of simulated samples are $S = 1$ and $S = 5$. The moments used in the simulations are:

$$\widehat{\psi}_n(\tau) = \frac{1}{nT}\sum_{t=2}^{T}\sum_{j=1}^{n} e^{i\tau'(y_t, y_{t-1}, x_t, x_{t-1})}, \; \widehat{\psi}_n^s(\tau) = \frac{1}{nT}\sum_{t=2}^{T}\sum_{j=1}^{n} e^{i\tau'(y_t^s, y_{t-1}^s, x_t, x_{t-1})}.$$

Table 3.3: Dynamic Tobit: SMM vs. Sieve-SMM Estimates

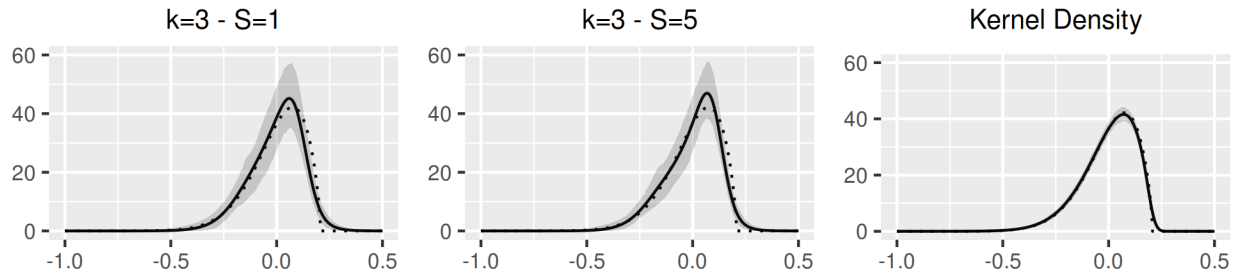
Parameter		$S = 1$		$S = 5$		
		SMM	Sieve-SMM	SMM	Sieve-SMM	True
ρ	Mean	0.796	0.801	0.796	0.796	0.80
	Std. Deviation	(0.042)	(0.039)	(0.031)	(0.031)	-
$\boldsymbol{\theta}_1$	Mean	-1.259	-1.230	-1.250	-1.233	-1.25
	Std. Deviation	(0.234)	(0.200)	(0.178)	(0.169)	-
$\boldsymbol{\theta}_2$	Mean	1.002	1.002	1.000	0.997	1.00
	Std. Deviation	(0.059)	(0.052)	(0.045)	(0.043)	-

Table 3.3 compares the parametric SMM and the Sieve-SMM estimates. The numbers are comparable except for $\boldsymbol{\theta}_1$ which has a small bias for the Sieve-SMM estimates. Additional results for misspecified SMM estimates with simulated samples use Gaussian shocks instead of the true GEV distribution also show bias for $\boldsymbol{\theta}_1$, the average estimate is higher than -1.1. The other estimates were found to have negligible bias.[35]

Figure 3.4 shows the Sieve-SMM estimates of the distribution of the shocks and the infeasible kernel density estimates of the unobserved e_t. Because of the censoring in the sample, note that the effective sample size for the Sieve-SMM estimates is smaller than

[35] Li & Zheng (2008) consider an alternative design where ρ displays more significant bias.

Figure 3.4: Dynamic Tobit: Sieve-SMM vs. Kernel Density Estimates

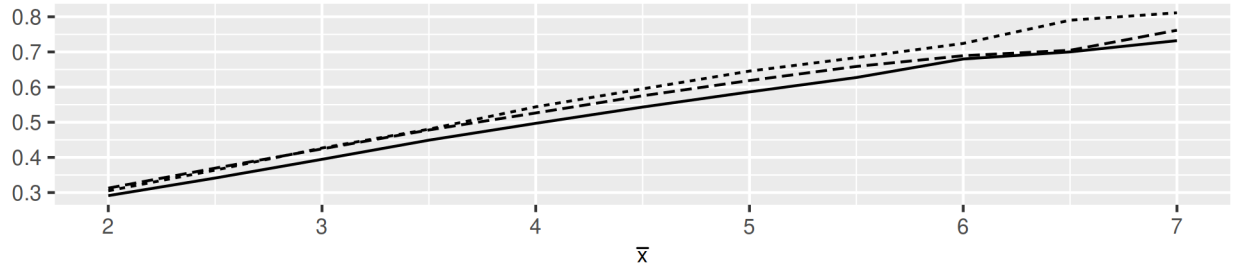


Note: dotted line: true density, solid line: average estimate, bands: 95% pointwise interquantile range.

for the kernel density estimates in this model. The left and middle plots show the sieve estimates when $S = 1, 5$; the right plot corresponds to the kernel density estimates.

Figure 3.5 illustrates the differences between SMM and Sieve-SMM for a counterfactual that involves the full density f. It shows the estimates of the probability of re-entering the market $\mathbb{P}(y_{j,5} > 0 | y_{j,4} = 0, x_5 = \cdots = x_1 = \overline{x})$ using the true value $(\boldsymbol{\theta}_0, f_0)$, the SMM estimates $\widehat{\boldsymbol{\theta}}_n^{SMM}$ with Gaussian shocks and the Sieve-SMM estimates $(\widehat{\boldsymbol{\theta}}_n, \widehat{f}_n)$. The true distribution is the GEV density which differs from the Gaussian density in the tails which implies a larger difference in the counterfactual when $\overline{x}$ is large, as shown in figure 3.5. For this particular counterfactual, the Sieve-SMM estimates are much closer to the true value for larger values of $\overline{x}$.

Figure 3.5: Dynamic Tobit: SMM vs. Sieve-SMM Estimates of the Counterfactual



Note: Estimated counterfactual: $\mathbb{P}(y_{j,5} > 0 | y_{j,4} = 0, x_5 = \cdots = x_1 = \overline{x})$ *- solid line: true probability, dashed line: Sieve-SMM estimate, dotted line: SMM estimate with Gaussian shocks, 1 Monte-Carlo estimate for SMM, Sieve-SMM, probabilities computed using* 10^6 *Simulated Samples.*

The Monte-Carlo simulations show the good finite sample behavior of the Sieve-SMM estimator with a non-smooth DGP. Indeed, the indicator function implies that the DGP

is Lipschitz but not continuously differentiable. It also illustrates the extension to short panels in section 3.4.

3.6 Empirical Applications

This section considers two empirical examples of the Sieve-SMM estimator. The first example illustrates the importance of non-Gaussian shocks for welfare analysis and asset pricing using US monthly output data. The shocks are found to display both asymmetry and tails after controlling for time-varying volatility. As a result, the Sieve-SMM estimates imply welfare costs that are 25% greater than with the Gaussian SMM estimates. Furthermore, the effect of uncertainty on risk-free is nearly 3 times as large for the Sieve-SMM estimates compared to the Gaussian SMM estimates. The second one uses daily GBP/USD exchange rate data and highlights the bias and sensitivity implications of fat tails on parametric SV volatility estimates.

Welfare and Asset Pricing Implications of Non-Gaussian Shocks

The first example considers a simplified form of the DGP for output in the Long-Run Risks (LRR) model of Bansal & Yaron (2004). The data consists of monthly growth rate of US industrial production (IP), as a proxy for monthly consumption, from January 1960 to March 2017 for a total of 690 observations, from the FRED[36] database and downloaded via the R package Quandl.[37] IP is modeled using a stochastic volatility model with AR(1) mean dynamics:

$$\begin{aligned}\Delta c_t &= \mu_c + \rho_c \Delta c_{t-1} + z_t e_{t,1} \\ \sigma_t^2 &= \mu_\sigma + \rho_\sigma \sigma_{t-1}^2 + \kappa_\sigma [e_{t,2} - 1]\end{aligned}$$

where $e_{t,2} \overset{iid}{\sim} \chi_1^2$ and $e_{t,1} \overset{iid}{\sim} f$ to be estimated assuming mean zero and unit variance. The stochastic volatility literature has mainly focused on the distribution of the shocks to the mean $e_{t,1}$ rather than the volatility[38] hence the volatility shocks are modelled parametrically in this application. Using the chi-squared distribution ensures that the volatility is

[36] `https://fred.stlouisfed.org/`.

[37] `https://www.quandl.com/tools/r`

[38] See Fridman & Harris (1998); Mahieu & Schotman (1998); Liesenfeld & Jung (2000); Jacquier et al. (2004); Comte (2004); Jensen & Maheu (2010); Chiu et al. (2017) for instance.

non-negative. This DGP is a simplification of the one considered in Bansal & Yaron (2004). They assume that consumption is the sum of an AR(1) process and iid shocks with a common SV component. The DGP above only estimates the AR(1) component for simplicity given that the focus is of this example is on the shocks and the volatility rather than the mean dynamics. The volatility shocks are also assumed to be χ^2_1 rather than Gaussian to ensure non-negativity.

Empirical Estimates

The model is estimated using a Gaussian mixture and is compared with parametric SMM estimates. $S = 10$ simulated samples are used to perform the estimation. As in the Monte-Carlo an auxiliary GARCH(1,1) model is used. The empirical CF uses 2 lagged observations:

$$\widehat{\psi}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} e^{i\tau'(\Delta c_t, \Delta c_{t-1}, \Delta c_{t-2}, \log(\widehat{\sigma}_t^{aux}), \log(\widehat{\sigma}_{t-1}^{aux}))}$$

$$\widehat{\psi}_n^s(\tau) = \frac{1}{n}\sum_{t=1}^{n} e^{i\tau'(\Delta c_t^s, \Delta c_{t-1}^s, \Delta c_{t-2}^s, \log(\widehat{\sigma}_t^{s,aux}), \log(\widehat{\sigma}_{t-1}^{s,aux}))}.$$

Table 3.4 shows the point estimates and the 95% confidence intervals for the parametric SMM, assuming Gaussian shocks, and the Sieve-SMM estimates using $k = 3$ mixture components. For reference, the OLS point estimate for ρ_c is 0.34 and the 95% confidence interval using HAC standard errors is $[0.23, 0.46]$ which is very similar to the SMM and Sieve-SMM estimates.[39]

Table 3.4: Industrial Production: Parametric and Sieve-SMM Estimates

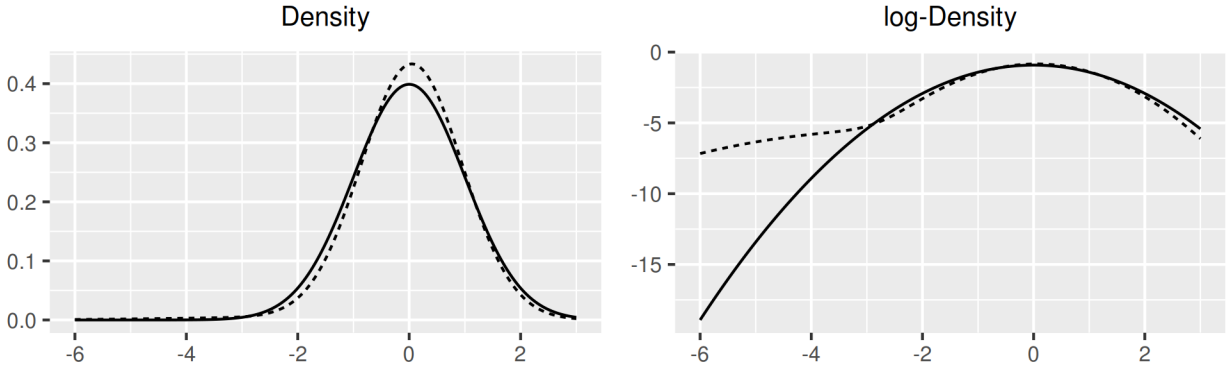
		ρ_c	μ_σ	ρ_σ	κ_σ
SMM	Estimate	0.33	0.39	0.65	0.15
	95% CI	[0.22, 0.43]	[0.34, 0.45]	[0.22, 0.86]	[0.08, 0.26]
Sieve-SMM	Estimate	0.32	0.43	0.75	0.13
	95% CI	[0.20, 0.42]	[0.34, 0.55]	[0.35, 0.92]	[0.06, 0.29]

Figure 3.6 compares the densities estimated using the parametric SMM and Sieve-SMM. The log-density reveals a larger left tail for the sieve estimates and potential asym-

[39] HAC standard errors are computed using the R package *sandwich* (Zeileis, 2004).

metry: conditional on the volatility regime, large negative shocks are more likely than the Gaussian SV estimates suggest. For instance, the log-difference at $e = -4$ is about 5 so that the ratio of densities is nearly 150 and the log-difference for $e = -5$ is roughly 10 so the density ratio is more than $20,000$.

Figure 3.6: Industrial Production: Sieve-SMM Density Estimate vs. Normal Density



Note: dotted line: Sieve-SMM density estimate, solid line: standard Normal density.

Table 3.5 shows that sieve estimated shocks have significant skewness and large kurtosis. It also shows the first four moments of the data compared to those implied by the estimates. Both sets of estimates match the first two moments similarly. The Sieve-SMM estimates provide a better fit for the skewness and kurtosis.

Table 3.5: Industrial Production: Moments of Δc_t, Δc_t^s and e_t^s

		Mean	Std Dev	Skewness	Kurtosis
Data	y_t	0.21	0.75	-0.92	7.56
SMM	y_t^s	0.25	0.66	0.06	4.39
Sieve-SMM	y_t^s	0.24	0.67	-0.35	6.65
SMM	e_t^s	0.00	1.00	0.00	3.00
Sieve-SMM	e_t^s	0.00	1.00	-0.75	7.74

Altogether, these results suggest significant non-Gaussian features in the shocks with both negative skewness and excess kurtosis. The welfare implications and the impact on the risk-free rate are now discussed.

Welfare Implications

The first implication considered here is the welfare effect of the fluctuations implied by each set of estimates. The approach considered here is based on the simple calculation approach of Lucas (1991, 2003).[40] The main advantage of this approach is that it does not require a full economic model: only a statistical model for output and a utility function are needed. To set the framework, a brief overview of his setting is now given. Lucas (1991) considers a setting where consumption is iid log-normal with constant growth rate $C_t = e^{\mu t + \sigma e_t}$ where $e_t \overset{iid}{\sim} \mathcal{N}(0,1)$ and has a certainty equivalent $C_t^\star = e^{\mu t + \sigma^2/2}$.

For a given level of risk-aversion $\gamma \geq 0$ and time preference $e^{-a} \in (0,1)$, he defines the welfare cost of business cycle fluctuations as the proportion λ by which the C_ts increase to achieve the same lifetime utility as under $C_t^\star$. This implies the following equation:

$$(1+\lambda)^{1-\gamma} \sum_{t \geq 0} e^{-at} \mathbb{E}_0 \left(\frac{C_t^{1-\gamma} - 1}{1-\gamma} \right) = \sum_{t \geq 0} e^{-at} \frac{C_t^{\star 1-\gamma} - 1}{1-\gamma}.$$

The estimates for the cost of business cycle fluctuations depends only on γ and σ in the Gaussian case: $\log(1+\lambda) = \gamma \frac{\sigma^2}{2}$. Lucas estimates this cost to be very small in the US.

Combining the SMM and Sieve-SMM with Monte-Carlo simulations[41], the welfare cost of business cycle fluctuations is now computed under Gaussian and mixture SV dynamics. Table 3.6 compares the two welfare costs for different levels of risk aversion with the baseline iid Gaussian case of Lucas.[42] For the full range of risk aversion considered here the welfare cost is estimated to be above 1% of monthly consumption. As a comparison Lucas (1991) estimates the welfare cost to be very small, a fraction of a percent, while Krusell et al. (2009) estimates it to be around 1%.[43] Both SV models imply much larger costs for business cycle fluctuations compared to the iid results: for $\gamma = 4$ and an annual income of \$55,000 the estimated welfare cost is \$990, \$800 and \$7 for Sieve-SMM, SMM and Gaussian iid estimates respectively. The Sieve-SMM estimates imply a welfare

[40] A number of alternative methods to estimate the welfare effect of business cycle fluctuations exist in the literature using, to cite only a few, models with heterogeneous agents (Krusell & Smith, Jr., 1999; Krusell et al., 2009), asset pricing models (Alvarez & Jermann, 2004; Barro, 2006a) and RBC models (Cho et al., 2015).

[41] Expectations are taken over 1,000 Monte-Carlo samples for an horizon of 5,000 months or about 420 years.

[42] The iid case is calibrated to match the mean and standard deviation of monthly IP growth. The monthly time preference parameter is chosen to match a quarterly rate of 0.99.

[43] Additional calculations and results under an AR(1) process and using linearized DSGE models are also given in Reis (2009).

Table 3.6: Welfare Cost of Business Cycle Fluctuations λ (%)

Risk Aversion γ	2	4	6	10
Gaussian iid	0.01	0.01	0.02	0.03
SMM	1.32	1.46	1.53	1.65
Sieve-SMM	1.54	1.80	1.93	2.12

cost that is nearly \$200, or 25%, higher than the parametric SMM welfare estimates. This difference is quite large highlighting the non-negligible role of asymmetry in welfare.

Implications for the risk-free rate

The second implication considers the effect of uncertainty on the risk-free rate. As discussed in the introduction, the Euler equation implies that the risk-free rate r_t satisfies: $e^{-r_t} = e^{-a}\mathbb{E}_t\left(\left(C_{t+1}/C_t\right)^{-\gamma}\right)$ where e^{-a} and γ are the time preference and risk aversion parameters. To explain the low-level of the risk-free rate observed in the data (Weil, 1989) a number of resolutions have been proposed including the long-run risks model of Bansal & Yaron (2004), which involves stochastic volatility and a recursive utility, and the rare disasters literature which involves very low frequency, high impact shocks and a power utility (Rietz, 1988; Barro, 2006b). This empirical application considers a simple power utility together with the higher frequency of shocks (monthly) over a recent period (since 1960) to achieve a similar result.

Given the AR(1) mean dynamics and volatility process postulated for IP growth, the risk-free rate can be written as:

$$r_t = a + \underbrace{\gamma\mu_c + \gamma\rho_c\Delta c_t}_{\text{Predictable Component}} \underbrace{-\log\left(\int e^{-\gamma e_{t+1,1}\sqrt{\mu_\sigma+\rho_\sigma\sigma_t^2+\kappa_\sigma[e_{t+1,2}-1]}} f(e_{t+1,1}) f_{\chi_1^2}(e_{t+1,2}) de_{t+1,1} de_{t+1,2}\right)}_{\text{Effect of uncertainty}}$$

where $f_{\chi_1^2}$ is the density of a χ_1^2 distribution.

Other than time preference a, there are two components in the risk-free rate: a predictable component $\gamma\mu_c + \gamma\rho_c\Delta c_t$ and another factor which only depends on the distribution of the shocks, it is the effect of uncertainty. In the second term, the integral over $e_{t+1,1}$ is the moment generating function of $e_{t+1,1}$ evaluated at $-\gamma\sqrt{\mu_\sigma + \rho_\sigma\sigma_t^2 + \kappa_\sigma[e_{t+1,2}-1]}$

and has closed-form when the distribution is either a Gaussian or a Gaussian mixture:

$$\int e^{-\gamma e_{t+1,1}\sqrt{\mu_\sigma+\rho_\sigma\sigma_t^2+\kappa_\sigma[e_{t+1,2}-1]}} f(e_{t+1,1}) f_{\chi_1^2}(e_{t+1,2}) de_{t+1,1} de_{t+1,2}$$
$$= \sum_{j=1}^{k} \omega_j \int e^{-\gamma\mu_j\sqrt{\mu_\sigma+\rho_\sigma\sigma_t^2+\kappa_\sigma[e_{t+1,2}-1]}+\frac{\gamma^2}{2}\sigma_j^2(\mu_\sigma+\rho_\sigma\sigma_t^2+\kappa_\sigma[e_{t+1,2}-1])} f_{\chi_1^2}(e_{t+1,2}) de_{t+1,2}.$$

The integral over $e_{t+1,2}$ is computed using Gaussian quadrature. Using this formula, table 3.7 computes the effect of uncertainty on the risk-free rate over a range of values for risk aversion γ for a Gaussian AR(1) model as well as the parametric SMM and Sieve-SMM SV estimates. The effect of uncertainty is estimated to be nearly 3 times as large under the Sieve-SMM estimates compared to the Gaussian SMM estimates. Given that the risk free-

Table 3.7: Effect of uncertainty on the risk-free rate (% annualized)

Risk aversion γ	2	4	6	10
Gaussian AR(1)	-0.12	-0.24	-0.35	-0.59
SMM	-0.09	-0.37	-0.84	-2.34
Sieve-SMM	-0.25	-1.02	-2.32	-6.59

rate is predicted to be much lower with the Sieve-SMM estimates, the results suggest that the non-Gaussian features in the shocks matter for precautionary savings. Altogether, the results suggest that the choice of distribution f matters in computing both welfare effects and the risk-free rate.

GBP/USD Exchange Rate Data

The second example highlights the effect of fat tails and outliers on SV estimates for GBP/USD exchange rate data. The results highlight the presence of heavy tails even after controlling for time-varying volatility. Similar findings were also documented with parametric methods (see e.g. Fridman & Harris, 1998; Liesenfeld & Jung, 2000). This paper also finds significant asymmetry in the distribution of the shocks. Furthermore, comparing the estimates with common Bayesian estimates shows that parametric estimates severely underestimate the persistence of the volatility. Mahieu & Schotman (1998) also consider a mixture approximation for the distribution of the shocks in a SV model, using quasi-MLE for weekly exchange rate data. However, they do not provide asymptotic

theory for their estimator and quasi-MLE does not estimate asymmetries in the density which turns out to be significant in this setting.

The data consists of a long series of daily exchange rate data between the British Pound and the US Dollar (GBP/USD) downloaded using the R package Quandl. The data begins in January 2000 and ends in December 2016 for a total of $5,447$ observations. The exchange rate is modeled using a log-normal stochastic volatility model with no mean dynamics:

$$y_t = \mu_y + \sigma_t e_{t,1}, \quad \log(\sigma_t) = \rho_\sigma \log(\sigma_{t-1}) + \kappa_\sigma e_{t,2}$$

where $e_{t,2} \overset{iid}{\sim} \mathcal{N}(0,1)$ and $e_{t,1} \overset{iid}{\sim} f$ to be estimated assuming mean zero and unrestricted variance. This allows to model extreme events associated with volatility clustering, when σ_t is large, as well as more isolated extreme events, represented by the tails of f. For this empirical application, μ_σ is set to 0 and f is only constrained to have unit variance. This illustrates the type of flexibility allowed when using mixtures for estimation. The data y_t consists of the daily log-growth rate of the GBP/USD exchange rate:

$$y_t = 100 \times \log\left(\frac{GBP/USD_t}{GBP/USD_{t-1}}\right).$$

Sieve-SMM estimates are compared to a common Gibbs sampling Bayesian estimate using the R package *stochvol* (Kastner, 2016). Two sets of Sieve-SMM estimates are computed: the first uses a Gaussian mixture with $k = 5$ components and the second a Gaussian and tails mixture with $k = 5$ components: 3 Gaussians and 2 tails. The two Sieve-SMM estimators have the same number of parameters to be estimated.

Table 3.8 shows the posterior mean and 95% credible interval for the Bayesian estimates as well as the point estimates and te 95% confidence interval for two Sieve-SMM estimators. The Bayesian estimate for the persistence of volatility ρ_z is much smaller than the SMM and Sieve-SMM estimates: it is outside their 95% confidence intervals. This reflects the bias issues discussed in the Monte-Carlo when f has large tails. As a robustness check, the estimates for the Sieve-SMM are similar when removing observations after the United Kingdom European Union membership referendum, that is between June 23rd and December 31st 2016: $(\widehat{\rho}_n, \widehat{\sigma}_z) = (0.96, 0.23)$ for the Gaussian mixture and $(0.97, 0.20)$ for the Gaussian and tails mixture. The Bayesian estimates are also of the same order of magniture $(0.26, 1.27)$. The density estimates $\widehat{f}_n$ are also very similar when removing these observations.

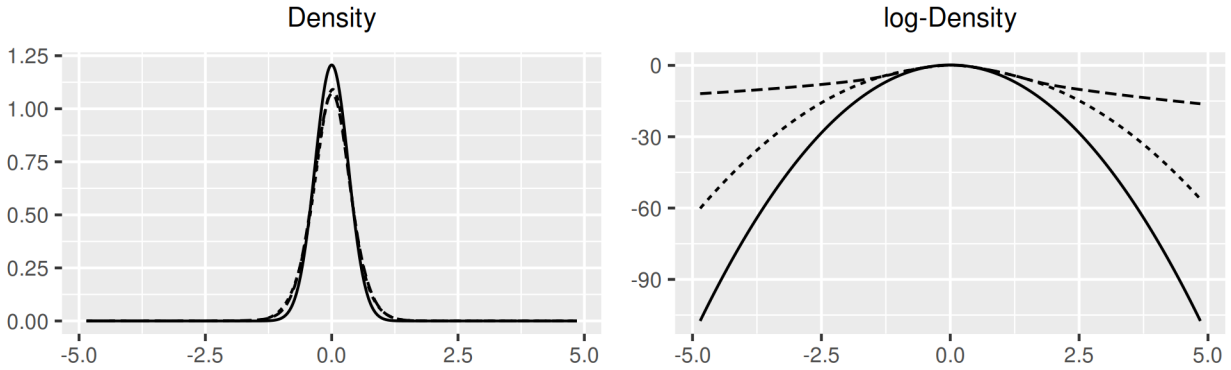
Figure 3.7 compares the density $\widehat{f}_n$ of $e_{t,1}$ for the Bayesian and Sieve-SMM estimates. The log-density $\log[\widehat{f}_n]$ is also computed as it higlights the differences in the tails. The

Table 3.8: Exchange Rate: Bayesian and Sieve-SMM Estimates

		ρ_z	σ_z
Bayesian	Estimate	0.24	1.31
	95% CI	[0.16, 0.34]	[1.21, 1.41]
Sieve-SMM	Estimate	0.96	0.22
	95% CI	[0.59, 0.99]	[0.06, 0.83]
Sieve-SMM Tails	Estimate	0.97	0.19
	95% CI	[0.62, 0.99]	[0.05, 0.79]

Note: CI is the credible interval for the Bayesian and the confidence interval for the frequentist estimates.

Figure 3.7: Exchange Rate: Density and log-Density Estimates



Note: solid line: Gaussian density, dotted line: Gaussian mixture, dashed: Gaussian and tails mixture.

Bayesian assumes Gaussian shocks, so the log-density is quadratic, the density declines faster in the tails compared to the other two estimates. For the mixture with tail components, the density decays much slower than for both the Bayesian and Gaussian mixture estimates.

Table 3.9 compares the first four moments in the data to those implied by the estimates.[44] The Bayesian estimates fit the fourth moment of the full dataset best. Note that for time series data, estimates of kurtosis can be very unprecise (Bai & Ng, 2005). Hence a robustness check can be important: when removing the observation corresponding to

[44]The moments for the Bayesian and Sieve-SMM estimates are computed using numerical simulations.

United Kingdom European Union membership referendum on June 23rd 2016 which is the largest variation in the sample,[45] the kurtosis drops to about 10. Furthermore, when removing all observations between June 23rd and December 31st 2016, the kurtosis declines further to about 9. As discussed above, the point estimates remain similar when removing these observations. The Sieve-SMM estimates match the fourth moment of the restricted sample more closely but the Gaussian mixture fits the third moment poorly. The Gaussian and tails mixture fits all four moments of the restricted sample best. It also has the lowest value for the sample objective function. The Gaussian and tails mixture is thus the preferred specifications for this dataset.

Table 3.9: Exchange Rate: Moments of y_t, y_t^s and e_t^s

		Mean	Std Dev	Skewness	Kurtosis
Data	y_t	0.00	0.49	-1.15	21.05
Data*	y_t	0.00	0.47	-0.32	8.92
Bayesian	y_t^s	0.00	0.52	0.00	18.47
Sieve-SMM	y_t^s	0.00	0.85	0.10	5.88
Sieve-SMM tails	y_t^s	0.00	0.45	-0.28	7.74
Bayesian	e_t^s	0.00	1.00	0.00	3.00
Sieve-SMM	e_t^s	0.00	1.00	-0.06	3.68
Sieve-SMM tails	e_t^s	0.00	1.00	-0.17	4.83

Note: Data corresponds to the full sample: January 1st 2000-December 31st 2016. Data is a restricted sample: January 1st 2000-June 22nd 2016. Sieve-SMM: Gaussian mixture, Sieve-SMM tails: mixture with tail components.*

In terms of forecasting, there are three main implications. First, the Bayesian estimates severely underestimate the persistence of the volatility: as a result, forecasts would underestimate the persistence of a high volatility episode. Second, $\widehat{f_n}$ displays a significant amount of tails: a non-negligible amount of large shocks are isolated rather than associated with high volatility regimes. Third, there is evidence of asymmetry in $\widehat{f_n}$: large depreciations in the GBP relative to the USD are historically more likely than large appreciations.

[45] It is associated with a depreciation of the the GBP of more than 8 log percentage points. This is much larger than typical daily fluctuations.

3.7 Conclusion

Simulation-based estimation is a powerful approach to estimate intractable models. This paper extends the existing parametric literature to a semi-nonparametric setting using a Sieve-SMM estimator. General asymptotic results are given using the mixture sieve for the distribution of the shocks and the empirical characteristic function as a moment function. On the theoretical side, this paper provides new and more general results for static models and allows for a new class of dynamics in the Sieve-GMM literature. Monte-Carlo simulations illustrate the range of applications of the method and its finite sample properties. Extensions to a larger class of moments and short panels are given.

Two empirical applications highlight the importance of the density in the shocks in practice. The first one shows asymmetry and tail behavior in output shocks. Welfare estimates suggest that the cost of business cycle fluctuations are larger under these non-Gaussian shocks. The risk-free rate is also significantly lower, reflecting the greater downside risks in the estimated distribution and the additional precautionary savings it implies.

The second empirical example highlights the effect of misspecification on volatility estimates. Sieve-SMM estimation applied to daily GBP/USD exchange rate data reveals significant tail behavior and asymmetry, even after controlling for the time-varying volatility. The parametric Bayesian estimates are not robust to misspecification and large rare events.

Going forward, a number of extensions to this paper's results should be of interest. On the theoretical side, extending the inequality in this paper to unbounded moments would allow for more general Sieve-GMM settings as in Chen et al. (2013). The results could also be extended to a generalization of Indirect Inference with both infinite dimensional moments and parameters. The mixture sieve can be extended to accommodate heteroskedasticiy as in Norets (2010) or multivariate densities without the independence assumption as in De Jonge & Van Zanten (2010). On the empirical side, the results in this paper suggest that the distribution of the shocks is important in estimating welfare effects in DSGE models or risk-premia in asset pricing models. Also, using the results in this paper, the Sieve-SMM can be applied to estimate cross-sectional heterogeneity in short panels where fixed effects cannot be differenced out.

Bibliography

Ai, C. & Chen, X. (2003). Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica*, 71(6), 1795–1843.

Altonji, J., Smith, A., & Vidangos, I. (2013). Modeling Earnings Dynamics. *Econometrica*, 81(4), 1395–1454.

Alvarez, F. & Jermann, U. J. (2004). Using Asset Prices to Measure the Cost of Business Cycles. *Journal of Political Economy*, 112(6), 1223–1256.

Andersen, T. G., Chung, H.-J., & Sorensen, B. E. (1999). Efficient method of moments estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Econometrics*, 91(1), 61–87.

Andrews, D. W. (1993). An Introduction to Econometric Applications of Empirical Process Theory for Dependent Random Variables. *Econometric Reviews*, 12(2), 183–216.

Andrews, D. W. (1994). Chapter 37 Empirical process methods in econometrics. In *Handbook of Econometrics*, volume 4 (pp. 2247–2294).

Andrews, D. W. K. & Pollard, D. (1994). An Introduction to Functional Central Limit Theorems for Dependent Stochastic Processes. *International Statistical Review / Revue Internationale de Statistique*, 62(1), 119.

Arellano, M., Blundell, R., & Bonhomme, S. (2017). Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework. *Econometrica*, 85(3), 693–734.

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

Arellano, M. & Honoré, B. (2001). Chapter 53 Panel data models: some recent developments. In *Handbook of Econometrics*, volume 5 (pp. 3229–3296).

Bai, J. & Ng, S. (2005). Tests for Skewness, Kurtosis, and Normality for Time Series Data. *Journal of Business & Economic Statistics*, 23(1), 49–60.

Bansal, R. & Yaron, A. (2004). Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. *The Journal of Finance*, 59(4), 1481–1509.

Bao, Y. & Ullah, A. (2007). The second-order bias and mean squared error of estimators in time-series models. *Journal of Econometrics*, 140(2), 650–669.

Barro, R. J. (2006a). *On the Welfare Costs of Consumption Uncertainty*. Working Paper 12763, National Bureau of Economic Research.

Barro, R. J. (2006b). Rare disasters and asset markets in the twentieth century. *Quarterly Journal of Economics*, 121(3), 823–866.

Beaumont, M., Corneut, J., Marin, J., & Robert, C. (2009). Adaptive Approximate Bayesian Computation. *Biometrika*, 96(4), 983–990.

Ben Hariz, S. (2005). Uniform CLT for empirical process. *Stochastic Processes and their Applications*, 115(2), 339–358.

Ben-Israel, A. (1999). The Change of Variables Forumla Using Matrix Volume. *SIAM Journal of Matrix Analysis*, 21, 300–312.

Ben-Israel, A. (2001). An Application of the Matrix Volume in Probability. *Linear Algebra and Applications*, 312, 9–25.

Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. (2017). Inference in generative models using the Wasserstein distance. *arxiv:1701.05146*.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile Prices in Market Equilibrium. *Econometrica*, 63(4), 841.

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

Bierens, H. J. (1990). A Consistent Conditional Moment Test of Functional Form. *Econometrica*, 58(6), 1443.

Bierens, H. J. & Song, H. (2012). Semi-nonparametric estimation of independently and identically repeated first-price auctions via an integrated simulated moments method. *Journal of Econometrics*, 168(1), 108–119.

Bierens, H. J. & Song, H. (2017). Semi-Nonparametric Modeling and Estimation of First-Price Auctions Models with Auction-Specific Heterogeneity.

Blasques, F. (2011). Semi-Nonparametric Indirect Inference. *Unpublished Manuscript*.

Blum, M. (2010). Approximate Bayesian Computation: A Nonparametric Perspective. *Journal of the American Statistical Association*, 105(491), 1178–1187.

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

Blundell, R., Chen, X., & Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant engel curves. *Econometrica*, 75(6), 1613–1669.

Blundell, R., Costa Dias, M., Meghir, C., & Shaw, J. (2016). Female Labor Supply, Human Capital, and Welfare Reform. *Econometrica*, 84(5), 1705–1753.

Bollerslev, T. (2010). Glossary to ARCH (GARCH). In *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle* (pp. 137–163). Oxford University Press.

Bonassi, F. & West, M. (2015). Sequential Monte Carlo With Adaptive Weights for Approximate Bayesian Computation. *Bayesian Analysis*, 10(1), 171–187.

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

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

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

Carrasco, M., Chernov, M., Florens, J.-P., & Ghysels, E. (2007a). Efficient estimation of general dynamic models with a continuum of moment conditions. *Journal of Econometrics*, 140(2), 529–573.

Carrasco, M. & Florens, J.-P. (2000). Generalization of GMM to a Continuum of Moment Conditions. *Econometric Theory*, 16(6), 797–834.

Carrasco, M., Florens, J.-P., & Renault, E. (2007b). Chapter 77 Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization. In *Handbook of Econometrics* (pp. 5633–5751).

Carrasco, M. & Kotchoni, R. (2016). Efficient Estimation Using the Characteristic Function. *Econometric Theory*, 33(02), 479–526.

Chang, S.-K. (2011). Simulation estimation of two-tiered dynamic panel Tobit models with an application to the labor supply of married women. *Journal of Applied Econometrics*, 26(5), 854–871.

Chen, X. (2007). Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models. In *Handbook of Econometrics*, volume 6 (pp. 5549–5632).

Chen, X. (2011). Penalized Sieve Estimation and Inference of Seminonparametric Dynamic Models: A Selective Review. In D. Acemoglu, M. Arellano, & E. Dekel (Eds.), *Advances in Economics and Econometrics* (pp. 485–544). Cambridge: Cambridge University Press.

Chen, X., Chernozhukov, V., Lee, S., & Newey, W. K. (2014a). Local Identification of Nonparametric and Semiparametric Models. *Econometrica*, 82(2), 785–809.

Chen, X. & Christensen, T. M. (2017). Optimal Sup-norm Rates and Uniform Inference on Nonlinear Functionals of Nonparametric IV Regression. *Forthcoming in Quantitative Economics*.

Chen, X., Favilukis, J., & Ludvigson, S. C. (2013). An estimation of economic models with recursive preferences. *Quantitative Economics*, 4(1), 39–83.

Chen, X. & Liao, Z. (2015). Sieve semiparametric two-step GMM under weak dependence. *Journal of Econometrics*, 189(1), 163–186.

Chen, X., Linton, O., & Van Keilegom, I. (2003). Estimation of Semiparametric Models when the Criterion Function Is Not Smooth. *Econometrica*, 71(5), 1591–1608.

Chen, X. & Ludvigson, S. C. (2009). Land of addicts? An empirical investigation of habit-based asset pricing models. *Journal of Applied Econometrics*, 24(7), 1057–1093.

Chen, X., Ponomareva, M., & Tamer, E. (2014b). Likelihood inference in some finite mixture models. *Journal of Econometrics*, 182(1), 87–99.

Chen, X. & Pouzo, D. (2012). Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals. *Econometrica*, 80(1), 277–321.

Chen, X. & Pouzo, D. (2015). Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models. *Econometrica*, 83(3), 1013–1079.

Chen, X. & Shen, X. (1998). Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica*, 66(2), 289.

Chernozhukov, V. & Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, 73(1), 245–261.

Chernozhukov, V. & Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2), 293–346.

Chernozhukov, V., Imbens, G. W., & Newey, W. K. (2007). Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1), 4–14.

Chiu, C.-W. J., Mumtaz, H., & Pintér, G. (2017). Forecasting with VAR models: Fat tails and stochastic volatility. *International Journal of Forecasting*.

Cho, J.-O., Cooley, T. F., & Kim, H. S. E. (2015). Business cycle uncertainty and economic welfare. *Review of Economic Dynamics*, 18(2), 185–200.

Christensen, T. M. (2017). Nonparametric Stochastic Discount Factor Decomposition. *Forthcoming in Econometrica*.

Comte, F. (2004). Kernel deconvolution of stochastic volatility models. *Journal of Time Series Analysis*, 25(4), 563–582.

Coppejans, M. (2001). Estimation of the binary response model using a mixture of distributions estimator (MOD). *Journal of Econometrics*, 102(2), 231–269.

Creel, M., Gao, J., Hong, H., & Kristensen, D. (2016). Bayesian Indirect Inference and the ABC of GMM. *Unpublished Manuscript*.

Creel, M. & Kristensen, D. (2013). Indirect Likelihood Inference. *Mimeo, UCL*.

Creel, M. & Kristensen, D. (2015). On Selection of Statistics for Approximate Bayesian Computing or the Method of Simulated Moments. *Computational Statistics and Data Analysis*.

Darolles, S., Fan, Y., Florens, J. P., & Renault, E. (2011). Nonparametric Instrumental Regression. *Econometrica*, 79(5), 1541–1565.

Davydov, Y. A. (1968). Convergence of Distributions Generated by Stationary Stochastic Processes. *Theory of Probability & Its Applications*, 13(4), 691–696.

de Jong, R. M. (1997). Central Limit Theorems for Dependent Heterogeneous Random Variables. *Econometric Theory*, 13(03), 353.

De Jonge, R. & Van Zanten, J. H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Annals of Statistics*, 38(6), 3300–3320.

Dean, T. A., Singh, S. S., Jasra, A., & Peters, G. W. (2011). Parameter Estimation for Hidden Markov Models with Intractable Likelihoods. *arXiv:1103.5399*.

Deaton, A. (1991). Saving and Liquidity Constraints. *Econometrica*, 59(5), 1221–1248.

Deaton, A. & Laroque, G. (1992). On the behaviour of commodity prices. *Review of Economic Studies*, 59(1), 1–23.

Dedecker, J. & Louhichi, S. (2002). Maximal Inequalities and Empirical Central Limit Theorems. In *Empirical Process Techniques for Dependent Data* (pp. 137–159). Boston, MA: Birkhäuser Boston.

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

Doukhan, P., Massart, P., & Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Annales de l'Institut Henri Poincaré, section B*, tome 31(2), 393–427.

Dridi, R. & Renault, E. (2000). *Semi-parametric indirect inference*. Technical report, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.

Drovandi, C. C., Pettitt, A. N., & Faddy, M. J. (2011). Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3), 317–337.

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

Duffie, D. & Singleton, K. J. (1993). Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica*, 61(4), 929.

Eddelbuettel, D. & Fran, R. (2011a). Rcpp : Seamless R and C ++ Integration. *Journal Of Statistical Software*, 40(8), 1–18.

Eddelbuettel, D. & Fran, R. (2011b). *Rcpp : Seamless R and C ++ Integration*, volume 40. New York: Springer.

Eddelbuettel, D. & Sanderson, C. (2016). RcppArmadillo : Accelerating R with High-Performance C ++ Linear Algebra. *Computational Statistics and Data Analysis*, 71(2014), 1–16.

Fenton, V. M. & Gallant, A. R. (1996). Convergence Rates of SNP Density Estimators. *Econometrica*, 64(3), 719.

Fermanian, J.-D. & Salanié, B. (2004). A Nonparametric Simulated Maximum Likelihood Estimation Method. *Econometric Theory*, 20(04), 701–734.

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

Forneron, J. & Ng, S. (2018). The ABC of Simulation Estimation with Auxiliary Statistics. *Journal of Econometrics*.

Fridman, M. & Harris, L. E. (1998). A Maximum Likelihood Approach for Non-Gaussian Stochastic Volatility Models. *Journal of Business & Economic Statistics*, 16(3), 284–291.

Gach, F. & Pötscher, B. (2010). Non-Parametric Maximum Likelihood Density Estimation and Simulation-Based Minimum Distance Estimators. *MPRA Paper*, 1(2004), 1–46.

Gallant, A. R. & Nychka, D. W. (1987). Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica*, 55(2), 363–390.

Gallant, A. R. & Tauchen, G. (1993). A nonparametric approach to nonlinear time series analysis: estimation and simulation. In *New directions in time series analysis* (pp. 71–92). Springer.

Gallant, A. R. & Tauchen, G. (1996). Which Moments to Match? *Econometric Theory*, 12(04), 657.

Gao, J. & Hong, H. (2014). A Computational Implementation of GMM. *SSRN Working Paper 2503199*.

Geweke, J. & Keane, M. (2000). An empirical analysis of earnings dynamics among men in the PSID: 1968- 1989. *Journal of Econometrics*, 96, 293–356.

Ghalanos, A. (2017). Introduction to the rugarch package (Version 1.30-1).

Gospodinov, N., Komunjer, I., & Ng, S. (2017). Simulated minimum distance estimation of dynamic models with errors-in-variables. *Journal of Econometrics*, 200(2), 181–193.

Gospodinov, N. & Ng, S. (2015). Minimum Distance Estimation of Possibly Noninvertible Moving Average Models. *Journal of Business & Economic Statistics*, 33(3), 403–417.

Gouriéroux, C. & Monfort, A. (1996). *Simulation-based Econometric Methods*. CORE lectures. Oxford University Press.

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

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

Gouriéroux, C., Phillips, P. C., & Yu, J. (2010). Indirect inference for dynamic panel models. *Journal of Econometrics*, 157(1), 68–77.

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

Gourinchas, P.-O. & Parker, J. A. (2010). The empirical importance of precautionary saving in Turkey. *AEA Papers and Proceedings*, 91(2), 406–412.

Gourio, F. & Roys, N. (2014). Size-dependent regulations, firm size distribution, and reallocation. *Quantitative Economics*, 5(2), 377–416.

Guvenen, F., Karahan, F., Ozkan, S., & Song, J. (2015). What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Dynamics? *NBER working paper*, wp29013.

Guvenen, F. & Smith, A. A. (2014). Inferring Labor Income Risk and Partial Insurance from Economic Choices. *Econometrica*, 82(6), 2085–2129.

Hall, P. & Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics*, 33(6), 2904–2929.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica1*, 50, 1029–1054.

Hansen, L. P. & Richard, S. F. (1987). The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models. *Econometrica*, 55(3), 587–613.

Hansen, L. P. & Singleton, K. J. (1982). Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models. *1Econometrica*, 50, 1269–1296.

Heggland, K. & Frigessi, A. (2004). Estimating functions in indirect inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2), 447–462.

Heiss, F. & Winschel, V. (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, 144(1), 62–80.

Hentschel, L. (1995). All in the family Nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics*, 39(1), 71–104.

Holtz, M. (2011). *Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance*, volume 77 of *Lecture Notes in Computational Science and Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Horowitz, J. L. (2011). Applied Nonparametric Instrumental Variables Estimation. *Econometrica*, 79(2), 347–394.

Horowitz, J. L. (2014). Ill-Posed Inverse Problems in Economics. *Annual Review of Economics*, 6(1), 21–51.

Horowitz, J. L. & Lee, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, 75(4), 1191–1208.

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

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

Jacquier, E., Polson, N. G., & Rossi, P. E. (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics*, 122(1), 185–212.

Jensen, M. J. & Maheu, J. M. (2010). Bayesian semiparametric stochastic volatility modeling. *Journal of Econometrics*, 157(2), 306–316.

Jiang, W. & Turnbull, B. (2004). The Indirect Method: Inference Based on Intermediate Statistics- A Synthesis and Examples. *Statistical Science2*, 19(2), 239–263.

Johnson, S. G. (2014). The NLopt nonlinear-optimization package. *Version 2.4.2 http://ab-initio.mit.edu/nlopt*.

Jones, D. R., Perttunen, C. D., & Stuckman, B. E. (1993). Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1), 157–181.

Judd, K. L. (1996). Chapter 12 Approximation, perturbation, and projection methods in economic analysis. In *Handbook of Computational Economics* (pp. 509–585).

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

Kastner, G. (2016). Dealing with Stochastic Volatility in Time Series Using the R Package stochvol. *Journal of Statistical Software*, 69.

Kim, S., Shepherd, N., & Chib, S. (1998). Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies*, 65(3), 361–393.

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

Knight, J. L. & Yu, J. (2002). Empirical Characteristic Function in Time Series Estimation. *Econometric Theory*, 18(03), 691–721.

Kolmogorov, A. N. & Tikhomirov, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2), 3–86.

Kormiltsina, A. & Nekipelov, D. (2014). Consistent Variance of the Laplace Type Estimators. *Mimeo, SMU*.

Koul, H. L. (1986). Minimum Distance Estimation and Goodness-of-Fit Tests in First-Order Autoregression. *The Annals of Statistics*, 14(3), 1194–1213.

Kristensen, D. & Salanié, B. (2017). Higher-order properties of approximate estimators. *Journal of Econometrics*, 198(2), 189–208.

Kristensen, D. & Shin, Y. (2012). Estimation of dynamic models with nonparametric simulated maximum likelihood. *Journal of Econometrics*, 167(1), 76–94.

Kruijer, W., Rousseau, J., & van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4, 1225–1257.

Krusell, P., Mukoyama, T., Sahin, A., & Smith, A. A. (2009). Revisiting the welfare effects of eliminating business cycles. *Review of Economic Dynamics*, 12(3), 393–404.

Krusell, P. & Smith, Jr., A. a. (1999). On the Welfare Effects of Eliminating Business Cycles. *Review of Economic Dynamics*, 2(1), 245–272.

Lee, B.-S. & Ingram, B. F. (1991). Simulation estimation of time-series models. *Journal of Econometrics*, 47(2-3), 197–205.

Lenormand, M., Jabot, F., & Deffuant, G. (2013). Adaptive Approximate Bayesian Computation Computation for Complex Models. *Journal of Computational and Graphical Statistics*, 28(6), 2777–2796.

Li, T. & Zheng, X. (2008). Semiparametric Bayesian inference for dynamic Tobit panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 23(6), 699–728.

Liebscher, E. (2005). Towards a Unified Approach for Proving Geometric Ergodicity and Mixing Properties of Nonlinear Autoregressive Processes. *Journal of Time Series Analysis*, 26(5), 669–689.

Liesenfeld, R. & Jung, R. C. (2000). Stochastic volatility models: conditional normality versus heavy-tailed distributions. *Journal of applied Econometrics*, 15(2), 137–160.

Lise, J., Meghir, C., & Robin, J. M. (2015). Matching, Sorting, and Wages. *Review of Economic Dynamics*.

Lucas, R. E. (1991). *Models of Business Cycles*. Wiley.

Lucas, R. E. (2003). Macroeconomic Priorities. *American Economic Review*, 93(1), 1–14.

Mahieu, R. J. & Schotman, P. C. (1998). An empirical application of stochastic volatility models. *Journal of Applied Econometrics*, 13(4), 333–360.

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

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov Chain Monte Carlo Without Likelihoods. *Procedings of the National Academy of Science*, 100(23), 15324–15328.

McFadden, D. (1978). Modeling the Choice of Residential Location. *Transportation Research Record*.

McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica*, 57(5), 995.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc.

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

Michaelides, A. & 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.

Mitrovic, S., Sejdinovic, D., & Teh, Y. (2016). Dr-ABC: Approximate Bayesian Computation with Kernel-Based Distribution Regression. *ICML Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 48, 1482–1491.

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

Nelder, J. & Mead, R. (1965). A simplex method for function minimization. *The computer journal*.

Newey, W. & McFadden, D. (1994). Large Sample Estimation and Hypothesis Testing. *Handbook of Econometrics*, 36(4), 2111–2234.

Newey, W. K. (1994). The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62(6), 1349.

Newey, W. K. (2001). Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models. *Review of Economics and Statistics*, 83(4), 616–627.

Newey, W. K. & Smith, R. J. (2004). Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica*, 72(1), 219–255.

Nickl, R. & Pötscher, B. M. (2011). Efficient simulation-based minimum distance estimation and indirect inference. *Mathematical Methods of Statistics*, 19(4), 327–364.

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *Annals of Statistics*, 38(3), 1733–1766.

Owen, A. B. (2003). Quasi-monte carlo sampling. *Monte Carlo Ray Tracing: Siggraph*, 1, 69–88.

Pagan, A. & Ullah, A. (1999). Nonparametric Econometrics. *Themes in Modern Econometrics, Cambridge University Press*.

Pakes, A. (1986). Patents as Options: Some Estimates of the Value of Holding European Patent Stocks. *Econometrica*, 54(4), 755.

Pakes, A. & Pollard, D. (1989). Simulation and the Asymptotics of Optimization Estimators. *Econometrica*, 57(5), 1027.

Park, M., Jitkrittum, W., & Sejdinovic, D. (2016). Approximate Bayesian Computation with Kernel Embeddings. *Proceedings of the 19th International Confernce on Artifical Intelligence and Statistics*, 51(398-407).

Pisier, G. (1983). *Some applications of the metric entropy condition to harmonic analysis*, (pp. 123–154). Springer Berlin Heidelberg: Berlin, Heidelberg.

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

Reis, R. (2009). The Time-Series Properties of Aggregate Consumption: Implications for the Costs of Fluctuations. *Journal of the European Economic Association*, 7(4), 722–753.

Rietz, T. A. (1988). The equity risk premium a solution. *Journal of Monetary Economics*, 22(1), 117–131.

Rilstone, P., Srivastara, K., & Ullah, A. (1996). The Second-Order Bias and Mean Squared Error of Nonlinear Estimators. *Journal of Econometrics1*, 75, 369–385.

Rio, E. (2000). *Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants*, volume 31 of *Mathématiques et Applications*. Springer Berlin Heidelberg.

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

Ruge-Murcia, F. (2012). Estimating nonlinear DSGE models by the simulated method of moments: With an application to business cycles. *Journal of Economic Dynamics and Control*, 36(6), 914–938.

Ruge-Murcia, F. (2017). Skewness Risk and Bond Prices. *Journal of Applied Econometrics*, 32(2), 379–400.

Rust, J. (1987). Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica*, 55(5), 999.

Schennach, S. M. (2014). Entropic Latent Variable Integration via Simulation. *Econometrica*, 82(1), 345–385.

Sisson, S. & 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*, 12, 313–335.

Sisson, S. A., Fan, Y., & Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6), 1760–1765.

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

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

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

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

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31), 187–202.

Ushakov, N. G. (1999). *Selected Topics in Characteristic Functions*. Modern Probability and Statistics. Mouton De Gruyter.

van der Vaart, A. W. & Ghosal, S. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5), 1233–1263.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York, NY: Springer New York.

Vo, R., Drovandi, C., Pettitt, A., & Pettet, G. (2015). Melanoma Cell Colony Expansion Parameters Revealed by Approximate Bayesian Computation. *eprints.qut.edu/au/83824*.

Weil, P. (1989). The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics*, 24(3), 401–421.

Wong, W. H. & Severini, T. A. (1991). On Maximum Likelihood Estimation in Infinite Dimensional Parameter Spaces. *The Annals of Statistics*, 19(2), 603–632.

Wood, S. N. (2010). Statistical Inference for Noisy Nonlinear Ecological Dynamic Systems. *Nature*, 466(7310), 1102–1104.

Wooldridge, J. M. & White, H. (1988). Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes. *Econometric Theory*, 4(02), 210–230.

Yu, J. (1998). *Empirical Characteristic Function Estimation and its Applications*. PhD thesis, University of Western Ontario.

Zeileis, A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software*, 11(10), 1–17.

Appendix to Chapter 1

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

$$\widehat{\boldsymbol{\psi}} = \boldsymbol{\psi}(\boldsymbol{\theta}_0) + \frac{\mathbb{A}(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\mathbb{C}(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}). \tag{.0.1}$$

Now $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \frac{A(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})$. Thus expanding $\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}})$ around $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$:

$$\begin{aligned}\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}}) &= \boldsymbol{\psi}\left(\boldsymbol{\theta}_0 + \frac{A(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right) \\ &= \boldsymbol{\psi}(\boldsymbol{\theta}_0) + \boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\left(\frac{A(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right) + \frac{1}{2T}\sum_{j=1}^{K}\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A(\boldsymbol{\theta}_0)A_j(\boldsymbol{\theta}_0) + o_p(\frac{1}{T}).\end{aligned}$$

Equating with $\boldsymbol{\psi}(\boldsymbol{\theta}_0) + \frac{\mathbb{A}(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\mathbb{C}(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})$ and solving for A, C we get:

$$\begin{aligned}A(\boldsymbol{\theta}_0) &= \left[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\mathbb{A}(\boldsymbol{\theta}_0) \\ C(\boldsymbol{\theta}_0) &= \left[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\mathbb{C}(\boldsymbol{\theta}_0) - \frac{1}{2}\sum_{j=1}^{K}\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A(\boldsymbol{\theta}_0)A_j(\boldsymbol{\theta}_0)\right).\end{aligned}$$

For estimator specific A_d^b and a_d^b, define $a_d^b = \text{trace}([\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)]^{-1}[\sum_{j=1}^{K}\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A_{d,j}^b(\boldsymbol{\theta}_0) + \mathbb{A}_{d,\boldsymbol{\theta}}^b(\boldsymbol{\theta}_0)])$,

$$\begin{aligned}C_d^M(\boldsymbol{\theta}_0) &= 2\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\overline{A}_d(\boldsymbol{\theta}_0)\overline{a}_d(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 - \overline{a}_d(\boldsymbol{\theta}_0)^2\boldsymbol{\theta}_0 - \left[\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)'}{\pi(\boldsymbol{\theta}_0)^2}\right]\overline{A}_d(\boldsymbol{\theta}_0)'\overline{A}_d(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 \\ &\quad -\frac{1}{B}\sum_{b=1}^{B}(a_d^b(\boldsymbol{\theta}_0) - \overline{a}_d(\boldsymbol{\theta}_0))A_d^b(\boldsymbol{\theta}_0).\end{aligned} \tag{.0.2}$$

Where $\overline{a}_d = \frac{1}{B}\sum_{b=1}^{B}a_d^b$, $\overline{A}_d$ is defined analogously. Note that $\overline{a}(\boldsymbol{\theta}_0) \to 0$ as $B \to \infty$ if $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and the first two terms drop out.

Proof of Proposition 1, RS

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

$$\boldsymbol{\theta}^b = \boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}). \tag{.0.1}$$

i. Expansion of $\widehat{\psi}^b(\boldsymbol{\theta}_0)$ and $\widehat{\psi}_{\boldsymbol{\theta}^b(\boldsymbol{\theta}_0)}$:

$$\begin{aligned}
\widehat{\psi}^b(\boldsymbol{\theta}^b) &= \psi(\boldsymbol{\theta}^b) + \frac{\mathbb{A}^b(\boldsymbol{\theta}^b)}{\sqrt{T}} + \frac{\mathbb{C}^b(\boldsymbol{\theta}^b)}{T} + o_p(\frac{1}{T}) \\
&= \psi(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})) + \frac{\mathbb{A}^b(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}))}{\sqrt{T}} \\
&+ \frac{\mathbb{C}^b(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}))}{T} + o_p(\frac{1}{T}) \\
&= \psi(\boldsymbol{\theta}_0) + \frac{\mathbb{A}^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\mathbb{C}^b(\boldsymbol{\theta}_0)}{T} + \frac{\mathbb{A}_{\boldsymbol{\theta}^b(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)}}{T} \\
&+ \frac{1}{2}\sum_{j=1}^{K} \frac{\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}).
\end{aligned}$$

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

$$A^b(\boldsymbol{\theta}_0) = \left[\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\mathbb{A}(\boldsymbol{\theta}_0) - \mathbb{A}^b(\boldsymbol{\theta}_0)\right) \tag{.0.2}$$

$$C^b(\boldsymbol{\theta}_0) = \left[\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\mathbb{C}(\boldsymbol{\theta}_0) - \mathbb{C}^b(\boldsymbol{\theta}_0) - \frac{1}{2}\sum_{j=1}^{K}\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0) - \mathbb{A}_{\boldsymbol{\theta}^b(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)}\right), \tag{.0.3}$$

it follows that

$$\begin{aligned}
\widehat{\psi}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b) &= \widehat{\psi}^b_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right) \\
&= \psi_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right) + \frac{\mathbb{A}^b_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right)}{\sqrt{T}} \\
&\quad + \frac{\mathbb{C}^b_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right)}{T} + o_p(\frac{1}{T}) \\
&= \psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) + \sum_{j=1}^{K}\frac{\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\mathbb{A}^b a(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K}\frac{\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j,\boldsymbol{\theta}_k}(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0)A_k^b(\boldsymbol{\theta}_0)}{T} \\
&\quad + \sum_{j=1}^{K}\frac{\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)C_j^b(\boldsymbol{\theta}_0)}{T} + \sum_{j=1}^{K}\frac{\mathbb{A}^b_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0)}{T} + \frac{\mathbb{C}^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}).
\end{aligned}$$

To obtain the determinant of $\widehat{\psi}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b)$, let $a^b(\boldsymbol{\theta}_0) = \text{trace}(\mathcal{A}^b(\boldsymbol{\theta}_0))$, $a_2^b(\boldsymbol{\theta}_0) = \text{trace}(\mathcal{A}^b(\boldsymbol{\theta}_0)^2)$, $c^b(\boldsymbol{\theta}_0) = \text{trace}(\mathcal{C}^b(\boldsymbol{\theta}_0))$, where

$$\mathcal{A}^b(\boldsymbol{\theta}_0) = \left[\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\sum_{j=1}^{K}\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0) + \mathbb{A}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right)$$

$$\mathcal{C}^b(\boldsymbol{\theta}_0) = \left[\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\frac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K}\frac{\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j,\boldsymbol{\theta}_k}(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0)A_k^b(\boldsymbol{\theta}_0)}{T} + \sum_{j=1}^{K}\frac{\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)C_j^b(\boldsymbol{\theta}_0)}{T} + \sum_{j=1}^{K}\mathbb{A}^b_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0) + \mathbb{C}^b(\boldsymbol{\theta}_0)\right).$$

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:

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

Hence the required determinant is

$$\begin{aligned}\left|\widehat{\psi}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b)\right| &= \left|\widehat{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|\left|I + \frac{\mathcal{A}^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\mathcal{C}^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right| \\ &= \left|\widehat{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|\left(1 + \frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{a_2^b(\boldsymbol{\theta}_0)}{T} + \frac{c^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right).\end{aligned}$$

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

$$\begin{aligned}\left|\widehat{\psi}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b)\right|^{-1}\pi(\boldsymbol{\theta}^b) &= \left|\widehat{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\left(1 + \frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{a_2^b(\boldsymbol{\theta}_0)}{T} + \frac{c^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right)^{-1}\pi(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})) \\ &= \left|\widehat{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\left(1 - \frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}} - \frac{a_2^b(\boldsymbol{\theta}_0)}{T} - \frac{c^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right) \\ &\times \left(\pi(\boldsymbol{\theta}_0) + \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\frac{C^b(\boldsymbol{\theta}_0)}{T} + \frac{1}{2}\sum_{j=1}^{K}\frac{\pi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right) \\ &= \left|\widehat{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\pi(\boldsymbol{\theta}_0)\Bigg(1 - \frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} - \frac{a_2^b(\boldsymbol{\theta}_0)}{T} - \frac{c^b(\boldsymbol{\theta}_0)}{T} \\ &- \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{a^b(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)}{T} + \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{C^b(\boldsymbol{\theta}_0)}{T} + \frac{1}{2}\frac{A^b(\boldsymbol{\theta}_0)\pi_{\boldsymbol{\theta},\boldsymbol{\theta}'}(\boldsymbol{\theta}_0)A^{b\prime}(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\Bigg).\end{aligned}$$

Now $\overline{A}(\boldsymbol{\theta}_0) = \frac{1}{B}\sum_{b=1}^{B}A^b(\boldsymbol{\theta}_0)$. Similarly define $\overline{C}(\boldsymbol{\theta}_0) = \frac{1}{B}C^b(\boldsymbol{\theta}_0)$. Also, denote the term in $1/T$ by:

$$e^b(\boldsymbol{\theta}_0) = -a_2^b(\boldsymbol{\theta}_0) - c^b(\boldsymbol{\theta}_0) - \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}a^b(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0) + \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}C^b(\boldsymbol{\theta}_0) + \frac{1}{2}A^b(\boldsymbol{\theta}_0)\pi_{\boldsymbol{\theta},\boldsymbol{\theta}'}(\boldsymbol{\theta}_0)A^{b\prime}(\boldsymbol{\theta}_0).$$

The normalized weight for draw b is:

$$\begin{aligned}
\overline{w}^b(\boldsymbol{\theta}^b) &= \frac{\left|\widehat{\psi}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b)\right|^{-1}\pi(\boldsymbol{\theta}^b)}{\sum_{c=1}^B \left|\widehat{\psi}^c_{\boldsymbol{\theta}}(\boldsymbol{\theta}^c)\right|^{-1}\pi(\boldsymbol{\theta}^c)} = \frac{1}{B}\Big(\frac{1-\frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}(\boldsymbol{\theta}_0)}}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{e^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})}{1+\frac{1}{B}\sum_{c=1}^B -\frac{a^c(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^c(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{e^c(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})}\Big) \\
&= \frac{1}{B}\Big(\frac{1-\frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}(\boldsymbol{\theta}_0)}}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{e^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})}{1-\frac{\overline{a}(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}(\boldsymbol{\theta}_0)}}{\pi(\boldsymbol{\theta}_0)}\frac{\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\overline{e}(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})}\Big) \\
&= \frac{1}{B}\Big(1-\frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{e^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\Big)\times\Big(1+\frac{\overline{a}(\boldsymbol{\theta}_0)}{\sqrt{T}}-\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}}-\frac{\overline{e}(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\Big) \\
&= \frac{1}{B}\Big(1-\frac{a^b(\boldsymbol{\theta}_0)-\overline{a}(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)-\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{e^b(\boldsymbol{\theta}_0)-\overline{e}(\boldsymbol{\theta}_0)}{T}-\frac{a^b(\boldsymbol{\theta}_0)\overline{a}(\boldsymbol{\theta}_0)}{T}-\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)\overline{a}(\boldsymbol{\theta}_0)}{T} \\
&\quad -\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{\overline{A}(\boldsymbol{\theta}_0)a^b(\boldsymbol{\theta}_0)}{T}-\Big[\frac{\pi a(\boldsymbol{\theta}_0)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)'}{\pi(\boldsymbol{\theta}_0)^2}\Big]\frac{A^b(\boldsymbol{\theta}_0)'\overline{A}(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\Big).
\end{aligned}$$

The posterior mean is $\overline{\boldsymbol{\theta}}_{RS} = \sum_{b=1}^B \overline{w}^b(\boldsymbol{\theta}^b)\boldsymbol{\theta}^b$. Using $\boldsymbol{\theta}^b$ defined in (.0.1), A and C defined in (.0.2) and (.0.3):

$$\overline{\boldsymbol{\theta}}_{RS} = \boldsymbol{\theta}_0 + \frac{1}{B}\sum_{b=1}^B \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{1}{B}\sum_{b=1}^B \frac{C^b(\boldsymbol{\theta}_0)}{T} + \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{1}{B}\sum_{b=1}^B \frac{(A^b(\boldsymbol{\theta}_0)-\overline{A}(\boldsymbol{\theta}_0))A^b(\boldsymbol{\theta}_0)}{T} + C^M(\boldsymbol{\theta}_0) + o_p(\frac{1}{T}).$$

Proof of Results for LT

From

$$\boldsymbol{\theta}^b = \boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}),$$

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

$$\begin{aligned}
\widehat{\psi}^b(\boldsymbol{\theta}^b) &= \psi(\boldsymbol{\theta}^b) + \frac{\mathbb{A}^b_\infty(\boldsymbol{\theta}^b)}{\sqrt{T}} \\
&= \psi\Big(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\Big) + \frac{\mathbb{A}^b_\infty(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}))}{\sqrt{T}} \\
&= \psi(\boldsymbol{\theta}_0) + \frac{\mathbb{A}^b_\infty(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\mathbb{A}^b_{\infty,\boldsymbol{\theta}}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)}{T} + \frac{1}{2}\sum_{j=1}^K \frac{\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})
\end{aligned}$$

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

$$A^b(\boldsymbol{\theta}_0) = \Big[\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\Big]^{-1}\Big(\mathbb{A}(\boldsymbol{\theta}_0) - \mathbb{A}^b_\infty(\boldsymbol{\theta}_0)\Big) \tag{.0.1}$$

$$C^b(\boldsymbol{\theta}_0) = \Big[\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\Big]^{-1}\left(\mathbb{C}(\boldsymbol{\theta}_0) - \frac{1}{2}\sum_{j=1}^K \psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0) - \mathbb{A}^b_{\infty,\boldsymbol{\theta}}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)\right). \tag{.0.2}$$

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

$$\begin{aligned}
\widehat{\psi}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}^b) &= \psi_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right) + \frac{\mathbb{A}^b_{\infty,\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right)}{\sqrt{T}} \\
&= \psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) + \sum_{j=1}^{K}\frac{\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\mathbb{A}^b_{\infty,\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\sqrt{T}} + \sum_{j=1}^{k}\frac{\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)C^b_j(\boldsymbol{\theta}_0)}{T} + \sum_{j=1}^{K}\frac{\mathbb{A}^b_{\infty,\boldsymbol{\theta},\boldsymbol{\theta}_j}A^b_j(\boldsymbol{\theta}_0)}{T} \\
&\quad + \frac{1}{2}\sum_{j,k=1}^{K}\frac{\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j,\boldsymbol{\theta}_k}(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)A^b_k(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}).
\end{aligned}$$

Let

$$\mathcal{A}^b(\boldsymbol{\theta}_0) = \left[\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\mathbb{A}^b_{\infty,\boldsymbol{\theta}}(\boldsymbol{\theta}_0) + \sum_{j=1}^{K}\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)\right)$$

$$\mathcal{C}^b(\boldsymbol{\theta}_0) = \left[\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\sum_{j=1}^{K}\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)C^b_j(\boldsymbol{\theta}_0) + \sum_{j=1}^{K}\mathbb{A}^b_{\infty,\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K}\psi_{\boldsymbol{\theta},\boldsymbol{\theta}_j,\boldsymbol{\theta}_k}(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)A^b_k(\boldsymbol{\theta}_0)\right)$$

$$a^b(\boldsymbol{\theta}_0) = \text{trace}(\mathcal{A}^b(\boldsymbol{\theta}_0)), \quad a^b_2(\boldsymbol{\theta}_0) = \text{trace}(\mathcal{A}^b(\boldsymbol{\theta}_0)^2), \quad c^b(\boldsymbol{\theta}_0) = \text{trace}(\mathcal{C}^b(\boldsymbol{\theta}_0)).$$

The determinant is

$$\begin{aligned}
\left|\widehat{\psi}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1} &= \left|\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\left|I + \frac{\mathcal{A}^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\mathcal{C}^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right|^{-1} \\
&= \left|\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\left(1 + \frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{a^b_2(\boldsymbol{\theta}_0)}{T} + \frac{c^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right)^{-1} \\
&= \left|\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\left(1 - \frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}} - \frac{a^b_2(\boldsymbol{\theta}_0)}{T} - \frac{c^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right).
\end{aligned}$$

The prior is

$$\begin{aligned}
\pi(\boldsymbol{\theta}^b) &= \pi\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right) \\
&= \pi(\boldsymbol{\theta}_0) + \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\frac{C^b(\boldsymbol{\theta}_0)}{T} + \frac{1}{2}\frac{A^b(\boldsymbol{\theta}_0)\pi_{\boldsymbol{\theta},\boldsymbol{\theta}'}A^{b\prime}(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}).
\end{aligned}$$

Let: $e^b(\boldsymbol{\theta}_0) = -c^b(\boldsymbol{\theta}_0) - a^b_2(\boldsymbol{\theta}_0) + \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}C^b(\boldsymbol{\theta}_0) + A^b(\boldsymbol{\theta}_0)\frac{\pi_{\boldsymbol{\theta},\boldsymbol{\theta}'}}{\pi}(\boldsymbol{\theta}_0)A^{b\prime}(\boldsymbol{\theta}_0)$. After some simplification, the product is

$$\left|\widehat{\psi}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\pi(\boldsymbol{\theta}^b) = \left|\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\pi(\boldsymbol{\theta}_0)\left(1 - \frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{e^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right).$$

Hence, the normalized weight for draw b is

$$\begin{aligned}
\overline{w}^b(\boldsymbol{\theta}^b) &= \frac{\left|\widehat{\psi}^b_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\pi(\boldsymbol{\theta}^b)}{\sum_{c=1}^B \left|\widehat{\psi}^c_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\pi(\boldsymbol{\theta}^c)} = \frac{1}{B}\frac{1-\frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{e^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})}{1-\frac{\overline{a}(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\overline{e}(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})} \\
&= \frac{1}{B}\left(1-\frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{e^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\right)\left(1+\frac{\overline{a}(\boldsymbol{\theta}_0)}{\sqrt{T}}-\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}}-\frac{\overline{e}(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\right) \\
&= \frac{1}{B}\Big(1-\frac{a^b(\boldsymbol{\theta}_0)-\overline{a}(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)-\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{e^b(\boldsymbol{\theta}_0)-\overline{e}(\boldsymbol{\theta}_0)}{T}-\frac{a^b(\boldsymbol{\theta}_0)\overline{a}(\boldsymbol{\theta}_0)}{T}-\frac{\frac{\pi a(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}A^b(\boldsymbol{\theta}_0)\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\overline{A}(\boldsymbol{\theta}_0)}{T} \\
&+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{a^b(\boldsymbol{\theta}_0)\overline{A}(\boldsymbol{\theta}_0)}{T}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{\overline{a}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\Big).
\end{aligned}$$

Hence the posterior mean is $\overline{\boldsymbol{\theta}}_{\mathrm{LT}} = \sum_{b=1}^B \overline{w}^b(\boldsymbol{\theta}^b)\boldsymbol{\theta}^b$ and $\boldsymbol{\theta}^b = \left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right)$.
After simplification, we have

$$\begin{aligned}
\overline{\boldsymbol{\theta}}_{\mathrm{LT}} &= \boldsymbol{\theta}_0 + \frac{\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\overline{C}(\boldsymbol{\theta}_0)}{T} - \frac{1}{B}\sum_{b=1}^B \frac{(a^b(\boldsymbol{\theta}_0)-\overline{a}(\boldsymbol{\theta}_0))A^b(\boldsymbol{\theta}_0)}{T} - \frac{[\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\overline{A}(\boldsymbol{\theta}_0)]^2\boldsymbol{\theta}_0}{T} \\
&+ \frac{1}{B}\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\sum_{b=1}^B \frac{(A^b(\boldsymbol{\theta}_0)-\overline{A}(\boldsymbol{\theta}_0))A^b(\boldsymbol{\theta}_0)}{T} \\
&- \frac{\overline{a}(\boldsymbol{\theta}_0)^2\boldsymbol{\theta}_0}{T} + 2\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{\overline{a}(\boldsymbol{\theta}_0)\overline{A}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0}{T} + o_p(\frac{1}{T}) \\
&= \boldsymbol{\theta}_0 + \frac{\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\overline{C}(\boldsymbol{\theta}_0)}{T} + \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{1}{B}\sum_{b=1}^B \frac{(A^b(\boldsymbol{\theta}_0)-\overline{A}(\boldsymbol{\theta}_0))A^b(\boldsymbol{\theta}_0)}{T} + C^M(\boldsymbol{\theta}_0) + o_p(\frac{1}{T}),
\end{aligned}$$

where all terms are based on $A^b(\boldsymbol{\theta}_0)$ defined in (.0.1) and $C^b(\boldsymbol{\theta}_0)$ in (.0.2).

Results for SLT:

From

$$\begin{aligned}
\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}) &= \frac{1}{S}\sum_{s=1}^S \widehat{\boldsymbol{\psi}}^s(\boldsymbol{\theta}) + \frac{\mathbb{A}^b_\infty(\boldsymbol{\theta})}{\sqrt{T}} \\
\widehat{\boldsymbol{\psi}}^s(\boldsymbol{\theta}) &= \boldsymbol{\psi}(\boldsymbol{\theta}) + \frac{\mathbb{A}^s(\boldsymbol{\theta})}{\sqrt{T}} + \frac{\mathbb{C}^s(\boldsymbol{\theta})}{T} + o_p(\frac{1}{T}) \\
\boldsymbol{\theta}^b &= \boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}),
\end{aligned}$$

we have

$$\begin{aligned}\widehat{\psi}^s(\boldsymbol{\theta}^b) &= \frac{1}{S}\sum_{s=1}^{S}\widehat{\psi}^s\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right) + \frac{\mathbb{A}^b_\infty(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}))}{\sqrt{T}} \\ &= \boldsymbol{\psi}(\boldsymbol{\theta}_0) + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{A}^s(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\mathbb{A}^b_\infty(\boldsymbol{\theta}_0)}{\sqrt{T}} + \boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{A}^s_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)}{T} + \frac{\mathbb{A}^b_{\infty,\boldsymbol{\theta}}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)}{T} \\ &+ \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{C}^s(\boldsymbol{\theta}_0)}{T} + \frac{1}{2}\sum_{j=1}^{K}\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)\frac{A^b(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)}{T} + \boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}).\end{aligned}$$

Thus,

$$A^b(\boldsymbol{\theta}_0) = \left[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\mathbb{A}(\boldsymbol{\theta}_0) - \frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s(\boldsymbol{\theta}_0) - \mathbb{A}^b_\infty(\boldsymbol{\theta}_0)\right) \tag{.0.1}$$

$$\begin{aligned}C^b(\boldsymbol{\theta}_0) &= \left[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\mathbb{C}(\boldsymbol{\theta}_0) - \frac{1}{S}\sum_{s=1}^{S}\mathbb{C}^s(\boldsymbol{\theta}_0) - \frac{1}{2}\sum_{j=1}^{K}\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)\right) \\ &\quad - \left[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left[\frac{1}{S}\sum_{s=1}^{S}\mathbb{A}_{\boldsymbol{\theta}^s(\boldsymbol{\theta}_0)+\mathbb{A}^b_{\infty,\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}\right]A^b(\boldsymbol{\theta}_0).\end{aligned} \tag{.0.2}$$

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

$$\begin{aligned}\widehat{\psi}^b(\boldsymbol{\theta}^b) &= \boldsymbol{\psi}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right) + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{A}^s\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right)}{\sqrt{T}} \\ &+ \frac{\mathbb{A}^b_\infty\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right)}{\sqrt{T}} + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{C}^s\left(\boldsymbol{\theta}_0 + \frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{C^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right)}{T} + o_p(\frac{1}{T}) \\ &= \boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) + \sum_{j=1}^{K}\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)\frac{A^b_j(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{A}^s_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{\mathbb{A}^b_{\infty,\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\sqrt{T}} + \frac{1}{S}\sum_{s=1}^{S}\frac{\mathbb{C}^s(\boldsymbol{\theta}_0)}{T} + \sum_{j=1}^{K}\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)\frac{C^b_j(\boldsymbol{\theta}_0)}{T} \\ &+ \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{K}\frac{\mathbb{A}^s_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)}{T} + \sum_{j=1}^{K}\frac{\mathbb{A}^b_{\infty,\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)}{T} + \frac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K}\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j,\boldsymbol{\theta}_k}(\boldsymbol{\theta}_0)\frac{A^b_k(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T}).\end{aligned}$$

Let:

$$\begin{aligned}\mathcal{A}^b(\boldsymbol{\theta}_0) &= \left[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) + \mathbb{A}^b_{\infty,\boldsymbol{\theta}}(\boldsymbol{\theta}_0) + \sum_{j=1}^{K}\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j}A^b_j(\boldsymbol{\theta}_0)\right) \\ \mathcal{C}^b(\boldsymbol{\theta}_0) &= \left[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\frac{1}{S}\sum_{s=1}^{S}\mathbb{C}^s(\boldsymbol{\theta}_0) + \sum_{j=1}^{K}\left[\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)C^b_j(\boldsymbol{\theta}_0) + \frac{1}{S}\sum_{s=1}^{S}\mathbb{A}^s_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0) + \mathbb{A}^b_{\infty,\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)\right]\right) \\ &+ \left[\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1}\left(\frac{1}{2}\sum_{j,k=1}^{K}\boldsymbol{\psi}_{\boldsymbol{\theta},\boldsymbol{\theta}_j,\boldsymbol{\theta}_k}(\boldsymbol{\theta}_0)A^b_k(\boldsymbol{\theta}_0)A^b_j(\boldsymbol{\theta}_0)\right) \\ a^b(\boldsymbol{\theta}_0) &= \text{trace}(\mathcal{A}^b(\boldsymbol{\theta}_0)), \quad a^b_2(\boldsymbol{\theta}_0) = \text{trace}(\mathcal{A}^b(\boldsymbol{\theta}_0)^2), \quad c^b(\boldsymbol{\theta}_0) = \text{trace}(\mathcal{C}^b(\boldsymbol{\theta}_0)).\end{aligned}$$

The determinant is

$$\left|\widehat{\boldsymbol{\psi}}^b(\boldsymbol{\theta}^b)\right|^{-1} = \left|\boldsymbol{\psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\left(1 - \frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}} - \frac{a^b_2(\boldsymbol{\theta}_0)}{T} - \frac{c^b(\boldsymbol{\theta}_0)}{T} + o_p(\frac{1}{T})\right).$$

Hence

$$\begin{aligned}\left|\widehat{\psi}^b(\boldsymbol{\theta}^b)\right|^{-1}\pi(\boldsymbol{\theta}^b) &= \left|\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\pi(\boldsymbol{\theta}_0)\left(1-\frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}}-\frac{a_2^b(\boldsymbol{\theta}_0)}{T}-\frac{c^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\right)\\ &\times\left(1+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{C^b(\boldsymbol{\theta}_0)}{T}+\frac{1}{2}\sum_{j=1}^{K}\frac{\pi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\right)\\ &= \left|\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right|^{-1}\pi(\boldsymbol{\theta}_0)\left(1-\frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{e^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\right)\end{aligned}$$

where $e^b(\boldsymbol{\theta}_0)=-a^b(\boldsymbol{\theta}_0)\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}A^b(\boldsymbol{\theta}_0)-a_2^b(\boldsymbol{\theta}_0)-c^b(\boldsymbol{\theta}_0)+\frac{\pi a(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}C^b(\boldsymbol{\theta}_0)+\frac{1}{2}\sum_{j=1}^{K}\frac{\pi_{\boldsymbol{\theta},\boldsymbol{\theta}_j}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}A^b(\boldsymbol{\theta}_0)A_j^b(\boldsymbol{\theta}_0)$. The normalized weights are

$$\begin{aligned}\overline{w}^b(\boldsymbol{\theta}^b) &= \frac{\left|\widehat{\psi}^b(\boldsymbol{\theta}^b)\right|^{-1}\pi(\boldsymbol{\theta}^b)}{\sum_{c=1}^{B}\left|\widehat{\psi}^c(\boldsymbol{\theta}^c)\right|^{-1}\pi(\boldsymbol{\theta}^c)}\\ &= \frac{1}{B}\left(1-\frac{a^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{e^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\right)\left(1+\frac{\overline{a}(\boldsymbol{\theta}_0)}{\sqrt{T}}-\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}}-\frac{\overline{e}(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})\right).\end{aligned}$$

The posterior mean $\overline{\boldsymbol{\theta}}_{\mathrm{SLT}}=\sum_{b=1}^{B}\overline{w}^b(\boldsymbol{\theta}^b)\boldsymbol{\theta}^b$ with $\boldsymbol{\theta}^b=\boldsymbol{\theta}_0+\frac{A^b(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{C^b(\boldsymbol{\theta}_0)}{T}+o_p(\frac{1}{T})$. After some simplification,

$$\begin{aligned}\overline{\boldsymbol{\theta}}_{SLT} &= \boldsymbol{\theta}_0+\frac{\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\overline{C}(\boldsymbol{\theta}_0)}{T}+\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{1}{B}\sum_{B=1}^{B}\frac{(A^b(\boldsymbol{\theta}_0)-\overline{A}(\boldsymbol{\theta}_0))A^b(\boldsymbol{\theta}_0)}{T}-\frac{1}{B}\sum_{b=1}^{B}\frac{(a^b(\boldsymbol{\theta}_0)-\overline{a}(\boldsymbol{\theta}_0))A^b(\boldsymbol{\theta}_0)}{T}\\ &\quad+2\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\frac{\overline{a}(\boldsymbol{\theta}_0)\overline{A}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0}{T}-\frac{\overline{a}^2(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0}{T}-[\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_0)}\overline{A}(\boldsymbol{\theta}_0)]^2\frac{\boldsymbol{\theta}_0}{T}+o_p(\frac{1}{T})\\ &= \boldsymbol{\theta}_0+\frac{\overline{A}(\boldsymbol{\theta}_0)}{\sqrt{T}}+\frac{\overline{C}(\boldsymbol{\theta}_0)}{T}+\frac{\pi_{\boldsymbol{\theta}(\boldsymbol{\theta}_0)}}{\pi(\boldsymbol{\theta}_0)}\frac{1}{B}\sum_{B=1}^{B}\frac{(A^b(\boldsymbol{\theta}_0)-\overline{A}(\boldsymbol{\theta}_0))A^b(\boldsymbol{\theta}_0)}{T}+C^M(\boldsymbol{\theta}_0)+o_p(\frac{1}{T})\end{aligned}$$

where terms in A and C are defined from (.0.1) and (.0.2).

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, $\bar{e}_S = \frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T} e_t^s = \frac{1}{S}\sum_{s=1}^{S}\bar{e}^s$.

MLE: Define $\bar{e} = \frac{1}{T}\sum_{t=1}^{T} e_t$. Then the mean estimator is $\widehat{m} = m_0 + \sigma_0\bar{e} \sim N(0, \sigma_0^2/T)$. For the variance estimator, $\widehat{e} = y - \widehat{m} = \sigma_0(e - \bar{e}) = \sigma_0 Me$, $M = I_T - 1(1'1)^{-1}1'$ 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

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

The mean of an invG(α, β) 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 $\widehat{m} = \widehat{m}_S + \sigma\frac{1}{S}\sum_{s=1}^{S}\bar{e}^s$. 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 $y_t^s - \overline{y}_t^s = \sigma(e_t^s - \bar{e}^s)$, the variance estimator $\widehat{\sigma}^2_S$ is the σ^2 that solves $\widehat{\sigma}^2 = \sigma^2(\frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T}(e_t^s - \bar{e}^s)^2)$ Hence

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

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 $\mathbb{E}_{\text{LT}}(\sigma^2|\widehat{\sigma}^2) = \widehat{\sigma}^2\left(1+\kappa_{\text{LT}}\right)$. The expectation of the estimator is

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

from which we deduce the bias of the estimator

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

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

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

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

SLT: The SLT is defined as

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

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

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

and the posterior mean becomes

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

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

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

and the bias

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

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

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

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

The MSE is

$$\begin{aligned}&\sigma^4\left[\frac{2}{S(T-1)-2} + \frac{T-1}{T}\mathbb{E}(\kappa_{\text{SLT}})\right]^2 + 2\sigma^4\kappa_1\frac{1}{T-1} + 2\sigma^4\mathbb{V}(\kappa_{\text{SLT}}) + 4\sigma^4\frac{T-1}{T^2}\text{Cov}(\kappa_{\text{SLT}}, \frac{S}{\chi^2_{S(T-1)}}) \\ &= \underbrace{2\sigma^4\left[\frac{2}{[S(T-1)-2]^2} + \kappa_1\frac{1}{T-1}\right]}_{\text{MSE of SMD}} + \frac{(T-1)^2}{T^2}\mathbb{E}(\kappa^2_{\text{SLT}} + \frac{4\sigma^4}{S(T-1)-2}\frac{T-1}{T}\mathbb{E}(\kappa_{\text{SLT}}) \\ &\quad + 2\sigma^4\mathbb{V}(\kappa_{\text{SLT}}) + 4\sigma^4\frac{T-1}{T^2}\text{Cov}(\kappa_{\text{SLT}}, \frac{S}{\chi^2_{S(T-1)}}).\end{aligned}$$

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, $\widehat{\sigma}^{2,b} = \sigma^{2,b} \sum_t (e_t^b - \overline{e}^b)^2/T$. Hence

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

Note that $p_{BC}(\sigma^2|\widehat{\sigma}^2) \sim \text{invG}\left(\frac{T-3}{2}, \frac{T\widehat{\sigma}^2}{2}\right)$ 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: $|\psi_{\boldsymbol{\theta}|^{-1}=\frac{\partial \sigma^{2,s}}{\partial \widehat{\sigma}^2}}$[46]. Since $\sigma^{2,b} = \frac{T\widehat{\sigma}^2}{\sum_t (e_t^b - \overline{e}^b)^2}$, $|\psi_{\boldsymbol{\theta}|^{-1}=\frac{T}{\sum_t (e_t^b - \overline{e}^b)^2}}$.

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

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

The posterior mean after adjusting for the Jacobian transformation is

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

where $1/z^b = \sum_t (e_t^b - \overline{e}^b)^2$. As $B \to \infty$, $\frac{1}{B}\sum_b (z^b)^2 \xrightarrow{p} E[(z^b)^2]$ and $\frac{1}{S}\sum_b z^b \xrightarrow{p} E[z^b]$. Now $z^b \sim \text{invG}\chi^2_{T-1}$ 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, $\nabla_{\boldsymbol{\theta}\boldsymbol{\psi}(\boldsymbol{\theta})}$ and $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'}\boldsymbol{\psi}(\boldsymbol{\theta}) = 0$. For $Z \sim \mathcal{N}(0,1), A(v;\sigma^2) = \sqrt{2}\sigma^2(1-\frac{1}{T})Z$, Thus $\partial_{\sigma^2} A(v;\sigma^2) = \sqrt{2}(1-\frac{1}{T})Z, a^s = \sqrt{2}\sigma^2(1-\frac{1}{T})(Z - Z^s)$. The terms in the asymptotic expansion are therefore

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

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

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

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

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 $\pi(\sigma^{2,s}) = 1/\sigma^{2,s}$:

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

Further Results for Dynamic Panel Model with Fixed Effects

Table .0.1: 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.

Appendix to Chapter 3

Background Material

The Characteristic Function and Some of its Properties

The joint characteristic function (CF) of $(\mathbf{y}_t, \mathbf{x}_t)$ is defined as

$$\psi : \tau \to \mathbb{E}\left(e^{i\tau'(\mathbf{y}_t,\mathbf{x}_t)}\right) = \mathbb{E}\left(\cos(\tau'(\mathbf{y}_t,\mathbf{x}_t)) + i\sin(\tau'(\mathbf{y}_t,\mathbf{x}_t))\right).$$

An important result for the CF is that the mapping between distribution and CF is bijective: two CFs are equal if, and only if they come from the same distribution $f_1 = f_2 \Leftrightarrow \psi_{f_1} = \psi_{f_2}$. The characteristic function has several other attractive features:

i. Existence: The CF is well defined for any probability distribution: it can be computed even if no moment of (y_t, x_t) exist.
ii. Boundedness: The CF is bounded $|\psi(\tau)| \leq 1$ for any distribution. As a result, the objective function $\widehat{Q}_n^S$ is always well defined assuming the density π is integrable.
iii. Continuity in f: The CF is continuous in the distribution $f_n \to f_0$ implies $\psi_{f_n} \to \psi_{f_0}$.
iv. Continuity in τ: The CF is continuous in τ.

The continuity properties are very useful when the data y_t does not have a continuous density, e.g. discrete, but the density of the shocks f is continuous as in Example 2. For instance, the data generated by:

$$y_t = \mathbb{1}_{x_t'\boldsymbol{\theta}+e_t \geq 0}$$

is discrete but its conditional characteristic function is continuous in both f and $\boldsymbol{\theta}$:

$$\mathbb{E}\left(e^{i\tau_y y_t}|x_t\right) = 1 - F(x_t'\boldsymbol{\theta}) + F(x_t'\boldsymbol{\theta})e^{i\tau_y},$$

where F is the CDF of $e_t \sim f$. As a result, the joint CF is also continuous:

$$\mathbb{E}\left(e^{i\tau(y_t,x_t)}\right) = \mathbb{E}\left(e^{i\tau_x x_t}[1 - F(x_t'\boldsymbol{\theta}) + F(x_t'\boldsymbol{\theta})e^{i\tau_y}]\right).$$

The empirical CDF however is not continuous. As a result, a population objective Q based on the CF is continuous but the one based on a CDF is not.

Computing the Sample Objective Function $\widehat{Q}_n^S$

This section discusses the numerical implementation of the Sieve-SMM estimator. First, several transformations are used to normalize the weights ω and impose restrictions such as mean zero $\sum_j \omega_j \mu_j = 0$ and unit variance $\sum_j \omega_j(\mu_j^2 + \sigma_j^2) = 1$ without requiring constrained optimization. For the weights, take $k-1$ unconstrained parameters $\widetilde{\omega}$ and apply the transformation:

$$\omega_1 = \frac{1}{1 + \sum_{\ell=1}^{k-1} e^{\widetilde{\omega}_\ell}}, \quad \omega_j = \frac{e^{\widetilde{\omega}_{j-1}}}{1 + \sum_{\ell=1}^{k-1} e^{\widetilde{\omega}_\ell}} \text{ for } j = 2, \ldots, k.$$

The resulting $\omega_1, \ldots, \omega_k$ are positive and sum to one. To impose a mean zero restriction take $\mu_2, \ldots, \mu_k$ unconstrained and compute:

$$\mu_1 = -\frac{\sum_{j=2}^{k} \omega_j \mu_j}{\omega_1}$$

The mixture has mean zero by construction. In practice, it is assumed that $\sigma_j \geq \underline{\sigma}_k$. Take unconstrained $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_k$ and compute:

$$\sigma_j = \underline{\sigma}_k + e^{\widetilde{\sigma}_j}.$$

The resulting σ_j are greater or equal than the lower bound $\underline{\sigma}_k \geq 0$. To impose unit variance, restrict $\widetilde{\sigma}_1 = 0$ and then divide μ, σ by $\sqrt{\sum_j \omega_j(\mu_j^2 + \sigma_j^2)}$: standardized this way, the mixture has unit variance.

Once the parameters ω, μ, σ are appropriately transformed and normalized, the mixture draws e_t^s can be simulated, and then y_t^s itself is simulated. Numerical integration is used to approximate the sample objective function $\widehat{Q}_n^S$. For an integration grid $\tau_1, \ldots, \tau_m$ with weights $\pi_1, \ldots, \pi_m$ compute the vectors:

$$\widehat{\boldsymbol{\psi}}_n = (\widehat{\boldsymbol{\psi}}_n(\tau_1), \ldots, \widehat{\boldsymbol{\psi}}_n(\tau_m))', \quad \widehat{\boldsymbol{\psi}}_n^S = (\widehat{\boldsymbol{\psi}}_n^S(\tau_1), \ldots, \widehat{\boldsymbol{\psi}}_n^S(\tau_m))'$$

and the objective:

$$\widehat{Q}_n^S(\beta) = (\widehat{\boldsymbol{\psi}}_n - \widehat{\boldsymbol{\psi}}_n^S)' \text{diag}(\pi_1, \ldots, \pi_m)(\widehat{\boldsymbol{\psi}}_n - \widehat{\boldsymbol{\psi}}_n^S).$$

In practice, the objective function is computed the same as for a parametric SMM estimator. If a linear operator B is used to weight the moments, then the finite matrix approximation B_m is computed on $\tau_1, \ldots, \tau_m$ and the objective becomes $(\widehat{\boldsymbol{\psi}}_n - \widehat{\boldsymbol{\psi}}_n^S)' B' \text{diag}(\pi_1, \ldots, \pi_m)(\widehat{\boldsymbol{\psi}}_n - \widehat{\boldsymbol{\psi}}_n^S)'$; a detailed overview on computing the objective function with a linear operator B, using quadrature, is given in the appendix of Carrasco & Kotchoni (2016).

Local Measure of Ill-Posedness

The following provides the derivations for Remark 1. Recall that the simple model consists of:

$$f_{1,k(n)}(e) = \underline{\sigma}_{k(n)}^{-1}\phi(\frac{e}{\underline{\sigma}_{k(n)}}), \quad f_{2,k(n)}(e) = \underline{\sigma}_{k(n)}^{-1}\phi(\frac{e-\mu_{k(n)}}{\underline{\sigma}_{k(n)}}).$$

The only difference between the two densities is the location parameter $\mu_{k(n)}$ in $f_{2,k(n)}$. The total variance, weak and supremum distances between $f_{1,k(n)}$ and $f_{1,k(n)}$ are given below:

i. **Distance in the Weak Norm**
The distance between f_1 and f_2 in the weak norm is:

$$\|f_1 - f_2\|_{weak}^2 = 2\int e^{-\underline{\sigma}_{k(n)}^2\tau^2}\sin(\tau\mu_{k(n)})^2\pi(\tau)d\tau.$$

When $\mu_{k(n)} \to 0$, $\sin(\tau\mu_{k(n)})^2 \to 0$ as well. By the dominated convergence theorem this implies that $\|f_{1,k(n)} - f_{2,k(n)}\|_{weak} \to 0$ as $\mu_{k(n)} \to 0$ regardless of the sequence $\underline{\sigma}_{k(n)} > 0$. The rate at which the distance in weak norm goes to zero when $\mu_{k(n)} \to 0$ can be approximated using the power series for the sine function $\|f_1 - f_2\|_{weak} = |\mu_{k(n)}|\sqrt{2\int e^{-\underline{\sigma}_{k(n)}^2\tau^2}\tau^2\pi(\tau)d\tau} + o(|\mu_{k(n)}|)$. For $\mu_{k(n)} \to 0$, the distance in weak norm declines linearly in $\mu_{k(n)}$. For a specific choice of sequence $(\mu_{k(n)})$ the total variation and supremum distances can be shown to be bounded below. As a result, the ratio with the distance in weak norm is proportional to $|\mu_{k(n)}|^{-1} \to +\infty$.

ii. **Total Variation Distance**
The total variation distance between $f_{1,k(n)}$ and $f_{2,k(n)}$ is bounded below and above by[47]:

$$1 - e^{-\frac{\mu_{k(n)}^2}{8\underline{\sigma}_{k(n)}}} \leq \|f_1 - f_2\|_{TV} \leq \sqrt{2}\left(1 - e^{-\frac{\mu_{k(n)}^2}{8\underline{\sigma}_{k(n)}}}\right)^{1/2}.$$

For any $\varepsilon > 0$, one can pick $\mu_{k(n)} = \pm\underline{\sigma}_{k(n)}\sqrt{-8\log(1-\varepsilon^2)}$ so that $\|f_{1,k(n)} - f_{2,k(n)}\|_{TV} \in [\varepsilon^2/2, \varepsilon]$. However, for the same choice of $\mu_{k(n)}$, the paragraph above

[47]The bounds make use of the relationship between the Hellinger distance $H(f_1, f_2)$: $H(f_1, f_2)^2 \leq \|f_1 - f_2\|_{TV} \leq \sqrt{2}H(f_1, f_2)$. The Hellinger distance between two univariate Gaussian densities is available in closed-form: $H(f,g)^2 = 1 - \sqrt{\frac{2\sigma_f\sigma_g}{\sigma_f^2+\sigma_g^2}}e^{-\frac{1}{4}\frac{(\mu_f-\mu_g)^2}{(\sigma_f^2+\sigma_g^2)}}$.

implies that $\|f_{1,k(n)} - f_{2,k(n)}\|_{weak} \to 0$ as $\underline{\sigma}_{k(n)} \to 0$. The ratio goes to infinity when $\underline{\sigma}_{k(n)} \to 0$:

$$\frac{\|f_{1,k(n)} - f_{2,k(n)}\|_{TV}}{\|f_{1,k(n)} - f_{2,k(n)}\|_{weak}} \geq \underline{\sigma}_{k(n)}^{-1} \frac{1}{\sqrt{2}\varepsilon\sqrt{-8\log(1-\varepsilon^2)}}$$

iii. **Distance in the Supremum Norm**

Using the intermediate value theorem the supremum distance can be computed as:

$$\begin{aligned}\|f_{1,k(n)} - f_{2,k(n)}\|_\infty &= \sup_{e\in\mathbb{R}} \frac{1}{\underline{\sigma}_{k(n)}} \left|\phi\left(\frac{e}{\underline{\sigma}_{k(n)}}\right) - \phi\left(\frac{e-\mu_{k(n)}}{\underline{\sigma}_{k(n)}}\right)\right| \\ &= \sup_{\widetilde{e}\in\mathbb{R}} \frac{|\mu_{k(n)}|}{\underline{\sigma}^2_{k(n)}} \left|\phi'\left(\frac{\widetilde{e}}{\underline{\sigma}_{k(n)}}\right)\right| = \frac{|\mu_{k(n)}|}{\underline{\sigma}^2_{k(n)}}\|\phi'\|_\infty\end{aligned}$$

For any $\varepsilon > 0$, pick $\mu_k = \pm\varepsilon\underline{\sigma}^2_{k(n)}/\|\phi'\|_\infty$ then the distance is supremum norm is fixed, $\|f_{1,k(n)} - f_{2,k(n)}\|_\infty = \varepsilon$, for any strictly positive sequence $\underline{\sigma}_{k(n)} \to 0$. However, the distance in weak norm goes to zero, again the ratio goes to infinity when $\underline{\sigma}_{k(n)} \to 0$:

$$\frac{\|f_{1,k(n)} - f_{2,k(n)}\|_\infty}{\|f_{1,k(n)} - f_{2,k(n)}\|_{weak}} \geq \underline{\sigma}^{-2}_{k(n)}\varepsilon\|\phi'\|_\infty$$

The degree of ill-posedness depends on the bandwidth $\underline{\sigma}_{k(n)}$ in both cases. In order to achieve the approximation rate in Lemma 2, the bandwidth $\underline{\sigma}_{k(n)}$ must be $O(\log[k(n)]^{2/b}/k(n))$. As a result the local measures of ill-posedness for the total variation and supremum distances are:

$$\tau_{TV,n} = O\left(\frac{k(n)}{\log[k(n)]^{2/b}}\right), \quad \tau_{\infty,n} = O\left(\frac{k(n)^2}{\log[k(n)]^{4/b}}\right).$$

Identification in the Stochastic Volatility Model

This section provides an identification result for the SV model in the first empirical application:

$$\begin{aligned}y_t &= \mu_y + \rho_y y_{t-1} + \sigma_t e_{t,1}, \quad e_{t,1} \overset{iid}{\sim} f \\ \sigma_t^2 &= \mu_\sigma + \rho_\sigma \sigma^2_{t-1} + \kappa_\sigma e_{t,2}\end{aligned}$$

with the restriction $e_{t,1} \sim (0,1)$, $|\rho_y|, |\rho_\sigma| < 1$ and $e_{t,2}$ follows a known distribution standardized to have mean zero and unit variance.[48] Suppose the CF $\widehat{\psi}_n$ includes y_t

[48]This assumption makes the derivations easier in terms of notation.

and two lagged observations (y_{t-1}, y_{t-2}) and that the moment generating functions of (y_t, y_{t-1}, y_{t-2}) and $e_{t,1}$ are analytic so that all the moments are finite and characterise the density. Suppose that for two sets of parameters β_1, β_2 we have: $Q(\beta_1) = Q(\beta_2) = 0$. This implies that π almost surely:

$$\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^s(\tau, \beta_1)) = \mathbb{E}(\widehat{\boldsymbol{\psi}}_n^s(\tau, \beta_2)), \quad \forall \tau \in \mathbb{R}^3. \tag{.0.1}$$

Using the notation $\tau = (\tau_1, \tau_2, \tau_3)$ this implies that for any integers $\ell_1, \ell_2, \ell_3 \geq 0$:

$$\begin{aligned} i^{\ell_1+\ell_2+\ell_3}\mathbb{E}_{\beta_1}(y_t^{\ell_1} y_{t-1}^{\ell_2} y_{t-2}^{\ell_3}) &= \frac{d^{\ell_1+\ell_2+\ell_3}\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^s(\tau, \beta_1))}{d\tau_1^{\ell_1} d\tau_2^{\ell_2} d\tau_3^{\ell_3}}\Big|_{\tau=0} \\ &= \frac{d^{\ell_1+\ell_2+\ell_3}\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^s(\tau, \beta_2))}{d\tau_1^{\ell_1} d\tau_2^{\ell_2} d\tau_3^{\ell_3}}\Big|_{\tau=0} = i^{\ell_1+\ell_2+\ell_3}\mathbb{E}_{\beta_2}(y_t^{\ell_1} y_{t-1}^{\ell_2} y_{t-2}^{\ell_3}) \end{aligned}$$

In particular for $\ell_1 = 1, \ell_2 = 0, \ell_3 = 0$, it implies $\mu_{y,1} = \mu_{y,2}$ so that the mean is identified. Then, taking $\ell_1 = 2, \ell_2 = 0, \ell_3 = 0$ implies that $\mathbb{E}_{\beta_1}(\sigma_t^2)/(1-\rho_{y,1}^2) = \mathbb{E}_{\beta_2}(\sigma_t^2)/(1-\rho_{y,2}^2)$. For $\ell_1 = \ell_2 = 1, \ell_3 = 0$ it implies $\rho_{y,1}\mathbb{E}_{\beta_1}(\sigma_t^2)/(1-\rho_{y,1}^2) = \rho_{y,2}\mathbb{E}_{\beta_2}(\sigma_t^2)/(1-\rho_{y,2}^2)$ which, given the result above implies $\rho_{y,1} = \rho_{y,2}$ and then $\mathbb{E}_{\beta_1}(\sigma_t^2) = \mathbb{E}_{\beta_2}(\sigma_t^2)$. The latter implies $\mu_{\sigma,1}/(1-\rho_{\sigma,1}) = \mu_{\sigma,2}/(1-\rho_{\sigma,2})$. Taking $\ell_1 = 2, \ell_2 = 2, \ell_3 = 0$ and $\ell_1 = 2, \ell_2 = 0, \ell_3 = 0$ implies two additional moment conditions (after de-meaning):[49] $\rho_{\sigma,1}\kappa_{\sigma,1}^2/(1-\rho_{\sigma,1}^2) = \rho_{\sigma,2}\kappa_{\sigma,2}^2/(1-\rho_{\sigma,2}^2)$ and $\rho_{\sigma,1}^2\kappa_{\sigma,1}^2/(1-\rho_{\sigma,1}^2) = \rho_{\sigma,2}^2\kappa_{\sigma,2}^2/(1-\rho_{\sigma,2}^2)$. If $\rho_{\sigma,1}, \rho_{\sigma,2} \neq 0$ this imples $\rho_{\sigma,1} = \rho_{\sigma,2}$ and $\kappa_{\sigma,1}, \kappa_{\sigma,2}$ and also $\mu_{\sigma,1} = \mu_{\sigma_2}$.

Overall if $\rho_\sigma \neq 0$, then condition (.0.1) implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, the parametric component is identified. Since $\boldsymbol{\theta}$ is identified, all the moments of σ_t are known. After recentering, this implies that for all $\ell_1 \geq 3$ if $\mathbb{E}_{\boldsymbol{\theta}(\sigma_t^{\ell_1})\neq 0}$:

$$\mathbb{E}_{f_1}(e_{t,1}^{\ell_1}) = \mathbb{E}_{f_1}(e_{t,2}^{\ell_1}). \tag{.0.2}$$

If σ_t is non-negative, which is implied by e.g. $e_{t,2} \sim \chi_1^2$ and parameter constraints, then all moments are stictly positive so that (.0.2) holds. Since the moment generating function is analytic and the first two moments are fixed, (.0.2) implies $f_1 = f_2$. Altogether, if $\rho_\sigma \neq 0$ and $\sigma_t > 0$ then the joint CF of (y_t, y_{t-1}, y_{t-2}) identifies β.

Additional Results on Asymptotic Normality

The following provides two additional results on the root-n asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$. A positive result is given in Proposition .0.1 and a negative result is given in Remark .0.1.

[49] Since μ_y, ρ_y are identified, it is possible to compute $\mathbb{E}([y_t - \mu_y - \rho_y y_{t-1}]^2[y_{t-1} - \mu_y - \rho_y y_{t-2}]^2) = \mathbb{E}(\sigma_t^2\sigma_{t-1}^2)$ from the information given by the CF.

The results apply to DGPs of the form:[50]

$$y_t = g_{obs}(y_{t-1}, \boldsymbol{\theta}, u_t)$$
$$u_t = g_{latent}(u_{t-1}, \boldsymbol{\theta}, e_t)$$

where g_{obs}, g_{latent} are smooth in $\boldsymbol{\theta}$. In this class of models, the data depends on f only through e_t. Examples 1 and 2 satisfy this restriction but dynamic programming models typically don't. The smoothness restriction holds in Example 1 but not Example 2.

Proposition .0.1 (Sufficient Conditions for Asymptotic Normality of $\widehat{\boldsymbol{\theta}}_n$)**.** *If $\mathbb{E}_{\boldsymbol{\theta}_0,f}(\mathbf{y}_t^s)$ and $\mathbb{V}_{\boldsymbol{\theta}_0,f}(\mathbf{y}_t^s)$ do not depend on f then $\widehat{\boldsymbol{\theta}}_n$ is root-n asymptotically normal if:*

$$\mathbb{E}_{\boldsymbol{\theta}_0,f_0}\left(\frac{d\mathbf{y}_t^s}{d\boldsymbol{\theta}'}\left[\begin{pmatrix} 1 & \mathbf{y}_t^{s\prime} \end{pmatrix} \otimes I_{d_{\mathbf{y}}}\right]\right)$$

has rank greater or equal than $d_{\boldsymbol{\theta}}$ when $t \to \infty$.

Proposition .0.1 provides some sufficient conditions for models where the mean and the variance of y_t^s do not vary with f, this holds for Example 1 but not Example 2. This condition requires that $\mathbf{y}_t^s$ varies sufficiently with $\boldsymbol{\theta}$ on average to affect the draws. The proof of the proposition is given at the end of this subsection.

Example 1 (Continued) (Stochastic Volatility)**.** *Recall the DGP for th stochastic volatility model:*

$$y_t = \sum_{j=0}^{t} \rho_y^j \sigma_{t-j} e_{t-j,1} \quad \sigma_t^2 = \sum_{j=0}^{t} \rho_\sigma^j (\mu_\sigma + \kappa_\sigma e_{t-j,2}).$$

It is assumed that the initial condition is $y_0 = \sigma_0 = 0$ in the following. To reduce the number of derivatives to compute, suppose $\mu_\sigma, \kappa_\sigma$ are known and $e_{t-j,2}$ is normalized so that it has mean zero and unit variance. During the estimation $e_{t,1}$ is also restricted to have mean zero, unit variance which implies that the mean of y_t^s and its variance do not depent on f. First, compute the derivatives of y_t^s with respect to ρ_y, ρ_σ:

$$\frac{dy_t^s}{d\rho_y} = \sum_{j=1}^{\infty} j\rho_y^{j-1}\sigma_{t-j}e_{t-j,1}$$
$$\frac{dy_t^s}{d\rho_\sigma} = 0.5\sum_{j=0}^{\infty} \rho_y^j \frac{d\sigma_{t-j}^2}{d\rho_\sigma} e_{t-j,1}/\sigma_{t-j} \quad \textit{where} \quad \frac{d\sigma_{t-j}^2}{d\rho_\sigma} = \sum_{\ell=1}^{t-j} \ell\rho_\sigma^{\ell-1}(\mu_\sigma + \kappa_\sigma e_{\ell,2}).$$

[50] The regressors x_t are omitted here to simplify notation in the proposition and the proof, results with x_t can be derived in a similar way as in this section.

Both derivatives have mean zero, the derivatives of the lags are zero as well. Hence, $\mathbb{E}\left(\frac{d\mathbf{y}_t^s}{d\boldsymbol{\theta}'}\mathbf{y}_t^s\right)$ *must have rank greater than* 2 *for Proposition .0.1 to apply. Now, compute a first set of expectations:*

$$\begin{aligned}
\mathbb{E}(\frac{dy_t^s}{d\rho_y}y_t^s) &= \sum_{j=1}^{t} j\rho_y^{2j-1}\mathbb{E}(\sigma_{t-j}^2)\\
\mathbb{E}(\frac{dy_t^s}{d\rho_y}y_{t-1}^s) &= \sum_{j=0}^{t-1}(j+1)\rho_y^{2j}\mathbb{E}(\sigma_{t-j-1}^2)\\
\mathbb{E}(\frac{dy_t^s}{d\rho_y}y_{t-2}^s) &= \sum_{j=0}^{t-2}(j+2)\rho_y^{2j+1}\mathbb{E}(\sigma_{t-j-2}^2)\\
\mathbb{E}(\frac{dy_{t-1}^s}{d\rho_y}y_t^s) &= \sum_{j=1}^{t-1} j\rho_y^{2j}\mathbb{E}(\sigma_{t-j-1})\\
\mathbb{E}(\frac{dy_{t-2}^s}{d\rho_y}y_t^s) &= \sum_{j=1}^{t-2} j\rho_y^{2j+1}\mathbb{E}(\sigma_{t-j-2}).
\end{aligned}$$

The remaining expectation for ρ_y *can be deduced from the expectations above. Since* $\mathbb{E}(\frac{dy_t^s}{d\rho_y}y_{t-1}^s) > 0$*, these expectations are not all equal to zero as long as* $\mathbb{E}(\sigma_t^2) > 0$*. If* ρ_σ *was known then the rank condition would hold. For the second set of expectations:*

$$\begin{aligned}
\mathbb{E}(\frac{dy_t^s}{d\rho_\sigma}y_t^s) &= \sum_{j=0}^{t}\rho_y^j\mathbb{E}(\frac{d\sigma_{t-j}^2}{d\rho_\sigma}) = \sum_{j=0}^{t}\rho_y^j\sum_{\ell=1}^{t-j}\ell\rho_\sigma^{2\ell-1}\mu_\sigma\\
\mathbb{E}(\frac{dy_t^s}{d\rho_\sigma}y_{t-1}^s) &= \sum_{j=1}^{t}\rho_y^{j+1}\mathbb{E}(\frac{d\sigma_{t-j}^2}{d\rho_\sigma}) = \sum_{j=1}^{t}\rho_y^{j+1}\sum_{\ell=1}^{t-j}\ell\rho_\sigma^{2\ell-1}\mu_\sigma\\
\mathbb{E}(\frac{dy_t^s}{d\rho_\sigma}y_{t-2}^s) &= \sum_{j=2}^{t}\rho_y^{j+2}\mathbb{E}(\frac{d\sigma_{t-j}^2}{d\rho_\sigma}) = \sum_{j=1}^{t}\rho_y^{j+1}\sum_{\ell=1}^{t-j}\ell\rho_\sigma^{2\ell-1}\mu_\sigma.
\end{aligned}$$

The remaining derivatives can be computed similarly. The calculations above imply that the matrix is full rank only if $\rho_\sigma \neq 0$ *and* $\mu_\sigma \neq 0$ *since all the expectations above are zero when either* $\rho_\sigma = 0$ *or* $\mu_\sigma = 0$*.*

Remark .0.1 ($\widehat{\boldsymbol{\theta}}_n$ is generally not an adaptive estimator of $\boldsymbol{\theta}_0$)**.** *For the estimator* $\widehat{\boldsymbol{\theta}}_n$ *to be adaptive*[51] *an orthogonality condition is required, namely:*

$$\frac{d^2Q(\beta_0)}{d\boldsymbol{\theta}df}[f - f_0] = 0, \text{ for all } f \in \mathcal{F}_{osn}.$$

[51] If the estimator is adaptive then $\widehat{\boldsymbol{\theta}}_n$ is root-n asymptotically normal and its asymptotic variance does not depend on $\widehat{f}_n$, i.e. it has the same asymptotic variance as the CF based parametric SMM estimator with f_0 known.

For the CF, this amounts to:

$$\lim_{n\to\infty}\int Real\left(\frac{d\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{d\boldsymbol{\theta}}\overline{\frac{d\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{df}[f-f_0]}\pi(\tau)d\tau\right)=0.$$

Given the restrictions on the DGP and using the notation in the proof of Proposition .0.1, it implies:

$$\lim_{t\to\infty}\int Real\left(i\tau'\frac{dg_t(\boldsymbol{\theta}_0,e_1)}{d\boldsymbol{\theta}}e^{i\tau'[g_t(\boldsymbol{\theta}_0,e_1)-g_t(\boldsymbol{\theta}_0,e_2)]}f_0(e_1)\Delta f(e_2)\pi(\tau)d\tau de_1de_2\right)=0.$$

After some simplification, the orthogonality condition can be re-written as:

$$\lim_{t\to\infty}\int \tau'\frac{dg_t(\boldsymbol{\theta}_0,e_1)}{d\boldsymbol{\theta}}sin\left(\tau'[g_t(\boldsymbol{\theta}_0,e_1)-g_t(\boldsymbol{\theta},e_2)]\right)f_0(e_1)\Delta f(e_2)\pi(\tau)d\tau de_1de_2=0.$$

This function is even in τ so that it does not average out over τ in general when π is chosen to be the Gaussian or the exponential density with mean-zero. Hence, the orthogonality condition holds if the integral of $\frac{dg_t(\boldsymbol{\theta}_0,e_1)}{d\boldsymbol{\theta}}sin\left(\tau'[g_t(\boldsymbol{\theta}_0,e_1)-g_t(\boldsymbol{\theta},e_2)]\right)f_0(e_1)\Delta f(e_2)$ over e_1 and e_2 is zero. This is the case if $g_t(\boldsymbol{\theta}_0,e_1)$ is separable in e_1 and f_0, f are symmetric densities which is quite restrictive.

Proof of Proposition .0.1. Chen & Pouzo (2015), pages 1031-1033 and their Remark A.1, implies that $\widehat{\boldsymbol{\theta}}_n$ is root-n asymptotically normal if:

$$\lim_{n\to\infty}\text{diag}_{v\in\overline{V},v_{\boldsymbol{\theta}}\neq 0}\frac{1}{\|v_{\boldsymbol{\theta}}\|_1^2\int\left|\frac{d\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{d\boldsymbol{\theta}}v_{\boldsymbol{\theta}}+\frac{d\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{df}[v_f]\right|^2\pi(\tau)d\tau>0.}$$

By definition of $\overline{V}$ the vector $v=(v_{\boldsymbol{\theta},v_f})$ has the form $v_{\boldsymbol{\theta}\in\mathbb{R}^{d_{\boldsymbol{\theta}}}}$ and $v_f=\sum_{j=0}^{\infty}a_j[f_j-f_0]$ for a sequence $(a_1,a_2,\dots)$ in $\mathbb{R}$ and $(f_1,f_2,\dots)$ such that $(\boldsymbol{\theta}_j,f_j)\in\mathcal{B}_{osn}$ for some $\boldsymbol{\theta}_j$. To prove the result, we can proceed by contradiction suppose that for some non-zero $v_{\boldsymbol{\theta}}$ and a v_f:

$$\int\left|\frac{d\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{d\boldsymbol{\theta}}v_{\boldsymbol{\theta}}+\frac{d\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{df}[v_f]\right|^2\pi(\tau)d\tau=0. \tag{.0.3}$$

This implies that $\frac{d\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{d\boldsymbol{\theta}}v_{\boldsymbol{\theta}}+\frac{d\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{df}[v_f]=0$ for all τ (π almost surely). This implies that the following holds:

$$\frac{d\mathbb{E}(\widehat{\psi}_n^s(0,\beta_0))}{d\boldsymbol{\theta}}v_{\boldsymbol{\theta}}+\frac{d\mathbb{E}(\widehat{\psi}_n^s(0,\beta_0))}{df}[v_f]=0 \tag{.0.4}$$

$$\frac{d^2\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{d\boldsymbol{\theta}d\tau}\Big|_{\tau=0}v_{\boldsymbol{\theta}}+\frac{d^2\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{dfd\tau}[v_f]\Big|_{\tau=0}=0 \tag{.0.5}$$

$$\frac{d^3\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{d\boldsymbol{\theta}d\tau d\tau_\ell}\Big|_{\tau=0}v_{\boldsymbol{\theta}}+\frac{d^3\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta_0))}{dfd\tau d\tau_\ell}\Big|_{\tau=0}[v_f]=0 \tag{.0.6}$$

for all $\ell = 1, \ldots, d_{\mathbf{y}}$. To simplify notation the following will be used: $f(e) = f(e_1) \times \cdots \times f(e_t)$ and $\Delta f_j(e) = [f_k(e_1) - f_0(e_1)]f_0(e_2) \times \cdots \times f_0(e_t) + f_0(e_1)[f_j(e_2) - f_0(e_2)]f_0(e_3) \times \cdots \times f_0(e_t) + \cdots + f_0(e_1) \ldots f_0(e_{t-1})[f_j(e_t) - f_0(e_t)]$ and $\mathbf{y}_t^s = g_t(\boldsymbol{\theta}, e_t^s, \ldots, e_1^s)$ (the dependence on x is removed for simplicity). The first order derivatives can be written as:

$$\frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_t^s(\tau, \beta_0))}{d\boldsymbol{\theta}} = \int i\tau' \frac{dg_t(\boldsymbol{\theta}_0, e)}{d\boldsymbol{\theta}} e^{i\tau' g_t(\boldsymbol{\theta}_0, e)} f_0(e) de$$
$$\frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_t^s(\tau, \beta_0))}{df}[v_f] = \sum_{j=0}^{\infty} a_j \int e^{i\tau' g_t(\boldsymbol{\theta}_0, e)} \Delta f_j(e) de$$

For $\tau = 0$ this yields $\frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_t^s(0, \beta_0))}{d\boldsymbol{\theta}} = 0$ and $\frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_t^s(0, \beta_0))}{df}[v_f] = 0$, so equality (.0.4) holds automatically. Taking derivatives and setting $\tau = 0$ again implies:

$$\frac{d^2\mathbb{E}(\widehat{\boldsymbol{\psi}}_t^s(\tau, \beta_0))}{d\boldsymbol{\theta} d\tau}\Big|_{\tau=0} = i \int \frac{dg_t(\boldsymbol{\theta}_0, e)}{d\boldsymbol{\theta}'} f_0(e) de$$
$$\frac{d^2\mathbb{E}(\widehat{\boldsymbol{\psi}}_t^s(\tau, \beta_0))}{df d\tau}[v_f]\Big|_{\tau=0} = i \sum_{j=0}^{\infty} a_j \int g_t(\boldsymbol{\theta}_0, e) \Delta f_j(e) de$$

If $\mathbb{E}(\mathbf{y}_t^s)$ does not depend on f then $\int g_t(\boldsymbol{\theta}_0, e) \Delta f_j(e) de = 0$ for all j and $\frac{d^2\mathbb{E}(\widehat{\boldsymbol{\psi}}_t^s(\tau, \beta_0))}{df d\tau}[v_f]\Big|_{\tau=0} = 0$ holds automatically. This implies that condition (.0.5) becomes:

$$\mathbb{E}\left(\frac{d\mathbf{y}_t^s}{d\boldsymbol{\theta}}\right) v_{\boldsymbol{\theta}=0} \tag{.0.7}$$

If $\mathbb{E}\left(\frac{d\mathbf{y}_t^s}{d\boldsymbol{\theta}}\right)$ has rank greater or equal than $d_{\boldsymbol{\theta}}$ then condition (.0.7) holds only if $v_{\boldsymbol{\theta}\neq 0}$; this is a contradiction. If the rank is less than $d_{\boldsymbol{\theta}}$, then taking derivatives with respect to τ again yields $\frac{d^3\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^s(0, \beta_0))}{df d\tau d\tau'}\Big|_{\tau=0}[v_f] = -\sum_{j=0}^{\infty} a_j \int g_t(\boldsymbol{\theta}, e) g_t(\boldsymbol{\theta}, e)' \Delta f_j(e) de = 0$ assuming $\mathbb{E}(\mathbf{y}_t^s \mathbf{y}_t^{s\prime})$ does not depend on f. Computing the other derivatives imply that condition (.0.6) becomes $-v_{\boldsymbol{\theta}'} \int \frac{dg(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}'} g(\boldsymbol{\theta}_0, e) f_0(e) de$ i.e.:

$$v_{\boldsymbol{\theta}'} \mathbb{E}\left(\frac{d\mathbf{y}_t^s}{d\boldsymbol{\theta}'} \mathbf{y}_{t,\ell}^s\right) = 0 \text{ for all } \ell = 1, \ldots, d_{\mathbf{y}}. \tag{.0.8}$$

Then, stacking conditions (.0.7)-(.0.8) together implies:

$$v_{\boldsymbol{\theta}'} \mathbb{E}\left(\frac{d\mathbf{y}_t^s}{d\boldsymbol{\theta}'} \left[\begin{pmatrix} 1 & \mathbf{y}_t^{s\prime} \end{pmatrix} \otimes I_{d_{\mathbf{y}}}\right]\right) = 0. \tag{.0.9}$$

If the matrix has rank greater or equal to $d_{\boldsymbol{\theta}}$ then it implies $v_{\boldsymbol{\theta}=0}$ which is a contradiction. Hence (.0.3) holds only if $v_{\boldsymbol{\theta}=0}$ which proves the result. □

Proofs for the Main Results

The proofs for the main results allow for a bounded linear operator B, as in Carrasco & Florens (2000), to weight the moments. In the appendices, the operator is assumed to be fixed:

$$\widehat{Q}_n^S(\beta) = \int \left| B\widehat{\psi}_n(\tau) - B\widehat{\psi}_n^S(\tau,\beta) \right|^2 \pi(\tau) d\tau.$$

Since B is bounded linear there exists a $M_B > 0$ such that for any two CFs:

$$\int \left| B\widehat{\psi}_n(\tau) - B\widehat{\psi}_n^S(\tau,\beta) \right|^2 \pi(\tau) d\tau \leq M_B^2 \int \left| \widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau,\beta) \right|^2 \pi(\tau) d\tau.$$

As a result, the rate of convergence for the objective function with the weighting B is the same as the rate of convergence without.[52]

Properties of the Mixture Sieve

Lemma .0.1 (Kruijer et al. (2010))**.** *Suppose that f is a continuous univariate density satisfying:*

i. Smoothness: f is r-times continuously differentiable with bounded r-th derivative.

ii. Tails: f has exponential tails, i.e. there exists $\bar{e}, M_{f_1}, a, b > 0$ such that:

$$f_1(e) \leq M_{f_1} e^{-a|e|^b}, \forall |e| \geq \bar{e}.$$

iii. Monotonicity in the Tails: f is strictly positive and there exists $\underline{e} < \bar{e}$ such that f_S is weakly decreasing on $(-\infty, \underline{e}]$ and weakly increasing on $[\bar{e}, \infty)$.

Let $\mathcal{F}_k$ be the sieve space consisting of Gaussian mixtures with the following restrictions:

iv. Bandwidth: $\sigma_j \geq \underline{\sigma}_k = O(\frac{\log[k(n)]^{2/b}}{k})$.

v. Location Parameter Bounds: $\mu_j \in [-\overline{\mu}_k, \overline{\mu}_k]$.

vi. Growth Rate of Bounds: $\overline{\mu}_k = O\left(\log[k]^{1/b}\right)$.

Then there exists $\Pi_k f \in \mathcal{F}_k$, a mixture sieve approximation of f, such that as $k \to \infty$:

$$\|f - \Pi_k f\|_{\mathcal{F}} = O\left(\frac{\log[k(n)]^{2r/b}}{k(n)^r}\right)$$

where $\|\cdot\|_{\mathcal{F}} = \|\cdot\|_{TV}$ or $\|\cdot\|_\infty$.

[52] For results on estimating the optimal B see Carrasco & Florens (2000); Carrasco et al. (2007a). Using their method would lead to $M_{\widehat{B}} \to \infty$ as $n \to \infty$ resulting in a slower rate of convergence for $\widehat{\beta}_n$. Further investigation of this effect and the resulting rate of convergence are left to future research.

Proof of Lemma 3. :

The difference between e_t^s and $\widetilde{e}_t^s$ can be split into two terms:

$$\sum_{j=1}^{k(n)} \left(\mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \omega_l, \sum_{l=0}^{j} \omega_l]} - \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \right) \left(\mu_j + \sigma_j Z_{t,j}^s \right) \tag{.0.1}$$

$$\sum_{j=1}^{k(n)} \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \left(\mu_j - \widetilde{\mu}_j + [\sigma_j - \widetilde{\sigma}_j] Z_{t,j}^s \right). \tag{.0.2}$$

To bound the term (.0.1) in expectation, combine the fact that $|\mu_j| \leq \overline{\mu}_{k(n)}$, $|\sigma_j| \leq \overline{\sigma}$ and ν_t^s and $Z_{t,j}^s$ are independent so that:

$$\left[\mathbb{E} \left(\sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left| \sum_{j=1}^{k(n)} \left(\mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \omega_l, \sum_{l=0}^{j} \omega_l]} - \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \right) \left(\mu_j + \sigma_j Z_{t,j}^s \right) \right|^2 \right) \right]^{1/2}$$
$$\leq \sum_{j=1}^{k(n)} \left[\mathbb{E} \left(\sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left| \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \omega_l, \sum_{l=0}^{j} \omega_l]} - \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \right|^2 \right) \right]^{1/2} \left(\overline{\mu}_{k(n)} + \overline{\sigma} \mathbb{E} \left(|Z_{t,j}^s|^2 \right)^{1/2} \right).$$

The last term is bounded above by $\overline{\mu} + \overline{\sigma} C_Z$. Next, note that

$$\mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \omega_l, \sum_{l=0}^{j} \omega_l]} - \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \in \{-1, 0, 1\}$$

so that:

$$\mathbb{E} \left(\sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left| \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \omega_l, \sum_{l=0}^{j} \omega_l]} - \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \right|^2 \right)$$
$$= \mathbb{E} \left(\sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left| \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \omega_l, \sum_{l=0}^{j} \omega_l]} - \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \right| \right).$$

Also, for any j: $|\sum_{l=0}^{j} \widetilde{\omega}_l - \sum_{l=0}^{j} \omega_l| \leq \sum_{l=0}^{j} |\widetilde{\omega}_l - \sum_{l=0}^{j} \omega_l| \leq \left(\sum_{l=0}^{j} |\widetilde{\omega}_l - \omega_l|^2 \right)^{1/2} \leq \|\widetilde{\omega} - \omega\|_2 \leq \delta$. Following a similar approach to Chen et al. (2003):

$$\left[\mathbb{E} \left(\sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left| \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \omega_l, \sum_{l=0}^{j} \omega_l]} - \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \right| \right) \right]^{1/2}$$
$$\leq \left[\mathbb{E} \left(\sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left| \mathbb{1}_{\nu_t^s \in [(\sum_{l=0}^{j-1} \widetilde{\omega}_l) - \delta, (\sum_{l=0}^{j} \widetilde{\omega}_l) + \delta]} - \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \right| \right) \right]^{1/2}$$
$$= \left[\left([(\sum_{l=0}^{j} \widetilde{\omega}_l) + \delta] - [(\sum_{l=0}^{j-1} \widetilde{\omega}_l) - \delta] - [(\sum_{l=0}^{j} \widetilde{\omega}_l) - (\sum_{l=0}^{j-1} \widetilde{\omega}_l)] \right) \right]^{1/2} = \sqrt{2\delta}.$$

Overall the term (.0.1) is bounded above by $\sqrt{2}(1 + C_Z) \left(\overline{\mu}_{k(n)} + \overline{\sigma} + k(n) \right) \sqrt{\delta}$. The term

(.0.2) can be bounded above by using the simple fact that $0 \leq \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \leq 1$ and:

$$\begin{aligned}
&\left[\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left| \sum_{j=1}^{k(n)} \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \widetilde{\omega}_l, \sum_{l=0}^{j} \widetilde{\omega}_l]} \left(\mu_j - \widetilde{\mu}_j + [\sigma_j - \widetilde{\sigma}_j] Z_{t,j}^s\right) \right|^2 \right)\right]^{1/2} \\
&\leq \sum_{j=1}^{k(n)} \left[\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left|(\mu_j - \widetilde{\mu}_j) + [\sigma_j - \widetilde{\sigma}_j] Z_{t,j}^s\right|^2\right)\right]^{1/2} \\
&\leq \sum_{j=1}^{k(n)} \sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left(|\mu_j - \widetilde{\mu}_j| + |\sigma_j - \widetilde{\sigma}_j| C_Z\right) \\
&\leq (1 + C_Z) \sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left(\sum_{j=1}^{k(n)} |\mu_j - \widetilde{\mu}_j|^2 + |\sigma_j - \widetilde{\sigma}_j|^2\right)^{1/2} \leq (1 + C_Z)\delta.
\end{aligned}$$

Without loss of generality assume that $\delta \leq 1$ so that:

$$\left[\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\widetilde{\omega},\widetilde{\mu},\widetilde{\sigma})\|_2 \leq \delta} \left|e_t^s - \widetilde{e}_t^s\right|^2\right)\right]^{1/2} \leq 2\sqrt{2}(1 + C_Z)\left(1 + \overline{\mu}_{k(n)} + \overline{\sigma} + k(n)\right)\delta^{1/2}.$$

which concludes the proof. □

Lemma .0.2 (Properties of the Tails Distributions)**.** *Let $\overline{\xi} \geq \xi_1, \xi_2 \geq \underline{\xi} > 0$. Let $\nu_{t,1}^s$ and $\nu_{t,2}^s$ be uniform $\mathcal{U}_{[0,1]}$ draws and:*

$$e_{t,1}^s = -\left(\frac{1}{\nu_{t,1}^s} - 1\right)^{\frac{1}{2+\xi_1}}, \quad e_{t,2}^s = \left(\frac{1}{1-\nu_{t,2}^s} - 1\right)^{\frac{1}{2+\xi_2}}.$$

The densities of $e_{t,1}^s, e_{t,2}^s$ satisfy $f_{e_{t,1}^s}(e) \sim e^{-3-\xi_1}$ as $e \to -\infty$, $f_{e_{t,2}^s}(e) \sim e^{-3-\xi_2}$ as $e \to +\infty$. There exists a finite C bounding the second moments $\mathbb{E}\left(|e_{t,1}^s|^2\right) \leq C < \infty$ and $\mathbb{E}\left(|e_{t,2}^s|^2\right) \leq C < \infty$. Furthermore, the draws $y_{t,1}^s$ and $y_{t,2}^s$ are L^2-smooth in ξ_1 and ξ_2 respectively:

$$\begin{aligned}
\left[\mathbb{E}\left(sup_{|\xi_1 - \widetilde{\xi}_1| \leq \delta} |e_{t,1}^s(\xi_1) - e_{t,1}^s(\widetilde{\xi}_1)|^2\right)\right]^{1/2} &\leq C\delta \\
\left[\mathbb{E}\left(sup_{|\xi_2 - \widetilde{\xi}_2| \leq \delta} |e_{t,2}^s(\xi_2) - e_{t,2}^s(\widetilde{\xi}_2)|^2\right)\right]^{1/2} &\leq C\delta
\end{aligned}$$

Where the constant C only depends on $\underline{\xi}$ and $\overline{\xi}$.

Proof of Lemma .0.2. :
To reduce notation, the t and s subscripts will be dropped in the following. The proof is similar for both e_1 and e_2 so the proof is only given for e_1.

First, the densities of e_1 and e_2 are derived, the first two results follow. Noting that the draws are defined using quantile functions, inverting the formula yields: $\nu_1 = \frac{1}{1 - e_1^{2+\xi_1}}$.

This is a proper CDF on $(-\infty, 0]$ since $e_1 \to \frac{1}{1-e_1^{2+\xi_1}}$ is increasing and has limits 0 at $-\infty$ and 1 at 0. Its derivative is the density function: $(2+\xi_1)\frac{e_1^{1+\xi_1}}{(1-e_1^{2+\xi_1})^2}$ which is continuous on $(-\infty, 0]$ and has an asymptote at $-\infty$: $(2+\xi_1)\frac{e_1^{1+\xi_1}}{(1-e_1^{2+\xi_1})^2} \times e_1^{3+\xi_1} \to (2+\xi_1)$ as $e_1 \to -\infty$. Since $\xi_1 \in [\underline{\xi}, \overline{\xi}]$ with $0 < \underline{\xi}$ then $\mathbb{E}|e_1|^2 \leq C < \infty$ for some finite $C > 0$. Similar results hold for e_2 which has density $(2+\xi_2)\frac{e_2^{1+\xi_2}}{(1+e_2^{2+\xi_2})^2}$ on $[0, +\infty)$.

Second, $\xi_1 \to e_1(\xi_1)$ is shown to be L^2-smooth. Let $|\xi_1 - \widetilde{\xi}_1| \leq \delta$, using the mean value theorem, for each ν_1 there exists an intermediate value $\check{\xi}_1 \in [\xi_1, \widetilde{\xi}_1]$ such that:

$$\left(\frac{1}{\nu_1} - 1\right)^{\frac{1}{2+\xi_1}} - \left(\frac{1}{\nu_1} - 1\right)^{\frac{1}{2+\widetilde{\xi}_1}} = \frac{1}{2+\check{\xi}_1} \log(\frac{1}{\nu_1} - 1) \left(\frac{1}{\nu_1} - 1\right)^{\frac{1}{2+\check{\xi}_1}} (\xi_1 - \widetilde{\xi}_1).$$

The first part is bounded above by $1/(2+\underline{\xi})$, the second part is bounded above by:

$$\log(\frac{1}{\nu_1} + 1) \left(\frac{1}{\nu_1} + 1\right)^{\frac{1}{2+\underline{\xi}}}$$

and the last term is bounded above, in absolute value, by δ.

Finally, in order to conclude the proof, the following integral needs to be finite:

$$\int_0^1 \log(\frac{1}{\nu_1} + 1) \left(\frac{1}{\nu_1} + 1\right)^{\frac{2}{2+\underline{\xi}}} d\nu_1.$$

By a change of variables, it can be re-written as:

$$\int_2^\infty \log(\nu) \nu^{\frac{2}{2+\underline{\xi}} - 2} d\nu.$$

Since $\frac{2}{2+\underline{\xi}} - 2 < -1$, the integral is finite and thus:

$$\left[\mathbb{E}\left(\sup_{|\xi_1 - \widetilde{\xi}_1| \leq \delta} |e_{t,1}^s(\xi_1) - e_{t,1}^s(\widetilde{\xi}_1)|^2\right)\right]^{1/2} \leq \frac{\delta}{2+\underline{\xi}} \sqrt{\int_2^\infty \log(\nu) \nu^{\frac{2}{2+\underline{\xi}} - 2} d\nu}.$$

□

Proof of Lemma 2. The proof proceeds by recursion. Denote $\pi_{k(n)} f_j \in \mathcal{B}\mathcal{B}_{k(n)}$ the mixture approximation of f_j from Lemma .0.1. For $d_e = 1$, Lemma .0.1 implies

$$\|f_1 - \Pi_{k(n)} f_1\|_{TV} = O(\frac{\log[k(n)]^{r/b}}{k(n)^r}), \quad \|f_1 - \Pi_{k(n)} f_1\|_\infty = O(\frac{\log[k(n)]^{r/b}}{k(n)^r}).$$

Suppose the result holds for $f_1 \times \cdots \times f_{d_e}$. Let $f = f_1 \times \cdots \times f_{d_e} \times f_{d_e+1}$; let:

$$d_{t+1} = f_1 \times \cdots \times f_{d_e} \times f_{d_e+1} - \Pi_{k(n)} f_1 \times \cdots \times \Pi_{k(n)} f_{d_e} \times \Pi_{k(n)} f_{d_e+1}$$
$$d_t = f_1 \times \cdots \times f_{d_e} - \Pi_{k(n)} f_1 \times \cdots \times \Pi_{k(n)} f_{d_e}.$$

The difference can be re-written as a recursion:

$$d_{t+1} = d_t f_{d_e+1} + \Pi_{k(n)} f_1 \times \cdots \times \Pi_{k(n)} f_{d_e} \left(f_{d_e+1} - \Pi_{k(n)} f_{d_e+1} \right).$$

Since $\int f_{d_e+1} = \int \Pi_{k(n)} f_1 \times \cdots \times \Pi_{k(n)} f_{d_e} = 1$, the total variation distance is:

$$\|d_{t+1}\|_{TV} \leq \|d_t\|_{TV} + \|f_{d_e+1} - \Pi_{k(n)} f_{d_e+1}\|_{TV} = O\left(\frac{\log[k(n)]^{r/b}}{k(n)^r} \right).$$

And the supremum distance is:

$$\|d_{t+1}\|_\infty \leq \|d_t\|_\infty \|f_{d_e+1}\|_\infty + \|\Pi_{k(n)} f_1 \times \cdots \times \Pi_{k(n)} f_{d_e}\|_\infty \|f_{d_e+1} - \Pi_{k(n)} f_{d_e+1}\|_\infty$$
$$\leq \|d_t\|_\infty \left(\|f_{d_e+1}\|_\infty + \|f_1 \times \cdots \times f_{d_e}\|_\infty \|f_{d_e+1} - \Pi_{k(n)} f_{d_e+1}\|_\infty \right) = O\left(\frac{\log[k(n)]^{r/b}}{k(n)^r} \right).$$

□

Definition .0.1 (Pseudo-Norm $\|\cdot\|_m$ on $\mathcal{B}_{k(n)}$)**.** *Let $\beta_1, \beta_2 \in \mathcal{B}_{k(n)}$ where $\beta_l = (\boldsymbol{\theta}_l, f_l), l = 1, 2$ with $f_j = f_{1,j} \times \ldots f_{d_e,j}$, each $f_{l,j}$ as in definition 1. The pseudo-norm $\|\cdot\|_m$ is the ℓ^2 norm on $(\boldsymbol{\theta}, \omega, \mu, \sigma, \xi)$, the associated distance is:*

$$\|\beta_1 - \beta_2\|_m = \|(\boldsymbol{\theta}_1, \omega_1, \mu_1, \sigma_1, \xi_1) - (\boldsymbol{\theta}_2, \omega_2, \mu_2, \sigma_2, \xi_2)\|_2$$

using the vector notation $\omega_1 = (\omega_{1,1}, \ldots, \omega_{1,k(n)+2}, \ldots, \omega_{d_e,1}, \ldots, \omega_{d_e,k(n)+2})$ for $\boldsymbol{\theta}, \omega, \mu, \sigma, \xi$.

Remark .0.1. *Using lemma 6 in Kruijer et al. (2010), for any two mixtures f_1, f_2 in $\mathcal{B}_{k(n)}$:*

$$\|f_1 - f_2\|_\infty \leq C_\infty \frac{\|f_1 - f_2\|_m}{\underline{\sigma}^2_{k(n)}}, \quad \|f_1 - f_2\|_{TV} \leq C_{TV} \frac{\|f_1 - f_2\|_m}{\underline{\sigma}_{k(n)}}$$

for some constants $C_\infty, C_{TV} > 0$. The result extends to $d_e > 1$, for instance when $d_e = 2$:

$$f_1^1 f_1^2 - f_2^1 f_2^2 = f_1^1 (f_1^2 - f_2^2) + (f_1^2 - f_2^1) f_2^2$$

In total variation distance the difference becomes:

$$\|f_1^1 f_1^2 - f_2^1 f_2^2\|_{TV} \leq = \|f_1^2 - f_2^2\|_{TV} + \|f_1^2 - f_2^1\|_{TV}$$
$$\leq C_{TV} \frac{\|f_1^2 - f_2^2\|_m + \|f_1^2 - f_2^1\|_m}{\underline{\sigma}_{k(n)}} \leq C_{TV,2} \frac{\|f_1 - f_2\|_m}{\underline{\sigma}_{k(n)}}.$$

A recursive argument yields the result for arbitrary $d_e > 1$. In supremum distance a similar result holds assuming $\|f_1^j\|_\infty, \|f_2^j\|_\infty$, with $j = 1, 2$, are bounded above by a constant.

Consistency

Assumption 2′ (Data Generating Process - L^2-Smoothness). *y_t^s is simulated according to the dynamic model (3.1)-(3.2) where g_{obs} and g_{latent} satisfy the following L^2-smoothness conditions for some $\gamma \in (0,1]$ and any $\delta \in (0,1)$:*

$y(i)'$. *For some $0 \leq \overline{C}_1 < 1$:*

$$\left[\mathbb{E}\left(sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta}\|g_{obs}(y_t^s(\beta_1),x_t,\beta_1,u_t^s(\beta_1)) - g_{obs}(y_t^s(\beta_2),x_t,\beta_1,u_t^s(\beta_1))\|^2\middle|y_t^s(\beta_1),y_t^s(\beta_2)\right)\right]^{1/2}$$
$$\leq \overline{C}_1\|y_t^s(\beta_1) - y_t^s(\beta_2)\|$$

$y(ii)'$. *For some $0 \leq \overline{C}_2 < \infty$:*

$$\left[\mathbb{E}\left(sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta}\|g_{obs}(y_t^s(\beta_1),x_t,\beta_1,u_t^s(\beta_1)) - g_{obs}(y_t^s(\beta_1),x_t,\beta_2,u_t^s(\beta_1))\|^2\right)\right]^{1/2} \leq \overline{C}_2\delta^{\gamma}$$

$y(iii)'$. *For some $0 \leq \overline{C}_3 < \infty$:*

$$\left[\mathbb{E}\left(sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta}\|g_{obs}(y_t^s(\beta_1),x_t,\beta_1,u_t^s(\beta_1)) - g_{obs}(y_t^s(\beta_1),x_t,\beta_1,u_t^s(\beta_2))\|^2\middle|u_t^s(\beta_1),u_t^s(\beta_2)\right)\right]^{1/2}$$
$$\leq \overline{C}_3\|u_t^s(\beta_1) - u_t^s(\beta_2)\|^{\gamma}$$

$u(i)'$. *For some $0 \leq \overline{C}_4 < 1$*

$$\left[\mathbb{E}\left(sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta}\|g_{latent}(u_{t-1}^s(\beta_1),\beta,e_t^s(\beta_1)) - g_{latent}(u_{t-1}^s(\beta_2),\beta,e_t^s(\beta_1))\|^2\right)\right]^{1/2}$$
$$\leq \overline{C}_4\|u_{t-1}^s(\beta_1) - u_{t-1}^s(\beta_2)\|$$

$u(ii)'$. *For some $0 \leq \overline{C}_5 < \infty$:*

$$\mathbb{E}\left(sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta}\|g_{latent}(u_{t-1}^s(\beta_1),\beta_1,e_t^s(\beta_1)) - g_{latent}(u_{t-1}^s(\beta_1),\beta_2,e_t^s(\beta_1))\|^2\right) \leq \overline{C}_5\delta^{\gamma}$$

$u(iii)'$. *For some $0 \leq \overline{C}_5 < \infty$:*

$$\mathbb{E}\left(sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta}\|g_{latent}(u_{t-1}^s(\beta_1),\beta_1,e_t^s(\beta_1)) - g_{latent}(u_{t-1}^s(\beta_1),\beta_1,e_t^s(\beta_2))\|^2\middle|e_t^s(\beta_1),e_t^s(\beta_2)\right)$$
$$\leq \overline{C}_6\|e_1 - e_2\|$$

for $\|\beta_1 - \beta_2\|_{\mathcal{B}} = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \|f_1 - f_2\|_{\infty}$ or $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \|f_1 - f_2\|_{TV}$.

Proof of Lemma 4: First note that the cosine and sine functions are uniformly Lispchitz on the real line with Lipschitz coefficient 1. This implies for any two $(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x})$ and any $\tau \in \mathbb{R}^{d_\tau}$:

$$|\cos(\tau'(\mathbf{y}_1,\mathbf{x})) - \cos(\tau'(\mathbf{y}_2,\mathbf{x}))| \leq |\tau'(\mathbf{y}_1 - \mathbf{y}_2, 0)| \leq \|\tau\|_{\infty}\|\mathbf{y}_1 - \mathbf{y}_2\|$$
$$|\sin(\tau'(\mathbf{y}_1,\mathbf{x})) - \sin(\tau'(\mathbf{y}_2,\mathbf{x}))| \leq |\tau'(\mathbf{y}_1 - \mathbf{y}_2, 0)| \leq \|\tau\|_{\infty}\|\mathbf{y}_1 - \mathbf{y}_2\|.$$

As a result, the moment function is also Lipschitz in $\mathbf{y}, \mathbf{x}$:

$$|e^{i\tau'(\mathbf{y}_1,\mathbf{x})} - e^{i\tau'(\mathbf{y}_2,\mathbf{x})}|\pi(\tau)^{\frac{1}{4}} \leq 2\|\tau\|_\infty \pi(\tau)^{\frac{1}{4}}\|\mathbf{y}_1 - \mathbf{y}_2\|.$$

Since π is chosen to be the Gaussian density, it satisfies $\sup_\tau \|\tau\|_\infty \phi(\tau)^{\frac{1}{4}} \leq C_\pi < \infty$ and $\phi(\tau)^{\frac{1}{2}} \propto \phi(\tau/\sqrt{2})$ which has finite integral.

The Lispschitz properties of the moments combined with the conditions properties of π imply that the L^2-smoothness of the moments is implied by the L^2-smoothness of the simulated data itself. As a result, the remainder of the proof establishes the L^2-smoothness of $\mathbf{y}_t^s$.

First note that since $\mathbf{y}_t = (y_t, \ldots, y_{t-L})$:

$$\|\mathbf{y}_t(\beta_1) - \mathbf{y}_t(\beta_2)\| \leq \sum_{j=1}^{L} \|y_{t-j}(\beta_1) - y_{t-j}(\beta_2)\|.$$

To bound the term in $\mathbf{y}$ above, it suffices to bound the expression for each term y_t with arbitrary $t \geq 1$. Assumptions 2, 2′ imply that, for some $\gamma \in (0, 1]$:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m} \|y_t(\beta_1) - y_t(\beta_2)\|^2\right)\right]^{1/2} \leq \overline{C}_1 \left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m} \|y_{t-1}(\beta_1) - y_{t-1}(\beta_2)\|^2\right)\right]^{1/2} + \overline{C}_2 \frac{\delta^\gamma}{\underline{\sigma}_{k(n)}^{2\gamma}}$$
$$+ \overline{C}_3 \left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m} \|u_t(\beta_1) - u_t(\beta_2)\|^2\right)\right]^{\gamma/2}.$$

The term $\frac{\delta^\gamma}{\underline{\sigma}_{k(n)}^{2\gamma}}$ comes from the fact that $\|\beta_1 - \beta_2\|_\infty \leq \frac{\|\beta_1-\beta_2\|_m}{\underline{\sigma}_{k(n)}^2}$ and $\|\beta_1 - \beta_2\|_{TV} \leq \frac{\|\beta_1-\beta_2\|_m}{\underline{\sigma}_{k(n)}}$ on $\mathcal{B}_{k(n)}$. Without loss of generality, suppose that $\underline{\sigma}_{k(n)} \leq 1$.[53] Applying this inequality recursively, and using the fact that y_0^s, u_0^s are the same regardless of β, yields:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m} \|y_t(\beta_1) - y_t(\beta_2)\|^2\right)\right]^{1/2}$$
$$\leq \frac{\overline{C}_2}{1-\overline{C}_1} \frac{\delta^\gamma}{\underline{\sigma}_{k(n)}^{2\gamma}} + \overline{C}_3 \sum_{l=0}^{t-1} \overline{C}_1^l \left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m} \|u_{t-l}(\beta_1) - u_{t-l}(\beta_2)\|^2\right)\right]^{\gamma/2}.$$

Using Lemmas 3 and .0.2 and the same approach as above:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m} \|u_t(\beta_1) - u_t(\beta_2)\|^2\right)\right]^{1/2}$$
$$\leq \overline{C}_4 \left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m} \|u_{t-1}(\beta_1) - u_{t-1}(\beta_2)\|^2\right)\right]^{1/2} + \overline{C}_5 \frac{\delta^\gamma}{\underline{\sigma}_{k(n)}^{2\gamma}}$$
$$+ \overline{C}_6 C \left(k(n) + \overline{\mu}_{k(n)} + \overline{\sigma}\right) \delta^{\gamma/2}.$$

[53] Recall that by assumption $\underline{\sigma}_{k(n)} = O(\frac{log[k(n)]^{2/b}}{k(n)})$ goes to zero.

Again, applying this inequality recursively yields:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m}\|u_t(\beta_1)-u_t(\beta_2)\|^2\right)\right]^{1/2}$$
$$\leq \frac{\overline{C}_5}{1-\overline{C}_4}\frac{\delta^\gamma}{\underline{\sigma}^{2\gamma}_{k(n)}}+\frac{\overline{C}_6}{1-\overline{C}_4}C\left(k(n)+\overline{\mu}_{k(n)}+\overline{\sigma}\right)\delta^{\gamma/2}.$$

Putting everything together:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m}\|y_t(\beta_1)-y_t(\beta_2)\|^2\right)\right]^{1/2}$$
$$\leq \frac{\overline{C}_2}{1-\overline{C}_1}\frac{\delta^\gamma}{\underline{\sigma}^{2\gamma}_{k(n)}}+\frac{\overline{C}_3}{1-\overline{C}_1}\left(\frac{\overline{C}_5}{1-\overline{C}_4}\frac{\delta^\gamma}{\underline{\sigma}^{2\gamma}_{k(n)}}+\frac{\overline{C}_6}{1-\overline{C}_4}C\left(k(n)+\overline{\mu}_{k(n)}+\overline{\sigma}\right)\delta^{\gamma/2}\right)^\gamma.$$

Without loss of generality, suppose that $\delta \leq 1$. Then, for some positive constant $\overline{C}$:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m}\|y_t(\beta_1)-y_t(\beta_2)\|^2\right)\right]^{1/2} \leq \overline{C}\max\left(\frac{\delta^{\gamma^2}}{\underline{\sigma}^{2\gamma^2}_{k(n)}}, [k(n)+\overline{\mu}_{k(n)}+\overline{\sigma}]^\gamma\delta^{\gamma^2/2}\right).$$

□

Lemma .0.3 (Covering Numbers)**.** *Under the L^2-smoothness of the DGP (as in Lemma 4), the bracketing number satisfies for $x \in (0,1)$ and some $\overline{C}$:*

$$N_{[\,]}(x, \Psi_{k(n)}(\tau), \|\cdot\|_{L^2})$$
$$\leq (3[k(n)+2]+d_{\boldsymbol{\theta}})\left(2\max(\overline{\mu}_{k(n)},\underline{\sigma})\overline{C}^{2/\gamma^2}\frac{\left(k(n)+\overline{\mu}_{k(n)}+\overline{\sigma}\right)^{2/\gamma}+\underline{\sigma}^4_{k(n)}}{x^{2/\gamma^2}}+1\right)^{3[k(n)+2]+d_{\boldsymbol{\theta}}}.$$

For $\tau \in \mathbb{R}^{d_\tau}$, let $\Psi_{k(n)}(\tau)$ be the set of functions $\Psi_{k(n)}(\tau) = \left\{\beta \to e^{i\tau'(\mathbf{y}_t(\beta),\mathbf{x}_t)}\pi(\tau)^{1/2},\ \beta \in \mathcal{B}_{k(n)}\right\}$. The bracketing entropy of each set $\Psi_{k(n)}(\tau)$ satisfies for some $\widetilde{C}$:

$$\log\left(N_{[\,]}(x, \Psi_{k(n)}(\tau), \|\cdot\|_{L^2}\right) \leq \widetilde{C}k(n)\log[k(n)])|\log\delta|.$$

Using the above, for some $\widetilde{C}_2 < \infty$:

$$\int_0^1 \log^2\left(N_{[\,]}(x, \Psi_k(n), \|\cdot\|_{L^2}\right)dx \leq \widetilde{C}_2 k(n)^2\log[k(n)]^2.$$

Proof of Lemma .0.3: Since $\mathcal{B}_{k(n)}$ is contained in a ball of radius $\max(\overline{\mu}_{k(n)}, \overline{\sigma}, \|\boldsymbol{\theta}\|_\infty)$ in $\mathbb{R}^{3[k(n)+2]+d_{\boldsymbol{\theta}}}$ under $\|\cdot\|_m$, the covering number for $\mathcal{B}_{k(n)}$ can be computed under the $\|\cdot\|_m$ norm using a result from Kolmogorov & Tikhomirov (1959).[54] As a result, the covering number

[54]See also Fenton & Gallant (1996) for an application of this result for the sieve estimation of a density and Coppejans (2001) for a CDF.

$N(x, \mathcal{B}_{k(n)}, \|\cdot\|_m)$ satisfies:

$$N(x, \mathcal{B}_{k(n)}, \|\cdot\|_m) \leq 2\left(3[k(n)+2]+d_{\boldsymbol{\theta}}\right)\left(\frac{2\max(\overline{\mu}_{k(n)}, \overline{\sigma})}{x}+1\right)^{3[k(n)+2]+d_{\boldsymbol{\theta}}}.$$

The rest follows from Lemma 4 and Appendix 3.7. □

Proof of Theorem 1: If the assumptions of Corollary .0.1 hold then the result of Theorem 1 holds as well. The following relates the previous lemmas and assumptions to the required assumption for the corollary.

Assumption 1 implies Assumptions .0.1 and .0.2. Furthermore, by Lemmas 4 and .0.3, Assumptions 1 with 2 (or 2′) imply Assumption .0.4 with $\sqrt{C_n/n} = O(\frac{k(n)^2 \log^2[k(n)]}{\sqrt{n}})$ using the norm $\|\cdot\|_m$. The order of $Q_n(\Pi_{k(n)}\beta_0)$ is given in Lemma 5. This implies that all the assumptions for Corollary .0.1 so that the estimator is consistent if $\sqrt{C_n/n} = o(1)$ which concludes the proof. □

Rate of Convergence

Proof of Lemma 5: First, using the assumption that B is a bounded linear operator:

$$\begin{aligned} Q_n(\Pi_{k(n)}\beta_0) &\leq M_B^2 \int \left|\mathbb{E}\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0)\right)\right|^2 \pi(\tau)d\tau \\ &\leq 3M_B^2\left(\int \left|\mathbb{E}\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \beta_0)\right)\right|^2 \pi(\tau)d\tau + \int \left|\mathbb{E}\left(\psi_n^S(\tau, \beta_0) - \widehat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0)\right)\right|^2 \pi(\tau)d\tau\right) \end{aligned}$$

Each term can be bounded above individually. Re-write the first term in terms of distribution:

$$\left|\mathbb{E}\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \beta_0)\right)\right| = \left|\frac{1}{n}\sum_{t=1}^{n}\int e^{i\tau'(\mathbf{y}_t, \mathbf{x_t})}[f_t^*(\mathbf{y}_t, \mathbf{x_t}) - f_t(\mathbf{y}_t, \mathbf{x_t})]d\mathbf{y}_t d\mathbf{x}_t\right|$$

where f_t is the distribution of $(\mathbf{y}_t(\beta_0), \mathbf{x}_t)$ and f_t the stationary distribution of $(\mathbf{y}_t(\beta_0), \mathbf{x}_t)$. Using the geometric ergodicity assumption, for all τ:

$$\begin{aligned} &\left|\frac{1}{n}\sum_{t=1}^{n}\int e^{i\tau'(\mathbf{y}_t, \mathbf{x_t})}[f_t^*(\mathbf{y}_t, \mathbf{x_t}) - f_t(\mathbf{y}_t, \mathbf{x_t})]d\mathbf{y}_t d\mathbf{x}_t\right| \leq \frac{1}{n}\sum_{t=1}^{n}\int \left|f_t^*(\mathbf{y}_t, \mathbf{x_t}) - f_t(\mathbf{y}_t, \mathbf{x_t})\right| d\mathbf{y}_t d\mathbf{x}_t \\ &= \frac{2}{n}\sum_{t=1}^{n}\|f_t^* - f_t\|_{TV} \leq \frac{2C_\rho}{n}\sum_{t=1}^{n}\rho^t \leq \frac{2C_\rho}{(1-\rho)n} \end{aligned}$$

for some $\rho \in (0,1)$ and $C_\rho > 0$. This yields a first bound:

$$\int \left|\mathbb{E}\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \beta_0)\right)\right|^2 \pi(\tau)d\tau \leq \frac{4C_\rho^2}{(1-\rho)^2}\frac{1}{n^2} = O\left(\frac{1}{n^2}\right).$$

The mixture norm $\|\cdot\|_m$ is not needed here to bound the second term since it involves population CFs. Some changes to the proof of Lemma 4 allows to find bounds in terms of $\|\cdot\|_{\mathcal{B}}$ and $\|\cdot\|_{TV}$ for which Lemma 2 gives the approximation rates.

To bound the second term, re-write the simulated data as:

$$y_t^s = g_{obs,t}(x_t, \ldots, x_1, \beta, e_t^s, \ldots, e_1^s), \quad u_t^s = g_{latent,t}(\beta, e_t^s, \ldots, e_1^s)$$

with $\beta = (\boldsymbol{\theta}, f)$ and $e_t^s \sim f$. Under Assumption 2 or 2′, using the same sequence of shocks (e_t^s):

$$\mathbb{E}\left(\left\|g_{obs,t}(x_t, \ldots, x_1, \beta_0, e_t^s, \ldots, e_1^s) - g_{obs,t}(x_t, \ldots, x_1, \Pi_{k(n)}\beta_0, e_t^s, \ldots, e_1^s)\right\|\right) \leq \overline{C}\|\Pi_{k(n)}f_0 - f_0\|_{\mathcal{B}}^{\gamma}.$$

This is similar to the proof of Lemma 4, first re-write the difference as:

$$\begin{aligned}
&\mathbb{E}\Big(\Big\|g_{obs}(g_{obs,t-1}(x_{t-1}, \ldots, x_1, \beta_0, e_{t-1}^s, \ldots, e_1^s), x_t, \beta_0, g_{latent}(g_{latent,t-1}(\beta_0, e_{t-1}^s, \ldots, e_1^s), \beta_0, e_t^s)) \\
&- g_{obs}(g_{obs,t-1}(x_{t-1}, \ldots, x_1, \Pi_{k(n)}\beta_0, e_{t-1}^s, \ldots, e_1^s), x_t, \Pi_{k(n)}\beta_0, g_{latent}(g_{latent,t-1}(\Pi_{k(n)}\beta_0, e_{t-1}^s, \ldots, e_1^s), \Pi_{k(n)}\beta_0, e_t^s)\Big\|\Big).
\end{aligned}$$

Using Assumptions 2-2′, there is a recursive relationship:

$$\begin{aligned}
&\mathbb{E}\Big(\Big\|g_{obs}(g_{obs,t-1}(x_{t-1}, \ldots, x_1, \beta_0, e_{t-1}^s, \ldots, e_1^s), x_t, \beta_0, g_{latent}(g_{latent,t-1}(\beta_0, e_{t-1}^s, \ldots, e_1^s), \beta_0, e_t^s)) \\
&- g_{obs}(g_{obs,t-1}(x_{t-1}, \ldots, x_1, \Pi_{k(n)}\beta_0, e_{t-1}^s, \ldots, e_1^s), x_t, \Pi_{k(n)}\beta_0, g_{latent}(g_{latent,t-1}(\Pi_{k(n)}\beta_0, e_{t-1}^s, \ldots, e_1^s), \Pi_{k(n)}\beta_0, e_t^s)\Big\|\Big) \\
&\leq \Big[\mathbb{E}\Big(\Big\|g_{obs}(g_{obs,t-1}(x_{t-1}, \ldots, x_1, \beta_0, e_{t-1}^s, \ldots, e_1^s), x_t, \beta_0, g_{latent}(g_{latent,t-1}(\beta_0, e_{t-1}^s, \ldots, e_1^s), \beta_0, e_t^s)) \\
&- g_{obs}(g_{obs,t-1}(x_{t-1}, \ldots, x_1, \Pi_{k(n)}\beta_0, e_{t-1}^s, \ldots, e_1^s), x_t, \Pi_{k(n)}\beta_0, g_{latent}(g_{latent,t-1}(\Pi_{k(n)}\beta_0, e_{t-1}^s, \ldots, e_1^s), \Pi_{k(n)}\beta_0, e_t^s)\Big\|^2\Big)\Big]^{1/2} \\
&\leq \overline{C}_1\Big[\mathbb{E}\Big(\Big\|g_{obs,t-1}(x_{t-1}, \ldots, x_1, \beta_0, e_{t-1}^s, \ldots, e_1^s) - g_{obs,t-1}(x_{t-1}, \ldots, x_1, \Pi_{k(n)}\beta_0, e_{t-1}^s, \ldots, e_1^s)\Big\|^2\Big)\Big]^{1/2} \\
&+ \overline{C_2}\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}^{\gamma} + \overline{C}_3\Big[\mathbb{E}\Big(\Big\|g_{latent,t}(\beta_0, e_t^s, \ldots, e_1^s) - g_{latent,t}(\Pi_{k(n)}\beta_0, e_t^s, \ldots, e_1^s)\Big\|^2\Big)\Big]^{\gamma/2}.
\end{aligned}$$

The last term also has a recursive structure:

$$\begin{aligned}
&\Big[\mathbb{E}\Big(\Big\|g_{latent,t}(\beta_0, e_t^s, \ldots, e_1^s) - g_{latent,t}(\Pi_{k(n)}\beta_0, e_t^s, \ldots, e_1^s)\Big\|^2\Big)\Big]^{1/2} \\
&\leq \overline{C}_4\left[\mathbb{E}\Big(\Big\|g_{latent,t-1}(\beta_0, e_{t-1}^s, \ldots, e_1^s) - g_{latent,t-1}(\Pi_{k(n)}\beta_0, e_{t-1}^s, \ldots, e_1^s)\Big\|^2\Big)\right]^{1/2} + \overline{C}_5\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}^{\gamma}.
\end{aligned}$$

Together these inequalities imply:

$$\begin{aligned}
&\mathbb{E}\Big(\Big\|g_{obs}(g_{obs,t-1}(x_{t-1}, \ldots, x_1, \beta_0, e_{t-1}^s, \ldots, e_1^s), x_t, \beta_0, g_{latent}(g_{latent,t-1}(\beta_0, e_{t-1}^s, \ldots, e_1^s), \beta_0, e_t^s)) \\
&- g_{obs}(g_{obs,t-1}(x_{t-1}, \ldots, x_1, \Pi_{k(n)}\beta_0, e_{t-1}^s, \ldots, e_1^s), x_t, \Pi_{k(n)}\beta_0, g_{latent}(g_{latent,t-1}(\Pi_{k(n)}\beta_0, e_{t-1}^s, \ldots, e_1^s), \Pi_{k(n)}\beta_0, e_t^s)\Big\|\Big) \\
&\leq \frac{1}{1-\overline{C}_1}\left(\overline{C}_2\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}^{\gamma} + \overline{C}_3\frac{\overline{C}_5^{\gamma}}{(1-\overline{C_4})^{\gamma}}\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}^{\gamma^2}\right).
\end{aligned}$$

Recall that $\|\tau\|_\infty \sqrt{\pi(\tau)}$ is bounded above and $\pi(\tau)^{1/4}$ is integrable so that:

$$\int \left| \mathbb{E}\left(e^{i\tau'(\mathbf{y}_t(\beta_0, x_t, \ldots, x_1))} - e^{i\tau'(\mathbf{y}_t(\Pi_{k(n)}\beta_0, x_t, \ldots, x_1))} \right) \right|^2 \pi(\tau) d\tau$$
$$\leq \frac{1}{1-\overline{C}_1} \left(\overline{C}_2 \|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}^{\gamma} + \overline{C}_3 \frac{\overline{C}_5^{\gamma}}{(1-\overline{C}_4)^{\gamma}} \|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}^{\gamma^2} \right) \sup_\tau [\|\tau\|_\infty \sqrt{\pi(\tau)}] \int \pi(\tau)^{1/4} d\tau.$$

To conclude the proof, the difference due to e_t^s needs to be bounded. In order to do so, it suffice to bound the following integral:

$$\int e^{i\tau'(\mathbf{y}_t(y_0, u_0, x_t, \ldots, x_1, \beta_0, e_t^s, \ldots, e_1^s), \mathbf{x}_t)} \left(f_0(e_t^s) \times \cdots \times f_0(e_1^s) - \Pi_{k(n)} f_0(e_t^s) \times \cdots \times \Pi_{k(n)} f_0(e_1^s) \right) f_{\mathbf{x}}(\mathbf{x}_t) de_t^s \ldots de_1^s d\mathbf{x}_t.$$

A direct bound on this integral yields a term of order of $t\|f_0 - \Pi_{k(n)} f_0\|_{TV}$ which increases too fast with t to generate useful rates. Rather than using a direct bound, consider Assumptions 2-2′. The time-series y_t^s can be approximated by another time-series term which only depends on a fixed and finite $(e_t^s, \ldots, e_{t-m}^s)$ for a given integer $m \geq 1$. Making m grow with n at an appropriate rate allows to balance the bias $m\|f_0 - \Pi_{k(n)} f_0\|_{TV}$ (computed from a direct bound) and the approximation due to $m < t$.

The m-approximation rate of y_t is now derived. Let $\beta = (\boldsymbol{\theta}, f) \in \mathcal{B}$, $e_t^s, \ldots, e_1^s \sim f$ and $\widetilde{y}_t^s$ such that $\widetilde{y}_{t-m}^s = 0, \widetilde{u}_{t-m}^s = 0$ and then $\widetilde{y}_j^s = g_{obs}(\widetilde{y}_{j-1}^s, x_j, \beta, \widetilde{u}_j^s), \widetilde{u}_j^s = g_{latent}(\widetilde{u}_{j-1}^s, \beta, e_j^s)$ for $t - m + 1 \leq j \leq t$. Each observation t is approximated by its own time-series. For observation $t - m$, by construction:

$$\mathbb{E}\left(\left\| y_{t-m}^s - \widetilde{y}_{t-m}^s \right\| \right) = \mathbb{E}\left(\left\| y_{t-m}^s \right\| \right) \leq \left[\mathbb{E}\left(\left\| y_{t-m}^s \right\|^2 \right) \right]^{1/2}$$
$$\mathbb{E}\left(\left\| u_{t-m}^s - \widetilde{u}_{t-m}^s \right\| \right) = \mathbb{E}\left(\left\| u_{t-m}^s \right\| \right) \leq \left[\mathbb{E}\left(\left\| u_{t-m}^s \right\|^2 \right) \right]^{1/2}$$

Then, for any $t \geq \widetilde{t} \geq t - m$:

$$\mathbb{E}\left(\left\| u_{\widetilde{t}}^s - \widetilde{u}_{\widetilde{t}}^s \right\| \right) \leq \overline{C}_4 \left[\mathbb{E}\left(\left\| u_{\widetilde{t}-1}^s - \widetilde{u}_{\widetilde{t}-1}^s \right\|^2 \right) \right]^{1/2}$$
$$\mathbb{E}\left(\left\| y_{\widetilde{t}}^s - \widetilde{y}_{\widetilde{t}}^s \right\| \right) \leq \overline{C}_3 \overline{C}_4^{\gamma} \left[\mathbb{E}\left(\left\| u_{\widetilde{t}-1}^s - \widetilde{u}_{\widetilde{t}-1}^s \right\|^2 \right) \right]^{\gamma/2} + \overline{C}_1 \left[\mathbb{E}\left(\left\| y_{\widetilde{t}-1}^s - \widetilde{y}_{\widetilde{t}-1}^s \right\|^2 \right) \right]^{1/2}.$$

The previous two results and a recursion arguments leads to the following inequality:

$$\mathbb{E}\left(\left\| u_t^s - \widetilde{u}_t^s \right\| \right) \leq \overline{C}_4^m \left[\mathbb{E}\left(\left\| u_{t-m}^s \right\|^2 \right) \right]^{1/2} \tag{.0.3}$$
$$\mathbb{E}\left(\left\| y_t^s - \widetilde{y}_t^s \right\| \right) \leq \overline{C}_3 \overline{C}_4^{\gamma m} \left[\mathbb{E}\left(\left\| u_{t-m}^s \right\|^2 \right) \right]^{\gamma/2} + \overline{C}_1^m \left[\mathbb{E}\left(\left\| y_{t-m}^s \right\|^2 \right) \right]^{1/2}. \tag{.0.4}$$

For $\beta = \beta_0, \Pi_{k(n)}\beta_0$ since the expectations are finite and bounded by assumption, $\mathbb{E}\left(\left\|y_t^s - \widetilde{y}_t^s\right\|\right) \leq \overline{C}\max(\overline{C}_1, \overline{C}_4)^{\gamma m}$ with $0 \leq \max(\overline{C}_1, \overline{C}_4) < 1$ and some $\overline{C} > 0$. For the first observations $t \leq m$ the data is unchanged, $y_t^s = \widetilde{y}_t^s$, so that the bound still holds. The integral can be split and bounded:

$$\begin{aligned}
&\left| \int e^{i\tau'(\mathbf{y}_t(y_0,u_0,x_t,\ldots,x_1,\beta_0,e_t^s,\ldots,e_1^s),\mathbf{x}_t)} \left(f_0(e_t^s) \times \cdots \times f_0(e_1^s) - \Pi_{k(n)} f_0(e_t^s) \times \cdots \times \Pi_{k(n)} f_0(e_1^s) \right) f_{\mathbf{x}}(\mathbf{x}_t) de_t^s \ldots de_1^s d\mathbf{x}_t \right| \\
&\leq \left| \mathbb{E}\left([\widehat{\psi}_n^S(\tau, \beta_0) - \widehat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0)] - [\widetilde{\psi}_n^S(\tau, \beta_0) - \widetilde{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0)] \right) \right| \\
&+ \int \left| \left(f_0(e_t^s) \times \cdots \times f_0(e_{t-m+1}^s) - \Pi_{k(n)} f_0(e_t^s) \times \cdots \times \Pi_{k(n)} f_0(e_{t-m+1}^s) \right) f_{\mathbf{x}}(\mathbf{x}_t) de_t^s \ldots de_{t-m+1}^s d\mathbf{x}_t \right| \\
&\leq 4\overline{C}\max(\overline{C}_1, \overline{C}_4)^{\gamma m} + 2m\|\Pi_{k(n)} f_0 - f_0\|_{TV}.
\end{aligned}$$

The last inequality is due to the cosine, and sine function being uniformly Lipschitz continuous and equations (.0.3)-(.0.4). Recall that $\|\Pi_{k(n)} f_0 - f_0\|_{TV} = O(\frac{\log[k(n)]^{2r/b}}{k(n)^r})$. To balance the two terms, choose:

$$m = -\frac{r}{\gamma \log \max(\overline{C}_1, \overline{C}_4)} \log[k(n)] > 0$$

so that $\max(\overline{C}_1, \overline{C}_4)^{\gamma m} = k(n)^{-r}$ and

$$\overline{C}\max(\overline{C}_1, \overline{C}_4)^{\gamma m} + 2m\|\Pi_{k(n)} f_0 - f_0\|_{TV} = O\left(\frac{\log[k(n)]^{2r/b+1}}{k(n)^r} \right).$$

Combining all the bounds above yields:

$$Q_n(\Pi_{k(n)}\beta_0) = O\left(\max\left[\frac{\log[k(n)]^{4r/b+2}}{k(n)^{2r}}, \frac{\log[k(n)]^{4\gamma^2 r/b}}{k(n)^{2\gamma^2 r}}, \frac{1}{n^2} \right] \right)$$

where $\|\cdot\|_{\mathcal{B}} = \|\cdot\|_\infty$ or $\|\cdot\|_{TV}$ so that $\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}^{\gamma^2} = O(\frac{\log[k(n)]^{4\gamma^2 r/b}}{k(n)^{2\gamma^2 r}})$. The term due to the non-stationarity is of order $1/n^2 = o\left(\max\left[\frac{\log[k(n)]^{4r/b+2}}{k(n)^{2r}}, \frac{\log[k(n)]^{4\gamma^2 r/b}}{k(n)^{2\gamma^2 r}} \right] \right)$ so it can be ignored. This concludes the proof. □

Proof of Theorem 2: The theorem is a corollary of Theorem .0.2 with a mixture sieve. Lemma 5 gives an explicit derivation of $\sqrt{Q_n(\Pi_{k(n)}\beta_0)}$ in this setting. □

Asymptotic Normality

Remark .0.2. *Note that for each τ the matrix $B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0))}{d(\boldsymbol{\theta},\omega,\mu,\sigma)}'\overline{B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0))}{d(\boldsymbol{\theta},\omega,\mu,\sigma)}}$ is singular - the requirement is that the average, over τ, of this matrix is invertible. Lemma 6 states that $\widehat{\beta}_n$ and the approximation $\Pi_{k(n)}\beta_0$ have a representation that are at a distance $\delta_n \underline{\lambda}_n^{-1/2}$ of each other in $\|\cdot\|_m$ norm.*

Proof of Lemma 6: Using the simple inequality $1/2|a|^2 \leq |a-b|^2 + |b|^2$ for any $a, b \in \mathbb{R}$:

$$0 \leq 1/2 \int \left| B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta} [\widehat{\beta}_n - \Pi_{k(n)}\beta_0] \right|^2 \pi(\tau) d\tau$$

$$\leq \int \left| B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta_0))}{d\beta} [\widehat{\beta}_n - \beta_0] \right|^2 \pi(\tau) d\tau$$

$$+ \int \left| B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta_0))}{d\beta} [\widehat{\beta}_n - \beta_0] - B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta} [\widehat{\beta}_n - \Pi_{k(n)}\beta_0] \right|^2 \pi(\tau) d\tau$$

$$\leq \int \left| B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta_0))}{d\beta} [\widehat{\beta}_n - \beta_0] \right|^2 \pi(\tau) d\tau + \int \left| B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta} [\Pi_{k(n)}\beta_0 - \beta_0] \right|^2 \pi(\tau) d\tau$$

$$+ \int \left| B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta_0))}{d\beta} [\widehat{\beta}_n - \beta_0] - B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta_0))}{d\beta} [\widehat{\beta}_n - \Pi_{k(n)}\beta_0] \right|^2 \pi(\tau) d\tau.$$

By assumption the term on the left is $O_p(\delta_n^2)$, by assumption ii. the middle term is $O_p(\delta_n^2)$ and assumption i. implies that the term on the right is also $O_p(\delta_n^2)$. It follows that:

$$\int \left| B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta} [\widehat{\beta}_n - \Pi_{k(n)}\beta_0] \right|^2 \pi(\tau) d\tau = O_p(\delta_n^2). \tag{.0.5}$$

Now note that both $\widehat{\beta}_n$ and $\Pi_{k(n)}\beta_0$ belong to the finite dimensional space $\mathcal{B}_{k(n)}$ parameterized by $(\boldsymbol{\theta}, \omega, \mu, \sigma)$. To save space, $\widehat{\beta}_n$ will be represented by $\widehat{\varphi}_n = (\widehat{\boldsymbol{\theta}}_n, \widehat{\omega}_n, \widehat{\mu}_n, \widehat{\sigma}_n)$ and $\Pi_{k(n)}\beta_0$ by $\varphi_{k(n)} = (\boldsymbol{\theta}_{k(n)}, \omega_{k(n)}, \mu_{k(n)}, \sigma_{k(n)})$. Using this notation, equation (.0.5) becomes:

$$\int \left| B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta} [\widehat{\beta}_n - \Pi_{k(n)}\beta_0] \right|^2 \pi(\tau) d\tau = \int \left| B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d(\boldsymbol{\theta}, \omega, \mu, \sigma)} [\widehat{\varphi}_n - \varphi_{k(n)}] \right|^2 \pi(\tau) d\tau$$

$$= trace \left([\widehat{\varphi}_n - \varphi_{k(n)}]' \int B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d(\boldsymbol{\theta}, \omega, \mu, \sigma)}' \overline{B \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d(\boldsymbol{\theta}, \omega, \mu, \sigma)}} \pi(\tau) d\tau [\widehat{\varphi}_n - \varphi_{k(n)}] \right)$$

$$\geq \underline{\lambda}_n \|\widehat{\varphi}_n - \varphi_{k(n)}\|^2 = \underline{\lambda}_n \|\widehat{\beta}_n - \Pi_{k(n)}\beta_0\|_m^2.$$

It follows that $0 \leq \underline{\lambda}_n \|\widehat{\beta}_n - \Pi_{k(n)}\beta_0\|_m^2 \leq O_p(\delta_n^2)$ so that the rate of convergence in mixture norm is:

$$\|\widehat{\beta}_n - \Pi_{k(n)}\beta_0\|_m = O_p\left(\delta_n \underline{\lambda}_n^{-1/2}\right).$$

□

Lemma .0.4 (Stochastic Equicontinuity). *Let $M_n = \log\log(n+1)$ and $\delta_{mn} = \delta_n / \sqrt{\underline{\lambda}_n}$. Suppose that the assumptions of Lemma 6 and Assumption .0.4 hold then for any $\eta > 0$, uniformly over $\beta \in \mathcal{B}_{k(n)}$:*

$$\left[\mathbb{E} \left(sup_{\|\beta - \Pi_{k(n)}\beta_0\|_m \leq M_n\delta_{mn}} \left| [\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta) - \widehat{\boldsymbol{\psi}}_n^S(\tau, \Pi_{k(n)}\beta_0)] - \mathbb{E}[\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta) - \widehat{\boldsymbol{\psi}}_n^S(\tau, \Pi_{k(n)}\beta_0)] \right|^2 \pi(\tau)^{\frac{2}{2+\eta}} \right) \right]^{1/2}$$

$$\leq C \frac{(M_n\delta_{mn})^{\frac{\gamma^2}{2}}}{\sqrt{n}} \int_0^1 \left(x^{-\boldsymbol{\vartheta}/2} \sqrt{\log N([xM_n\delta_{mn}]^{\frac{2}{\gamma^2}}, \mathcal{B}_{k(n)}, \|\cdot\|_m)} + \log^2 N([xM_n\delta_{mn}]^{\frac{2}{\gamma^2}}, \mathcal{B}_{k(n)}, \|\cdot\|_m) \right) dx$$

For the mixture sieve the integral is a $O(k(n)\log[k(n)] + k(n)|\log(M_n\delta_{mn})|)$ so that:

$$\left[\mathbb{E}\left(\int sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m\leq M_n\delta_{mn}}\left|[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]-\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]\right|^2\pi(\tau)d\tau\right)\right]^{1/2}$$
$$= O\left((M_n\delta_{mn})^{\frac{\gamma^2}{2}}\max(\log[k(n)]^2,|\log[M_n\delta_{mn}]|^2)\frac{k(n)^2}{\sqrt{n}}\right)$$

Now suppose that $(M_n\delta_{mn})^{\frac{\gamma^2}{2}}\max(\log[k(n)]^2,|\log[M_n\delta_{mn}]|^2)k(n)^2 = o(1)$. The first stochastic equicontinuity result is:

$$\left[\mathbb{E}\left(\int sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m\leq M_n\delta_{mn}}\left|[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]-\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]\right|^2\pi(\tau)d\tau\right)\right]^{1/2}$$
$$= o(1/\sqrt{n}).$$

Also, suppose that $\beta\to\int\mathbb{E}\left|\widehat{\psi}_t^s(\tau,\beta_0)-\widehat{\psi}_t^s(\tau,\beta)\right|^2\pi(\tau)d\tau$ is continuous at $\beta=\beta_0$ under the norm $\|\cdot\|_{\mathcal{B}}$, uniformly in $t\geq 1$. Then, the second stochastic equicontinuity result is:

$$\left[\mathbb{E}\left(\int sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m\leq M_n\delta_{mn}}\left|[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\beta_0)]-\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\beta_0)]\right|^2\pi(\tau)d\tau\right)\right]^{1/2}$$
$$= o(1/\sqrt{n}).$$

Proof of Lemma .0.4. This proof relies on the results in Lemma 4 together with Lemma .0.5. First, Lemma 4 implies that, after simplifying the bounds, for some $C>0$:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m\leq\delta,\|\beta_j-\Pi_{k(n)}\beta_0\|_m\leq M_n\delta_{m,n},j=1,2}\left|\widehat{\psi}_t^s(\tau,\beta_1)-\widehat{\psi}_t^s(\tau,\beta_2)\right|^2\right)\right]^{1/2}\frac{\sqrt{\pi(\tau)}}{(M_n\delta_{m,n})^{\gamma^2/2}}$$
$$\leq Ck(n)^{2\gamma^2}\left(\frac{\delta}{M_n\delta_{m,n}}\right)^{\gamma^2/2}.$$

Next, apply the inequality of Lemma .0.4 to generate the bound:

$$\left[\mathbb{E}\left(\sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m\leq M_n\delta_{m,n}}\left|[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]-\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]\right|^2\right)\right]^{1/2}\sqrt{\pi(\tau)}$$
$$\leq\overline{C}\frac{(M_n\delta_{m,n})^{\gamma^2/2}}{\sqrt{n}}\int_0^1\left(x^{-\vartheta/2}\sqrt{\log N(\left[\frac{xM_n\delta_{mn}}{k(n)^{2\gamma^2}}\right]^{\frac{2}{\gamma^2}},\mathcal{B}_{k(n)},\|\cdot\|_m)+\log^2 N(\left[\frac{xM_n\delta_{mn}}{k(n)^{2\gamma^2}}\right]^{\frac{2}{\gamma^2}},\mathcal{B}_{k(n)},\|\cdot\|_m)}\right)dx$$

for some $\overline{C}>0,\vartheta\in(0,1)$. Since $\int\sqrt{\pi(\tau)}d\tau<\infty$, the term on the left-hand side ca be squared and multiplied by $\sqrt{\pi(\tau)}$. Then, taking the integral:

$$\left[\mathbb{E}\left(\int\sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m\leq M_n\delta_{m,n}}\left|[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]-\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]\right|^2\pi(\tau)d\tau\right)\right]^{1/2}$$
$$\leq\overline{C}_\pi\frac{(M_n\delta_{m,n})^{\gamma^2/2}}{\sqrt{n}}\int_0^1\left(x^{-\vartheta/2}\sqrt{\log N(\left[\frac{xM_n\delta_{mn}}{k(n)^{2\gamma^2}}\right]^{\frac{2}{\gamma^2}},\mathcal{B}_{k(n)},\|\cdot\|_m)+\log^2 N(\left[\frac{xM_n\delta_{mn}}{k(n)^{2\gamma^2}}\right]^{\frac{2}{\gamma^2}},\mathcal{B}_{k(n)},\|\cdot\|_m)}\right)dx$$

where $\overline{C}_\pi = \overline{C}\int\sqrt{\pi(\tau)}d\tau$. The integral on the right-hand side is a

$$O(k(n)^2\max(\log[k(n)]^2, \log[M_n\delta_{m,n}]^2)).$$

To prove the final statement, notation will be shortened using $\Delta\widehat{\psi}_t^s(\tau,\beta) = \widehat{\psi}_t^s(\tau,\beta_0) - \widehat{\psi}_t^s(\tau,\beta)$. Note that, by applying Davydov (1968)'s inequality:

$$\begin{aligned}
&n\mathbb{E}\left|\Delta\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]\right|^2 \\
&\leq \frac{1}{n}\sum_{t=1}^n \mathbb{E}\left|\Delta\widehat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\widehat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0)]\right|^2 \\
&+ \frac{24}{n}\sum_{m=1}^n (n-m)\alpha(m)^{1/3}\max_{1\leq t\leq n}\left(\mathbb{E}\left|\Delta\widehat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\widehat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0)]\right|^6\right)^{2/3} \\
&\leq \left(1 + 24\sum_{m\geq 1}\alpha(m)^{1/3}\right)\max_{1\leq t\leq n}\left(\mathbb{E}\left|\Delta\widehat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\widehat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0)]\right|^6\right)^{2/3} \\
&\leq 4^{8/3}\left(1 + 24\sum_{m\geq 1}\alpha(m)^{1/3}\right)\max_{1\leq t\leq n}\left(\mathbb{E}\left|\Delta\widehat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\widehat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0)]\right|^2\right)^{2/3}.
\end{aligned}$$

The last inequality is due to $|\Delta\widehat{\psi}_t^s(\tau,\beta)| \leq 2$. By the continuity assumption the last term is a $o(1)$ when $\|\beta_0 - \Pi_{k(n)}\|_{\mathcal{B}} \to 0$. As a result:

$$\int \mathbb{E}\left|\Delta\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]\right|^2\pi(\tau)d\tau = o(1/n).$$

To conclude the proof, apply a triangular inequality and the results above:

$$\begin{aligned}
&\left[\mathbb{E}\left(\int \sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m\leq M_n\delta_{mn}}\left|[\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\beta_0)] - \mathbb{E}[\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\beta_0)]\right|^2\pi(\tau)d\tau\right)\right]^{1/2} \\
&\leq \left[\mathbb{E}\left(\int \sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m\leq M_n\delta_{mn}}\left|[\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)] - \mathbb{E}[\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]\right|^2\pi(\tau)d\tau\right)\right]^{1/2} \\
&+ \left(\int \mathbb{E}\left|\Delta\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]\right|^2\pi(\tau)d\tau\right)^{1/2} = o(1/\sqrt{n}).
\end{aligned}$$

□

Remark .0.3. *Note that* $\delta_n = \frac{k(n)^2\log[k(n)]^2}{\sqrt{n}} = o(1)$ *by assumption so that* $\log[\delta_n]^2 = O(\log(n)^2)$. *Furthermore, it is assumed that* $\delta_n = o\left(\sqrt{\underline{\lambda}_n}\right)$ *and* $\delta_{m,n} = o(1)$, *so that* $\max(\log[k(n)]^2, \log[M_n\delta_{m,n}]^2))$ *is dominated by a* $O(\log(n))$. *The condition* $k(n)^2\max(\log[k(n)]^2, \log[M_n\delta_{m,n}]^2)$ *can thus be re-written as:*

$$(M_n\delta_{mn})^{\frac{\gamma^2}{2}}[k(n)\log(n)]^2 = o(1)$$

which is equivalent to:

$$\delta_n = o\left(\frac{\sqrt{\underline{\lambda}_n}}{M_n[k(n)\log(n)]^{\frac{4}{\gamma^2}})}\right).$$

Furthermore, since $\delta_n = \frac{k(n)^2\log[k(n)]^2}{\sqrt{n}}$*, this condition can be re-written in terms of* $k(n)$*:*

$$k(n) = o\left(\left(\frac{\sqrt{\underline{\lambda}_n}}{M_n\log(n)^{\frac{4}{\gamma^2}}}\right)^{\frac{1}{2+4/\gamma^2}} n^{\frac{1}{2(2+4/\gamma^2)}}\right).$$

Proof of Theorem 3: Theorem 3 mostly follows from Theorem .0.3 with two differences: the rate of convergence and the stochastic equicontinuity results in mixture norm. Lemmas 6 and .0.4 provide these results for the mixture sieve. Hence, given these results, Theorem 3 is a corollary of Theorem .0.3. □

Extension 1: Using Auxiliary Variables

Proof of Corollary 2: Since the proof of Corollary 2 is very similar to the main proofs, only the differences in the steps are highlighted.

i. **Consistency**: The objective function with auxiliary variables is:

$$Q_n(\beta) = \int \left|\mathbb{E}\left(\widehat{\psi}_n(\tau,\widehat{\eta}_n^{aux}) - \widehat{\psi}_n^s(\tau,\widehat{\eta}_n^{aux},\beta\right)\right|^2 \pi(\tau)d\tau.$$

To derive its rate of convergence consider:

$$\begin{aligned}\int \left|\widehat{\psi}_n(\tau,\widehat{\eta}_n^{aux}) - \mathbb{E}\left(\widehat{\psi}_n(\tau,\widehat{\eta}_n^{aux})\right)\right|^2\pi(\tau)d\tau \leq\ & 9\int \left|\widehat{\psi}_n(\tau,\eta^{aux}) - \mathbb{E}\left(\widehat{\psi}_n(\tau,\eta^{aux})\right)\right|^2\pi(\tau)d\tau \\ & + 9\int \left|\widehat{\psi}_n(\tau,\widehat{\eta}_n^{aux}) - \widehat{\psi}_n(\tau,\eta^{aux})\right|^2\pi(\tau)d\tau \\ & + 9\int \left|\mathbb{E}\left(\widehat{\psi}_n(\tau,\widehat{\eta}_n^{aux}) - \widehat{\psi}_n(\tau,\eta^{aux})\right)\right|^2\pi(\tau)d\tau.\end{aligned}$$

The first term is $O_p(1/n)$. By the Lipschitz condition, the second term satisfies:

$$\begin{aligned}\int \left|\widehat{\psi}_n(\tau,\widehat{\eta}_n^{aux}) - \widehat{\psi}_n(\tau,\eta^{aux})\right|^2\pi(\tau)d\tau &\leq \|\widehat{\eta}_n^{aux} - \eta^{aux}\|^2|C_n^{aux}|^2\int\|\tau\|_\infty\pi(\tau)d\tau \\ &= O_p(1/n)O_p(1).\end{aligned}$$

C_n^{aux} is an average of the Lipschitz constants in the assumptions. The third term can be bounded using the Lipschitz assumption and the Cauchy-Schwarz inequality:

$$\begin{aligned}\int \left|\widehat{\psi}_n(\tau,\widehat{\eta}_n^{aux}) - \widehat{\psi}_n(\tau,\eta^{aux})\right|^2\pi(\tau)d\tau &\leq \mathbb{E}\|\widehat{\eta}_n^{aux} - \eta^{aux}\|^2\mathbb{E}|C_n^{aux}|^2\int\|\tau\|_\infty\pi(\tau)d\tau \\ &= O_p(1/n^2)O_p(1).\end{aligned}$$

Altogether, these inequalities imply:

$$\int \left| \widehat{\psi}_n(\tau, \widehat{\eta}_n^{aux}) - \mathbb{E}\left(\widehat{\psi}_n(\tau, \widehat{\eta}_n^{aux}) \right) \right|^2 \pi(\tau) d\tau = O_p(1/n^2).$$

The L^2-smoothness result still holds given the summability condition:

$$\begin{aligned}
&\left[\mathbb{E}\left(\sup_{\|\beta_1 - \beta_2\|_{\mathcal{B}} \leq \delta, \eta \in E} \| g_{aux}(y_t^s(\beta_1), \ldots, y_1^s(\beta_1), x_t, \ldots, x_1; \eta) - g_{aux}(y_t^s(\beta_2), \ldots, y_1^s(\beta_2), x_t, \ldots, x_1; \eta) \|^2 \right) \right]^{1/2} \\
&\qquad \leq \sum_{j=1}^{t} \rho_j \left[\mathbb{E}\left(\sup_{\|\beta_1 - \beta_2\|_{\mathcal{B}} \leq \delta, \eta \in E} \| y_j^s(\beta_1) - y_j^s(\beta_2) \|^2 \right) \right]^{1/2} \\
&\qquad \leq \left(\sum_{j=1}^{\infty} \rho_j \right) \sup_{t \geq 1} \left[\mathbb{E}\left(\sup_{\|\beta_1 - \beta_2\|_{\mathcal{B}} \leq \delta, \eta \in E} \| y_t^s(\beta_1) - y_t^s(\beta_2) \|^2 \right) \right]^{1/2} \\
&\qquad \leq \overline{C} \left(\sum_{j=1}^{\infty} \rho_j \right) \max \left(\frac{\delta^{\gamma^2}}{\underline{\sigma}_{k(n)}^{2\gamma^2}}, [k(n) + \overline{\mu}_{k(n)} + \overline{\sigma}]^{\gamma} \delta^{\gamma^2/2} \right)
\end{aligned}$$

The last inequality is a consequence of Lemma 4.

$$\begin{aligned}
\int \left| \widehat{\psi}_n^s(\tau, \widehat{\eta}_n^{aux}) - \mathbb{E}\left(\widehat{\psi}_n^s(\tau, \widehat{\eta}_n^{aux}) \right) \right|^2 \pi(\tau) d\tau \leq\ & 9 \int \left| \widehat{\psi}_n^s(\tau, \eta^{aux}) - \mathbb{E}\left(\widehat{\psi}_n^s(\tau, \eta^{aux}) \right) \right|^2 \pi(\tau) d\tau \\
& + 9 \int \left| \widehat{\psi}_n^s(\tau, \widehat{\eta}_n^{aux}) - \widehat{\psi}_n^s(\tau, \eta^{aux}) \right|^2 \pi(\tau) d\tau \\
& + 9 \int \left| \mathbb{E}\left(\widehat{\psi}_n^s(\tau, \widehat{\eta}_n^{aux}) - \widehat{\psi}_n^s(\tau, \eta^{aux}) \right) \right|^2 \pi(\tau) d\tau.
\end{aligned}$$

The first term is a $O_p(\delta_n^2)$ given the L^2-smoothness above and the main results. The last two terms are $O_p(1/n^2)$ as in the calculations above.

Together, these results imply that the rate of convergence for the objective function is $O_p(\delta_n^2)$ as before. As a result, given that the other assumptions hold, the estimator is consistent.

ii. **Rate of Convergence**: The variance term is still $O_p(\delta)$ as discussed above. The only term remaining to discuss if the bias accumulation term.

Recall that the first part of the bias term involves changing f in g_{obs}, g_{latent} while keeping the shocks e_t^s unchanged. Using the same method of proof as for the L^2-smoothness it can be shown that the first bias term is only inflated by $\sum_{j=1}^{\infty} \rho_j < \infty$: a finite factor.

The second part involves changing the shocks keeping g_{obs}, g_{latent} unaffected. An alternative simulated sequence $\widetilde{y}_t^s$ where part of the history is changed $\widetilde{y}_{t-j}^s = \widetilde{u}_{t-j}^s = 0$ for $j \geq m$. For a well chosen sequence m, the difference between y_t^s and $\widetilde{y}_t^s$ declines exponentially in m. Here $\widetilde{z}_t^s$ only depends on a finite number of shocks since $\widetilde{y}_{t-m}^s =$

$\cdots = \widetilde{y}_1^s = 0$. The difference between z_t^s and $\widetilde{z}_t^s$ becomes:

$$\mathbb{E}\left(\|z_t^s - \widetilde{z}_t^s\|\right) \leq \sum_{j=1}^{t} \rho_j \mathbb{E}\left(\|y_j^s - \widetilde{y}_j^s\|\right) \leq \left(\sum_{j=1}^{\infty} \rho_j\right) \overline{C} \max(\overline{C}_1, \overline{C}_4)^{\gamma m}$$

where the last inequality comes from Lemma 5. To apply this lemma, the bounded moment condition $v.$ is required.

Overall, the bias term is unchanged. As a result, the rate of convergence is the same as in the main proofs.

iii. **Asymptotic Normality**: The L^2-smoothness result was shown above to be unchanged. As a result, stochastic equicontinuity can be proved the same way as before. The Lipschitz condition also implies stochastic equicontinuity in η^{aux} using the same approach as for the rate of convergence of the objective. The asymptotic expansion can be proved the same way as in the main results where $\widehat{\psi}_n(\tau)$ and $\widehat{\psi}_n^s(\tau, \beta_0)$ are replaced with $\widehat{\psi}_n(\tau, \widehat{\eta}_n^{aux})$ and $\widehat{\psi}_n^s(\tau, \widehat{\eta}_n^{aux}, \beta_0)$. Eventually, the expansion implies:

$$\frac{\sqrt{n}}{\sigma_n^*}\left(\phi(\widehat{\beta}_n) - \phi(\beta_0)\right) = \sqrt{n} Real\left(\int \psi_\beta(\tau, u_n^*, \eta^{aux}) \overline{\left(\widehat{\psi}_n(\tau, \widehat{\eta}_n^{aux}) - \widehat{\psi}_n^s(\tau, \widehat{\eta}_n^{aux}, \beta_0)\right)} \pi(\tau) d\tau\right) + o_p(1)$$

where the term on the right is asymptotically normal by assumption.

□

Extension 2: Using Short Panels

Proof of Lemma 8. The second part of the lemma is implied by using $y_{j,1}^s = y_{j,1}$ for all j.

For the first part of Lemma 8, using the notation for the proof of Proposition .0.4: f is the distribution for the simulated $\mathbf{y}_{j,t}^s$ and $\mathbf{u}_{j,t}^s$ and f^* is the stationary distribution. Note that $f(\mathbf{y}_{j,t}^s, \mathbf{x}_{j,t} | \mathbf{u}_{j,t}^s) = f^*(\mathbf{y}_{j,t}^s, \mathbf{x}_{j,t} | \mathbf{u}_{j,t}^s)$ for $\beta = \beta_0$ and $\|f_{\mathbf{u}} - f_{\mathbf{u}}^*\|_{TV} \leq C_u \overline{\rho}_u^m$ for some $C_u > 0$ and $\overline{\rho}_u \in (0,1)$.

$$\begin{aligned}
\sqrt{Q_n(\beta_0)} &\leq M_B \left(\int \left|\mathbb{E}\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^s(\tau, \beta_0)\right)\right|^2 \pi(\tau) d\tau\right)^{1/2} \\
&= M_B \left(\int \left|\frac{1}{n}\sum_{j=1}^{n} \int e^{i\tau'(\mathbf{y}_{j,t}^s, \mathbf{x}_{j,t})} \left(f(\mathbf{y}_{j,t}^s, \mathbf{x}_{j,t}) - f^*(\mathbf{y}_{j,t}^s, \mathbf{x}_{j,t})\right) d\mathbf{y}_{j,t}^s d\mathbf{x}_{j,t}\right|^2 \pi(\tau) d\tau\right)^{1/2} \\
&= M_B \left(\int \left|\frac{1}{n}\sum_{j=1}^{n} \int e^{i\tau'(\mathbf{y}_{j,t}^s, \mathbf{x}_{j,t})} f^*(\mathbf{y}_{j,t}^s, \mathbf{x}_{j,t} | \mathbf{u}_{j,t}^s) \left(f(\mathbf{u}_{j,t}^s) - f^*(\mathbf{u}_{j,t}^s)\right) d\mathbf{y}_{j,t}^s d\mathbf{x}_{j,t} d\mathbf{u}_{j,t}^s\right|^2 \pi(\tau) d\tau\right)^{1/2} \\
&\leq M_B \int f^*(\mathbf{y}_{j,t}^s, \mathbf{x}_{j,t} | \mathbf{u}_{j,t}^s) \left|f(\mathbf{u}_{j,t}^s) - f^*(\mathbf{u}_{j,t}^s)\right| d\mathbf{y}_{j,t}^s d\mathbf{x}_{j,t} d\mathbf{u}_{j,t}^s.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality implies:

$$\int f^*(\mathbf{y}^s_{j,t}, \mathbf{x}_{j,t}|\mathbf{u}^s_{j,t}) \left| f(\mathbf{u}^s_{j,t}) - f^*(\mathbf{u}^s_{j,t}) \right| d\mathbf{y}^s_{j,t} d\mathbf{x}_{j,t} d\mathbf{u}^s_{j,t}$$
$$\leq \left(\int f^*(\mathbf{y}^s_{j,t}, \mathbf{x}_{j,t}|\mathbf{u}^s_{j,t})^2 \left| f(\mathbf{u}^s_{j,t}) - f^*(\mathbf{u}^s_{j,t}) \right| d\mathbf{y}^s_{j,t} d\mathbf{x}_{j,t} d\mathbf{u}^s_{j,t} \right)^{1/2} \left(\int \left| f(\mathbf{u}^s_{j,t}) - f^*(\mathbf{u}^s_{j,t}) \right| d\mathbf{u}^s_{j,t} \right)^{1/2}.$$

By assumption the first term is finite and bounded while the second term is a $O(\overline{\rho}_u^{m/2})$. Taking squares on both sides on the inequality concludes the proof. □

Proof of Corollary 3: As discussed in section 3.4 asymptotic are conducted over the cross-sectional dimension n for the moments:

$$\widehat{\psi}_j(\tau) = \frac{1}{T} \sum_{t=1}^{T} e^{i\tau'(\mathbf{y}_{j,t}, \mathbf{x}_{j,t})}, \quad \widehat{\psi}^s_j(\tau) = \frac{1}{T} \sum_{t=1}^{T} e^{i\tau'(\mathbf{y}^s_{j,t}, \mathbf{x}_{j,t})}$$

which are iid under the stated assumptions. The bias can accumulate dynamically for DGP (3.12), as in the time-series case, but it accumulates with m instead of sample size. Assumption 2 or $2'$ ensure that the bias does not accumulate too much when $m \to \infty$. Lemma 8 shows how the assumed DGPs handle the initial condition problem in the panel setting. Note that:

$$n\overline{\rho}_u^m = e^{\log[n] + m\log[\overline{\rho}_u]} = e^{m(\log[n]/m + \log[\overline{\rho}_u])} \to 0$$

as $m, n \to \infty$ if $\lim_{m,n\to\infty} \log[n]/m < -\log[\overline{\rho}_u] > 0$. Given, this result and the dynamic bias accumulation the results for the iid case apply with an inflation bias term for DGP (3.12). □

Additional Asymptotic Results

This appendix provides general results for Sieve-SMM estimates for other sieve bases and bounded moment functions. It adapts existing results from the sieve literature to a continuum of bounded complex-valued moments and extends them to a more general class of dynamic models. The following definition gives the two measures of dependence used in the results.

Definition .0.1 (α-Mixing and Uniform α-Mixing)**.** *For the sample observations $(y_t)_{t\geq 1}$, the α-mixing coefficients are defined as:*

$$\alpha(m) = 2sup_{t\geq 1} sup_{y_1, y_2 \in \mathbb{R}^{d_y}} \left| \mathbb{P}\left(y_t \geq y_1, y_{t+m} \geq y_2\right) - \mathbb{P}\left(y_t \geq y_1\right) \mathbb{P}\left(y_{t+m} \geq y_2\right) \right|.$$

$(y_t)_{t\geq 1}$ *is* α*-mixing if* $\alpha(m) \to 0$ *when* $m \to \infty$.

For the simulated samples $(y(\beta)_t^s)_{t\geq 1}$ *indexed by* $\beta \in \mathcal{B}$ *the uniform* α*-mixing coefficients are defined as:*

$$\alpha^*(m) = 2sup_{t\geq 1,\beta\in\mathcal{B}} sup_{y_1,y_2\in\mathbb{R}^{d_y}} \left| \mathbb{P}\left(y_t^s(\beta) \geq y_1, y_{t+m}^s(\beta) \geq y_2\right) - \mathbb{P}\left(y_t^s(\beta) \geq y_1\right) \mathbb{P}\left(y(\beta)_{t+m}^s \geq y_2\right) \right|.$$

$(y_t^s(\beta))_{t\geq 1}$ *is uniformly* α*-mixing if* $\alpha^*(m) \to 0$ *when* $m \to \infty$.

Consistency

Recall that the Sieve-SMM estimator $\widehat{\beta}_n$ satisfies:

$$\widehat{Q}_n^S(\widehat{\beta}_n) \leq \text{diag}_{\beta\in\mathcal{B}_{k(n)}} \widehat{Q}_n^S(\beta) + O_p(\eta_n)$$

where $\eta_n = o(1)$. The sample objective function is:

$$\widehat{Q}_n^S(\beta) = \int \left| B\widehat{\psi}_n(\tau) - B\widehat{\psi}_n^S(\tau,\beta) \right|^2 \pi(\tau) d\tau$$

As in the main results, there is a sequence of population objective functions:

$$Q_n(\beta) = \int \left| \mathbb{E}\left(B\widehat{\psi}_n(\tau) - B\widehat{\psi}_n^S(\tau,\beta) \right) \right|^2 \pi(\tau) d\tau.$$

Q_n may depend on n when y_t^s is non-stationary. The following three assumptions are adapted from the sufficient high-level conditions in Chen (2007, 2011) and Chen & Pouzo (2012) to a continuum of of moments (Carrasco & Florens, 2000; Carrasco et al., 2007a).

Assumption .0.1 (Sieves)**.** $\{\mathcal{B}_k, k \geq 1\}$ *is a sequence of non-empty compact subsets of* $\mathcal{B}$ *such that* $\mathcal{B}_k \subseteq \mathcal{B}_{k+1} \subseteq \mathcal{B}$, $\forall k \geq 1$. *There exists an approximating sequence* $\Pi_k\beta_0 \in \mathcal{B}_k$ *such that* $\|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}} = o(1)$ *as* $k(n) \to \infty$.

Assumption .0.2 (Identification)**.** *i)* $\lim_{n\to\infty} \mathbb{E}\left(\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n(\tau)\right) = 0 \quad \pi$ *a.s.* $\Leftrightarrow \|\beta - \beta_0\|_{\mathcal{B}} = 0$. *The null space of B is the singleton* $\{0\}$. *ii)* $Q_n(\Pi_{k(n)}\beta_0) = o(1)$ *as* $n \to \infty$. *iii) There exists a function g such that for all* $\varepsilon > 0$*:* $g(k(n), n, \varepsilon) = diag_{\beta\in\mathcal{B}_{k(n)}, \|\beta-\beta_0\|_{\mathcal{B}}\geq\varepsilon} Q_n(\beta)$, *g is decreasing in the first and last argument and* $g(k(n), n, \varepsilon) > 0$ *for all* $k(n), n, \varepsilon > 0$.

Assumption .0.3 (Convergence Rate over Sieves)**.** *There exists two constants* $C_1, C_2 > 0$ *such that, uniformly over* $h \in \mathcal{B}_{k(n)}$*:* $\widehat{Q}_n^S(\beta) \leq C_1 Q_n(\beta) + O_p(\delta_n^2)$, $Q_n(\beta) \leq C_2\widehat{Q}_n(\beta) + O_p(\delta_n^2)$ *and* $\delta_n^2 = o(1)$.

Theorem .0.1 (Consistency)**.** *Suppose Assumptions .0.1-.0.3 hold. Furthermore, suppose that* $h \to Q_n(\beta)$ *is continuous on* $\left(\mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}}\right)$*. If* $k(n) \overset{n\to\infty}{\longrightarrow} \infty$ *and for all* $\varepsilon > 0$ *the following holds:*

$$\max\left(\eta_n, Q_n(\Pi_{k(n)}\beta_0), \delta_n^2\right) = o\left(g\left(k(n), \varepsilon\right)\right).$$

Then the estimator $\widehat{\beta}_n$ *is consistent:* $\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} = o_p(1)$*.*

Theorem .0.1 is a direct consequence of the general consistency lemma in Chen & Pouzo (2012) reproduced as Lemma .0.1 in the next appendix. Assumption .0.1 is standard and satisfied by the mixture sieve, the Hermite polynomial basis of Gallant & Nychka (1987) or the cosine basis as in Bierens & Song (2012). See e.g. Chen (2007) for further examples of sieve bases and their approximation properties. The choice of moments $\widehat{\psi}_n$ and the restrictions on the parameter space $\mathcal{B}$ are assumed to ensure identification in Assumption .0.2. Verifying Assumption .0.3 is more challenging in this setting because of the dynamics and the continuum of moments. Furthermore, the rate for $Q_n(\Pi_{k(n)}\beta_0)$ needs to be derived. The following proposition derives the rate for iid data under an additional restriction.[55]

Proposition .0.1. *If* y_t^s *is iid and depends on* f *only through* e_t^s*, i.e.* $y_t^s = g_{obs}(x_t, \boldsymbol{\theta}, e_t^s)$ *with* $e_t^s \sim f$*, then for* Q_n *based on the CF:*

$$Q_n(\Pi_{k(n)}\beta_0) \leq 2M_B^2 \|\Pi_{k(n)}f_0 - f_0\|_{TV}^2$$

where TV is the total variation norm: $\|\Pi_{k(n)}f_0 - f_0\|_{TV} = \int |\Pi_{k(n)}f_0(\varepsilon) - f_0(\varepsilon)| d\varepsilon$*.*

Remark .0.1. *Proposition .0.1 can be restated in terms of Hellinger distance by the inequality* $\|\Pi_{k(n)}f_0 - f_0\|_{TV} \leq 2d_H(\Pi_{k(n)}f_0, f_0)$*. Pinsker's inequality gives a similar relationship for the Kullback-Leibler divergence:* $\|\Pi_{k(n)}f_0 - f_0\|_{TV} \leq \sqrt{2KL(\Pi_{k(n)}f_0|f_0)}$*.*

Assumption .0.4 (Smoothness, Dependence, Complexity)**.** *Suppose that:*

i. (Smoothness) For $P \geq 2$*,* $\beta \to \psi_t^s(\tau, \beta)$ *is* L^p*-smooth. That is, there exists* $C > 0, \eta > 0$ *and* $\gamma \in (0, 1]$ *such that for all* $\tau \in \mathbb{R}^d$ *and all* $\delta > 0$*:*

$$sup_{t\geq 1}\left[\mathbb{E}\left(sup_{\beta_1,\beta_2\in\mathcal{B}, \|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta}\left|[\psi_t^s(\tau,\beta_1) - \psi_t^s(\tau,\beta_2)]\pi(\tau)^{1/(2+\eta)}\right|^P\right)\right]^{1/P} \leq C\delta^\gamma$$

and $\int \pi(\tau)^{1-2/(2+\eta)} d\tau < \infty$*.*

[55] A more general rate for $Q_n(\Pi_{k(n)}\beta_0)$ will be given in Proposition .0.3.

ii. (Dependence) (y_t^s, x_t) *and* (y_t, x_t) *are either iid or uniformly* $\alpha-$*mixing with* $\alpha^*(m) \leq C\exp(-am)$ *for all* $m \geq 1$ *with* $C, a > 0$.

iii. (Complexity) The moment function is uniformly bounded: $|\widehat{\psi}_t^s(\tau, \beta)| \leq M$ *for all* τ, β *and some* $M > 0$. *One of the following holds:*

a. if (y_t, x_t) *is iid, the integral*

$$\sqrt{C_n} := \int_0^1 \sqrt{1 + \log N\left(x^{1/\gamma}, \mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}}\right)}dx$$

is such that $C_n/n \to 0$.

b. if (y_t^s, x_t) *is dependent, the integral*

$$\sqrt{C_n} := \int_0^1 \left(x^{-\vartheta/2}\sqrt{\log N\left(x^{1/\gamma}, \mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}}\right)} + \log^2 N\left(x^{1/\gamma}, \mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}}\right)\right)dx$$

is such that $C_n/n \to 0$.

Where the covering number $N(x, \mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}})$ *is the minimal number of balls of radius* x *in* $\|\cdot\|_{\mathcal{B}}$ *norm needed to cover the space* $\mathcal{B}_{k(n)}$.

Assumption .0.4 provides conditions on the moments $\widehat{\psi}_n^s$, the weights π, the dependence and the sieve space to ensure Assumption .0.3 holds. Condition *i.* assumes that the moments are L^p-smooth. Note that the condition involves π, the moments themselves need not be uniformly L^p-smooth. An additional requirement is given for π to handle the continuum of moments. Giving the condition on the moments rather than the DGP itself as in the main results in more common (Duffie & Singleton, 1993, see e.g.) in the literature. The two are actually related, as shown in the following remark.

Remark .0.2 (L^p-Smoothness of the Moments and the DGP)**.** *For the empirical CF, smoothness of the moment function directly relates to smoothness of the data generating process: i.e.* L^p*-smoothness of* $\beta \to y_t^s(\beta)$ *implies Assumption .0.4 i. It is a direct implication of the sine and cosine functions being uniformly Lipschitz on the real line:*

$$\begin{aligned}\left|\psi_t^s(\tau, \beta_1) - \psi_t^s(\tau, \beta_2)\right|\pi(\tau)^{1/(2+\eta)} &\leq 2\|\tau'(\mathbf{y}_t^s(\beta_1), \mathbf{x}_t) - \tau'(\mathbf{y}_t^s(\beta_2), \mathbf{x}_t)\|\pi(\tau)^{1/(2+\eta)} \\ &\leq 2sup_{\tau\in\mathbb{R}^{d_\tau}}\left(\|\tau\|_\infty\pi(\tau)^{1/(2+\eta)}\right) \times \|\mathbf{y}_t^s(\beta_1) - \mathbf{y}_t^s(\beta_2)\|.\end{aligned}$$

This is the basis for the main results presented in section 3.3.

Examples of DGPs and moments satisfying condition *i.* are given in Appendix 3.7.

Assumption .0.4, condition *ii.* is satisfied under the geometric ergodicity condition of Duffie & Singleton (1993) as shown in Liebscher (2005)'s propositions 2 and 4. Note that Liebscher's result holds whether (y_t, x_t) is stationary or not.

Assumption .0.4, condition *iii.* hold for linear sieves with $k(n)/n \to 0$ in the iid case and $k(n)^4/n \to 0$ in the dependent case. For non-linear sieves such as mixtures and neural networks the condition becomes $k(n)\log[k(n)]/n \to 0$ in the iid case and $(k(n)\log[k(n)])^4/n \to 0$ in the dependent case. The following Proposition .0.2 relates the low-level conditions in Assumption .0.4 to Assumption .0.3.

Proposition .0.2. *Suppose that Assumption .0.4 holds, then Assumption .0.3 holds with* $\delta_n^2 = C_n/n$.

Given this proposition, Corollary .0.1 is a direct consequence of Theorem .0.1.

Corollary .0.1. *Suppose Assumptions .0.1-.0.2 and .0.4 hold. Furthermore, suppose that* $\beta \to Q_n(\beta)$ *is continuous on* $\left(\mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}}\right)$. *If* $k(n) \overset{n\to\infty}{\longrightarrow} \infty$ *and for all* $\varepsilon > 0$ *the following holds:*

$$\max\left(\eta_n, Q_n(\Pi_{k(n)}\beta_0), \delta_n^2\right) = o\left(g\left(k(n), \varepsilon\right)\right)$$

then the estimator $\widehat{\beta}_n$ *is consistent:*

$$\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} = o_p(1).$$

Proposition .0.3. *Suppose that the* L^p*-smoothness in Assumption .0.4 i. is satisfied, then there exists* $K > 0$ *which only depends on* C *and* η*, defined in Assumption .0.4 i.,* M_B *and* π *such that:*

$$Q_n(\Pi_{k(n)}\beta_0) \leq K\left(\|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{2\gamma} + Q_n(\beta_0)\right).$$

The rate in Proposition .0.3 is different from the main results because the L^p-smoothness assumption is given on the moments rather than the DGP itself. Also, in Assumption .0.3 the L^p-smoothness constant does not increase with $k(n)$ so that the decay condition is not required to derive the rate.

For iid and stationary $(y_t^s)_{t\geq 1}$, $Q_n(\beta_0) = 0$ should generally hold so the rate at which $Q_n(\Pi_{k(n)}\beta_0)$ goes to zero only depends on the smoothness γ and the approximation rate of β_0. When the L^p-smoothness coefficient is $\gamma = 1$, the rate is similar to Proposition .0.1 while for $\gamma \in (0,1)$ the rate is slower. In the non-stationary case $Q_n(\beta_0)$ will depend on the rate at which $f_{y_t^s, x_t}$ convergences to the stationary distribution.

Rate of Convergence

This section establishes the rate of convergence of the estimator in the weak norm of Ai & Chen (2003) and the strong norm $\|\cdot\|_{\mathcal{B}}$. As in Chen & Pouzo (2012), assuming consistency holds, the parameter space can be restricted to a local neighborhood $\mathcal{B}_{os} = \{\beta \in \mathcal{B}, \|\beta - \beta_0\|_{\mathcal{B}} \leq \varepsilon\}$ with $\varepsilon > 0$ small such that $\mathbb{P}\left(\widehat{\beta}_n \notin \mathcal{B}_\varepsilon\right) < \varepsilon$. Similarly, let $\mathcal{B}_{osn} = \mathcal{B}_{os} \cap \mathcal{B}_{k(n)}$.

Assumption .0.5 (Differentiability)**.** *Suppose that for all $\beta_1, \beta_2 \in \mathcal{B}_{os}$, the pathwise derivative:*

$$\lim_{\varepsilon \in (0,1), \varepsilon \to 0} \int \left| B \frac{\mathbb{E}\left(\widehat{\psi}_n^S(\tau, (1-\varepsilon)\beta_1 + \varepsilon\beta_2) - \widehat{\psi}_n^S(\tau, \beta_1)\right)}{\varepsilon} \right|^2 \pi(\tau) d\tau = \int \left| B \frac{d\mathbb{E}\left(\widehat{\psi}_n^S(\tau, \beta_1)\right)}{d\beta} [\beta_2] \right|^2 \pi(\tau) d\tau$$

exists and is finite.

Following Ai & Chen (2003), the weak norm measure uses the norm of the pathwise derivative of the moments at β_0:

$$\|\beta_1 - \beta_2\|_{weak} = \left(\int \left| B \frac{d\mathbb{E}[\psi_n^s(\tau, \beta)}{d\beta} \Big|_{\beta = \beta_0} [\beta_1 - \beta_2] \right|^2 \pi(\tau) d\tau \right)^{1/2}.$$

Suppose that there exists a $C > 0$ such that for all $\beta \in \mathcal{B}_{os}$ and all $n \geq 1$:

$$\|\beta - \beta_0\|_{weak}^2 \leq C Q_n(\beta).$$

Assumption .0.5 implies that $\|\cdot\|_{weak}$ is Lipschitz continuous with respect to the population criterion Q_n as in Chen & Pouzo (2012)'s assumption 4.1. Under Assumption .0.5, the rate of convergence is easier to derive in $\|\cdot\|_{weak}$ than in the stronger norm $\|\cdot\|_{\mathcal{B}}$. However, a sufficiently fast rate of convergence in the stronger norm will be required for the stochastic equicontinuity results, since the strong norm $\|\cdot\|_{\mathcal{B}}$ appears in L^p-smoothness Assumption .0.4. The two convergence rates are related by the local measure of ill-posedness of Blundell et al. (2007).

Definition .0.2 (Local Measure of Ill-Posedness of Blundell et al. (2007))**.** *The local measure of ill-posedness τ_n is:*

$$\tau_n = sup_{\beta \in \mathcal{B}_{osn}: \|\beta - \Pi_{k(n)}\beta_0\| \neq 0} \frac{\|\beta - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}}{\|\beta - \Pi_{k(n)}\beta_0\|}.$$

The following theorem adapts the results of Chen & Pouzo (2012) to the continuum of moments with simulated data.

Theorem .0.2 (Rate of Convergence)**.** *Suppose that Assumptions .0.1, .0.2, .0.4 and .0.5 are satisfied and suppose that $\eta_n = o(\delta_n^2)$. Let $\beta_0, \Pi_{k(n)}\beta_0 \in \mathcal{B}_{os}$, then we have the rate of convergence in weak and strong norm:*

$$\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p\left(\max\left(\delta_n, \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{\gamma}, \sqrt{Q_n(\beta_0)}\right)\right) \text{ and}$$

$$\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} = O_p\left(\|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}} + \tau_n \max\left(\delta_n, \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{\gamma}, \sqrt{Q_n(\beta_0)}\right)\right).$$

The rate δ_n is derived in Proposition .0.2: for linear sieves with iid data $\delta_n = \sqrt{k(n)/n}$ and $\delta_n = k(n)^2/\sqrt{n}$ in the dependent case. The rate $\|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{\gamma}$ depends on the approximation rate $\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}$ and the L^p-smoothness of the objective function. In the iid and stationary case, $Q_n(\beta_0) = 0$ is not a concern for the rate of convergence.

Proposition .0.4. *Suppose that $(\mathbf{y}_t^s, \mathbf{x}_t)_{t\geq 1}$ is geometrically ergodic for $\beta = \beta_0$ and the moments are bounded $|\widehat{\psi}_t^s(\tau, \beta_0)| \leq M$ for all τ then $Q_n(\beta_0) = O(1/n^2)$.*

Proposition .0.4 shows that $Q_n(\beta_0)$ is negligible under the geometric ergodicity condition of Duffie & Singleton (1993): since δ_n is typically larger than a $O(1/\sqrt{n})$ term, $Q_n(\beta_0) = o(\delta_n^2)$.

Corollary .0.2. *Suppose that the assumptions of Theorem .0.2 and the (y_t^s, x_t) are iid, stationary or geometrically ergodic then the rate of convergence is:*

$$\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} = O_p\left(\|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}} + \tau_n \max\left(\delta_n, \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{\gamma}\right)\right).$$

The rate of convergence can be further improved for static models with iid data under the assumptions of Proposition .0.1, as shown in the corollary below.

Corollary .0.3. *Suppose that the assumptions of Theorem .0.2 and Proposition .0.1 are satisfied then:*

$$\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} = O_p\left(\|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}} + \tau_n \max\left(\delta_n, \|\Pi_{k(n)}f_0 - f_0\|_{TV}\right)\right).$$

Asymptotic Normality

As in Chen & Pouzo (2015), this section gives asymptotic normality results for functionals ϕ of the estimates $\widehat{\beta}_n$. In order to conduct inferences, standard errors σ_n^* are derived such that:

$$\frac{\sqrt{n}}{\sigma_n^*}\left(\phi(\widehat{\beta}_n) - \phi(\beta_0)\right) \xrightarrow{d} \mathcal{N}(0,1). \tag{.0.1}$$

As in the main results, to reduce notation the following will be used:

$$\begin{aligned}\boldsymbol{\psi}_\beta(\tau, v) &= \frac{d\mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta_0))}{d\beta}[v] \\ Z_t^S(\tau) &= \widehat{\boldsymbol{\psi}}_t(\tau) - \frac{1}{S}\sum_{s=1}^{S} \widehat{\boldsymbol{\psi}}_t^s(\tau, \beta_0) \\ Z_n^S(\tau) &= \widehat{\boldsymbol{\psi}}_n(\tau) - \widehat{\boldsymbol{\psi}}_n^S(\tau, \beta_0)\end{aligned}$$

where v is a vector in $\overline{V}$ or $\overline{V}_n$ defined as in the main results. The sieve representer v_n^* is also defined as in the main results.

Definition .0.3 (Sieve Score, Sieve Variance, Scaled Sieve Representer)**.** *The Sieve Score S_n^* is defined as:*

$$S_n^* = \frac{1}{2}\int \left[B\boldsymbol{\psi}_\beta(\tau, v_n^*)\overline{BZ_n^S(\tau)} + \overline{B\boldsymbol{\psi}_\beta(\tau, v_n^*)}BZ_n^S(\tau)\right]\pi(\tau)d\tau.$$

The sieve variance is $\sigma_n^{*2} = n\mathbb{E}\left(|S_n^*|^2\right)$. The scaled sieve representer is $u_n^* = \frac{v_n^*}{\sigma_n^*}$.

As in the main results, the equivalence condition below is required.

Assumption .0.6 (Equivalence Condition)**.** *There exists $\underline{a} > 0$ such that $\forall n \geq 1$:*

$$\underline{a}\|v_n^*\|_{weak} \leq \sigma_n^*.$$

Furthermore assume that $\sigma_n^* = o(\sqrt{n})$.

An discussion of this condition is given in Appendix 3.7. The last part imposes that $k(n)$ does not increase too fast with n to control the variance of the sieve score.

Remark .0.3 (On the equivalence condition)**.** *Since $\widehat{\boldsymbol{\psi}}_t$ is bounded, the data is α-mixing and the simulations are geometrically ergodic there also exists a $\overline{a} > 0$ such that $\sigma_n^* \leq \overline{a}\|v_n^*\|_{weak}$. Hence under Assumption .0.6 the following holds $\sigma_n^* \asymp \|v_n^*\|_{weak}$. To prove this statement, note that the Cauchy-Schwarz inequality implies:*

$$\begin{aligned}\sigma_n^* &\leq \sqrt{2n}\left[\mathbb{E}\left(\left[\int \left|B\boldsymbol{\psi}_\beta(\tau, v_n^*)\right| \times \left|B\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta_0) - B\widehat{\boldsymbol{\psi}}_n(\beta_0)\right|\pi(\tau)d\tau\right]^2\right)\right]^{1/2} \\ &\leq \sqrt{2}\left[\int \left|B\boldsymbol{\psi}_\beta(\tau, v_n^*)\right|^2\pi(\tau)d\tau\right]^{1/2}\left[\mathbb{E}\left(\left[\int n\left|B\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta_0) - B\widehat{\boldsymbol{\psi}}_n(\beta_0)\right|^2\pi(\tau)d\tau\right]\right)\right]^{1/2}\end{aligned}$$

The first term in the product is $\|v_n^*\|_{weak}$. The second term is bounded by noting that for all $\tau \in \mathbb{R}^{d_\tau}$:

$$n\mathbb{E}\left|B\widehat{\psi}_n^S(\tau,\beta_0) - \mathbb{E}\left(B\widehat{\psi}_n^S(\tau,\beta_0)\right)\right|^2 \leq 1 + 24\sum_{m\geq 1}\alpha(m)^{1/p} < \infty.$$

for any $p > 1$ by Lemma .0.2, picking $p = 1/2$ implies:

$$\mathbb{E}\left(\int n\left|B\widehat{\psi}_n^S(\tau,\beta_0) - B\widehat{\psi}_n(\beta_0)\right|^2 \pi(\tau)d\tau\right) \leq \left(1 + 24\sum_{m\geq 1}\sqrt{\alpha(m)}\right)$$

which yields $\overline{a} = \sqrt{4 + 96\sum_{m\geq 1}\sqrt{\alpha(m)}}$.

Assumption .0.7 (Undersmoothing, Convergence Rate)**.** *Let $\delta_{sn} = \|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}}$ the convergence rate in strong norm.*

i. Undersmoothing: $\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p(\delta_n)$ and $\delta_{sn} = \|\Pi_{k(n)}\beta - \beta_0\|_{\mathcal{B}} + \tau_n\delta_n$.

ii. Sufficient Rate: $\delta_n = o(n^{-1/4})$.

iii. The convergence rate in weak norm δ_n and in strong norm δ_{sn} are such that:

$$(M_n\delta_{sn})^\gamma\sqrt{C_{sn}} = o(1) \tag{.0.2}$$

$$\sqrt{n}M_n^{1+\gamma}\delta_n^\gamma\sqrt{C_{sn}}\max\left(M_n\delta_n, \frac{1}{\sqrt{n}}\right) = o(1) \tag{.0.3}$$

where

$$\sqrt{C_{sn}} = \int_0^1 \left(x^{-\vartheta/2}\sqrt{\log N([xM_n\delta_{sn}]^{1/\gamma}, \mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}})} + \log^2 N([xM_n\delta_{sn}]^{1/\gamma}, \mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}})\right)dx$$

and $M_n = \log\log(n+1)$ for all all $n \geq 1$.

Assumptions .0.7 *i.*, *ii.* are common in the (semi)-nonparametric literature. Assumption .0.7 *iii.* ensures that a stochastic equicontinuity holds. It is needed several time throughout the proofs (see Lemma .0.6), in most cases the less demanding condition (.0.2) is sufficient. Condition (.0.3) is similar to Chen & Pouzo (2015)'s assumption A.5 *(iii)*, it ensures that when $\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\beta_0)$ is substituted under the integral with its smoothed version, the difference is negligible for $\sqrt{n}$-asymptotics.

Assumption .0.8 (Local Linear Expansion of ϕ)**.** *ϕ is continuously differentiable and $\frac{d\phi(\beta_0)}{d\beta}[\cdot]$ is a non-zero linear functional such that as $n \to \infty$:*

i. A linear expansion is locally uniformly valid

$$sup_{\|\beta-\beta_0\|_{weak}\leq M_n\delta_n}\frac{\sqrt{n}}{\sigma_n^*}\left|\phi(\beta) - \phi(\beta_0) - \frac{d\phi(\beta_0)}{d\beta}[\beta - \beta_0]\right| \to 0.$$

ii. The approximation bias is negligible

$$\frac{\sqrt{n}}{\sigma_n^*}\frac{d\phi(\beta_0)}{d\beta}[\beta_{0,n}-\beta_0]\to 0.$$

Remark .0.4 (Sufficient Conditions for Assumption .0.8 *i.*)**.** *If ϕ is twice continuously differentiable then for some $v\in\overline{V}$ and $h\in[-1,1]$, $\beta=\beta_0+hM_n\delta_n v$. Using a Mean Value Expansion:*

$$\begin{aligned}\left|\phi(\beta_0+hM_n\delta_n v)-\phi(\beta_0)-\frac{d\phi(\beta_0)}{d\beta}[hM_n\delta_n v]\right| &= \left|\frac{1}{2}\frac{d^2\phi(\beta_0+\widetilde{h}M_n\delta_n v)}{d\beta d\beta}[hM_n\delta_n v, hM_n\delta_n v]\right| \\ &= h^2(M_n\delta_n)^2\left|\frac{1}{2}\frac{d^2\phi(\beta_0+\widetilde{h}M_n\delta_n v)}{d\beta d\beta}[v,v]\right|.\end{aligned}$$

Hence Assumption .0.8 i. holds under the following two conditions:

i. The second derivative is locally uniformly bounded:

$$sup_{\|v\|_{weak}=1,h\in(-1,1)}\left|\frac{1}{2}\frac{d^2\phi(\beta_0+hM_n\delta_n v)}{d\beta d\beta}[v,v]\right|=O(1).$$

ii. The rate of convergence satisfies:

$$\frac{\sqrt{n}}{\sigma_n^*}(M_n\delta_n)^2=o(1).$$

This condition holds if $\delta_n=o(M_n^{-1}n^{-1/4})$ which is slightly stronger than Assumption .0.7 ii.

Remark .0.5 (Sufficient Conditions for Assumption .0.8 *ii.*)**.** *By definition of $\beta_{0,n}$, Assumptions .0.4, .0.5 and under geometric ergodicity:*

$$\begin{aligned}\|\beta_{0,n}-\beta_0\|_{weak} &\le \|\Pi_{k(n)}\beta_0-\beta_0\|_{weak} \\ &\le C\sqrt{Q_n(\Pi_{k(n)}\beta_0)} \\ &\le \widetilde{C}\|\Pi_{k(n)}\beta_0-\beta_0\|_{\mathcal{B}}^{\gamma}.\end{aligned}$$

The approximation rate is typically $\|\Pi_{k(n)}\beta_0-\beta_0\|_{\mathcal{B}}^{\gamma}=O(k(n)^{-r})$ where r is the smoothness of the density f_0 to be estimated. Rewriting $\beta_{0,n}=\beta_0+h_nk(n)^{-r}v_n$ with $\|v_n\|_{weak}=1$, $|h_n|\le\overline{h}$ bounded, the expression can be bounded using:

$$\frac{\sqrt{n}}{\sigma_n^*}\left|\frac{d\phi(\beta_0)}{d\beta}[\beta_{0,n}-\beta_0]\right|\le\overline{h}\frac{\sqrt{n}}{\sigma_n^*}k(n)^{-r}\left|\frac{d\phi(\beta_0)}{d\beta}[v_n]\right|.$$

Hence Assumption .0.8 ii. is satisfied under the following two conditions:

i. The first derivative is uniformly bounded on the unit circle:

$$sup_{\|v\|_{weak}=1}\left|\frac{1}{2}\frac{d\phi(\beta_0)}{d\beta}[v]\right| < \infty.$$

ii. The approximation rate satisfies:

$$\frac{\sqrt{n}}{\sigma_n^*}k(n)^{-\gamma r} = o(1).$$

With the undersmoothing assumption the $k(n)$ must satisfy $k(n)^{-\gamma r} = o(\delta_n) = o(n^{-1/4})$. aA sufficient condition on the bias/variance relation is $k(n)^{-\gamma r} = o(\delta_n^2\sigma_n^*)$.

The last condition is strong and can be weakened if for instance $\delta_n^2 \ll 1/\sqrt{n}$, replacing δ_n^2 with $1/\sqrt{n}$. Sharper bounds on the bias can also be found in the iid case (see Corollary .0.3) or under assumptions on the DGP itself as in the main results (see Lemma 5).

Assumption .0.9 (Local Behaviour of $\mathbb{E}(\widehat{\psi}(\tau,\beta))$)**.** *The mapping $\beta \to \mathbb{E}(\widehat{\psi}_n^S(\tau,\beta)$ is twice continuously differentiable for all τ and satisfies:*

i. A linear expansion is locally uniformly valid

$$\left(sup_{\|\beta-\beta_0\|_{weak}\leq M_n\delta_n}\int\left|\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta)) - \mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0)) - \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\beta-\beta_0]\right|^2\pi(\tau)d\tau\right)^{1/2}$$
$$= O\left((M_n\delta_n)^2\right)$$

ii. The second derivative in direction u_n^* is locally uniformly bounded

$$sup_{\|\beta-\beta_0\|_{weak}\leq M_n\delta_n}\int\left|\frac{d^2\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta))}{d\beta d\beta}[u_n^*,u_n^*]\right|^2\pi(\tau)d\tau = O(1)$$

Remark .0.6 (Sufficient Conditions for Assumption .0.9)**.** *Assumption .0.9 i. holds if $\mathbb{E}(\widehat{\psi}_n^S(\tau,\cdot))$ is twice continuously differentiable around β_0 with locally uniformly bounded second derivative since for some $\|v\|_{weak} = 1$ and $h \in [-1,1]$: $\beta = \beta_0 + hM_n\delta_n v$. A Mean Value Expansion yields:*

$$\left(\int\left|\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta)) - \mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0)) - \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\beta-\beta_0]\right|^2\pi(\tau)d\tau\right)^{1/2}$$
$$= h^2M_n^2\delta_n^2\left(\int\left|\frac{d^2\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0+\widetilde{h}M_n\delta_n v))}{d\beta d\beta}[v,v]\right|^2\pi(\tau)d\tau\right)^{1/2}$$

Since $\widetilde{h} \in (-1,1)$, the expression above is $O\left((M_n\delta_n)^2\right)$ if:

$$sup_{\|v\|_{weak}=1, h\in(-1,1)} \left(\int \left| \frac{d^2\mathbb{E}(\widehat{\psi}_n^S(\tau, \beta_0 + \widetilde{h}M_n\delta_n v))}{d\beta d\beta}[v,v] \right|^2 \pi(\tau)d\tau \right)^{1/2} = O(1).$$

Hence Assumptions .0.9 i. and ii. could be nested under the following condition:

$$sup_{\|v\|_{weak}=1, \|\beta-\beta_0\|_{weak}\leq M_n\delta_n} \left(\int \left| \frac{d^2\mathbb{E}(\widehat{\psi}_n^S(\tau, \beta))}{d\beta d\beta}[v,v] \right|^2 \pi(\tau)d\tau \right)^{1/2} = O(1).$$

The following theorem establishes the asymptotic normality of $\phi(\widehat{\beta}_n) - \phi(\beta_0)$ under the assumptions given above. Note that when $\sigma_n^* \to \infty$ the estimates converge at a slower than $\sqrt{n}$-rate.

Theorem .0.3 (Asymptotic Normality)**.** *Suppose Assumptions .0.6-.0.9 hold then:*

$$\frac{\sqrt{n}}{\sigma_n^*}\left(\phi(\widehat{\beta}_n) - \phi(\beta_0)\right) = \frac{\sqrt{n}}{\sigma_n^*}\left(\phi(\widehat{\beta}_n) - \phi(\beta_0)\right) S_n^* + o_p(1).$$

Furthermore, if the data (y_t, x_t) is stationary α-mixing, the simulated data is geometrically ergodic, the moments are bounded $|\widehat{\psi}_t^s(\tau, \beta)| \leq 1$ and B is bounded linear then S_n^*/σ_n^* satisfies a Central Limit Theorem so that:

$$\frac{\sqrt{n}}{\sigma_n^*}\left(\phi(\widehat{\beta}_n) - \phi(\beta_0)\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Examples of L^p-smooth models

The following provides examples of DGP and moment combinations which satisfy Assumption .0.4 condition *i.*

1. iid data without covariates: $y_t^s = u_t^s$, $u_t^s \sim F$. The moment function is the empirical CDF:

   $$\widehat{\psi}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{y_t \leq \tau}.$$

   Using the supremum distance, $\|\beta_1 - \beta_2\|_{\mathcal{B}} = \sup_y |F(y) - \widetilde{F}(y)|$, the following holds:

   $$\left[\mathbb{E}\left(\sup_{\|F_1 - F_2\|_\infty \leq \delta} \left| \mathbb{1}_{y_t^s(F_1)\leq\tau} - \mathbb{1}_{y_t^s(F_2)\leq\tau} \right|^2\right)\right]^{1/2} \leq 2\delta^{1/2}.$$

   Assumption .0.4, condition *i.* is satisfied with π equal to the normal density function for any $\eta > 0$.

2. Single Index Model: $y_t^s = \mathbb{1}_{x_t'\boldsymbol{\theta}+u_t^s \leq 0}$, $u_t^s \sim F$. The moment function is the empirical CF:

$$\widehat{\boldsymbol{\psi}}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} \exp\left(i\tau'\left(y_t, x_t\right)\right).$$

The metric is the supremum distance between CDFs $\|\beta_1 - \beta_2\|_{\mathcal{B}} = \sup_y |F_1(y) - F_2(y)|$ and $\mathcal{B} = \{\beta = (\boldsymbol{\theta}, F), \|F'\|_\infty \leq C_1 < \infty, \|\boldsymbol{\theta}\| \leq C_2 < \infty\}$, a space with CDFs having continuous and bounded densities. Also, suppose that $\mathbb{E}\|x_t\| < \infty$, then:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_\infty \leq \delta} \left|\mathbb{1}_{y_t^s(\beta_1)\leq\tau} - \mathbb{1}_{y_t^s(\beta_2)\leq\tau}\right|^2\right)\right]^{1/2} \leq 2\sqrt{1 + C_1\mathbb{E}\|x_t\|}\delta^{1/2}\|\tau\|_\infty.$$

Condition *i.* is satisfied with π equal to the normal density function for any $\eta > 0$.

3. MA(1) model: $y_t^s = u_t^s + \boldsymbol{\theta} u_{t-1}^s$, $u_t^s \sim F$. The moment function is the empirical CF:

$$\widehat{\boldsymbol{\psi}}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} \exp\left(i\tau'\left(y_t, y_{t-1}\right)\right).$$

The metric is the supremum distance between quantile functions:
$\|F_1^{-1} - {F_2}^{-1}\|_{\mathcal{B}} = \sup_{0\leq\nu\leq 1}|F_1^{-1}(\nu) - {F_2}^{-1}(\nu)|$. The parameter space $\mathcal{B} = \{\beta = (\boldsymbol{\theta}, F), \|F^{-1}\|_\infty \leq C_1 < \infty, |\boldsymbol{\theta}| \leq C_2 < \infty\}$ is the space of distributions with bounded quantile functions. The following holds:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_\infty\leq\delta} \left|\exp\left(i\tau'\left(y_t^s(\beta_1), y_{t-1}^s(\beta_1)\right)\right) - \exp\left(i\tau'\left(y_t^s(\beta_2), y_{t-1}^s(\beta_2)\right)\right)\right|^2\right)\right]^{1/2}$$
$$\leq 2(1 + C_1 + C_2)\delta\|\tau\|_\infty.$$

Condition *i.* is satisfied with π equal to the normal density function for any $\eta > 0$.

4. AR(1) model: $y_t^s = \boldsymbol{\theta} y_{t-1}^s + u_t^s$, $u_t^s \sim F$. The moment function is the empirical CF:

$$\widehat{\boldsymbol{\psi}}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} \exp\left(i\tau'\left(y_t, y_{t-1}\right)\right).$$

The metric is the supremum distance between quantile functions:

$$\|F^{-1} - \widetilde{F}^{-1}\|_{\mathcal{B}} = \sup_{0\leq\nu\leq 1}|F^{-1}(\nu) - \widetilde{F}^{-1}(\nu)|.$$

The parameter space $\mathcal{B} = \{\beta = (\boldsymbol{\theta}, F), \|F^{-1}\|_\infty \leq C_1 < \infty, |\boldsymbol{\theta}| \leq C_2 < 1\}$ is the space f distributions with bounded quantile functions. The following holds:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_\infty\leq\delta} \left|\exp\left(i\tau'\left(y_t^s(\beta_1), y_{t-1}^s(\beta_1)\right)\right) - \exp\left(i\tau'\left(y_t^s(\beta_2), y_{t-1}^s(\beta_2)\right)\right)\right|^2\right)\right]^{1/2}$$
$$\leq \frac{2}{1-C_2}\left(1 + \frac{C_1}{1-C_2}\right)\delta\|\tau\|_\infty.$$

Condition *i.* is satisfied with π equal to the normal density function for any $\eta > 0$.

5. Non-linear autoregressive model: $y_t^s = g_{obs}(y_{t-1}^s, \boldsymbol{\theta}) + u_t^s$, $u_t^s \sim F$. The moment function is the empirical CF:

$$\widehat{\boldsymbol{\psi}}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} \exp\left(i\tau'\left(y_t, y_{t-1}\right)\right).$$

The metric is the supremum distance between quantile functions:

$$\|F^{-1} - \widetilde{F}^{-1}\|_{\mathcal{B}} = \sup_{0\leq\nu\leq 1}|F^{-1}(\nu) - \widetilde{F}^{-1}(\nu)|.$$

The parameter space $\mathcal{B} = \{\beta = (\boldsymbol{\theta}, F), \|F^{-1}\|_\infty \leq C_1 < \infty, |\boldsymbol{\theta}| \leq C_2 < \infty\}$ is the space f distributions with bounded quantile functions. Furthermore, suppose that $|g_{obs}(y, \boldsymbol{\theta}) - g_{obs}(\widetilde{y}, \boldsymbol{\theta})| \leq C_3|y - \widetilde{y}| < |y - \widetilde{y}|$ for all $\boldsymbol{\theta}$ and $|g_{obs}(y, \boldsymbol{\theta}) - g_{obs}(y, \widetilde{\boldsymbol{\theta}})| \leq C_4|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}|$, then:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_\infty\leq\delta}\left|\exp\left(i\tau'\left(y_t^s(\beta_1), y_{t-1}^s(\beta_1)\right)\right) - \exp\left(i\tau'\left(y_t^s(\beta_2), y_{t-1}^s(\beta_2)\right)\right)\right|^2\right)\right]^{1/2}$$
$$\leq 2\frac{1+C_4}{1-C_3}\delta\|\tau\|_\infty.$$

Condition *i.* is satisfied with π equal to the normal density function for any $\eta > 0$.

The derivations for these examples are given below.

1. iid data without covariates: $y_t^s = u_t^s$, $u_t^s \sim F$. The moment function is the empirical CDF:

$$\widehat{\boldsymbol{\psi}}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{y_t\leq\tau}.$$

The metric is the supremum distance between CDFs: $\|F_1 - F_2\|_{\mathcal{B}} = \sup_y|F_1(y) - F_2(y)|$.
If $\sup_y\|F_1(y) - F_2(y)\|_{\mathcal{B}} \leq \delta$ then $F_1(y) - \delta \leq F_2(y) \leq F_1(y) + \delta$. Hence for $\tau \in \mathbb{R}$:

$$\begin{aligned}
|\mathbb{1}_{y_t^s\leq\tau} - \mathbb{1}_{\widetilde{y}_t^s\leq\tau}|^2 &\leq 2|\mathbb{1}_{y_t^s\leq\tau} - \mathbb{1}_{\widetilde{y}_t^s\leq\tau}| \\
&= 2|\mathbb{1}_{\nu_t^s\leq F(\tau)} - \mathbb{1}_{\nu_t^s\leq\widetilde{F}(\tau)}| \\
&\leq 2\left(\mathbb{1}_{\nu_t^s\leq F_1(\tau)+\delta} - \mathbb{1}_{\nu_t^s\leq F_1(\tau)-\delta}\right)
\end{aligned}$$

Taking expectations with respect to $\nu_t^s \sim \mathcal{U}_{[0,1]}$, for all $\tau \in \mathbb{R}$:

$$\mathbb{E}\left(\sup_{\sup_y|F_1(y)-F_2(y)|\leq\delta}|\mathbb{1}_{y_t^s\leq\tau} - \mathbb{1}_{\widetilde{y}_t^s\leq\tau}|^2\right) \leq 4\delta.$$

2. Single Index Model: $y_t^s = \mathbb{1}_{x_t'\boldsymbol{\theta}+u_t^s\leq 0}$, $u_t^s \sim F$. The moment function is the empirical CF:

$$\widehat{\boldsymbol{\psi}}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} \exp\left(i\tau'\left(y_t, x_t\right)\right).$$

The metric is the supremum distance between CDFs: $\|\beta_1 - \beta_2\|_{\mathcal{B}} = \sup_y |F_1(y) - F_2(y)|$ and the parameter space is $\mathcal{B} = \{\beta = (\boldsymbol{\theta}, F), \|F'\|_\infty \leq C_1 < \infty, \|\boldsymbol{\theta}\| \leq C_2 < \infty\}$, a space with CDFs with continuous and bounded densities. Also assume that $\mathbb{E}\|x_t\| < \infty$.

Proceeding similarly to example *i.*:

$$\begin{aligned}
|\mathbb{1}_{y_t^s(\beta_1)\leq\tau} - \mathbb{1}_{y_t^s(\beta_2)\leq\tau}|^2 &\leq 2|\mathbb{1}_{y_t^s(\beta_1)\leq\tau} - \mathbb{1}_{y_t^s(\beta_2)\leq\tau}| \\
&= 2|\mathbb{1}_{v_t^s\leq F_1(\tau - x_t'\boldsymbol{\theta}_1)} - \mathbb{1}_{v_t^s\leq F_2(\tau - x_t'\boldsymbol{\theta}_2)}| \\
&\leq 2|\mathbb{1}_{v_t^s\leq F_1(\tau - x_t'\boldsymbol{\theta}_1)} - \mathbb{1}_{v_t^s\leq F_2(\tau - x_t'\boldsymbol{\theta}_1)}| \\
&+ 2|\mathbb{1}_{v_t^s\leq F_2(\tau - x_t'\boldsymbol{\theta}_1)} - \mathbb{1}_{v_t^s\leq F_2(\tau - x_t'\boldsymbol{\theta}_2)}| \\
&\leq 2\left(\mathbb{1}_{v_t^s\leq F_1(\tau - x_t'\boldsymbol{\theta}_1)+\delta} - \mathbb{1}_{v_t^s\leq F_2(\tau - x_t'\boldsymbol{\theta}_1)-\delta}\right) \\
&+ 2|\mathbb{1}_{v_t^s\leq F_2(\tau - x_t'\boldsymbol{\theta})} - \mathbb{1}_{v_t^s\leq F_2(\tau - x_t'\boldsymbol{\theta}_2)}|
\end{aligned}$$

Without loss of generality, assume that $x_t \geq 0$ so that:

$$\begin{aligned}
|\mathbb{1}_{y_t^s(\beta_1)\leq\tau} - \mathbb{1}_{y_t^s(\beta_2)\leq\tau}|^2 &\leq 2|\mathbb{1}_{y_t^s(\beta_1)\leq\tau} - \mathbb{1}_{y_t^s(\beta_2)\leq\tau}| \\
&\leq 2\left(\mathbb{1}_{v_t^s\leq F_1(\tau - x_t'\boldsymbol{\theta}_1)+\delta} - \mathbb{1}_{v_t^s\leq F_2(\tau - x_t'\boldsymbol{\theta}_1)-\delta}\right) \\
&+ 2|\mathbb{1}_{v_t^s\leq F_2(\tau - x_t'[\boldsymbol{\theta}_1-\delta])} - \mathbb{1}_{v_t^s\leq F_2(\tau - x_t'[\boldsymbol{\theta}_1+\delta])}|.
\end{aligned}$$

Taking expectations with respect to $v_t^s \sim \mathcal{U}_{[0,1]}$, for all $\tau \in \mathbb{R}$:

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\beta=(\boldsymbol{\theta},F),\,\|\beta_1-\beta_2\|\leq\delta} |\mathbb{1}_{y_t^s(\beta_1)\leq\tau} - \mathbb{1}_{y_t^s(\beta_2)\leq\tau}|^2 \middle| x_t\right) \\
&\leq 2\left([F_1(\tau - x_t'\boldsymbol{\theta}_1) + \delta] - [F_1(\tau - x_t'\boldsymbol{\theta}_1) - \delta]\right) + 2\left(F_2(\tau - x_t'[\boldsymbol{\theta}_1 - \delta]) - F_2(\tau - x_t'[\boldsymbol{\theta}_1 + \delta])\right) \\
&\leq 4\delta + 4C_1\|x_t\|\delta.
\end{aligned}$$

And then, taking expectations with respect to x_t:

$$\mathbb{E}\left(\sup_{\beta=(\boldsymbol{\theta},F),\,\|\beta_1-\beta_2\|\leq\delta} |\mathbb{1}_{y_t^s(\beta_1)\leq\tau} - \mathbb{1}_{y_t^s(\beta_2)\leq\tau}|^2\right) \leq 4\left(1 + C_1\mathbb{E}\|x_t\|\right)\delta.$$

3. MA(1) model: $y_t^s = u_t^s + \boldsymbol{\theta}u_{t-1}^s$, $u_t^s \sim F$. The moment function is the empirical CF:

$$\widehat{\boldsymbol{\psi}}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} \exp\left(i\tau'\left(y_t, y_{t-1}\right)\right).$$

The metric is the supremum distance on quantiles: $\|F^{-1} - \widetilde{F}^{-1}\|_{\mathcal{B}} = \sup_{0\leq\nu\leq 1}|F^{-1}(\nu) - \widetilde{F}^{-1}(\nu)|$. The parameter space is $\mathcal{B} = \{\beta = (\boldsymbol{\theta}, F), \|F^{-1}\|_\infty \leq C_1 < \infty, |\boldsymbol{\theta}| \leq C_2 < \infty\}$, a space with bounded quantile functions.
As discussed in section 3.7, because the sine and cosine functions are Lipschitz continuous, the following holds for all $\tau = (\tau_1, \tau_2) \in \mathbb{R}^2$:

$$\begin{aligned}&\left|\exp\left(i\tau'\left(y_t^s(\beta_1), y_{t-1}^s(\beta_1)\right)\right) - \exp\left(i\tau'\left(y_t^s(\beta_2), y_{t-1}^s(\beta_2)\right)\right)\right| \\ &\leq \|\tau\|_\infty\left(|y_t^s(\beta_1) - y_t^s(\beta_2)| + |y_{t-1}^s(\beta_1) - y_{t-1}^s(\beta_2)|\right).\end{aligned}$$

Consider the case of $|y_t^s(\beta_1) - y_t^s(\beta_2)|$:

$$\begin{aligned}|y_t^s(\beta_1) - y_t^s(\beta_2)| &= |[F_1^{-1}(\nu_t^s) + \boldsymbol{\theta}_1 F_1^{-1}(\nu_{t-1}^s)] - [F_2^{-1}(\nu_t^s) + \boldsymbol{\theta}_2 F_2^{-1}(\nu_{t-1}^s)]| \\ &\leq |[F_1^{-1}(\nu_t^s) - F_2^{-1}(\nu_t^s)| + |\boldsymbol{\theta}_1||F_1^{-1}(\nu_{t-1}^s) - F_2^{-1}(\nu_{t-1}^s)| + |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||F_2^{-1}(\nu_{t-1}^s)| \\ &\leq (1 + C_2 + C_1)\delta.\end{aligned}$$

The same bound applies for $|y_{t-1}^s(\beta_1) - y_{t-1}^s(\beta_2)|$. Together with the previous inequalities it implies:

$$\begin{aligned}&\left|\exp\left(i\tau'\left(y_t^s(\beta_1), y_{t-1}^s(\beta_1)\right)\right) - \exp\left(i\tau'\left(y_t^s(\beta_2), y_{t-1}^s(\beta_2)\right)\right)\right|^2 \\ &\leq [2(1 + C_2 + C_1)\delta\|\tau\|_\infty]^2.\end{aligned}$$

4. AR(1) model: $y_t^s = \boldsymbol{\theta} y_{t-1}^s + u_t^s$, $u_t^s \sim F$. The moment function is the empirical CF:

$$\widehat{\psi}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n}\exp\left(i\tau'\left(y_t, y_{t-1}\right)\right).$$

The metric is the supremum distance on quantile functions:

$$\|F_1^{-1} - {F_2}^{-1}\|_{\mathcal{B}} = \sup_{0\leq\nu\leq 1}|F_1^{-1}(\nu) - {F_2}^{-1}(\nu)|.$$

The parameter space is $\mathcal{B} = \{\beta = (\boldsymbol{\theta}, F), \|F^{-1}\|_\infty \leq C_1 < \infty, |\boldsymbol{\theta}| \leq C_2 < 1\}$, a space with bounded quantile functions.
Similarly to the MA(1), only $|y_t^s(\beta) - y_t^s(\widetilde{\beta})|$ needs to be bounded:

$$\begin{aligned}|y_t^s(\beta_1) - y_t^s(\beta_2)| &= |[\boldsymbol{\theta}_1 y_{t-1}^s(\beta_1) + F_1^{-1}(\nu_t^s)] - [\boldsymbol{\theta}_2 y_{t-1}^s(\beta_2) + F_2^{-1}(\nu_t^s)]| \\ &\leq |F_1^{-1}(\nu_t^s) - F_2^{-1}(\nu_t^s)| + |\boldsymbol{\theta}_1||y_{t-1}^s(\beta_1) - y_{t-1}^s(\beta_2)| + |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||y_{t-1}^s(\beta_2)| \\ &\leq \delta\left(1 + \frac{C_1}{1 - C_2}\right) + |C_2||y_{t-1}^s(\beta_1) - y_{t-1}^s(\beta_2)|.\end{aligned}$$

The last inequality comes from the fact that $|\boldsymbol{\theta}_1| \leq C_2 < 1$ and $|F_1^{-1}| \leq C_2$ combined with the fact that $y_t^s(\beta) = \sum_{k=0}^{t-1} \boldsymbol{\theta}^k F^{-1}(\nu_t^s) + \boldsymbol{\theta}^t y_0$. The initial condition y_0 is fixed, so by iterating the previous inequality:

$$|y_t^s(\beta_1) - y_t^s(\beta_2)| \leq \delta \left(1 + \frac{C_1}{1 - C_2}\right) \frac{1}{1 - C_2}.$$

Applying this inequality and the Lipschitz continuity of the sine and cosine functions:

$$\begin{aligned} &\left| \exp\left(i\tau' \left(y_t^s(\beta_1), y_{t-1}^s(\beta_1)\right)\right) - \exp\left(i\tau' \left(y_t^s(\beta_2), y_{t-1}^s(\beta_2)\right)\right) \right|^2 \\ &\leq [2 \left(1 + \frac{C_1}{1 - C_2}\right) \frac{1}{1 - C_2} \delta \|\tau\|_\infty]^2. \end{aligned}$$

5. Non-linear autoregressive model: $y_t^s = g_{obs}(y_{t-1}^s, \boldsymbol{\theta}) + u_t^s$, $u_t^s \sim F$. The moment function is the empirical CF:

$$\widehat{\psi}_n(\tau) = \frac{1}{n} \sum_{t=1}^{n} \exp\left(i\tau' \left(y_t, y_{t-1}\right)\right).$$

The metric is the supremum distance on quantile functions:

$$\|F_1^{-1} - {F_2}^{-1}\|_{\mathcal{B}} = \sup_{0 \leq \nu \leq 1} |F_1^{-1}(\nu) - {F_2}^{-1}(\nu)|.$$

The parameter space is $\mathcal{B} = \{\beta = (\boldsymbol{\theta}, F), \|F^{-1}\|_\infty \leq C_1 < \infty, |\boldsymbol{\theta}| \leq C_2 < \infty\}$, a space with bounded quantile functions. Furthermore, assume $|g_{obs}(y, \boldsymbol{\theta}) - g_{obs}(\widetilde{y}, \boldsymbol{\theta})| \leq C_3 |y - \widetilde{y}| < |y - \widetilde{y}|$ for all $\boldsymbol{\theta}$ and $|g_{obs}(y, \boldsymbol{\theta}) - g_{obs}(y, \widetilde{\boldsymbol{\theta}})| \leq C_4 |\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}|$.
The proof is similar to the AR(1) example, first y_t^s needs to be bounded:

$$\begin{aligned} |y_t^s(\beta_1) - y_t^s(\beta_2)| &= |[g_{obs}(y_{t-1}^s(\beta_1), \boldsymbol{\theta}) + F_1^{-1}(\nu_t^s)] - [g_{obs}(y_{t-1}^s(\beta_2), \boldsymbol{\theta}_2) + F_2^{-1}(\nu_t^s)]| \\ &\leq |F_1^{-1}(\nu_t^s) - F_2^{-1}(\nu_t^s)| + |g_{obs}(y_{t-1}^s(\beta_1), \boldsymbol{\theta}_1) - g_{obs}(y_{t-1}^s(\beta_2), \boldsymbol{\theta}_1)| \\ &+ |g_{obs}(y_{t-1}^s(\beta_1), \boldsymbol{\theta}_2) - g_{obs}(y_{t-1}^s(\beta_2), \boldsymbol{\theta}_2)| \\ &\leq (1 + C_4)\delta + C_3 |y_{t-1}^s(h) - y_{t-1}^s(\beta_2)|. \end{aligned}$$

Iterating this inequality up to $t = 0$ where the initial condition is fixed implies:

$$|y_t^s(\beta_1) - y_t^s(\beta_2)| \leq \frac{1 + C_4}{1 - C_3} \delta.$$

Similarly to the MA(1) and AR(1) models:

$$\begin{aligned} &\left| \exp\left(i\tau' \left(y_t^s(\beta_1), y_{t-1}^s(\beta_1)\right)\right) - \exp\left(i\tau' \left(y_t^s(\beta_2), y_{t-1}^s(\beta_2)\right)\right) \right|^2 \\ &\leq \left(2 \frac{1 + C_4}{1 - C_3} \delta \|\tau\|_\infty\right)^2. \end{aligned}$$

Interpretation of the Equivalence Conditions

To prove the existence of an $\underline{a} > 0$ in Assumption .0.6, Chen & Pouzo (2015) use an eigenvalue condition on the variance of the moments. Since they have a bounded support the smallest eigenvalue can be bounded below. Here, the variance operator is infinite dimensional (see Carrasco & Florens, 2000, for a discussion) so that the eigenvalues may not be bounded below. However, an interpretation in terms of the eigenvalues and eigenvectors of the variance opertator is still possible. First, note that $\sigma_n^*, \|v_n^*\|_{weak}$ can be written as:

$$\|v_n^*\|_{weak}^2 = \int \left[Real\left(B\psi_\beta(\tau, v_n^*)\right)^2 + Im\left(B\psi_\beta(\tau, v_n^*)\right)^2 \right] \pi(\tau) d\tau$$

$$\sigma_n^{*2} = \mathbb{E}\left[\int Real\left(B\psi_\beta(\tau, v_n^*)\right) Real\left(BZ_n^S(\tau)\right) + Im\left(B\psi_\beta(\tau, v_n^*)\right) Im\left(BZ_n^S(\tau)\right) \pi(\tau) d\tau \right]^2$$

The sieve variance can be expanded into three terms:

$$\begin{aligned}
\sigma_n^{*2} = & \int Real\left(B\psi_\beta(\tau_1, v_n^*)\right) Real\left(B\psi_\beta(\tau_2, v_n^*)\right) \mathbb{E}\left[Real\left(BZ_n^S(\tau_1)\right) Real\left(BZ_n^S(\tau_2)\right) \right] \pi(\tau_1)\pi(\tau_2) d\tau_1 d\tau_2 \\
& + \int Im\left(B\psi_\beta(\tau_1, v_n^*)\right) Im\left(B\psi_\beta(\tau_2, v_n^*)\right) \mathbb{E}\left[Im\left(BZ_n^S(\tau_1)\right) Im\left(BZ_n^S(\tau_2)\right) \right] \pi(\tau_1)\pi(\tau_2) d\tau_1 d\tau_2 \\
& + 2 \int Real\left(B\psi_\beta(\tau_1, v_n^*)\right) Im\left(B\psi_\beta(\tau_2, v_n^*)\right) \mathbb{E}\left[Real\left(BZ_n^S(\tau_1)\right) Im\left(BZ_n^S(\tau_2)\right) \right] \pi(\tau_1)\pi(\tau_2) d\tau_1 d\tau_2.
\end{aligned}$$

This expansion can be re-written more compactly in matrix form:

$$\sigma_n^{*2} = \int \begin{pmatrix} Real\left(B\psi_\beta(\tau_1, v_n^*)\right) \\ Im\left(B\psi_\beta(\tau_1, v_n^*)\right) \end{pmatrix}' \Sigma_n(\tau_1, \tau_2) \begin{pmatrix} Real\left(B\psi_\beta(\tau_2, v_n^*)\right) \\ Im\left(B\psi_\beta(\tau_2, v_n^*)\right) \end{pmatrix} \pi(\tau_1)\pi(\tau_2) d\tau_1 d\tau_2$$

where

$$\Sigma_n(\tau_1, \tau_2) = n\mathbb{E} \begin{pmatrix} Real\left(BZ_n^S(\tau_1)\right) Real\left(BZ_n^S(\tau_2)\right) & Real\left(BZ_n^S(\tau_1)\right) Im\left(BZ_n^S(\tau_2)\right) \\ Im\left(BZ_n^S(\tau_2)\right) Im\left(BZ_n^S(\tau_1)\right) & Im\left(BZ_n^S(\tau_1)\right) Im\left(BZ_n^S(\tau_2)\right) \end{pmatrix}.$$

Before comparing this expression with $\|v_n^*\|_{weak}$ further simplifications are possible. Let K_n be the operator satisfying:

$$K_n B\psi_\beta(\tau, v_n^*) = \int \Sigma_n(\tau, \tau_2) \begin{pmatrix} Real\left(B\psi_\beta(\tau_2, v_n^*)\right) \\ Im\left(B\psi_\beta(\tau_2, v_n^*)\right) \end{pmatrix} \pi(\tau_2) d\tau_2$$

Then the sieve variance can be expressed in terms of the operator K_n:

$$\sigma_n^{*2} = \int B\psi_\beta(\tau, v_n^*) K_n \begin{pmatrix} Real\left(B\psi_\beta(\tau, v_n^*)\right) \\ Im\left(B\psi_\beta(\tau, v_n^*)\right) \end{pmatrix} \pi(\tau) d\tau$$

The term $\|v_n^*\|_{weak}$ can also be re-written in a similar notation:

$$\|v_n^*\|_{weak}^2 = \int \begin{pmatrix} Real\left(B\boldsymbol{\psi}_\beta(\tau, v_n^*)\right) \\ Im\left(B\boldsymbol{\psi}_\beta(\tau, v_n^*)\right) \end{pmatrix}' \begin{pmatrix} Real\left(B\boldsymbol{\psi}_\beta(\tau, v_n^*)\right) \\ Im\left(B\boldsymbol{\psi}_\beta(\tau, v_n^*)\right) \end{pmatrix} \pi(\tau)d\tau$$

Now note that these integrals are associated with an inner product in the Hilbert space $\left(L^2(\pi), \langle\cdot,\cdot\rangle_{L^2(\pi)}\right)$ with for all complex valued $\varphi_1, \varphi_2 \in L^2(\pi)$:

$$\langle\varphi_1, \varphi_2\rangle_{L^2(\pi)} = \int \begin{pmatrix} Real(\varphi_1(\tau)) \\ Im(\varphi_1(\tau)) \end{pmatrix}' \begin{pmatrix} Real(\varphi_1(\tau)) \\ Im(\varphi_1(\tau)) \end{pmatrix} \pi(\tau)d\tau.$$

As a result, Assumption .0.6 can be re-written in terms of the covariance operator K_n:

$$\underline{a}\langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), \boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)} \leq \langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), K_n\boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)}.$$

Since $\sigma_n^* > 0$ by construction, K_n has positive eigenvalues. Let $(\varphi_{1,n}, \varphi_{2,n}, \dots)$ be the eigenvector associated with K_n and $(\lambda_{1,n}, \lambda_{2,n}, \dots)$ the associated eigenvalues (in decreasing modulus). Then $B\boldsymbol{\psi}_\beta(\cdot, v_n^*) = \sum_{j\geq 1} a_{j,n}\varphi_{j,n}$ and

$$\begin{aligned}\langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), K_n\boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)} &= \sum_{j\geq 1} a_{j,n}^2\lambda_{j,n} \\ \langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), \boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)} &= \sum_{j\geq 1} a_{j,n}^2.\end{aligned}$$

To go further, there are two cases:

i. $\|v_n^*\|_{weak} \to \infty$ (slower than $\sqrt{n}$ convergence rate): assume that there exists a pair $(a_{j,n}, \lambda_{j,n})$ such that $\lambda_{j,n} \geq \underline{\lambda}_j > 0$ and $a_{j,n} \to \infty$ at the same rate as $\|v_n^*\|_{weak}$: $\frac{a_{j,n}}{\|v_n^*\|_{weak}} \geq \underline{a}_j > 0$. In this case:

$$\begin{aligned}\langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), K_n\boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)} \geq a_{j,n}^2\underline{\lambda}_j &\geq \frac{a_{j,n}^2\underline{\lambda}_j}{\langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), \boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)}}\langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), \boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)} \\ &\geq \underline{a}_j\langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), \boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)}.\end{aligned}$$

Take for instance $\underline{a} = \underline{a}_j > 0$.

ii. $\|v_n^*\|_{weak} \not\to \infty$ ($\sqrt{n}$ convergence rate): it suffice that there exist a pair $(a_{j,n}, \lambda_{j,n})$ such that $\lambda_{j,n} \geq \underline{\lambda}_j > 0$ and $a_{j,n} \geq \underline{a}_j > 0$. In this case:

$$\langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), K_n\boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)} \geq \underline{a}_j^2\underline{\lambda}_j \geq \frac{\underline{a}_j^2\underline{\lambda}_j}{\langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), \boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)}}\langle\boldsymbol{\psi}_\beta(\cdot, v_n^*), \boldsymbol{\psi}_\beta(\cdot, v_n^*)\rangle_{L^2(\pi)}.$$

Let $\underline{a} = \text{diag}_{n\geq 1} \frac{\underline{a}_j^2 \lambda_j}{\langle \psi_\beta(\cdot, v_n^*), \psi_\beta(\cdot, v_n^*) \rangle_{L^2(\pi)}} > 0$ by assumption.

To satisfy the equivalence condition, the moments ψ_β must project on the covariance operator in directions where the variance increases at least as fast as the weak norm.

Proofs for the Additional Asymptotic Results

Consistency

The following lemma, taken from Chen & Pouzo (2012) (the notation is adapted for this paper's setting), gives sufficient conditions for consistency.

Lemma .0.1. *Let $\widehat{\beta}_n$ be such that $\widehat{Q}_n(\widehat{\beta}_n) \leq diag_{\beta \in \mathcal{B}_{k(n)}} + O_{p^*}(\eta_n)$, where $(\eta_n)_{n\geq 1}$ is a positive real-valued sequence such that $\eta_n = o(1)$. Let $\overline{Q}_n : \mathcal{B} \to [0, +\infty)$ be a sequence of non-random measurable functions and let the following conditions hold:*

a. i) $0 \leq \overline{Q}_n(\beta_0) = o(1)$; ii) there is a positive function $g_0(n, k, \varepsilon)$ such that:

$$diag_{h \in \mathcal{B}_k : \|\beta - \beta_0\|_{\mathcal{B}} > \varepsilon} \overline{Q}_n(\beta) \geq g_0(n, k, \varepsilon) > 0 \text{ for each } n, k \geq 1$$

and $\lim diag_{n\to\infty} g_0(n, k(n), \varepsilon) \geq 0$ for all $\varepsilon > 0$.

b. i) $\mathcal{B}$ is an infinite dimensional, possibly non-compact subset of a Banach space $(B, \|\|_{\mathcal{B}})$; ii) $\mathcal{B}_k \subseteq \mathcal{B}_{k+1} \subseteq \mathcal{B}$ for all $k \geq 1$, and there is a sequence $\{\Pi_{k(n)}\beta_0 \in \mathcal{B}_{k(n)}\}$ such that $\overline{Q}_n(\Pi_{k(n)}\beta_0) = o(1)$.

c. $\widehat{Q}_n(\beta)$ is jointly measurable in the data $(y_t, x_t)_{t\geq 1}$ and the parameter $h \in \mathcal{B}_{k(n)}$.

d. i) $\widehat{Q}_n(\Pi_{k(n)}\beta_0) \leq K_0 \overline{Q}_n(\Pi_{k(n)}\beta_0) + O_{p^*}(c_{0,n})$ for some $c_{0,n} = o(1)$ and a finite constant $K_0 > 0$; ii) $\widehat{Q}_n(\beta) \geq K\overline{Q}_n(\beta) - O_{p^*}(c_n)$ uniformly over $h \in \mathcal{B}_{k(n)}$ for some $c_n = o(1)$ and a finite constant $K > 0$; iii) $\max(c_{0,n}, c_n, \overline{Q}_n(\Pi_{k(n)}\beta_0), \eta_n) = o(g_0(n, k(n), \varepsilon))$ for all $\varepsilon > 0$.

Then for all $\varepsilon > 0$:

$$\mathbb{P}^* \left(\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} > \varepsilon \right) \to 0 \text{ as } n \to \infty.$$

Remark .0.1. *Condition a. is an identification conditions. Condition b. requires the sieve approximation to be valid for the objective function. Condition d. gives an asymptotic equivalence between $\widehat{Q}_n$ and $\overline{Q}_n$ up to a $O_{p^*}(\max(c_n, c_{0,n}))$ term; if one is close to zero, the other must be as well. It also requires that the sieve approximation rate, the rate at which $\overline{Q}_n$ and $\widehat{Q}_n$ become equivalent and the approximation error goes to zero faster than the ill-posedness of the problem as measured by g_0.*

Proof of Proposition .0.1. :
In the iid case, if y_t^s depends on f only via the shocks e_t^s, i.e. $y_t^s = g_{obs}(x_t, \boldsymbol{\theta}, e_t^s)$, then $\mathbb{E}(\widehat{\boldsymbol{\psi}}_t^s(\tau, \beta)) = \int \mathbb{E}\left(\exp(i\tau'(g_{obs}(x_t, \boldsymbol{\theta}, \boldsymbol{\varepsilon}), x_t)) f(\boldsymbol{\varepsilon}) d\boldsymbol{\varepsilon}\right)$ for each τ. First note that $\Pi_{k(n)}\beta_0 = (\boldsymbol{\theta}_0, \Pi_{k(n)} f_0)$ and:

$$\begin{aligned}\left|\mathbb{E}[\widehat{\boldsymbol{\psi}}_t^s(\tau, \Pi_{k(n)}\beta_0) - \widehat{\boldsymbol{\psi}}_t(\tau)]\right| &= \left|\mathbb{E}[\widehat{\boldsymbol{\psi}}_t^s(\tau, \Pi_{k(n)}\beta_0) - \widehat{\boldsymbol{\psi}}_t^s(\tau, \beta_0)]\right| \\ &= \left| \int \mathbb{E}\left(\exp(i\tau'(g_{obs}(x_t, \boldsymbol{\theta}_0, u), x_t))\right) [\Pi_{k(n)} f_0(u) - f_0(u)] du \right| \\ &\leq \int |\Pi_{k(n)} f_0(u) - f_0(u)| du = \|\Pi_{k(n)} f_0 - f_0\|_{TV}.\end{aligned}$$

Taking squares on both sides and integrating:

$$\int \left|\mathbb{E}[(\widehat{\boldsymbol{\psi}}_t^s(\tau, \Pi_{k(n)}\beta_0) - \widehat{\boldsymbol{\psi}}_t(\tau))]\right|^2 \pi(\tau) d\tau \leq \|\Pi_{k(n)} f_0 - f_0\|_{TV}^2.$$

To conclude the proof, use the assumption that B is bounded linear so that:

$$Q_n(\Pi_{k(n)}\beta_0) \leq M_B^2 \int \left|\mathbb{E}[(\widehat{\boldsymbol{\psi}}_t^s(\tau, \Pi_{k(n)}\beta_0) - \widehat{\boldsymbol{\psi}}_t(\tau))]\right|^2 \pi(\tau) d\tau \leq M_B^2 \|\Pi_{k(n)} f_0 - f_0\|_{TV}^2.$$

□

Proof of Proposition .0.2. :
To prove the proposition, proceed in four steps:

1. First, Assumption .0.4 implies:

$$\int |\widehat{\boldsymbol{\psi}}_n(\tau) - \mathbb{E}(\widehat{\boldsymbol{\psi}}_n(\tau))|^2 \pi(\tau) d\tau = O_p(1/n)$$

2. It also implies that, uniformly over $\beta \in \mathcal{B}_{k(n)}$:

$$\int |\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta) - \mathbb{E}(\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta))|^2 \pi(\tau) d\tau = O_p(C_n/n)$$

3. By the triangular inequality, the previous two results imply that, uniformly over $\beta \in \mathcal{B}_{k(n)}$:

$$\int \left|[\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta) - \widehat{\boldsymbol{\psi}}_n(\tau)] - \mathbb{E}[\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta) - \widehat{\boldsymbol{\psi}}_n(\tau)]\right|^2 \pi(\tau) d\tau = O_p(\max(1, C_n)/n).$$

And, because B is a bounded linear operator:

$$\begin{aligned}&\int \left|[B\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta) - B\widehat{\boldsymbol{\psi}}_n(\tau)] - \mathbb{E}[B\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta) - B\widehat{\boldsymbol{\psi}}_n(\tau)]\right|^2 \pi(\tau) d\tau \\ &\leq M_B^2 \int \left|[\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta) - \widehat{\boldsymbol{\psi}}_n(\tau)] - \mathbb{E}[\widehat{\boldsymbol{\psi}}_n^S(\tau, \beta) - \widehat{\boldsymbol{\psi}}_n(\tau)]\right|^2 \pi(\tau) d\tau = O_p(\max(1, C_n)/n).\end{aligned}$$

4. Using the inequality $|a-b|^2 \geq 1/2|a|^2 + |b|^2$ and the previous result, uniformly over $\beta \in \mathcal{B}_{k(n)}$:

$$1/2 \int |B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau)|^2 \pi(\tau) d\tau \leq \int |\mathbb{E}(B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau))|^2 \pi(\tau) d\tau + O_p(\max(1, C_n)/n)$$

and

$$1/2 \int |\mathbb{E}(B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau))|^2 \pi(\tau) d\tau \leq \int |B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau)|^2 \pi(\tau) d\tau + O_p(\max(1, C_n)/n).$$

The last step concludes the proof of the proposition with $\delta_n^2 = \max(1, C_n)/n = o(1)$ if $C_n/n \to 0$ as $n \to \infty$.

First, consider steps *1.* and *2*:

Step 1.: For $M > 0$, a convergence rate r_n and Markov's inequality:

$$\begin{aligned}
\mathbb{P}\left(\int |\widehat{\psi}_n(\tau) - \mathbb{E}(\widehat{\psi}_n(\tau))|^2 \pi(\tau) d\tau \geq M r_n \right) &\leq \frac{1}{M r_n} \mathbb{E}\left(\int |\widehat{\psi}_n(\tau) - \mathbb{E}(\widehat{\psi}_n(\tau))|^2 \pi(\tau) d\tau \right) \\
&= \frac{1}{M r_n} \int \mathbb{E}\left(|\widehat{\psi}_n(\tau) - \mathbb{E}(\widehat{\psi}_n(\tau))|^2 \right) \pi(\tau) d\tau \\
&\leq \frac{2}{M r_n} \frac{1 + 24 \sum_{m \geq 0} \alpha(m)^{1/p}}{n} \int \pi(\tau) d\tau \\
&\leq \frac{C_{\alpha,p}}{M r_n n}.
\end{aligned}$$

The last two inequalities come from Lemma .0.2. If the data is iid then the mixing coefficients $\alpha(m) = 0$ for all $m \geq 1$. $C_{\alpha,p}$ is a constant that only depends on the mixing rate α, p and the bound on $|\widehat{\psi}_t(\tau) - \mathbb{E}(\widehat{\psi}_t(\tau))| \leq 2$. For $r_n = 1/n$ and $M \to \infty$ the probability goes to zero. As a result: $\int |\widehat{\psi}_n(\tau) - \mathbb{E}(\widehat{\psi}_n(\tau))|^2 \pi(\tau) d\tau = O_p(1/n)$.

Step 2.: The proof is similar to the proof of lemma C.1 in Chen & Pouzo (2012). It also begins similarly to *Step 1*, for $M > 0$, a convergence rate r_n and Markov's inequality:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\beta \in \mathcal{B}_{k(n)}} \int |\widehat{\psi}_n^S(\tau,\beta) - \mathbb{E}(\widehat{\psi}_n^S(\tau,\beta))|^2 \pi(\tau) d\tau \geq M r_n \right) \\
&\quad \leq \frac{1}{M r_n} \mathbb{E}\left(\sup_{\beta \in \mathcal{B}_{k(n)}} \int |\widehat{\psi}_n^S(\tau,\beta) - \mathbb{E}(\widehat{\psi}_n^S(\tau,\beta))|^2 \pi(\tau) d\tau \right) \\
&\quad \leq \frac{1}{M r_n} \int \mathbb{E}\left(\sup_{\beta \in \mathcal{B}_{k(n)}} |\widehat{\psi}_n^S(\tau,\beta) - \mathbb{E}(\widehat{\psi}_n^S(\tau))|^2 \right) \pi(\tau) d\tau \\
&\quad \leq \frac{1}{M r_n} \int \mathbb{E}\left(\sup_{\beta \in \mathcal{B}_{k(n)}} |\widehat{\psi}_n^s(\tau,\beta) - \mathbb{E}(\widehat{\psi}_n^s(\tau))|^2 \right) \pi(\tau) d\tau
\end{aligned}$$

Suppose that there is an upper bound C_n such that for all τ:

$$\mathbb{E}\left(\sup_{\beta\in\mathcal{B}_{k(n)}}|[\widehat{\psi}_n^s(\tau,\beta)-\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta))]\pi(\tau)^{1/(2+\eta)}|^2\right)\leq C_n/n$$

If the following also holds $\int \pi(\tau)^{1-2/(2+\eta)}d\tau = C_\eta < \infty$ then:

$$\frac{1}{Mr_n}\int \mathbb{E}\left(\sup_{h\in\mathcal{B}_{k(n)}}|\widehat{\psi}_n^s(\tau,\beta)-\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta))|^2\right)\pi(\tau)d\tau \leq \frac{C_\eta C_n}{Mr_n n}.$$

Take $r_n = C_n/n = o(1)$, then for $M\to\infty$ the probability goes to zero. As a result:

$$\sup_{\beta\in\mathcal{B}_{k(n)}}\int|\widehat{\psi}_n^S(\tau,\beta)-\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta))|^2\pi(\tau)d\tau = O_p(C_n/n).$$

The bounds C_n are now computed, first in the iid case. By theorem 2.14.5 of van der Vaart & Wellner (1996):

$$\begin{aligned}&\mathbb{E}\left(\sup_{\beta\in\mathcal{B}_{k(n)}}\left|\sqrt{n}[\widehat{\psi}_n^s(\tau,\beta)-\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta))]\pi(\tau)^{1/(2+\eta)}\right|^2\right)\\&\leq\left(1+\mathbb{E}\left(\sup_{\beta\in\mathcal{B}_{k(n)}}\left|\sqrt{n}[\widehat{\psi}_n^s(\tau,\beta)-\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta))]\pi(\tau)^{1/(2+\eta)}\right|\right)\right)^2.\end{aligned}$$

Also, by theorem 2.14.2 of van der Vaart & Wellner (1996) there exists a universal constant $K>0$ such that for each $\tau\in\mathbb{R}^{d_\tau}$:

$$\mathbb{E}\left(\sup_{\beta\in\mathcal{B}_{k(n)}}\left|\sqrt{n}[\widehat{\psi}_n^s(\tau,\beta)-\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta))]\pi(\tau)^{1/(2+\eta)}\right|\right)\leq K\int_0^1\sqrt{1+\log N_{[]}(x,\Psi_{k(n)},\|\cdot\|)}dx$$

with $\Psi_{k(n)} = \{\psi : \mathcal{B}_{k(n)}\to\mathbb{C}, \beta\to\psi_t^S(\tau,\beta)\pi(\tau)^{1/(2+\eta)}\}$, $N_{[]}$ is the covering number with bracketing. Because of the L^p-smoothness, it is bounded above by:

$$N_{[]}(x,\Psi_{k(n)},\|\cdot\|)\leq N_{[]}(\frac{x^{1/\gamma}}{C^{1/\gamma}},\mathcal{B}_{k(n)},\|\cdot\|)\leq C'N_{[]}(x^{1/\gamma},\mathcal{B}_{k(n)},\|\cdot\|).$$

Let $\sqrt{C_n} = \sqrt{1+\log N_{[]}(x^{1/\gamma},\mathcal{B}_{k(n)},\|\cdot\|)}dx$, together with the previous inequality, it implies:

$$\mathbb{E}\left(\sup_{\beta\in\mathcal{B}_{k(n)}}\left|\sqrt{n}[\widehat{\psi}_n^s(\tau,\beta)-\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta))]\pi(\tau)^{1/(2+\eta)}\right|^2\right)\leq\left(1+K\sqrt{C_n}\right)^2\leq 4(1+K^2)C_n.$$

To conclude, divide by n on both sides to get the bound:

$$\mathbb{E}\left(\sup_{\beta\in\mathcal{B}_{k(n)}}\left|[\widehat{\psi}_n^s(\tau,\beta)-\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta))]\pi(\tau)^{1/(2+\eta)}\right|^2\right)\leq 4(1+K^2)C_n/n.$$

For the dependent case, Lemma .0.4 implies that if $\widehat{\psi}_t^s(\tau,\beta)$ is α-mixing at an exponential rate, the moments are bounded and the sieve spaces are compact:

$$\mathbb{E}\left(\sup_{\beta\in\mathcal{B}_{k(n)}}\left|\sqrt{n}[\widehat{\psi}_n^s(\tau,\beta)-\mathbb{E}(\widehat{\psi}_n^s(\tau,\beta))]\pi(\tau)^{1/(2+\eta)}\right|^2\right)\leq\left(1+K\sqrt{C_n}\right)^2\leq KC_n$$

with, for any $\vartheta \in (0,1)$ such that the integral exists:

$$C_n = \int_0^1 \left(x^{\vartheta/2-1} \sqrt{\log N_{[]}(x^{1/\gamma}, \mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}})} + \log^2 N_{[]}(x^{1/\gamma}, \mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}}) \right) dx$$

Step 3.: follows from the triangular inequality and the assumption that B is a bounded linear operator.

Step 4.: The following two inequalities can be derived from the inequality $|a-b|^2 \geq 1/2|a|^2 + |b|^2$, which is symmetric in a and b:

$$\begin{aligned}
&\left| [B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau)] - \mathbb{E}[B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau)] \right|^2 \\
&\geq 1/2 \left| B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau) \right|^2 + \left| \mathbb{E}[B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau)] \right|^2
\end{aligned}$$

and

$$\begin{aligned}
&\left| [B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau)] - \mathbb{E}[B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau)] \right|^2 \\
&\geq \left| B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau) \right|^2 + 1/2 \left| \mathbb{E}[B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau)] \right|^2.
\end{aligned}$$

Taking integrals on both sides and given that $\int \left| [B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau)] - \mathbb{E}[B\widehat{\psi}_n^S(\tau,\beta) - B\widehat{\psi}_n(\tau)] \right|^2 \pi(\tau) d\tau$ is $O_p(C_n/n)$ uniformly in $h \in \mathcal{B}_{k(n)}$, the desired result follows:

$$\begin{aligned}
1/2\widehat{Q}_n^S(\beta) &\leq Q_n(\beta) + O_p(C_n/n) \\
1/2Q_n(\beta) &\leq \widehat{Q}_n^S(\beta) + O_p(C_n/n).
\end{aligned}$$

With this, it follows that Assumption .0.3 is satisfied. □

Lemma .0.2. *Let $(Y_t)_{t\geq 1}$ mean zero, α-mixing with rate $\alpha(m)$ such that $\sum_{m\geq 1} \alpha(m)^{1/p} < \infty$ for some $p > 1$, and $|Y_t| \leq 1$ for all $t \geq 1$. Then we have:*

$$\mathbb{E}\left(n|\overline{Y}_n|^2 \right) \leq 1 + 24 \sum_{m\geq 1} \alpha(m)^{1/p}$$

Proof of Lemma .0.2: The proof follows from Davydov (1968)'s inequality: let $p, q, r \geq 0, 1/p + 1/q + 1/r = 1$, for any random variables X, Y:

$$|cov(X,Y)| \leq 12\alpha(\sigma(X), \sigma(Y))^{1/p} \mathbb{E}(|X|^q)^{1/q} \mathbb{E}(|Y|^r)^{1/r}$$

where $\alpha(\sigma(X),\sigma(Y))$ is the mixing coefficient between X and Y. As a result:

$$\begin{aligned}
\mathbb{E}\left(n|\overline{Y}_n|^2\right) &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}(|X_n|^2) + \frac{1}{n}\sum_{t\neq t'} cov(Y_t, Y_{t'}) \\
&\leq 1 + 2\times\frac{1}{n}\sum_{t>t'} cov(Y_t, Y_{t'}) \\
&\leq 1 + 24\times\frac{1}{n}\sum_{t>t'}\alpha(\sigma(Y_t),\sigma(Y_{t'}))^{1/p}(\mathbb{E}|Y_t|^q)^{1/q}(\mathbb{E}|Y_{t'}|^r)^{1/r} \\
&= 1 + 24\sum_{m=1}^{n}\frac{n-m}{n}\alpha(m)^{1/p} \\
&\leq 1 + 24\sum_{m=1}^{\infty}\alpha(m)^{1/p}.
\end{aligned}$$

□

The following lemma gives a Rosenthal type inequality for possibly non-stationary α-mixing random variables. As shown in van der Vaart & Wellner (1996) and Dedecker & Louhichi (2002) these inequalities are very important to bound the expected value of the supremum of an empirical process.

Lemma .0.3. *Let $(X_t)_{t>0}$ be a sequence of real-valued, centered random variables and $(\alpha_m)_{m\geq 0}$ be the sequence of strong mixing coefficients. Suppose that X_t is uniformly bounded and there exists $A, C > 0$ such that $\alpha(m) \leq A\exp(-Cm)$ then there exists $K > 0$ that depends only on the mixing coefficients such that for any $p \geq 2$:*

$$\mathbb{E}\left(|\sqrt{n}\overline{X}_n|^p\right)^{1/p} \leq K\left[\sqrt{p}\left(\int_0^1 \min(\alpha^{-1}(u), n)\sum_{t=1}^{n}\frac{Q_t^2(u)}{n}\right)^{1/2} + n^{1/p-1/2}p^2\|sup_{t>0}X_t\|_\infty\right]$$

where Q_t is the quantile function of X_t, $\min(\alpha^{-1}(u), n) = \sum_{i=k}^{n}\mathbb{1}_{u\leq\alpha_k}$.

Proof of Lemma .0.3: Theorem 6.3 Rio (2000) implies the following inequality:

$$\mathbb{E}\left(|\sum_{t=1}^{n}X_t|^p\right) \leq a_p s_n^p + nb_p\int_0^1 \min(\alpha^{-1}(u), n)^{p-1}Q^p(u)du$$

where $a_p = p4^{p+1}(p+1)^{p/2}$ and $b_p = \frac{p}{p-1}4^{p+1}(p+1)^{p-1}$, $Q = \sup_{t>0}Q_t$ and $s_n^2 = \sum_{t=1}^{n}\sum_{t'=1}^{n}|\text{cov}(X_t, X_{t'})|$.
Since X_t is uniformly bounded, using the results from appendix C of Rio (2000):

$$\int_0^1 \min(\alpha^{-1}(u), n)^{p-1}Q^p(u)du \leq 2\left[\sum_{k=0}^{n-1}(k+1)^{p-1}\alpha_k\right]\|\sup_{t>0}X_t\|_\infty.$$

Because the strong-mixing coefficients are exponentially decreasing, it implies:

$$\sum_{k=0}^{n-1}(k+1)^{p-1}\alpha_k \leq A\exp(C)\sum_{k\geq 1}k^{p-1}\exp(-Ck) \leq A\exp(C)(p-1)^{p-1}\frac{1}{(1-\exp(-C))^{p-1}}$$

And corollary 1.1 of Rio (2000) yields:

$$s_n^2 \leq 4\int_0^1 \min(\alpha^{-1}(u),n)\sum_{t=1}^{n}Q_k^2(u)du.$$

Altogether:

$$\begin{aligned}\mathbb{E}\left(|\sqrt{n}\overline{X}_n|^p\right)^{1/p} &\leq K_1(p+1)^{1/2}\left(\int_0^1 \min(\alpha^{-1}(u),n)\sum_{t=1}^{n}\frac{Q_t^2(u)}{n}\right)^{1/2}\\ &\quad + K_2 n^{1/p-1/2}(p-1)^{(p-1)/p}(p+1)^{(p-1)/p}\|\sup_{t>0}X_t\|_\infty\\ &\leq K\left(\sqrt{p}\left(\int_0^1 \min(\alpha^{-1}(u),n)\sum_{t=1}^{n}\frac{Q_t^2(u)}{n}\right)^{1/2} + n^{1/p-1/2}p^2\|\sup_{t>0}X_t\|_\infty\right).\end{aligned}$$

with $K_1 \geq 2^{1/p}p^{1/p}4^{(p+1)/p}$, $K_2 \geq (p/[p-1])^{1/p}4^{(p+1)/p}2^{1/p}A\exp(C)\frac{1}{(1-\exp(-C))^{(p-1)/p}}$. Note that since $p \geq 2$, $2^{1/p} \leq \sqrt{2}, p^{1/p} \leq 1, 4^{(p+1)/p} \leq 16$, etc. The constants K_1, K_2 do not depend on p. K only depends on the constants A and C. □

Lemma .0.4. *Suppose that $(X_t(\beta))_{t>0}$ is a real valued, mean zero random process for any $\beta \in \mathcal{B}$. Suppose that it is α-mixing with exponential decay: $\alpha(m) \leq A\exp(-Cm)$ for $A, C > 0$ and bounded $|X_t(\beta)| \leq 1$. Let $\mathcal{X} = \{X : \mathcal{B} \to \mathbb{C}, \beta \to X_t(\beta)\}$ and suppose that $\int_0^1 \log^2 N_{[]}(x,\mathcal{X},\|\cdot\|)dx < \infty$ then: $\int_0^1 x^{\vartheta/2-1}\sqrt{\log N_{[]}(x,\mathcal{X},\|\cdot\|)} + \log^2 N_{[]}(x,\mathcal{X},\|\cdot\|) < \infty$ for all $\vartheta \in (0,1)$ and:*

$$\begin{aligned}&\mathbb{E}\left(sup_{\beta\in\mathcal{B}}|\sqrt{n}[\widehat{\psi}_t^S(\beta) - \mathbb{E}(\widehat{\psi}_t^S(\beta))]|^2\right)\\ &\leq K\left(\int_0^1 x^{\vartheta/2-1}\sqrt{\log N_{[]}(x,\mathcal{X},\|\cdot\|)} + \log^2 N_{[]}(x,\mathcal{X},\|\cdot\|)dx\right).\end{aligned}$$

Proof of Lemma .0.4: The method of proof is adapted from the proof of theorem 3 of Ben Hariz (2005); he only considers the stationary case, the non-stationary case is permitted here. Let $Z_n(\beta) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}X_t(\beta)$, by Lemma .0.3:

$$\|Z_n(\beta)\|_p = \mathbb{E}\left(|Z_n(\beta)|^p\right)^{1/p} \leq K\left(\sqrt{p}\frac{1}{n}\sum_{t=1}^{n}\|X_t(\beta)\|^{\vartheta/2} + p^2 n^{-1/2+1/p}\|\sup_{t>0}X_t(\beta)\|_\infty\right).$$

The term $\frac{1}{n}\sum_{t=1}^{n}\|X_t(\beta)\|^{\vartheta}$ comes from Hölder's inequality, for any $\vartheta \in (0,1)$:

$$\begin{aligned}\left|\int_0^1 \min(\alpha^{-1}(u),n)\sum_{t=1}^{n}\frac{Q_t^2(u)}{n}\right|^{1/2} &\leq \left(\int_0^1 \min(\alpha^{-1}(u),n)^{1/(1-\vartheta)}\right)^{\frac{1-\vartheta}{2}}\left(\int_0^1 |\frac{1}{n}\sum_{t=1}^{n}Q_t(u)^2|^{1/\vartheta}\right)^{\frac{\vartheta}{2}}\\ &\leq \left(\frac{1}{1-\vartheta}\sum_{j=1}^{n}(1+j)^{1/(1-\vartheta)}\alpha(j)\right)^{\frac{1-\vartheta}{2}}\frac{1}{n}\sum_{t=1}^{n}\left(\int_0^1 |Q_t(u)|^{2/\vartheta}du\right)^{\frac{\vartheta}{2}}\\ &\leq \left(\frac{1}{1-\vartheta}\sum_{j=1}^{n}(1+j)^{1/(1-\vartheta)}\alpha(j)\right)^{\frac{1-\vartheta}{2}}\frac{1}{n}\sum_{t=1}^{n}\|Q_t\|_1^{\vartheta/2}.\end{aligned}$$

The last inequality follows from assuming $|Q_t| \leq 1$. To simplify notation, use $\frac{1}{n}\sum_{t=1}^{n}\|Q_t\|_1^{\vartheta}$ rather than $\frac{1}{n}\sum_{t=1}^{n}\|Q_t\|_1^{\vartheta/2}$. Also since $\alpha(j)$ has exponential decay, $\sum_{j=1}^{\infty}(1+j)^{1/(1-\vartheta)}\alpha(j) < \infty$ so the first term is a constant which only depends on $(\alpha(j))_j$ and ϑ. To derive the inequality, construct bracketing pairs $(\beta_j^k, \Delta_j^k)_{1\leq j\leq N(k)}$ need to be constructed with $N(k) = N_{[]}(2^{-k}, \mathcal{X}, \|\cdot\|_2)$ the minimal number of brackets needed to cover $\mathcal{X}$. By definition of $N(k)$ there exists brackets $(\Delta_{t,j}^k)_{j=1,\ldots,N(k)}$ such that:

1. $\mathbb{E}\left(|\Delta_{t,j}^k|^2\right)^{1/2} \leq 2^{-k}$ for all t, j, k.
2. For all $\beta \in \mathcal{B}$ and $k \geq 1$, there exists an index j such that $|X_t(\beta) - X_t(\beta_j^k) \leq \Delta_{t,j}^k$.

Remark .0.2. *Because of the dynamics, the dependence of X_t can vary with β, which is not the case in Ben Hariz (2005) or Andrews (1993). This remark, details the construction of the brackets $(\Delta_{t,j}^k)$ in the current setting. Suppose that $\beta \to X_t(\beta)$ is L^p-smooth as in Assumption .0.4. Let $\beta_1^k, \ldots, \beta_{N(k)}^k$ be such that $\mathcal{B}_{k_n} \subseteq \cup_{j=1}^{N(k)} B_{[\delta/C]^\gamma}(\beta_j^k)$ then for $j \leq N(k)$ and some $Q \geq 2$:*

$$\left[\mathbb{E}\left(sup_{\|\beta-\beta_j^k\|_{\mathcal{B}}\leq[\delta/C]^\gamma}|X_t(\beta) - X_t(\beta_j^k)|^Q\right)\right]^{1/Q} \leq \delta.$$

Let $\Delta_{t,j}^k = sup_{\|\beta-\beta_j^k\|_{\mathcal{B}}\leq[\delta/C]^\gamma}|X_t(\beta) - X_t(\beta_j^k)|$ then $\left[\mathbb{E}\left(\Delta_{t,j}^{2k}\right)\right]^{1/2} \leq \left[\mathbb{E}\left(\Delta_{t,j}^{Qk}\right)\right]^{1/Q}$ by Hölder's inequality which is smaller than δ by construction. $\left[\mathbb{E}(|\Delta_{t,j}^k|^2)\right]^{1/2} \leq \delta = 2^{-k}$ by construction.

However, there is no guarantee that $(\Delta_{t,j}^k)_{t\geq 1}$ as constructed above is α-mixing. Another construction for the bracket which preserves the mixing property is now suggested. Let $B \subseteq \mathcal{B}$ a non-empty compact set in $\mathcal{B}$. Note that since the (β_j^k) cover $\mathcal{B}$, they also cover B. Let $\widetilde{\Delta}_{t,j}^k$ be such that $|\frac{1}{n}\sum_{t=1}^{n}\widetilde{\Delta}_{t,j}^k| = sup_{\beta\in B,\,\|\beta-\beta_j^k\|\leq[\delta/C]^\gamma}|\frac{1}{n}\sum_{t=1}^{n}X_t(\beta) - X_t(\beta_j^k)|$. Because B is compact, the supremum is attained at some $\widetilde{\beta}_j^k \in B$. For all $t = 1, \ldots, n$, take $\widetilde{\Delta}_{t,j}^k = X_t(\widetilde{\beta}_j^k) - X_t(\beta_j^k)$. For each (j,k) the sequence $(\widetilde{\Delta}_{t,j}^k)_{t\geq 0}$ is α-mixing by construction. Furthermore, by construction:

$|\widetilde{\Delta}^k_{t,j}| \leq |\Delta^k_{t,j}|$ *and thus* $\left[\mathbb{E}(|\widetilde{\Delta}^k_{t,j}|^Q)\right]^{1/Q} \leq 2^{-k}$. *These brackets, built in B rather than* $\mathcal{B}$, *preserve the mixing properties. The rest of the proof applied to B implies:*

$$\begin{aligned}
&\mathbb{E}\left(sup_{\beta\in B}|\sqrt{n}[\widehat{\psi}^S_t(\beta) - \mathbb{E}(\widehat{\psi}^S_t(\beta))]|^2\right) \\
&\leq K\left(\int_0^1 x^{\vartheta/2-1}\sqrt{\log N_{[]}(x^{1/\gamma}, B, \|\cdot\|)} + \log^2 N_{[]}(x^{1/\gamma}, B, \|\cdot\|)dx\right) \\
&\leq K\left(\int_0^1 x^{\vartheta/2-1}\sqrt{\log N_{[]}(x^{1/\gamma}, \mathcal{B}, \|\cdot\|)} + \log^2 N_{[]}(x^{1/\gamma}, \mathcal{B}, \|\cdot\|)dx\right).
\end{aligned}$$

For an increasing sequence of compact sets $B_k \subseteq B_{k+1} \subseteq \mathcal{B}$ *dense in* $\mathcal{B}$, *there is an increasing and bounded sequence:*

$$\begin{aligned}
&\mathbb{E}\left(sup_{\beta\in B_k}|\sqrt{n}[\widehat{\psi}^S_t(\beta) - \mathbb{E}(\widehat{\psi}^S_t(\beta))]|^2\right) \\
&\leq \mathbb{E}\left(sup_{\beta\in B_{k+1}}|\sqrt{n}[\widehat{\psi}^S_t(\beta) - \mathbb{E}(\widehat{\psi}^S_t(\beta))]|^2\right) \\
&\leq K\left(\int_0^1 x^{\vartheta/2-1}\sqrt{\log N_{[]}(x^{1/\gamma}, \mathcal{B}, \|\cdot\|)} + \log^2 N_{[]}(x^{1/\gamma}, \mathcal{B}, \|\cdot\|)dx\right).
\end{aligned}$$

This sequence is thus convergent with limit less or equal than the upper-bound. Hence, it must be that the supremum over $\mathcal{B}$ *is also bounded. It can thus be assumed that* $(\Delta^k_{t,j})_{t\geq 1}$ *are* α*-mixing.*

Assume that, without loss of generality, $|\Delta^k_j| \leq 1$ for all j, k. Let $(\pi_k(\beta), \Delta_k(\beta))$ be a bracketing pair for $\beta \in \mathcal{B}$. Let q_0, k, q be positive integers such that $q_0 \leq k \leq q$ and let $T_k(\beta) = \pi_k \circ \pi_{k+1} \circ \cdots \circ \pi_q(\beta)$. Using the following identity:

$$\begin{aligned}
&\left[\mathbb{E}\left(\sup_{\beta\in\mathcal{B}}|Z_n(\beta)|^2\right)\right]^{1/2} \\
&= \left[\mathbb{E}\left(\sup_{\beta\in\mathcal{B}}|Z_n(\beta) - Z_n(T_q(\beta)) + \sum_{k=q_0+1}^{q}[Z_n(T_k(\beta)) - Z_n(T_{k-1}(\beta))] + Z_n(T_{q_0}(\beta))|^2\right)\right]^{1/2}
\end{aligned}$$

and the triangular inequality, decompose the identity into three groups:

$$\begin{aligned}
\left[\mathbb{E}\left(\sup_{\beta\in\mathcal{B}}|Z_n(\beta)|^2\right)\right]^{1/2} &\leq \left[\mathbb{E}\left(\sup_{\beta\in\mathcal{B}}|Z_n(\beta) - Z_n(T_q(\beta))|^2\right)\right]^{1/2} \\
&\quad + \sum_{k=q_0+1}^{q}\left[\mathbb{E}\left(\sup_{h\in\mathcal{B}}|Z_n(T_k(\beta)) - Z_n(T_{k-1}(\beta))|^2\right)\right]^{1/2} \\
&\quad + \left[\mathbb{E}\left(\sup_{\beta\in\mathcal{B}}|Z_n(T_{q_0}(\beta))|^2\right)\right]^{1/2} \\
&\leq E_{q+1} + \sum_{k=q_0+1}^{q} E_k + E_{q_0}.
\end{aligned}$$

The following inequality is due to Pisier (1983), for any $X_1, \ldots, X_N$ random variable:

$$\left[\mathbb{E}\left(\max_{1\leq t\leq N}|X_t|^p\right)\right]^{1/p} \leq N^{1/p}\max_{1\leq t\leq N}\left[\mathbb{E}\left(|X_t|^p\right)\right]^{1/p}.$$

Now that $\{T_k(\beta), \beta \in \mathcal{B}\}$ has at most $N(k)$ elements by construction. Some terms can be simplified $E_k = \mathbb{E}\left(\max_{g\in T_k(\mathcal{B})}|Z_n(g) - Z_n(T_{k-1}(g))|^2\right)^{1/2}$ for $q_0 + 1 \leq k \leq q$. For $p \geq 2$ using both Hölder's and Pisier's inequalities:

$$\begin{aligned}
E_k &\leq \left[\mathbb{E}\left(\sup_{\beta\in T_k(\mathcal{B})}|Z_n(\beta) - Z_n(T_{k-1}(\beta))|^p\right)\right]^{1/p} \\
&\leq N(k)^{1/p}\max_{g\in T_k(\mathcal{B})}\left[\mathbb{E}\left(|Z_n(g) - Z_n(T_{k-1}(g))|^p\right)\right]^{1/p}.
\end{aligned}$$

By the definition of Δ_j^k:

$$E_k \leq N(k)^{1/p}\max_{1\leq j\leq N(k)}\left[\mathbb{E}\left(|\Delta_j^k(g)|^p\right)\right]^{1/p}.$$

This is also valid for E_{q+1}. Using Rio's inequality for α-mixing dependent processes:

$$\begin{aligned}
E_k &\leq KN(k)^{1/p}\left(\sqrt{p}\max_{g\in T_k(\mathcal{B})}\|\Delta^k(g)\|_1^{\vartheta/2} + p^2 n^{-1/2+1/p}\max_{g\in T_k(\mathcal{B})}\|\Delta^k(g)\|_\infty\right) \\
&\leq KN(k)^{1/p}\left(\sqrt{p}2^{-\vartheta/2k} + p^2 n^{-1/2+1/p}\right) \\
&\leq KN(k)^{1/p}2^{-k}\left(\sqrt{p}2^{k-\vartheta/2k} + p^2[n^{-1/2}2^k]^{1-2/p}2^{2k/p}\right).
\end{aligned}$$

For $p > 2$ and $2^q/\sqrt{n} \geq 1$, the inequality becomes:

$$E_k \leq KN(k)^{1/p}2^{-k}\left(\sqrt{p}2^{k-\vartheta/2k} + p^2[n^{-1/2}2^q]2^{2k/p}\right).$$

Choosing $p = k + \log N(k)$ implies:

$$\begin{aligned}
&N(k)^{1/p} \leq \exp(1) \\
&\sqrt{p} \leq \sqrt{k} + \sqrt{\log N(k)} \\
&p^2 \leq 4[k^2 + \log^2 N(k)] \\
&2^{2k/p} \leq 4.
\end{aligned}$$

Applying these bounds to the previous inequality:

$$\begin{aligned}
E_k &\leq 16K\exp(1)2^{-k}\left([\sqrt{k} + \sqrt{\log N(k)}]2^{k-\vartheta/2k} + [k^2 + \log(N(k))^2]\frac{2^q}{\sqrt{n}}\right) \\
&\leq \frac{2^q}{\sqrt{n}}16K\exp(1)2^{-k}\left([\sqrt{k} + \sqrt{\log N(k)}]2^{k-\vartheta/2k} + k^2 + \log(N(k))^2\right).
\end{aligned}$$

Note that $\sum_{k\geq 1}(\sqrt{k}+k^2)2^{-k} \leq 2\sum_{k\geq 1} k^2 2^{-k} = 12$. Hence:

$$\sum_{k=q_0+1}^{q+1} E_k \leq \frac{2^{q+1}}{\sqrt{n}} 16K\exp(1)\left(12 + \int_0^1 [x^{\vartheta/2-1}\sqrt{\log N_{[]}(x,\mathcal{X},\|\cdot\|)} + \log^2 N_{[]}(x,\mathcal{X},\|\cdot\|)]dx\right).$$

Pick q to be the small integer such that $q \geq \log(n)/(2\log 2) - 1$ so that $4\sqrt{n} \geq 2^q \geq \sqrt{n}/2$ and $2^q/\sqrt{n} \in [1/2, 4]$. Only E_{q_0} remains to be bounded, using Rio's inequality again:

$$\left[\mathbb{E}\left(\sup_{\beta\in\mathcal{B}}|Z_n(T_{q_0}(\beta))|^2\right)\right]^{1/2} \leq KN(q_0)^{1/p}\left(\sqrt{p}\max_{h\in T_{q_0}(\mathcal{B})}\|X_1(\beta)\|^{\vartheta+p^2n^{-1/2+1/p}\|X_1(\beta)\|_\infty}\right).$$

For any $\varepsilon > 0$ pick $p = \max(2+\varepsilon, q_0 + \log N(q_0))$ then:

$$\begin{aligned} N(q_0)^{1/p} &\leq \exp(1) \\ n^{-1/2+1/p} &\leq n^{-1/2+1/(2+\varepsilon)} \leq 1. \end{aligned}$$

Then conclude that:

$$\begin{aligned} \left[\mathbb{E}\left(\sup_{\beta\in\mathcal{B}}|Z_n(T_{q_0}(\beta))|^2\right)\right]^{1/2} &\leq 4\exp(1)K\left(\sqrt{q_0} + \sqrt{\log N(q_0)} + q_0^2 + \log N(q_0)^2\right) \\ &\leq K'\log N(q_0)^2 \\ &\leq K'\int_0^1 \log^2 N_{[]}(x,\mathcal{X},\|\cdot\|)dx \end{aligned}$$

Hence, there exists a constant $K > 0$ which only depends on $(\alpha(m))_{m>0}$ such that:

$$\left[\mathbb{E}\left(\sup_{\beta\in\mathcal{B}}|Z_n(\beta)|^2\right)\right]^{1/2} \leq K\int_0^1 [x^{\vartheta/2-1}\sqrt{\log N_{[]}(x,\mathcal{X},\|\cdot\|)} + \log^2 N_{[]}(x,\mathcal{X},\|\cdot\|)]dx.$$

Let $\sqrt{C_n} = K\int_0^1 [x^{\vartheta/2-1}\sqrt{\log N_{[]}(x,\mathcal{X},\|\cdot\|)} + \log^2 N_{[]}(x,\mathcal{X},\|\cdot\|)]dx$, then $\mathbb{E}\left(\sup_{\beta\in\mathcal{B}}|Z_n(\beta)|^2\right) \leq C_n$ for all $n \geq 1$. □

Rate of Convergence

Proof of Proposition .0.3. : By Hölder's inequality and the L^p-smoothness assumption:

$$\left|\mathbb{E}\left(\widehat{\psi}_n^s(\tau,\Pi_{k(n)}\beta_0) - \widehat{\psi}_n^s(\tau,\beta_0)\right)\right|^2 \pi(\tau)^{1/(1+\eta/2)} \leq C^2\|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{2\gamma}.$$

Using the fact that $|a+b|^2 \leq 3[|a|^2 + |b|^2]$:

$$\begin{aligned} Q_n(\Pi_{k(n)}\beta_0) &\leq 3\left[Q_n(\beta_0) + \int |B\mathbb{E}\left(\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0) - \widehat{\psi}_n^S(\tau,\beta_0)\right)|^2\pi(\tau)d\tau\right] \\ &\leq 3\left[Q_n(\beta_0) + M_B^2\int |\mathbb{E}\left(\widehat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0) - \widehat{\psi}_n^S(\tau,\beta_0)\right)|^2\pi(\tau)d\tau\right] \\ &\leq 3\left[Q_n(\beta_0) + \left(C^2M_B^2\int \pi^{1-\frac{2}{2+\eta}}(\tau)d\tau\right)\|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{2\gamma}\right]. \end{aligned}$$

The last inequality comes from taking integrals on both sides of the first inequality. The integral on the right-hand side is finite by assumption. To conclude the proof, take $K = 3[1 + C^2 M_B^2 \int \pi^{1-\frac{2}{2+\eta}}(\tau) d\tau]$. □

Proof of Theorem .0.2: Let $\varepsilon > 0$ and $r_n = \max(\delta_n, \sqrt{\eta_n}, \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{\gamma}, \sqrt{Q_n(\beta_0)})$. To prove the result, it will be shown that there exists some $M > 0$ and $N > 0$ such that for all $n \geq N$:

$$\mathbb{P}\left(\|\widehat{\beta}_n - \beta_0\|_{weak} \geq M r_n\right) < \varepsilon. \tag{.0.1}$$

The approach to prove existence is similar to the proof of lemma B.1 in Chen & Pouzo (2012). First, under the stated assumptions, the following inequalities hold:

1. $\widehat{Q}_n^S(\beta) \leq 2Q_n(\beta) + O_p(\delta_n^2)$
2. $Q_n(\beta) \leq K\left(\|\beta - \beta_0\|^{2\gamma} + Q_n(\beta_0)\right)$
3. $\|\beta - \beta_0\|_{weak}^2 \leq C Q_n(\beta)$

Applying them in the same order, equation (.0.1) can be bounded above:

$$\begin{aligned}
\mathbb{P}\left(\|\widehat{\beta}_n - \beta_0\|_{weak} \geq M r_n\right) &\leq \mathbb{P}\left(\text{diag}_{\beta \in \mathcal{B}_{osn},\, \|\beta-\beta_0\|_{weak} \geq M r_n} \widehat{Q}_n^s(\beta) \leq \text{diag}_{\beta \in \mathcal{B}_{osn}} \widehat{Q}_n^S(\beta) + O_p(\eta_n)\right) \\
&\leq \mathbb{P}\left(\text{diag}_{\beta \in \mathcal{B}_{osn},\, \|\beta-\beta_0\|_{weak} \geq M r_n} Q_n(\beta) \leq \text{diag}_{\beta \in \mathcal{B}_{osn}} Q_n(\beta) + O_p(\max(\delta_n^s, \eta_n))\right) \\
&\leq \mathbb{P}\left(\text{diag}_{\beta \in \mathcal{B}_{osn},\, \|\beta-\beta_0\|_{weak} \geq M r_n} Q_n(\beta) \leq Q_n(\Pi_{k(n)}\beta_0) + O_p(\max(\delta_n^s, \eta_n))\right) \\
&\leq \mathbb{P}\left(\text{diag}_{\beta \in \mathcal{B}_{osn},\, \|\beta-\beta_0\|_{weak} \geq M r_n} Q_n(\beta) \leq O_p(\max(\|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{2\gamma}, Q_n(\beta_0), \delta_n^s, \eta_n))\right) \\
&\leq \mathbb{P}\left(M^2 r_n^2 \leq O_p(\max(\|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{2\gamma}, Q_n(\beta_0), \delta_n^s, \eta_n))\right)
\end{aligned}$$

For r_n defined above, this probability becomes:

$$\mathbb{P}\left(M^2 \leq O_p(1)\right) \to 0 \text{ as } M \to \infty.$$

This concludes the first part of the proof. Finally:

$$\begin{aligned}
&\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} \\
&\leq \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}} + \|\widehat{\beta}_n - \Pi_{k(n)}\beta_0\|_{\mathcal{B}} \frac{\|\widehat{\beta}_n - \Pi_{k(n)}\beta_0\|_{weak}}{\|\widehat{\beta}_n - \Pi_{k(n)}\beta_0\|_{weak}} \\
&\leq \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}} + \tau_n \|\widehat{\beta}_n - \Pi_{k(n)}\beta_0\|_{weak} \\
&\leq \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}} + \tau_n \left(\|\widehat{\beta}_n - \beta_0\|_{weak} + \|\beta_0 - \Pi_{k(n)}\beta_0\|_{weak}\right) \\
&\leq \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}} + \tau_n \left(\|\widehat{\beta}_n - \beta_0\|_{weak} + C Q_n(\Pi_{k(n)}\beta_0)\right) \\
&\leq \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}} + \tau_n \left(O_p\left(\max(\delta_n, \sqrt{\eta_n}, \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{\gamma}, \sqrt{Q_n(\beta_0)}, \|\Pi_{k(n)}\beta_0 - \beta_0\|^{2\gamma}, Q_n(\beta_0)\right)\right) \\
&= \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}} + \tau_n \left(O_p\left(\max(\delta_n, \sqrt{\eta_n}, \|\Pi_{k(n)}\beta_0 - \beta_0\|_{\mathcal{B}}^{\gamma}, \sqrt{Q_n(\beta_0)})\right)\right).
\end{aligned}$$

This concludes the proof. □

Proof of Proposition .0.4: Since $(\mathbf{y}_t^s, \mathbf{x}_t)$ is geometrically ergodic, the joint density converges to the stationary distribution at a geometric rate: $\|f_t(y,x) - f_t^*(y,x)\|_{TV} \leq C\rho^t$, $\rho < 1$. Because B is bounded linear and the moments $\widehat{\psi}_n, \widehat{\psi}_n^s$ are bounded above by M, uniformly in τ:

$$\begin{aligned}
Q_n(\beta_0) &\leq M_B^2 \int \left| \mathbb{E}\left(\widehat{\psi}_n^S(\tau,\beta_0)\right) - \lim_{n\to\infty} \mathbb{E}\left(\widehat{\psi}_n(\tau)\right) \right|^2 \pi(\tau) d\tau \\
&\leq M^2 M_B^2 \int \left| \frac{1}{n}\sum_{t=1}^{n} \int [f_t(y,x) - f_t^*(y,x)] dy dx \right|^2 \pi(\tau) d\tau \\
&\leq M^2 M_B^2 \left(\frac{1}{n}\sum_{t=1}^{n} \int |f_t(y,x) - f_t^*(y,x)| \, dy dx \right)^2 \\
&\leq CM^2 M_B^2 \left(\frac{1}{n}\sum_{t=1}^{n} \rho^t \right)^2 \\
&\leq \frac{CM^2 M_B^2}{(1-\rho)^2} \times \frac{1}{n^2} = O(1/n^2).
\end{aligned}$$

□

Asymptotic Normality

Lemma .0.5 (Stochastic Equicontinuity)**.** *Let $M_n = \log\log(n+1)$ as defined in Assumption .0.7. Also, $\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} = O_p(\delta_{sn})$. Suppose Assumption .0.4 holds then for any $\vartheta \in (0,1)$, there exists a $C > 0$ such that:*

$$\begin{aligned}
&\left[\mathbb{E}\left(sup_{\|\beta-\beta_0\|_{\mathcal{B}} \leq M_n\delta_{sn}} \left| [\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\beta_0)] - \mathbb{E}[\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\beta_0)] \right|^2 \pi(\tau)^{\frac{2}{2+\eta}} \right) \right]^{1/2} \\
&\leq C \frac{(M_n\delta_{sn})^\gamma}{\sqrt{n}} \int_0^1 \left(x^{-\vartheta/2} \sqrt{\log N([xM_n\delta_{sn}]^{1/\gamma}, \mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}})} + \log^2 N([xM_n\delta_{sn}]^{1/\gamma}, \mathcal{B}_{k(n)}, \|\cdot\|_{\mathcal{B}}) \right) dx \\
&:= \frac{(M_n\delta_{sn})^\gamma}{\sqrt{n}} \sqrt{C_{sn}}.
\end{aligned}$$

Now suppose that $\sqrt{C_{sn}}(M_n\delta_{sn})^\gamma = o(1)$ as in Assumption .0.7. For linear sieves, $\sqrt{C_{sn}}$ is proportional to:

$$\left(\log[M_n\delta_{sn}]k(n)\right)^2.$$

Hence, for linear sieves $\sqrt{C_{sn}}(M_n\delta_{sn})^\gamma = o(1)$ is implied by $(M_n\delta_{sn})^\gamma \log(M_n\delta_{sn})^2 = o(1/k(n)^2)$. Together with the previous inequality, this assumption implies a stochastic equicontinuity result:

$$\left(\int \left| [\widehat{\psi}_n^S(\tau,\widehat{\beta}_n) - \widehat{\psi}_n^S(\tau,\beta_0)] - \mathbb{E}[\widehat{\psi}_n^S(\tau,\widehat{\beta}_n) - \widehat{\psi}_n^S(\tau,\beta_0)] \right|^2 \pi(\tau) d\tau \right)^{1/2} = o_p(1/\sqrt{n}).$$

Proof of Lemma .0.5: Let $\Delta\widehat{\psi}_t^s(\tau,\beta) = \widehat{\psi}_t^s(\tau,\beta) - \widehat{\psi}_t^s(\tau,\beta_0)$. Under Assumption .0.4:

$$\left[\mathbb{E}\left(\sup_{\|\beta-\beta_0\|_{\mathcal{B}}\leq M_n\delta_{sn}}\left|\Delta\widehat{\psi}_t^s(\tau,\beta)\right|^2\pi(\tau)^{\frac{2}{2+\eta}}\right)\right]^{1/2} \leq C(M_n\delta_{sn})^{\gamma}$$

and

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta,\beta_1,\beta_2\in B_{M_n\delta_{sn}}(\beta_0)}\left|\Delta\widehat{\psi}_t^s(\tau,\beta_1)-\Delta\widehat{\psi}_t^s(\tau,\beta_2)\right|^2\frac{\pi(\tau)^{\frac{2}{2+\eta}}}{(M_n\delta_{sn})^{2\gamma}}\right)\right]^{1/2} \leq C\left(\frac{\delta}{M_n\delta_{sn}}\right)^{\gamma}.$$

Applying Lemma .0.4 to the empirical process $\Delta\widehat{\psi}_t^s(\tau,\beta)\frac{\pi(\tau)^{\frac{1}{2+\eta}}}{(M_n\delta_{sn})^{\gamma}}$ yields:

$$\begin{aligned}
&\left[\mathbb{E}\left(\sup_{\|\beta-\beta_0\|_{\mathcal{B}}\leq M_n\delta_{sn}}\left|\Delta\widehat{\psi}_n^S(\tau,\beta)-\mathbb{E}\left(\Delta\widehat{\psi}_n^S(\tau,\beta)\right)\right|^2\frac{\pi(\tau)^{\frac{2}{2+\eta}}}{(M_n\delta_{sn})^{2\gamma}}\right)\right]^{1/2}\\
&\leq \frac{C}{\sqrt{n}}\int_0^1\left(x^{-\vartheta/2}\sqrt{\log N([xM_n\delta_{sn}]^{1/\gamma},\mathcal{B}_{k(n)},\|\cdot\|_{\mathcal{B}})}+\log^2 N([xM_n\delta_{sn}]^{1/\gamma},\mathcal{B}_{k(n)},\|\cdot\|_{\mathcal{B}})\right)dx
\end{aligned}$$

for some constant $C > 0$ and any $\vartheta \in (0,1)$ such that the integral is finite. For finite dimensional linear sieves the integral is proportional to $k(n)^2\log(M_n\delta_{sn})^2$ and the bound becomes, after multiplying by $(M_n\delta_{sn})^{\gamma}$ on both sides:

$$\begin{aligned}
&\left[\mathbb{E}\left(\sup_{\|\beta-\beta_0\|_{\mathcal{B}}\leq M_n\delta_{sn}}\left|\Delta\widehat{\psi}_n^S(\tau,\beta)-\mathbb{E}\left(\Delta\widehat{\psi}_n^S(\tau,\beta)\right)\right|^2\pi(\tau)^{\frac{2}{2+\eta}}\right)\right]^{1/2}\\
&\leq \frac{C}{\sqrt{n}}(M_n\delta_{sn})^{\gamma}[\log(M_n\delta_{sn})k(n)]^2.
\end{aligned}$$

Note that $\mathbb{P}\left(\|\widehat{\beta}_n-\beta_0\|_{\mathcal{B}}\leq M_n\delta_{sn}\right)\to 1$ by construction of M_n and definition of δ_{sn}. The following inequalities can be used:

$$\begin{aligned}
&\mathbb{P}\left(\int\left|[\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)-\widehat{\psi}_n^S(\tau,\beta_0)]-\mathbb{E}[\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)-\widehat{\psi}_n^S(\tau,\beta_0)]\right|^2\pi(\tau)^{\frac{2}{2+\eta}}\pi(\tau)^{1-\frac{2}{2+\eta}}d\tau>\frac{\varepsilon}{n}\right)\\
&\leq \mathbb{P}\left(\sup_{\|\beta-\beta_0\|_{\mathcal{B}}\leq M_n\delta_{sn}}\int\left|[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\beta_0)]-\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\beta_0)]\right|^2\pi(\tau)^{\frac{2}{2+\eta}}\pi(\tau)^{1-\frac{2}{2+\eta}}d\tau>\frac{\varepsilon}{n}\right)\\
&+\mathbb{P}\left(\|\beta-\beta_0\|_{\mathcal{B}}>M_n\delta_{sn}\right)\\
&\leq \frac{n}{\varepsilon}\mathbb{E}\left(\int\left|[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\beta_0)]-\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\beta_0)]\right|^2\pi(\tau)^{\frac{2}{2+\eta}}\pi(\tau)^{1-\frac{2}{2+\eta}}d\tau\right)\\
&+\mathbb{P}\left(\|\beta-\beta_0\|_{\mathcal{B}}>M_n\delta_{sn}\right)\\
&=\int\frac{n}{\varepsilon}\mathbb{E}\left(\left|\Delta\widehat{\psi}_n^S(\tau,\beta)-\mathbb{E}[\Delta\widehat{\psi}_n^S(\tau,\beta)]\right|^2\pi(\tau)^{\frac{2}{2+\eta}}\right)\pi(\tau)^{1-\frac{2}{2+\eta}}d\tau+\mathbb{P}\left(\|\beta-\beta_0\|_{\mathcal{B}}>M_n\delta_{sn}\right)\\
&\leq C_{sn}(M_n\delta_{sn})^{2\gamma}\int\pi(\tau)^{1-\frac{2}{2+\eta}}d\tau+\mathbb{P}\left(\|\beta-\beta_0\|_{\mathcal{B}}>M_n\delta_{sn}\right)=o(1).
\end{aligned}$$

These inequalitites hold regardless of $\varepsilon > 0$ given the assumptions above and the definition of $M_n\delta_{sn}$. To conclude, the stochastic equicontinuity result holds:

$$\left(\int \left|[\widehat{\psi}_n^S(\tau,\widehat{\beta}_n) - \widehat{\psi}_n^S(\tau,\beta_0)] - \mathbb{E}[\widehat{\psi}_n^S(\tau,\widehat{\beta}_n) - \widehat{\psi}_n^S(\tau,\beta_0)]\right|^2 \pi(\tau)^{\frac{2}{2+\eta}}\pi(\tau)^{1-\frac{2}{2+\eta}}d\tau\right)^{1/2}$$
$$= o_p(1/\sqrt{n}).$$

□

Lemma .0.6. *Suppose that $\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p(\delta_n)$. Under Assumptions .0.4, .0.6, .0.7 and .0.9:*

a)

$$\int \psi_\beta(\tau,u_n^*)\left(\overline{B\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}_n) - \widehat{\psi}_n^S(\tau,\beta_0)) - B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\widehat{\beta}_n - \beta_0]}\right)\pi(\tau)d\tau = o(1/\sqrt{n}).$$

b)

$$\int \psi_\beta(\tau,u_n^*)\left(\overline{B\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}_n) - \widehat{\psi}_n^S(\tau,\beta_0)) - B[\widehat{\psi}_n^S(\tau,\widehat{\beta}_n) - \widehat{\psi}_n^S(\tau,\beta_0)]}\right)\pi(\tau)d\tau = o(1/\sqrt{n}).$$

c)

$$\int \left[\psi_\beta(\tau,u_n^*)\left(\overline{B[\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau,\widehat{\beta}_n)]}\right) + \overline{\psi_\beta(\tau,u_n^*)}\left(B[\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau,\widehat{\beta}_n)]\right)\right]\pi(\tau)d\tau = o(1/\sqrt{n}).$$

Proof of Lemma .0.6:

a) Since B bounded linear, the Cauchy-Schwarz inequality implies:

$$\left|\int \psi_\beta(\tau,u_n^*)\left(\overline{B\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}_n) - \widehat{\psi}_n^S(\tau,\beta_0)) - B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\widehat{\beta}_n - \beta_0]}\right)\pi(\tau)d\tau\right|$$
$$\leq M_B\left(\int |\psi_\beta(\tau,u_n^*)|^2\pi(\tau)d\tau\right)^{1/2}\left(\int\left|\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}_n) - \widehat{\psi}_n^S(\tau,\beta_0)) - \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\widehat{\beta}_n - \beta_0]\right|^2\pi(\tau)d\tau\right)^{1/2}$$

By definition of M_n and the inequality above:

$$\mathbb{P}\left(\left|\int \psi_\beta(\tau,u_n^*)\left(\overline{B\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}_n) - \widehat{\psi}_n^S(\tau,\beta_0)) - B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\widehat{\beta}_n - \beta_0]}\right)\pi(\tau)d\tau\right| > \frac{\varepsilon}{\sqrt{n}}\right)$$
$$\leq \mathbb{P}\Bigg[M_B\left(\int |\psi_\beta(\tau,u_n^*)|^2\pi(\tau)d\tau\right)^{1/2}$$
$$\times \sup_{\|\beta-\beta_0\|_{weak}\leq M_n\delta_n}\left(\int\left|\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta) - \widehat{\psi}_n^S(\tau,\beta_0)) - \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\beta - \beta_0]\right|^2\pi(\tau)d\tau\right)^{1/2} > \frac{\varepsilon}{\sqrt{n}}\Bigg]$$
$$+ \mathbb{P}\left(\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} > M_n\delta_n\right)$$

The term $\mathbb{P}\left(\|\widehat{\beta}_n - \beta_0\|_{\mathcal{B}} > M_n\delta_n\right) \to 0$ regardless of ε. Furthermore, Assumption .0.9 *i.* implies that

$$\sup_{\|\beta-\beta_0\|_{weak}\leq M_n\delta_n}\left(\int\left|\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta)-\widehat{\psi}_n^S(\tau,\beta_0))-\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\beta-\beta_0]\right|^2\pi(\tau)d\tau\right)^{1/2}$$
$$=O\left((M_n\delta_n)^2\right).$$

Furthermore Assumption .0.7 *iii.*, condition (.0.3) implies that $(M_n\delta_n)^{1+\gamma} = o(\frac{1}{\sqrt{nC_{sn}}})$. Since $\gamma \in (0,1]$ it implies $(M_n\delta_n)^2 = o(1/\sqrt{n})$ and thus:

$$\mathbb{P}\left(\left|\int\psi_\beta(\tau,u_n^*)\left(\overline{B\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)-\widehat{\psi}_n^S(\tau,\beta_0))-B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\widehat{\beta}_n-\beta_0]}\right)\pi(\tau)d\tau\right|>\frac{\varepsilon}{\sqrt{n}}\right)$$
$$=o(1)$$

regardless of $\varepsilon > 0$. Finally:

$$\int\psi_\beta(\tau,u_n^*)\left(\overline{B\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)-\widehat{\psi}_n^S(\tau,\beta_0))-B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\widehat{\beta}_n-\beta_0]}\right)\pi(\tau)d\tau$$
$$=o_p(1/\sqrt{n}).$$

b) By the stochastic equicontinuity result of Lemma .0.5 and the Cauchy-Schwarz inequality:

$$\left|\int\psi_\beta(\tau,u_n^*)\left(\overline{B\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)-\widehat{\psi}_n^S(\tau,\beta_0))-B[\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)-\widehat{\psi}_n^S(\tau,\beta_0)]}\right)\pi(\tau)d\tau\right|$$
$$\leq\left(\int|\psi_\beta(\tau,u_n^*)|^2\pi(\tau)d\tau\right)^{1/2}\left(\int\left|B\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)-\widehat{\psi}_n^S(\tau,\beta_0))-B[\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)-\widehat{\psi}_n^S(\tau,\beta_0)]\right|^2\pi(\tau)d\tau\right)^{1/2}$$
$$\leq M_B\left(\int|\psi_\beta(\tau,u_n^*)|^2\pi(\tau)d\tau\right)^{1/2}\left(\int\left|\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)-\widehat{\psi}_n^S(\tau,\beta_0))-[\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)-\widehat{\psi}_n^S(\tau,\beta_0)]\right|^2\pi(\tau)d\tau\right)^{1/2}$$
$$\leq M_B\left(\int|\psi_\beta(\tau,u_n^*)|^2\pi(\tau)d\tau\right)^{1/2}\left(\int\pi(\tau)^{1-\frac{2}{2+\eta}}d\tau\right)^{1/2}o_p(1/\sqrt{n})$$
$$=o_p(1/\sqrt{n}).$$

c) Let $\varepsilon_n = \pm\frac{1}{\sqrt{n}M_n} = o(\frac{1}{\sqrt{n}})$. For $h \in (0,1)$ define $\widehat{\beta}(h) = \widehat{\beta}_n + h\varepsilon_n u_n^*$. Since $\widehat{\beta}_n = \widehat{\beta}(0)$. Recall that $\widehat{\beta}_n$ is the approximate minimizer of $\widehat{Q}_n^s$ so that:

$$0 \leq \widehat{Q}_n^S(\widehat{\beta}_n) \leq \text{diag}_{\beta\in\mathcal{B}_{k(n)}}\widehat{Q}_n^S(\beta) + O_p(\eta_n).$$

Hence the following holds:

$$0 \leq \frac{1}{2}\left(\widehat{Q}_n^S(\widehat{\beta}(1)) - \widehat{Q}_n^S(\widehat{\beta}(0))\right) + O_p(\eta_n) \tag{.0.2}$$

$$= \frac{1}{2}\Big[\int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(0))\right)\overline{B\left(\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))\right)}\pi(\tau)d\tau \tag{.0.3}$$

$$+ \int \overline{B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(0))\right)}B\left(\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))\right)\pi(\tau)d\tau \tag{.0.4}$$

$$+ \int \left|B\left(\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))\right)\right|^2\pi(\tau)d\tau\Big] + O_p(\eta_n). \tag{.0.5}$$

To prove Lemma .0.6 c), (.0.3)-(.0.4) are expanded and shown to be $o_p(1/\sqrt{n})$ and (.0.5) is bounded, shown to be negligible under the assumptions.
The first step deals with (.0.5):

$$\left(\int \left|B\left(\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))\right)\right|^2\pi(\tau)d\tau\right)^{1/2}$$

$$\leq M_B\left(\int \left|\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))\right|^2\pi(\tau)d\tau\right)^{1/2}$$

$$\leq \left(\int \left|[\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))] - \mathbb{E}[\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))]\right|^2\pi(\tau)d\tau\right)^{1/2}$$

$$+ \left(\int \left|\mathbb{E}[\widehat{\psi}_n^S(\tau,\widehat{\beta}(t)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))]\right|^2\pi(\tau)d\tau\right)^{1/2}$$

By the triangular inequality and the stochastic equicontinuity results from Lemma .0.5:

$$\left(\int \left|[\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))] - \mathbb{E}[\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))]\right|^2\pi(\tau)d\tau\right)^{1/2}$$

$$= O_p\left(\frac{\sqrt{C_{sn}}(M_n\delta_{sn})^\gamma}{\sqrt{n}}\right).$$

Also, note that $\widehat{\beta}(1) = \widehat{\beta}(0) + \varepsilon_n u_n^*$, so that the Mean Value Theorem applies to last term:

$$\left(\int \left|\mathbb{E}[\widehat{\psi}_n^S(\tau,\widehat{\beta}(t)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))]\right|^2\pi(\tau)d\tau\right)^{1/2} = \left(\int \left|\frac{d\mathbb{E}[\widehat{\psi}_n^S(\tau,\widehat{\beta}(\widetilde{h}))}{d\beta}[\varepsilon_n u_n^*]\right|^2\pi(\tau)d\tau\right)^{1/2}$$

for some intermediate value $\widetilde{h} \in (0,1)$. Also, by assumption:

$$\left(\int \left|\frac{d\mathbb{E}[\widehat{\psi}_n^S(\tau,\widehat{\beta}(\widetilde{t}))}{d\beta}[u_n^*]\right|^2\pi(\tau)d\tau\right)^{1/2} = O_p(1).$$

Together these two elements imply:

$$\left(\int \left| \mathbb{E}[\widehat{\psi}_n^S(\tau, \widehat{\beta}(t)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1))] \right|^2 \pi(\tau) d\tau \right)^{1/2} = O(\varepsilon_n).$$

This yields the bound for (.0.5):

$$\int \left| B\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1)) \right) \right|^2 \pi(\tau) d\tau \leq O_p(\varepsilon_n^2) + O_p(\frac{(M_n \delta_{sn})^{2\gamma} C_{sn}}{n}).$$

The remaining terms, (.0.3)-(.0.4), are conjugates of each other. A bound for (.0.3) is also valid for (.0.4). Expanding (.0.3) yields:

$$\int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) \right) \overline{B\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1)) \right)} \pi(\tau) d\tau \tag{.0.3}$$

$$= \int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) \right) \left[\overline{B\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1)) \right) - B\mathbb{E}\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1)) \right)} \right] \pi(\tau) d\tau \tag{.0.6}$$

$$+ \int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) \right) \overline{B\mathbb{E}\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1)) \right)} \pi(\tau) d\tau. \tag{.0.7}$$

Applying the Cauchy-Schwarz inequality to (.0.6) implies:

$$\left| \int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) \right) \left[\overline{B\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1)) \right) - B\mathbb{E}\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1)) \right)} \right] \pi(\tau) d\tau \right| \tag{.0.6}$$

$$\leq M_B \left(\int \left| B\widehat{\psi}_n(\tau) - B\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) \right|^2 \pi(\tau) d\tau \right)^{1/2} \tag{.0.8}$$

$$\times \left(\int \left| \left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1)) \right) - \mathbb{E}\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1)) \right) \right|^2 \pi(\tau) d\tau \right)^{1/2} \tag{.0.9}$$

The term (.0.8) can be bounded above using the triangular inequality:

$$\left(\int \left| B\widehat{\psi}_n(\tau) - B\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) \right|^2 \pi(\tau) d\tau \right)^{1/2}$$
$$\leq M_B \left(\int \left| \widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \beta_0) \right|^2 \pi(\tau) d\tau \right)^{1/2} + \left(\int \left| B\widehat{\psi}_n^S(\tau, \beta_0) - B\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) \right|^2 \pi(\tau) d\tau \right)^{1/2}.$$

An application of Lemma .0.2 and the geometric ergodicity of $(\mathbf{y}_t^s, \mathbf{x}_t)$ yields:

$$\left(\int \left| \widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \beta_0) \right|^2 \pi(\tau) d\tau \right)^{1/2} = O_p(1/\sqrt{n}).$$

Expanding the term in $\widehat{\psi}_n^s$ yields:

$$\left(\int \left|B\widehat{\psi}_n^S(\tau,\beta_0) - B\widehat{\psi}_n^S(\tau,\widehat{\beta}(0))\right|^2 \pi(\tau)d\tau\right)^{1/2}$$
$$\leq \left(\int \left|B\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(0))]\right|^2 \pi(\tau)d\tau\right)^{1/2}$$
$$+ M_B\left(\int \left|[\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(0))] - \mathbb{E}[\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(0))]\right|^2 \pi(\tau)d\tau\right)^{1/2}$$
$$\leq \left(\int \left|B\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(0))]\right|^2 \pi(\tau)d\tau\right)^{1/2} + O_p(\frac{(M_n\delta_{sn})^\gamma\sqrt{C_{sn}}}{\sqrt{n}})$$
$$\leq M_B\left(\int \left|\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(0))] - \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\beta_0 - \widehat{\beta}(0)]\right|^2 \pi(\tau)d\tau\right)^{1/2}$$
$$+ \left(\int \left|B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\beta_0 - \widehat{\beta}(0)]\right|^2 \pi(\tau)d\tau\right)^{1/2} + O_p(\frac{(M_n\delta_{sn})^\gamma\sqrt{C_{sn}}}{\sqrt{n}}).$$

Note that:

$$\left(\int \left|\mathbb{E}[\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(0))] - \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\beta_0 - \widehat{\beta}(0)]\right|^2 \pi(\tau)d\tau\right)^{1/2} = O_p(M_n\delta_n)$$

by assumption and

$$\left(\int \left|B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\beta_0 - \widehat{\beta}(0)]\right|^2 \pi(\tau)d\tau\right)^{1/2} = \|\widehat{\beta}_n - \beta_0\|_{weak}$$

by definition. Furthermore, the rate is $\|\widehat{\beta}_n - \beta_0\|_{weak} = O_p(\delta_n)$ by assumption. Overall, the following bound holds for (.0.7):

$$\left(\int \left|B\widehat{\psi}_n(\tau) - B\widehat{\psi}_n^S(\tau,\widehat{\beta}(0))\right|^2 \pi(\tau)d\tau\right)^{1/2} \leq O_p(\frac{1}{\sqrt{n}}) + O_p(\delta_n) + O_p(\frac{(M_n\delta_n)^\gamma\sqrt{C_{sn}}}{\sqrt{n}}).$$

Re-arranging (.0.9) to apply the stochastic equicontinuity result again yields:

$$\left(\int \left|\left(\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))\right) - \mathbb{E}\left(\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))\right)\right|^2 \pi(\tau)d\tau\right)^{1/2}$$
$$\leq \left(\int \left|\left(\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))\right) - \mathbb{E}\left(\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(1))\right)\right|^2 \pi(\tau)d\tau\right)^{1/2}$$
$$+ \left(\int \left|\left(\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(0))\right) - \mathbb{E}\left(\widehat{\psi}_n^S(\tau,\beta_0) - \widehat{\psi}_n^S(\tau,\widehat{\beta}(0))\right)\right|^2 \pi(\tau)d\tau\right)^{1/2}$$
$$= O_p\left(\frac{(M_n\delta_{sn})^\gamma\sqrt{C_{sn}}}{\sqrt{n}}\right).$$

Using the bounds for (.0.7) and (.0.9) yields the bound for (.0.6):

$$\left| \int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0))\right) \left[\overline{B\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1))\right) - B\mathbb{E}\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1))\right)}\right] \pi(\tau) d\tau \right|$$
$$\leq O_p\left(\frac{(M_n\delta_{sn})^\gamma \sqrt{C_{sn}}}{\sqrt{n}}\right) O_p\left(\max\left(M_n\delta_n, \frac{1}{\sqrt{n}}, \frac{(M_n\delta_{sn})^\gamma \sqrt{C_{sn}}}{\sqrt{n}}\right)\right).$$

To bound (.0.7), apply the Mean Value theorem up to the second order:

$$\int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0))\right) \overline{B\mathbb{E}\left(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(1))\right)} \pi(\tau) d\tau$$
$$= \int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0))\right) \overline{\left[-B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)))}{d\beta}[\varepsilon_n u_n^*] + \frac{1}{2}B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \widehat{\beta}(\widetilde{h})))}{d\beta}[\varepsilon_n u_n^*, \varepsilon_n u_n^*]\right]} \pi(\tau) d\tau$$
$$= -\int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0))\right) \overline{B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \beta_0))}{d\beta}[\varepsilon_n u_n^*]} \pi(\tau) d\tau + O_p(\varepsilon_n^2)$$
$$+ \int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0))\right) \left[\overline{B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \widehat{\beta}(0)))}{d\beta}[\varepsilon_n u_n^*] - B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \beta_0))}{d\beta}[\varepsilon_n u_n^*]}\right] \pi(\tau) d\tau.$$

Where the $O_p(\varepsilon_n^2)$ term comes from the Cauchy-Schwarz inequality and the assumptions:

$$\left| \int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0))\right) \overline{\frac{1}{2}B\frac{d^2\mathbb{E}(\widehat{\psi}_n^S(\tau, \widehat{\beta}(\widetilde{t})))}{d\beta d\beta}[\varepsilon_n u_n^*, \varepsilon_n u_n^*]} \pi(\tau) d\tau \right|$$
$$\leq \left(\int \left| B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0))\right)\right|^2 \pi(\tau) d\tau\right)^{1/2} \frac{\varepsilon_n^2}{2} \left(\int \left| B\frac{d^2\mathbb{E}(\widehat{\psi}_n^S(\tau, \widehat{\beta}(\widetilde{t})))}{d\beta d\beta}[u_n^*, u_n^*]\right|^2 \pi(\tau) d\tau\right)^{1/2}.$$

It was shown above that:

$$\left(\int \left| B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}(0))\right)\right|^2 \pi(\tau) d\tau\right)^{1/2} = O_p\left(\max\left(M_n\delta_n, \frac{1}{\sqrt{n}}, \frac{(M_n\delta_{sn})^\gamma \sqrt{C_{sn}}}{\sqrt{n}}\right)\right).$$

Also, by Assumption .0.9 *ii.*:

$$\left(\int \left| B\frac{d^2\mathbb{E}(\widehat{\psi}_n^S(\tau, \widehat{\beta}(\widetilde{t})))}{d\beta d\beta}[u_n^*, u_n^*]\right|^2 \pi(\tau) d\tau\right)^{1/2} = O_p(1).$$

Finally, applying the Cauchy-Schwarz inequality to the last term of the expansion

of (.0.7) yields:

$$\int B\left(\widehat{\psi}_n(\tau)-\widehat{\psi}_n^S(\tau,\widehat{\beta}(0))\right)\left[\overline{B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)))}{d\beta}[\varepsilon_n u_n^*]-B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[\varepsilon_n u_n^*]}\right]\pi(\tau)d\tau$$
$$\leq\left(\int\left|B\widehat{\psi}_n(\tau)-\widehat{\psi}_n^S(\tau,\widehat{\beta}(0))\right|^2\pi(\tau)d\tau\right)^{1/2}$$
$$\times\varepsilon_n\left(\int\left|B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\widehat{\beta}(0)))}{d\beta}[u_n^*]-B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[u_n^*]\right|^2\pi(\tau)d\tau\right)^{1/2}$$
$$=O_p\left(\varepsilon_n\max\left(\delta_n,\frac{1}{\sqrt{n}},\frac{(M_n\delta_{sn})^\gamma\sqrt{C_{sn}}}{\sqrt{n}})\delta_n\right)\right).$$

Using inequality (.0.2) together with the bounds above and the expansions of (.0.3) and (.0.4) yields:

$$\begin{aligned}0\leq&-\varepsilon_n\int B\left(\widehat{\psi}_n(\tau)-\widehat{\psi}_n^S(\tau,\widehat{\beta}(0))\right)\overline{B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[u_n^*]}\pi(\tau)d\tau\\&-\varepsilon_n\int\overline{B\left(\widehat{\psi}_n(\tau)-\widehat{\psi}_n^S(\tau,\widehat{\beta}(0))\right)}B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[u_n^*]\pi(\tau)d\tau\\&+O_p\left(\varepsilon_n^2\right)+O_p\left(\frac{M_{sn}^\gamma C_{sn}}{\sqrt{n}}\max(\delta_{wn},\frac{1}{\sqrt{n}},\frac{M_{sn}^\gamma C_{sn}}{\sqrt{n}}))\right)\\&+O_p\left(\varepsilon_n\delta_{wn}\max(\delta_{wn},\frac{1}{\sqrt{n}},\frac{M_{sn}^\gamma C_{sn}}{\sqrt{n}})\right)+O_p\left(\frac{M_{sn}^{2\gamma}C_{sn}^2}{n}\right)\end{aligned}$$

Since $\varepsilon_n=\pm\frac{1}{\sqrt{n}M_n}$, dividing by ε_n both keeps and flips the inequality so that:

$$\begin{aligned}&\int B\left(\widehat{\psi}_n(\tau)-\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)\right)\overline{B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[u_n^*]}\pi(\tau)d\tau\\&+\int\overline{B\left(\widehat{\psi}_n(\tau)-\widehat{\psi}_n^S(\tau,\widehat{\beta}_n)\right)}B\frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau,\beta_0))}{d\beta}[u_n^*]\pi(\tau)d\tau\\&=O_p(\varepsilon_n)+O_p\left(\frac{(M_n\delta_{sn})^\gamma\sqrt{C_{sn}}}{\varepsilon_n\sqrt{n}}\max\left(M_n\delta_n,\frac{1}{\sqrt{n}},\frac{(M_n\delta_{sn})^\gamma\sqrt{C_{sn}}}{\sqrt{n}})\right)\right)\\&+O_p\left(\max\left(M_n\delta_n,\frac{1}{\sqrt{n}},\frac{(M_n\delta_{sn})^\gamma\sqrt{C_{sn}}}{\sqrt{n}}\right)\delta_n\right)+O_p\left(\frac{(M_n\delta_{sn})^{2\gamma}C_{sn}}{\varepsilon_n n}\right).\end{aligned}$$

By construction, $\varepsilon_n=o_p(1/\sqrt{n})$ and the assumptions imply that

$$M_n^{1+\gamma}\delta_{sn}^\gamma\sqrt{C_{sn}}\max\left(M_n\delta_n,\frac{1}{\sqrt{n}},\frac{(M_n\delta_{sn})^\gamma\sqrt{C_{sn}}}{\sqrt{n}}\right)=o(1/\sqrt{n})$$
$$\text{and}\quad\frac{M_n^{2\gamma+1}\delta_{sn}^{2\gamma}C_{sn}}{\sqrt{n}}=o(1/\sqrt{n}).$$

To conclude the proof, note that:

$$\begin{aligned}
&\int B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}_n)\right) \overline{B \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \beta_0))}{d\beta}[u_n^*]}\pi(\tau)d\tau \\
&+ \int \overline{B\left(\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}_n)\right)} B \frac{d\mathbb{E}(\widehat{\psi}_n^S(\tau, \beta_0))}{d\beta}[u_n^*]\pi(\tau)d\tau \\
&= \int [\psi_\beta(\tau, u_n^*)\left(\overline{B[\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}_n)]}\right) + \overline{\psi_\beta(\tau, u_n^*)}\left(B[\widehat{\psi}_n(\tau) - \widehat{\psi}_n^S(\tau, \widehat{\beta}_n)]\right)] \\
&= o_p(1/\sqrt{n}).
\end{aligned}$$

□

Proof of Theorem .0.3: Using Assumption .0.8, the difference between ϕ at $\widehat{\beta}_n$ and at the true value β_0 can be linearized:

$$\begin{aligned}
\frac{\sqrt{n}}{\sigma_n^*}\left(\phi(\widehat{\beta}_n) - \phi(\beta_0)\right) &= \frac{\sqrt{n}}{\sigma_n^*}\frac{d\phi(\beta_0)}{d\beta}[\widehat{\beta}_n - \beta_0] + o_p(1) \\
&= \frac{\sqrt{n}}{\sigma_n^*}\frac{d\phi(\beta_0)}{d\beta}[\widehat{\beta}_n - \beta_{0,n}] + o_p(1) \\
&= \sqrt{n}\langle u_n^*, \widehat{\beta}_n - \beta_{0,n}\rangle + o_p(1) \\
&= \sqrt{n}\langle u_n^*, \widehat{\beta}_n - \beta_0\rangle + o_p(1) \\
&= \frac{\sqrt{n}}{2}\left(\int \left[B\psi_\beta(\tau, u_n^*)\overline{B\psi_\beta(\tau, \widehat{\beta}_n - \beta_0)} + \overline{B\psi_\beta(\tau, u_n^*)}B\psi_\beta(\tau, \widehat{\beta}_n - \beta_0)\right]\right)\pi(\tau)d\tau.
\end{aligned}$$

Using Lemma .0.5 *a)* and *b)*, replace the term $B\psi_\beta(\tau, \widehat{\beta}_n - \beta_0)$ under the integral with $B\widehat{\psi}_n^S(\tau, \widehat{\beta}_n) - B\widehat{\psi}_n^S(\tau, \beta_0)$ so that:

$$\begin{aligned}
\frac{\sqrt{n}}{\sigma_n^*}\left(\phi(\widehat{\beta}_n) - \phi(\beta_0)\right) = \frac{1}{2}\Big(\int \Big[&B\psi_\beta(\tau, u_n^*)\overline{[B\widehat{\psi}_n^S(\tau, \widehat{\beta}_n) - B\widehat{\psi}_n^S(\tau, \beta_0)]} \\
&+ \overline{B\psi_\beta(\tau, u_n^*)}[B\widehat{\psi}_n^S(\tau, \widehat{\beta}_n) - B\widehat{\psi}_n^S(\tau, \beta_0)]\Big]\Big)\pi(\tau)d\tau + o_p(1).
\end{aligned}$$

Now Lemma .0.5 *c)* implies that $B\widehat{\psi}_n^S(\tau, \widehat{\beta}_n)$ can be replaced with $B\widehat{\psi}_n(\tau)$ up to a $o_p(1/\sqrt{n})$ so that the above becomes:

$$\frac{\sqrt{n}}{\sigma_n^*}\left(\phi(\widehat{\beta}_n) - \phi(\beta_0)\right) = \frac{\sqrt{n}}{2}\Big(\int \left[B\psi_\beta(\tau, u_n^*)\overline{BZ_n^S(\tau)} + \overline{B\psi_\beta(\tau, u_n^*)}BZ_n^S(\tau)\right]\Big)\pi(\tau)d\tau + o_p(1).$$

To conclude, apply a Central Limit Theorem to the scalar and real-valued random variable variable:

$$\frac{1}{2}\int [B\psi_\beta(\tau, u_n^*)\overline{BZ_t^S(\tau)} + \overline{B\psi_\beta(\tau, u_n^*)}BZ_t^S(\tau)]\pi(\tau)d\tau.$$

Because of u_n^* and the geometric ergodicity of the simulated data, a CLT for non-stationary mixing triangular arrays is required: results in Wooldridge & White (1988); de Jong (1997) can be applied. For any $\delta > 0$:

$$\begin{aligned}
&\mathbb{E}\left(\left|\int[\boldsymbol{\psi}_\beta(\tau,u_n^*)\overline{Z_t^S(\tau)}+\overline{\boldsymbol{\psi}_\beta(\tau,u_n^*)}Z_t^S(\tau)]\pi(\tau)d\tau\right|^{2+\delta}\right)\\
&\leq 2^{2+\delta}\left[\mathbb{E}\left(\int\left|\overline{\boldsymbol{\psi}_\beta(\tau,u_n^*)}Z_t^S(\tau)\right|\pi(\tau)d\tau\right)\right]^{2+\delta}\\
&\leq 2^{2+\delta}\left(\int\left|B\boldsymbol{\psi}_\beta(\tau,u_n^*)\right|^2\pi(\tau)d\tau\right)^{\frac{2+\delta}{2}}\left[\mathbb{E}\left(\int\left|BZ_t^S(\tau)\right|^2\pi(\tau)d\tau\right)\right]^{\frac{2+\delta}{2}}.
\end{aligned}$$

By definition of u_n^* and $\|\cdot\|_{weak}$:

$$\left(\int\left|B\boldsymbol{\psi}_\beta(\tau,u_n^*)\right|^2\pi(\tau)d\tau\right)^{1/2}=\|v_n^*\|_{weak}/\sigma_n^*\in[1/\overline{a},1/\underline{a}].$$

Because B is bounded linear and $|Z_t^S(\tau)|\leq 2$:

$$\left[\mathbb{E}\left(\int\left|BZ_t^S(\tau)\right|^2\pi(\tau)d\tau\right)\right]^{\frac{2+\delta}{2}}\leq[2M_B]^{2+\delta}.$$

Eventually, it implies:

$$\mathbb{E}\left(\left|\int[\boldsymbol{\psi}_\beta(\tau,u_n^*)\overline{Z_t^S(\tau)}+\overline{\boldsymbol{\psi}_\beta(\tau,u_n^*)}Z_t^S(\tau)]\pi(\tau)d\tau\right|^{2+\delta}\right)\leq\frac{[4M_B]^{2+\delta}}{\underline{a}}<\infty.$$

Given the mixing condition and the definition of σ_n^*:

$$\frac{\sqrt{n}}{2}\int[B\boldsymbol{\psi}_\beta(\tau,u_n^*)[\overline{BZ_t^S(\tau)-B\mathbb{E}(Z_t^S(\tau))}]+\overline{B\boldsymbol{\psi}_\beta(\tau,u_n^*)}[BZ_t^S(\tau)-B\mathbb{E}(Z_t^S(\tau))]]\pi(\tau)d\tau\xrightarrow{d}\mathcal{N}(0,1).$$

By geometric ergodicity and because the characteristic function is bounded $\sqrt{n}|\mathbb{E}(Z_t^S(\tau))|\leq C_\rho/\sqrt{n}=o(1)$, hence:

$$\frac{\sqrt{n}}{2}\int[B\boldsymbol{\psi}_\beta(\tau,u_n^*)\overline{BZ_t^S(\tau)}+\overline{B\boldsymbol{\psi}_\beta(\tau,u_n^*)}BZ_t^S(\tau)]\pi(\tau)d\tau\xrightarrow{d}\mathcal{N}(0,1).$$

This concludes the proof. □