A Sieve-SMM Estimator for Dynamic Models

Jean-Jacques Forneron*

February 4, 2019

Abstract

This paper proposes a Sieve Simulated Method of Moments (Sieve-SMM) estimator for the parameters and the distribution of the shocks in nonlinear dynamic models where the likelihood and the moments are not tractable. An important concern with SMM, which matches sample with simulated moments, is that a parametric distribution is required but economic quantities that depend on this distribution, such as welfare and asset-prices, can be sensitive to misspecification. The Sieve-SMM estimator addresses this issue by flexibly approximating the distribution of the shocks with a Gaussian and tails mixture sieve. The asymptotic framework provides consistency, rate of convergence and asymptotic normality results, extending existing sieve estimation theory to a new framework with more general dynamics and latent variables. Monte-Carlo simulations illustrate the finite sample properties of the estimator. Two empirical applications highlight the importance of the distribution of the shocks for estimates and counterfactuals.

JEL Classification: C14, C15, C32, C33.
Keywords: Simulated Method of Moments, Mixture Sieve, Semi-Nonparametric Estimation.

*Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215.
Email: jjmf@bu.edu. This paper is based on the third chapter of my doctoral dissertation at Columbia University. 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, Ivan Fernandez-Val, Ron Gallant, Eric Gautier, Hiro Kaido, Dennis Kristensen, Sokbae Lee, Kim Long, José Luis Montiel Olea, Zhongjun Qu, Christoph Rothe, Bernard Salanié and the participants of the Columbia Econometrics Colloquium as well as the participants of the econometrics seminar at Boston University, Chicago Booth, UC Berkeley, Bocconi, Georgetown, UPenn and participants at several conferences. All errors are my own.

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 in this paper have the form:

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

$$u_t = g_{latent}(u_{t-1}, \theta, f, e_t), \quad e_t \sim f. \tag{2}$$

The observed outcome variable is y_t, x_t are exogenous regressors and u_t is a vector of unobserved latent variables. The unknown parameters include θ, 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 θ 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 in this paper to satisfy the L^2-smoothness regularity conditions 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 also 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 $\hat{\psi}_n$, 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 $(\theta, f_{\omega,\mu,\sigma})$ using e_t^s.
  Compute $\hat{\psi}_n^s(\theta, f_{\omega,\mu,\sigma})$, the CF of the simulated data $(y_t^s, \ldots, y_{t-L}^s, x_t, \ldots, x_{t-L})$.
Compute the average simulated CF $\hat{\psi}_n^S(\theta, f_{\omega,\mu,\sigma}) = \frac{1}{S}\sum_{s=1}^{S} \hat{\psi}_n^s(\theta, f_{\omega,\mu,\sigma})$.
Compute the objective function $\hat{Q}_n^S(\theta, f_{\omega,\mu,\sigma}) = \int \left|\hat{\psi}_n(\tau) - \hat{\psi}_n^S(\theta, f_{\omega,\mu,\sigma})\right|^2 \pi(\tau) d\tau$.
Find the parameters $(\hat{\theta}_n, \hat{\omega}_n, \hat{\mu}_n, \hat{\sigma}_n)$ that minimize $\hat{Q}_n^S$.

To illustrate the class of models considered and the usefulness of the mixture sieve for economic analysis, consider the first empirical application in Section 5 where the growth rate of consumption $\Delta c_t = \log(C_t/C_{t-1})$ is assumed to follow the following process:

$$\Delta c_t = \mu_c + \rho_c \Delta c_{t-1} + \sigma_t e_{t,1}, \quad e_{t,1} \sim f \tag{3}$$

$$\sigma_t^2 = \mu_\sigma + \rho_\sigma \sigma_{t-1}^2 + \kappa_\sigma e_{t,2}, \quad e_{t,2} \sim \chi_1^2. \tag{4}$$

Compared to the general model (1)-(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 $\theta = (\mu_c, \rho_c, \mu_\sigma, \rho_\sigma, \kappa_\sigma)$. 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)-(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 main results only provide low-level conditions for a specific choice of moments and sieve basis, Appendix F provides high-level conditions which can be used for a larger class of bounded moments and sieve bases. These results also allow to nonparametrically estimate quantities other than the distribution of the shocks. 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 (1)-(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 holds under the geometric ergodicity condition found 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 for iid or strictly stationary data with limited dependence induced by the moments.[2]

The second challenge is that in the model (1)-(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

[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 review of sieve M-estimation with iid and dependent data.

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)-(4), this condition holds if both $|\rho_y|$ and $|\rho_\sigma|$ are strictly less than 1. The 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. Note that the problem of bias accumulation seems quite generic to sieve estimation of dynamic models: if the computation of the moments or likelihood involve a filtering step then the bias accumulates inside the prediction error of the filtered values.[3]

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 notable 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: sieve estimation and the Simulated Method of Moments (SMM). This section provide a non-exhaustive review of the existing methods and results to introduce the new challenges in the combined setting.

A key aspect to simulation-based estimation is the choice of moments $\hat{\psi}_n$. 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

[3]This is related to the accumulation of errors studied in the approximation of DSGE models (see e.g. Peralta-Alva & Santos, 2014). Note that in the present estimation context, the error in the moments involves the difference between n dimensional integral over the true and the approximated distribution of the shocks which complicates the analysis. This is also related to the propagation of prediction error in the filtering of unobserved latent variables using e.g. the Kalman or Particle filter.

(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 this paper.

To achieve parametric efficiency, a number of papers consider using nonparametric moments but assume the distribution f is known.[4] 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. In statistics, 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 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.[5] 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 consider specific static models and provide limited asymptotic results. Furthermore, they consider sampling methods for the simulations that are very computationally costly (see Section 2.3 for a discussion).[6]

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

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

[5]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 $\hat{\theta}_n$ on the covariates.

[6]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. 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.

of productivity and inflation shocks. Gospodinov & Ng (2015); Gospodinov et al. (2017) use the Generalized Lambda famility to flexibly model the first 4 moments of the shocks in a non-invertible moving avergage and a measurement error 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 $\hat{\theta}_n$ 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 IV regressions[7], quantile IV regressions,[8] and the semi-nonparametric estimation of asset pricing models,[9] 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. See e.g. Chen & Pouzo (2012, 2015) and Chen & Liao (2015) for recent results with iid data and 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. As in the Sieve-SMM setting, this has implications for bias accumulation in parameter dependent time-series properties. Exisinting results, based on coupling methods (see e.g. Doukhan et al., 1995; Chen & Shen, 1998), do not apply to this class of moments and the authors rely on Bootstrap inference without formal justification.

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

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

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

[9]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 $\|\beta_1 - \beta_2\|_{TV} = \|\theta_1 - \theta_2\| + \|f_1 - f_2\|_{TV}$ and $\|\beta_1 - \beta_2\|_\infty = \|\theta_1 - \theta_2\| + \|f_1 - f_2\|_\infty$, where $\|\theta\|$ and $\|e\|$ correspond the Euclidian norm of θ and e respectively. $\|\beta_1\|_m$ is a norm on the mixture components: $\|\beta_1\|_m = \|\theta\| + \|(\omega, \mu, \sigma)\|$ 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 $\frac{d\phi(\beta_1)}{d\beta}[\beta_2] = \left.\frac{d\phi(\beta_1 + \varepsilon\beta_2)}{d\varepsilon}\right|_{\varepsilon=0}$, 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 2 introduces the Sieve-SMM estimator, explains how to implement it in practice and provides important properties of the mixture sieve. Section 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 B provides two extensions, one to include auxiliary variables in the CF and another to allow for dynamic panels with small T. Section 4 provides Monte-Carlo simulations to illustrate the theoretical results. Section 5 gives empirical examples for the estimator. Section 6 concludes. Appendix A gives some information about the CF and details on how to compute the estimator in practice as well as identification and additional asymptotic normality results for the stochastic volatility model. Appendix B provides extensions of the main results to moments of auxiliary variables and short panel data. Appendix C provides additional Monte-Carlo simulations for short panels. Appendix D provide additional empirical results to the ones presented in the main text. Appendix E provides the proofs to the main results and the extensions. The online supplement includes:[10] Appendix F which provides results for more general moment functions and sieve bases and Appendix G which provides the proofs for these results.

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

[10] The online supplement can be found at http://jjforneron.com/SieveSMM/Supplement.pdf.

more details. It discusses the choice of mixture sieve, moments and objective function as well as some important properties of the mixture sieve. The simple running example used throughout the analysis is based on the empirical applications of Section 5.

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:

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

The second empirical application estimates a log-normal stochastic volatility process:

$$\log(\sigma_t) = \mu_\sigma + \rho_\sigma \log(\sigma_{t-1}) + \kappa_\sigma e_{t,2}, \quad e_{t,2} \overset{iid}{\sim} \mathcal{N}(0,1).$$

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. Using the notation given in (1)-(2), the latent variable is given by $u_t = (u_{t,1}, u_{t,2})$, where $u_{t,1} = e_{t,1}$ and $u_{t,2} = \sigma_t^2$ (or $u_{t,2} = \log(\sigma_t)$).

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 5). 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.

2.1 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 simple functions of uniform random variables.

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

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

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

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

with $f_L(e_t,\xi_L) = 0$ for $e_t \ge 0$ and $f_R(e_t,\xi_L) = 0$ for $e_t \le 0$. To simulate from the Gaussian and tails mixture, draw $Z_1,\dots,Z_k \overset{iid}{\sim} \mathcal{N}(0,1)$, $\nu,\nu_L,\nu_R \overset{iid}{\sim} \mathcal{U}_{[0,1]}$ 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 $\omega_0 = 0$:

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

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

For applications 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 get a Gaussian only 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:

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

When $\xi_L,\xi_R \ge 0$ then draws from the tail components have finite expectation, they also have finite variance if $\xi_L,\xi_R \ge 1$. More generally, for the j-th moment to be finite, $j \ge 1$, $\xi_L,\xi_R \ge j$ is necessary. Gallant & Nychka (1987) also add a parametric component to model fat tails by mixing a Hermite polynomial density with a Student density. Neither the Hermite polynomial nor the Student 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 tail indices ξ_L,ξ_R will be estimates along with the remaining parameters of the mixture distribution.

The indicator function $\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\omega_l,\sum_{l=0}^{j}\omega_l]}$ 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 4.[11]

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

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.[12] Using the characteristic function rather than the likelihood resolves the unbounded likelihood 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 density $f_{\omega,\mu,\sigma}$.

2.2 Continuunm of Moments 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 $\beta = (\theta, f)$ 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 $(\mathbf{y}_t, \mathbf{x}_t) = (y_t, \ldots, y_{t-L}, x_t, \ldots, x_{t-L})$:

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

where i is the imaginary number such that $i^2 = -1$.[13] The CF is one-to-one with the joint distribution of $(\mathbf{y}_t, \mathbf{x}_t)$, so that the model is identified by $\hat{\psi}_n(\cdot)$ 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 A.1.

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.[14]

[12] 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.

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

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

Since the moments are infinite dimensional, this paper adopts the approach of Carrasco & Florens (2000); Carrasco et al. (2007a) to handle the continuum of moment conditions:[15]

$$\hat{Q}_n^S(\beta) = \int \left| \hat{\psi}_n(\tau) - \hat{\psi}_n^S(\tau, \beta) \right|^2 \pi(\tau) d\tau. \tag{5}$$

The objective function is a weighted average of the square norm between the empirical $\hat{\psi}_n$ and the simulated $\hat{\psi}_n^S$ 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 4 and 5.[16] 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 $\hat{\beta}_n = (\hat{\theta}_n, \hat{f}_n)$ is defined as an approximate minimizer of $\hat{Q}_n^S$:

$$\hat{Q}_n^S(\hat{\beta}_n) \leq \inf_{\beta \in \mathcal{B}_{k(n)}} \hat{Q}_n^S(\beta) + O_p(\eta_n) \tag{6}$$

where $\eta_n \geq 0$ and $\eta_n = o(1)$ corresponds to numerical optimization and integration errors. Indeed, since the integral in (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.[17] All Monte-Carlo simulations and empirical results in this paper are based on quasi-Monte-Carlo integration. Additional computional details are given in Appendix A.2.

[15] 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.

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

[17] 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.

Example 1 (Continued) (Stochastic Volatility)**.** *The following illustrates the steps involved in Sieve-SMM Algorithm for the stochastic volatility model with a Gaussian only mixture:*

- *fix* $k(n) \geq 1$, $S \geq 1$ *and* $L \geq 1$,
- *construct a grid* $(\tau_1, \ldots, \tau_m)$, *e.g.* $m = 1,000$ *Box-Muller transformed Sobol sequence,*
- *compute the sample Characteristic Function over the grid*

$$\hat{\psi}_n = (\hat{\psi}_n(\tau_1), \ldots, \hat{\psi}_n(\tau_m)) = \frac{1}{n} \sum_{t=L+1}^{n} (e^{i\tau_1' \mathbf{y}_t}, \ldots, e^{i\tau_m' \mathbf{y}_t}), \quad \mathbf{y}_t = (y_t, \ldots, y_{t-L}),$$

- *draw* $Z_{t,j}^s \overset{iid}{\sim} \mathcal{N}(0,1)$, $\nu_t^s \overset{iid}{\sim} \mathcal{U}(0,1)$ *and,* $e_{t,2}^2 \overset{iid}{\sim} \chi_1^2$ *where* $j \in \{1, \ldots, k(n)\}$
- *minimize the objective* $\hat{Q}_n^S(\theta, \omega, \mu, \sigma)$, *computed as follows:*
  - *compute* $e_{t,1}^s = \sum_{j=1}^{k(n)} \mathbb{1}_{\nu \in [\sum_{l=0}^{j-1} \omega_l, \sum_{l=0}^{j} \omega_l]} (\mu_j + \sigma_j Z_{t,j}^s)$
  - *simulate* y_t^s, σ_t^s *using* $\theta = (\mu_y, \rho_y, \mu_\sigma, \rho_\sigma, \kappa_\sigma)$ *and* $e_t^s = (e_{t,1}^s, e_{t,2}^s)$
  - *compute* $\hat{\psi}_n^s$ *as above and* $\hat{Q}_n^s(\theta, \omega, \mu, \sigma) = \sum_{\ell=1}^{m} |\hat{\psi}_n(\tau_\ell) - \hat{\psi}_n^S(\tau_\ell)|^2$.

2.3 Approximation Rate and L^2-Smoothness 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 where exchange rate data may exhibit larger tails. The following lemma extends the approximation results of Kruijer et al. (2010) to multivariate densities with independent components and potentially fat tails.

Lemma 1 (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: $\mu_j \in [-\bar{\mu}_k, \bar{\mu}_k]$ with $\bar{\mu}_k = O\left(\log[k]^{1/b}\right)$

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. Then, a L^p-smoothness result for the mixture sieve is provided in Lemma 2 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 2 (L^2-Smoothness of Simulated Mixture Sieves)**.** *Suppose that*

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

with $|\mu_j|$ *and* $|\tilde{\mu}_j| \leq \bar{\mu}_{k(n)}$, $|\sigma_j|$ *and* $|\tilde{\sigma}_j| \leq \bar{\sigma}$. *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:*

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

Lemma 2 is key in proving the L^2-smoothness conditions of the moments $\hat{\psi}_n^s$ 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 $k(n)$.[18] 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 ξ_L, ξ_R. Hence, draws from the Gaussian and tails mixture are L^2-smooth in both (ω, μ, σ) and ξ.

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

3 Asymptotic Properties of the Estimator

This section provides conditions under which the Sieve-SMM estimator in (6) is consistent, derives its rate of convergence and asymptotic normality results for linear functionals of $\hat{\beta}_n$.

3.1 Consistency

Consistency results are given under low-level conditions on the DGP using the Gaussian and tails mixture sieve with the CF.[19] First, the population objective Q_n is:

$$Q_n(\beta) = \int \left| \mathbb{E}\left(\hat{\psi}_n(\tau) - \hat{\psi}_n^S(\tau, \beta) \right) \right|^2 \pi(\tau) d\tau. \tag{7}$$

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:[20]

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

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

[19] Consistency results allowing for non-mixture sieves and other moments are given in Appendix F.1.

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

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 comes from the non-linearity of the mixture sieve. The fourth power is due to the dependence: the inequality in Lemma G15 provides a bound 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 $\beta = (\theta, f)$. 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 ill-posedness.[21] 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 A.4.

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: $Q_n(\beta_0) = O(1/n^2)$.[22]

Assumption 2 (Data Generating Process)**.** *y_t^s is simulated according to the dynamic model (1)-(2) where g_{obs} and g_{latent} satisfy the following Hölder conditions for some $\gamma \in (0, 1]$, $\|\cdot\|_{\mathcal{B}}$ is either the total variation or supremum norm and:*

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

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

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

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

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

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

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 $(e_1, e_2) \in \mathbb{R}^{dim(e)}$.

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

[22] See Proposition F4 in the supplemental material for the second result.

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.[23]

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 the nonparametric approximation bias $\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}$ does not accumulate too much. These conditions are similar to Duffie & Singleton (1993)'s L^2-Unit Circle condition which they propose as an alternative to geometric ergodicity for uniform LLNs and CLTs. The decay conditions play a crucial role here since they 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 2. 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 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 E.2 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_1^2 volatility shocks.[24]

Example 1 (Continued) (Stochastic Volatility)**.** $|\rho_y| < 1$ *implies y(i) holds. 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 $\mathbb{E}(y_{t-1}^2)$ *is bounded. Since* f *has mean zero and unit variance,* $\mathbb{E}(y_{t-1}^2)$ *is bounded if* $|\mu_\sigma| \leq \bar{\mu}_\sigma < \infty$, $|\rho_\sigma| \leq \bar{\rho}_\sigma < 1$ *and* $\kappa_\sigma \leq \bar{\kappa}_\sigma < \infty$ *for some* $\bar{\mu}_\sigma, \bar{\rho}_\sigma, \bar{\kappa}_\sigma$*. For condition y(iii), take* $u_t = (\sigma_t^2, e_{t,1})$ *and* $\tilde{u}_t = (\tilde{\sigma}_t^2, \tilde{e}_{t,1})$*:*

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

The first inequality is due to the Hölder continuity of the square-root function.[25] σ_t *and* $\tilde{e}_{t,1}$

[23]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.

[24]Some additional examples are given in Appendix F.4. They are not tied to the use of mixtures, and as a result, impose stronger restrictions on the density f such as bounded support.

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

are independent, $\mathbb{E}(\sigma_t^2)$ *is bounded above under the previous parameter bounds and* $\mathbb{E}(e_{t,1}^2) = 1$ *and so condition y(iii) holds term by term. If the volatility* σ_t^2 *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_t^2 *follows an AR(1) process, assumptions u(i), u(ii) and u(iii) are satisfied.*

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 $\min(\gamma_1, \gamma_2 \times \gamma_3)$ instead of γ^2. This could lead to sharper rates of convergence in section 3.2 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 3 (Assumption 2/2′ implies L^2-Smoothness of the Moments)**.** *Under either Assumption 2 or 2′, if the assumptions of Lemma 2 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 $\|\beta\|_m = \|\theta\| + \|(\omega, \mu, \sigma)\|$ *is the pseudo-norm on* θ *and the mixture parameters* (ω, μ, σ) *from Lemma 2. Also, since* π *is the Gaussian density the integral* $\int \sqrt{\pi(\tau)} d\tau$ *is finite.*

Lemma 3 gives the first implication of Assumption 2. It shows that the moments $\hat{\psi}_t^s$ 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 1.[26] 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 4 (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* $\beta = \Pi_{k(n)}\beta_0$

[26] The assumptions of Lemma 1 imply: $\underline{\sigma}_{k(n)}^{-2\gamma^2} = O\left(k(n)^{2\gamma^2}/\log[(n)]^{4\gamma^2/b}\right)$ and $[k(n) + \overline{\mu}_{k(n)} + \overline{\sigma}]^\gamma = O\left(k(n)^\gamma\right)$. As a result, the maximum is bounded above by $\max\left(k(n)^{2\gamma^2}, k(n)^\gamma\right)\delta^{\gamma^2/2}$ (up to a constant).

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 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 1.

Lemma 4 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 $\hat{\beta}_n$. 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 $Q_n(\Pi_{k(n)}\beta_0) \asymp \|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}^2$. In comparison, the rate derived here is:

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

with $\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}} = O(\log[k(n)]^{2r/b}/k(n)^r)$ as given in Lemma 1. The $1/n^2$ term corresponds to the bias due to the nonstationarity. It is the result of geometric ergodicity 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 $(e_t^s, \ldots, e_1^s) \overset{iid}{\sim} \Pi_{k(n)} f_0 \neq f_0$, 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 $\|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}^{2\gamma^2}$ 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 $Q_n(\Pi_{k(n)}\beta_0) \asymp \|\beta_0 - \Pi_{k(n)}\beta_0\|_{\mathcal{B}}^2$, 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:*

$$\max\left(\frac{log[k(n)]^{4r/b+2}}{k(n)^{2\gamma^2 r}}, \frac{k(n)^4 \log[k(n)]^4}{n}, \frac{1}{n^2}\right) = o\left(\inf_{\beta \in \mathcal{B}_{k(n)}, \|\beta - \beta_0\|_{\mathcal{B}} \geq \varepsilon} Q_n(\beta)\right) \tag{8}$$

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:

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

Theorem 1 is a consequence of the general consistency lemma in Chen & Pouzo (2012) reproduced as Lemma G12 in the appendix. They provide high level conditions which Assumption 2 together with Lemmas 3 and 4 verify for simulation-based estimation of static and dynamic models. Condition (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 G15 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.

3.2 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 $\hat{\beta}_n$.[27] Ultimately, the norm of interest in the strong norm $\|\cdot\|_{\mathcal{B}}$ (i.e. the total variation or the supremum norm) which is generally not equivalent to the weak norm in infinite dimensional settings. 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.

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* $(\beta_1, \beta_2) \in \mathcal{B}_{osn}$*:*

$$\|\beta_1 - \beta_2\|_{weak} = \left[\int \left| \frac{d\mathbb{E}\left(\hat{\psi}_n^S(\tau, \beta_0)\right)}{d\beta} [\beta_1 - \beta_2] \right|^2 \pi(\tau) d\tau \right]^{1/2} \tag{9}$$

[27] 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 $\beta \in \mathcal{B}_{osn}$:

$$\underline{C}_w \|\beta - \beta_0\|_{weak}^2 \leq \int \left| \mathbb{E}\left(\hat{\psi}_n^S(\tau, \beta_0) - \hat{\psi}_n^S(\tau, \beta) \right) \right|^2 \pi(\tau) d\tau. \tag{10}$$

Assumption 3 adapts the weak norm of Ai & Chen (2003) to an objective with a continuum of complex-valued moments. Note that $\int |\mathbb{E}\left(\hat{\psi}_n^S(\tau, \beta_0) - \hat{\psi}_n^S(\tau, \beta) \right)|^2 \pi(\tau) d\tau = Q_n(\beta) + O_p(1/n^2)$ under geometric ergodicity. As a result, Assumption 3 implies that the weak norm is Lipschitz continuous with respect to $\sqrt{Q_n}$, up to a $1/n$ term. 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 4 to derive the bias of the estimator. The resulting bias is larger than in the usual sieve literature because of the dynamic accumulation of the bias discussed in the introduction.

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:*

$$\|\hat{\beta}_n - \beta_0\|_{weak} = O_p\left(\max\left(\frac{\log[k(n)]^{r/b+1}}{k(n)^{\gamma^2 r}}, \frac{k(n)^2 \log[k(n)]^2}{\sqrt{n}} \right) \right). \tag{11}$$

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

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

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 $\sqrt{k(n)/n}$ or $\sqrt{k(n)\log[k(n)]/n}$ in the iid case or strictly stationary case with fixed number of lags in the moments. This is because the inequality in Lemma G15 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. These do not apply in this simulation-based setting because the

dependence properties of y_t^s vary with $\theta \in \Theta$ so that a coupling approach may not apply unless y_t^s only depends on finitely many lags of e_t and x_t, which is quite restrictive. Determining whether this inequality, which leads to the larger variance, can be sharpened is 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 (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:

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

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:

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

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 least 3 bounded derivatives are needed compared to 12 when $\gamma = 1/2$. In comparison, in the iid case 2 and 8 bounded 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 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:

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

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

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

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 usually 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. Parallel computation can reduce some of this burden however.

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 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 A.3. 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.

3.3 Asymptotic Normality

This section derives asymptotic normality results for plug-in estimates $\phi(\hat{\beta}_n)$ 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:

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

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 $\hat{\theta}_n$ to be root-n asymptotically normal, that is $\lim_{n\to\infty} \sigma_n^* < \infty$, are given in Appendix A.5 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 5 (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) $\|\hat{\beta}_n - \beta_0\|_{weak} = O_p(\delta_n)$

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 $\beta \in \{\beta \in \mathcal{B}_{osn}, \|\beta - \beta_0\|_{weak} \leq M_n \delta_n\}$:

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

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

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

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

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

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

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

where $\|\beta\|_m = \|(\theta, \omega, \mu, \sigma)\|$ is the pseudo-norm on θ 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 $\underline{\lambda}_n^{-1/2}$. The relationship 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 1. 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 5. Chen & Christensen (2017) shows how to achieve the optimal rate of convergence using plug-in estimates of the measure of ill-posedness in nonparametric IV regression. A similar approach could be applicable here using these bounds. This is left to future research.

Lemma 6 (Stochastic Equicontinuity Results)**.** *Let* $\delta_{mn} = \delta_n \underline{\lambda}_n^{-1/2}$, $M_n = \log\log(n+1)$. *Suppose that the assumptions of Lemma 5 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:*

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

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

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

Lemma 6 uses the rate of convergence in mixture norm to establish stochastic equicontinuity results. With these results, the moments $\hat{\psi}_n^s(\tau,\beta) - \hat{\psi}_n^s(\tau,\beta_0)$ 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 6,* $k(n)$ *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 in Lemma 6 on the rate of convergence in mixture norm* $(M_n\delta_{mn})^{\frac{\gamma^2}{2}} \max(\log[k(n)]^2, |\log[M_n\delta_{mn}]|^2)k(n)^2 = o(1)$ *simplifies 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* $\|\beta_{0,n} - \beta_0\|_{weak} = \inf_{\beta\in\mathcal{B}_{osn}} \|\beta - \beta_0\|_{weak}$, *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 $v_n^* \in \overline{V}_{k(n)}$ *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:*

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

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

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

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.[28] Here, because the moments are bounded and the data is geometrically ergodic, the long-run variance of the moments is bounded above uniformly in τ.[29] However, since τ has unbounded support, the eigenvalues of the variance may not be bounded below. Assumption 4 plays the role of the lower bound on the eigenvalues.[30]

Assumption 5 (Convergence Rate, Smoothness, Bias)**.** *Suppose that the set $\mathcal{B}_{osn}$ 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:

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

The second derivative is bounded:

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

[28] See e.g. Assumption 3.1(iv) in Chen & Pouzo (2015).

[29] This is shown in Appendix F.3.

[30] A discussion of this assumption is given in Appendix F.5.

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 θ_0 is in the interior of Θ. Assumption 5 is standard in the literature. The first rate condition ensures a fast enough convergence of the nonparametric component for the central limit theorem to dominate the asymptotic distribution (Newey, 1994; Chen et al., 2003). The second rate condition is required for the stochastic equicontinuity result of Lemma 6. 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 5, 6 hold as well as Assumptions 4 and 5, then as n goes to infinity:*

$$r_n\times\left(\phi(\hat{\beta}_n)-\phi(\beta_0)\right)\xrightarrow{d}\mathcal{N}\left(0,1\right),\ \textit{where } r_n=\sqrt{n}/\sigma_n^*\to\infty.$$

Theorem 3 shows that under the previous assumptions, inferences on $\phi(\beta_0)$ can be conducted using the confidence interval $[\phi(\hat{\beta}_n)\pm 1.96\times\sigma_n^*/\sqrt{n}]$. 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 $\phi(\hat{\beta}_n)$ is $\sqrt{n}-$convergent. A result for $\hat{\theta}_n$ is given in Proposition A1 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. The notation is taken from Chen & Pouzo (2015), with sieve parameters $(\hat{\omega}_n,\hat{\mu}_n,\hat{\sigma}_n)$ the sieve variance can be estimated using:

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

where

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

$$\hat{\mho}_n=\int\hat{G}_n(\tau_1)'\hat{\Sigma}_n(\tau_1,\tau_2)\hat{G}_n(\tau_2)\pi(\tau_1)\pi(\tau_2)d\tau_1d\tau_2.$$

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

$$\hat{G}_n(\tau)=\left(\ Real\left(\tfrac{d\hat{\psi}_n^S(\tau,\hat{\theta}_n,\hat{\omega}_n,\hat{\mu}_n,\hat{\sigma}_n)}{d(\theta,\omega,\mu,\sigma)}\right)'\ \ Im\left(\tfrac{d\hat{\psi}_n^S(\tau,\hat{\theta}_n,\hat{\omega}_n,\hat{\mu}_n,\hat{\sigma}_n)}{d(\theta,\omega,\mu,\sigma)}\right)'\ \right)'.$$

Let $Z_n^S(\tau,\beta) = \hat{\psi}_n(\tau) - \hat{\psi}_n^S(\tau,\beta)$. The covariance operator $\hat{\Sigma}_n$ 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

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

is more convenient than computing the large matrices $\hat{G}_n, \hat{\Sigma}_n$. $\hat{\beta}_n$ 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.

4 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 in the optimization routine is a standard normal density. 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. Monte-Carlo simulations were conducted in R[31] for all examples but the AR(1) where 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 it has a sufficiently large approximation bias for both kernel and mixture sieve estimates, compared to smoother symmetric densities not reported here. This is useful when illustrating the effect of dynamics on the bias.

[31]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. Additional Monte-Carlo simulations are provided in Appendix C.

4.1 Basic Examples

The following basic tractable examples are used as benchmarks to illustrate the properties of the Sieve-SMM estimator derived in the previous section. As a benchmark, the estimates are compared to feasible kernel density and OLS estimates.

A Static Model

To illustrate Remark 1, the first example considers: $y_t = e_t \overset{iid}{\sim} f$, 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.[32] 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 $\hat{f}_n$ are compared to the feasible kernel density estimates $\hat{f}_{n,kde}$.

Figure 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 $n = 1,000$.

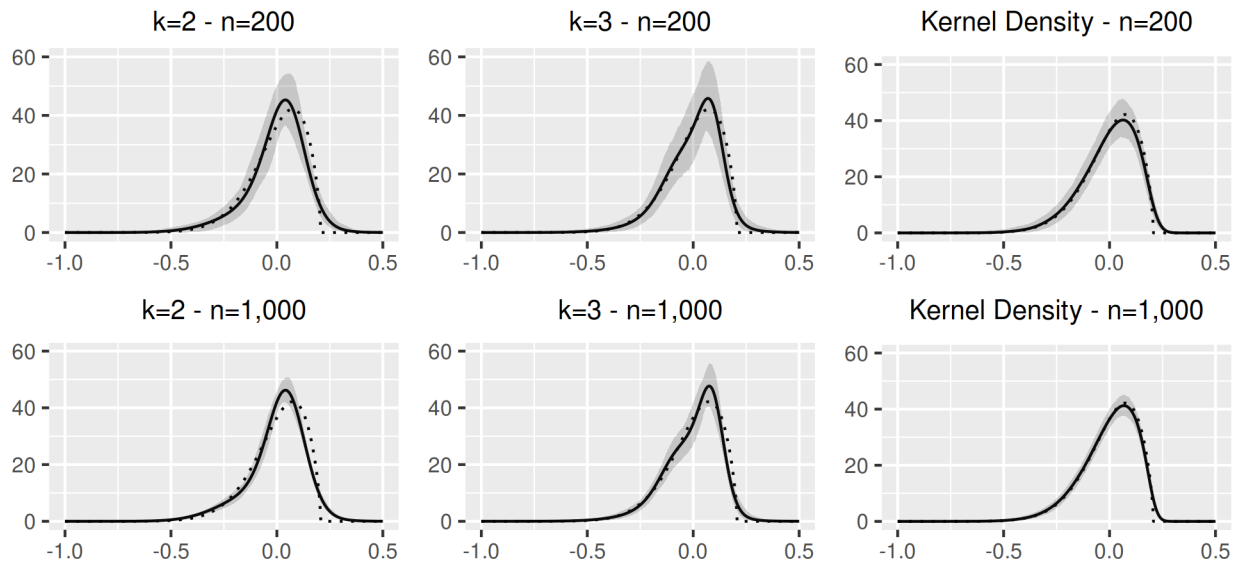
Autoregressive Dynamics

The second basic example considers AR(1) dynamics with an unknown shock distribution:

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

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

Figure 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.

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

$$\hat{\psi}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} e^{i\tau'(y_t, y_{t-1})}, \quad \hat{\psi}_n^s(\tau) = \frac{1}{n}\sum_{t=1}^{n} e^{i\tau'(y_t^s, y_{t-1}^s)}.$$

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 $\hat{f}_n$. Further increasing S has a much smaller effect on the variance of the estimates. Table 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.

Figure 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 3.2, the bias is more pronounced under AR(1) dynamics than in the static case. The variance is larger with

Table 1: Autoregressive Dynamics: Sieve-SMM vs. OLS Estimates

Parameter: ρ		Sieve-SMM $S=1$	Sieve-SMM $S=5$	Sieve-SMM $S=25$	OLS	True
$n=200$	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)	-

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.

4.2 Example 1: Stochastic Volatility

The stochastic volatility model of Example 1, illustrates the properties of the Sieve-SMM estimator for intractable, non-linear state-space models.

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

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

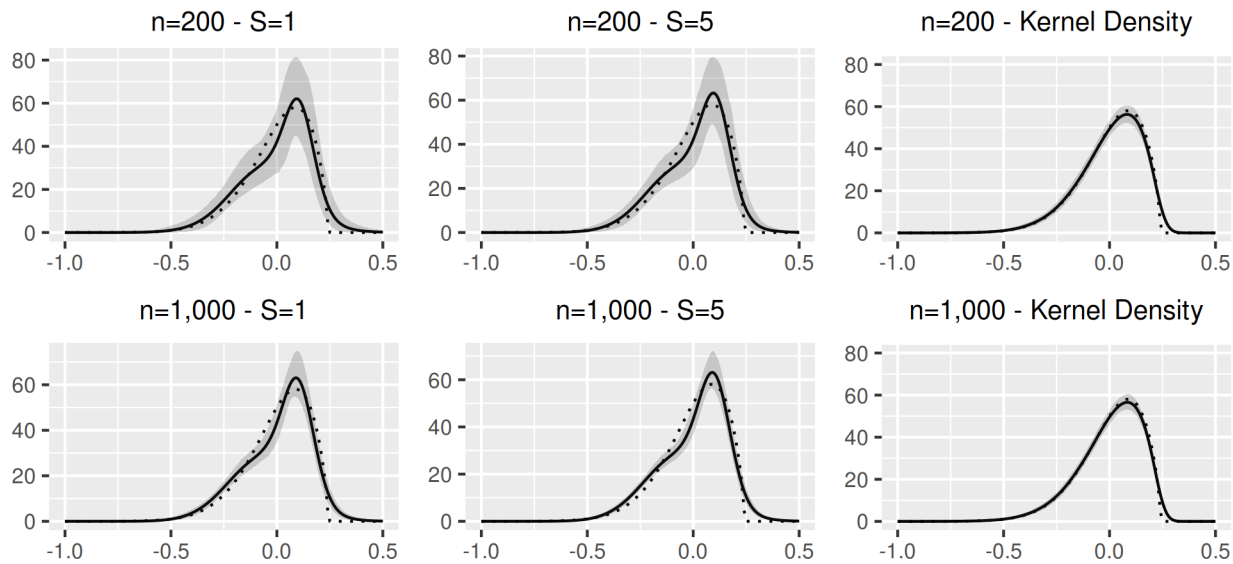
$$\hat{\psi}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} e^{\tau'(y_t, y_{t-1}, \hat{\sigma}_t^{aux}, \log(\hat{\sigma}_{t-1}^{aux}))}, \quad \hat{\psi}_n^s(\tau,\beta) = \frac{1}{n}\sum_{t=1}^{n} e^{\tau'(y_t^s, y_{t-1}^s, \hat{\sigma}_t^{s,aux}, \log(\hat{\sigma}_{t-1}^{s,aux}))}.$$

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.

[33] See Appendix B.1 for details.

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

Figure 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.*

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.[35] 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.

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 Appendix B which allows for auxiliary variables. Also not reported here, the bias and standard deviations of parametric estimates

[35]The Bayesian estimates are computed 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).

Table 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)

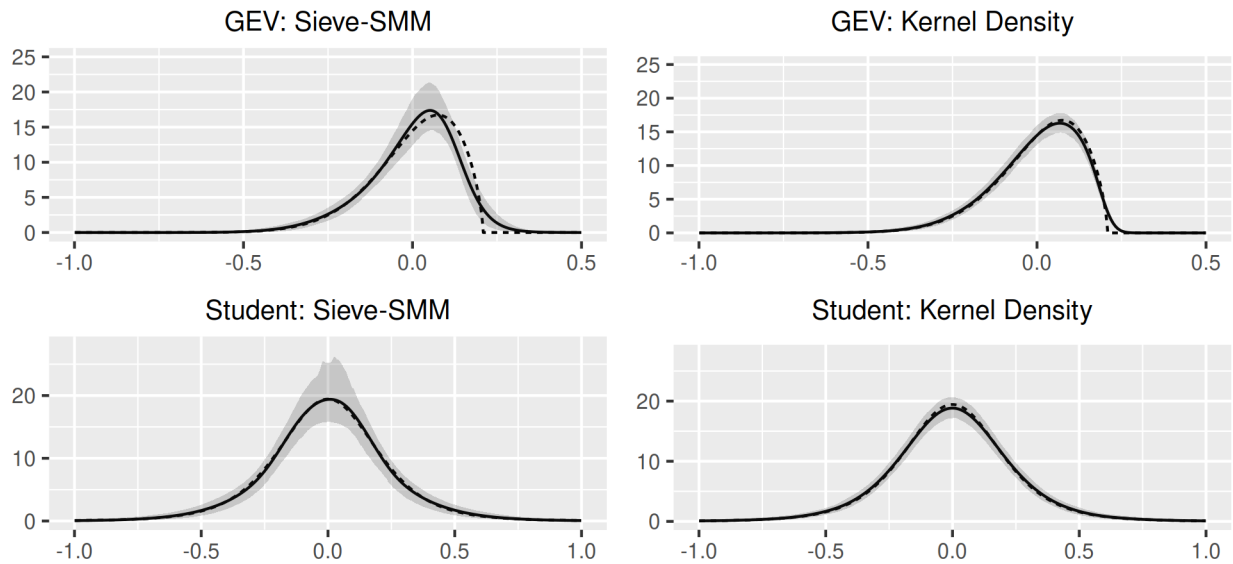
with f_0 are comparable to the GEV results.

Table 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 whereas the Sieve-SMM estimates are only slightly biased.

Figure 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 estimator, as one would expect. The density is also less precisely estimated than in the AR(1) case.

The Monte-Carlo simulations highlight the lack of robustness of the parametric Bayesian estimates to the tail behavior of the shocks. This is particularly relevant 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.

Figure 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.*

5 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. This has implications for the risk premium as discussed below. The second one uses daily GBP/USD exchange rate data to highlight the bias implications of fat tails on parametric SV volatility estimates.

5.1 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

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

package Quandl.[37] production is modeled using a SV model with AR(1) mean dynamics:

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

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. Allowing for a flexible distribution in $e_{t,1}$ in addition to the stochastic volatility process allows to capture shocks that are potentially large in magnitude relative to the volatility around the event. For instance, the volatility of output was generally lower between 2000-2007 though more sizeable, isolated,[38] shocks were still observed. The stochastic volatility literature has mainly focused on the distribution of the shocks to the mean $e_{t,1}$ rather than the volatility[39] hence the volatility shocks are modelled parametrically in this application. 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. Only the AR(1) component is estimated for simplicity given that the focus is of this example is on the shocks and the volatility rather than μ_c. The volatility shocks are also assumed to be chi-squared rather than Gaussian to ensure non-negativity.

5.1.1 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: $\hat{\psi}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n} e^{i\tau'(\Delta c_t, \Delta c_{t-1}, \Delta c_{t-2}, \log(\hat{\sigma}_t^{aux}), \log(\hat{\sigma}_{t-1}^{aux}))}$. Table 3 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.[40]

Figure 4 compares the densities estimated using the parametric and Sieve-SMM. The log-density reveals a larger left tail for the sieve estimates and potential asymmetry: conditional

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

[38] Isolated here is taken to mean not associated with a cluster of volatility as modelled by the stochastic volatility process. Indeed, the data generating process allows for one time shocks that are unconditionally moderate, or large, even in regimes of low volatility. This translates into two types of risk: one is persistent and modelled using the volatility process and the other is associated with one time tail behaviour and is modelled nonparametrically.

[39] 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.

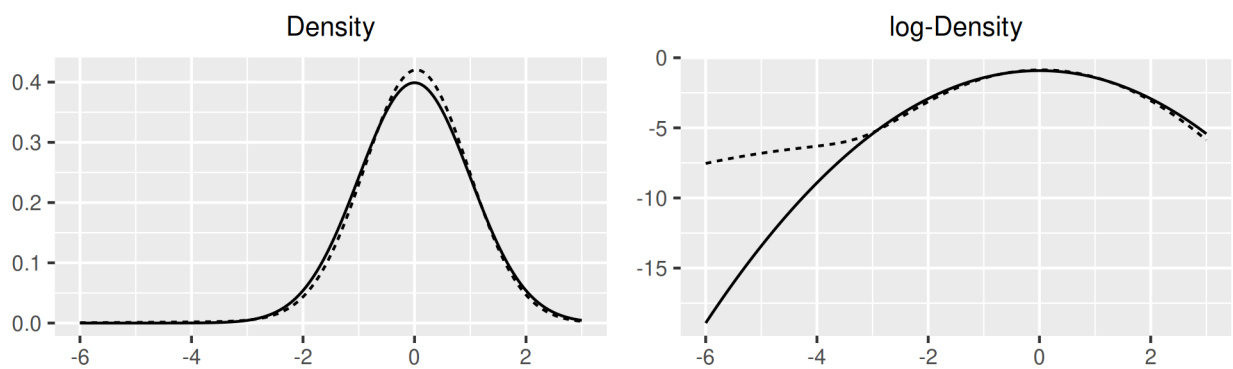
[40] HAC standard errors are computed using the R package *sandwich* (Zeileis, 2004).

Table 3: 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	$[0.20, 0.42]$	$[0.34, 0.55]$	$[0.35, 0.92]$	$[0.06, 0.29]$

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; at $e = -5$, the density ratio is greater than $20,000$. Note that $\hat{f}_n$ is a Gaussian mixture so that its tails remain Gaussian but with a decay driven by the components with the largest variance.

Figure 4: Industrial Production: Sieve-SMM Density Estimate vs. Normal Density



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

Table 4 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.

Altogether, these results suggest significant non-Gaussian features in the shocks with both negative skewness and excess kurtosis. The implications for the risk-free rate are now discussed; implication for welfare are given in Appendix D.

Table 4: Industrial Production: Moments of $\Delta c_t, \Delta c_t^s$ 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

5.1.2 Implications for the risk-free rate

As discussed in the introduction, the Euler equation implies that the risk-free rate r_t satisfies: $e^{-r_t} = e^{-a}\mathbb{E}_t\left((C_{t+1}/C_t)^{-\gamma}\right)$ 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 relies on very low frequency, high impact shocks and a power utility (Rietz, 1988; Barro, 2006b). In constrast, this empirical application considers a simple power utility with the higher frequency of shocks (monthly) over a recent period (post 1960).

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, this corresponds to the effect of uncertainty on the risk-free rate. 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 Gaus-

sian 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 5 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 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.

Table 5: Effect of uncertainty on the risk-free rate (% annualized)

Risk aversion γ	2	4	6	10
Gaussian AR(1)	-0.11	-0.47	-1.6	-2.94
	(0.01)	(0.04)	(0.10)	(0.27)
SMM	-0.09	-0.37	-0.84	-2.34
	(0.01)	(0.03)	(0.06)	(0.17)
Sieve-SMM	-0.25	-1.02	-2.32	-6.59
	(0.08)	(0.31)	(0.73)	(2.12)

Note: standard errors reported in parentheses.

Given that the risk free-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. As one would expect the standard errors are larger for the Sieve-SMM semi-nonparamatric estimates. This illustrates the bias/variance tradeoff from imposing a parametric distribution in the shocks: the parametric estimates are more precisely estimated but suggest much lower risk premia. Altogether, the results suggest that the choice of distribution f matters in computing both welfare effects and the risk-free rate.

5.2 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 documents asymmetry in the distribution of the shocks in Appendix D. 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.

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 process 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: $y_t = 100 \times [\log(GBP/USD_t) - \log(GBP/USD_{t-1})]$. 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. Both Sieve-SMM estimators have the same number of parameters.

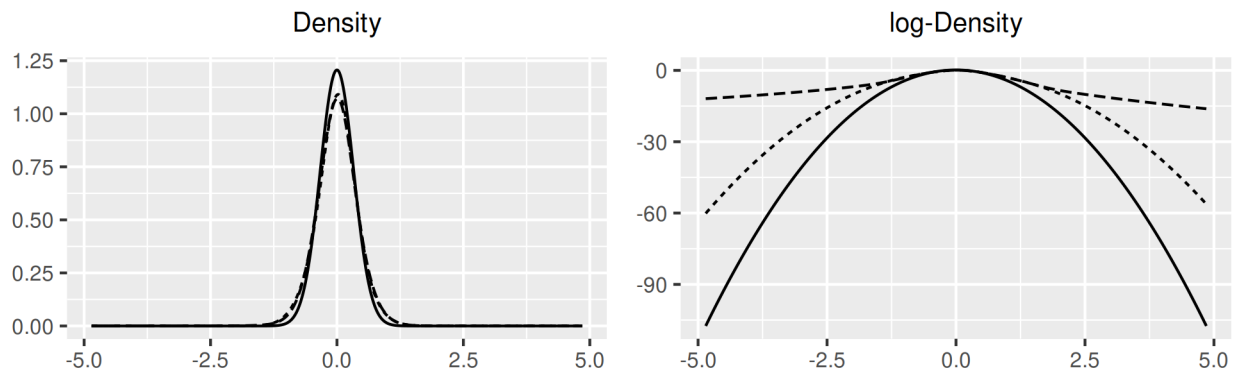
Table 6 shows the posterior mean and 95% credible interval for the Bayesian estimates as well as the point estimates and te 95% confidence interval for two Sieve-SMM estimators. The Bayesian estimate for the persistence of volatility ρ_z is much smaller than the SMM and Sieve-SMM estimates: it is outside 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: $(\hat{\rho}_n, \hat{\sigma}_z) = (0.96, 0.23)$ 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 $\hat{f}_n$ are also very similar when removing these observations. Figure 5

Table 6: Exchange Rate: Bayesian and Sieve-SMM Estimates

Parameter/Estimator	Bayesian	Sieve-SMM	Sieve-SMM tails
ρ_σ	0.24	0.96	0.92
	[0.16, 0.34]	[0.59,0.99]	[0.62,0.99]
κ_σ	1.31	0.22	0.19
	[1.21, 1.41]	[0.06,0.83]	[0.05,0.79]

Note: Bayesian credible intervals and confidence intervals reported below the point estimates.

Figure 5: Exchange Rate: Density and log-Density Estimates



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

compares the density $\hat{f}_n$ of $e_{t,1}$ for the Bayesian and Sieve-SMM estimates. The log-density $\log[\hat{f}_n]$ is also computed as it higlights the differences in the tails. The 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. The moments implied by the estimates of θ and f as well as a robustness check are given in Appendix D.

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, $\hat{f}_n$ 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 $\hat{f}_n$: large depreciations in the GBP relative to the USD are historically more likely than large appreciations.

6 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. The approach in the paper also extends to moments computed from transfomed observations, as shown in Appendix B; one can compute the joint CF of the data and filtered latent variables from a linearized model to estimate a more general non-linear state-space model for instance. Results for short-panels are also given in Appendix B under additional restrictions. Beyond the mixture sieve and the CF, the high-level conditions in Appendix F allow to approximate densities and other functions nonparametrically using more general sieve bases (e.g. splines, neural networks), as well as other bounded infinite dimensional moment functions such as the empirical CDF of the data.

Going forward, a number of extensions to this paper's results should be of interest. On the theoretical side, extending the results in this paper to a larger class of unbounded moments would allow to allow for more general dynamics in the Sieve-GMM settings. This would allow to derive asymptotic results for semi-nonparametric estimation of dynamic models with moments that involve filtering of unobserved latent variables. It would also allow for a more general Sieve-Indirect Inference estimation theory. The non-stationarity of the simulated data, for SMM and Indirect Inference, and of the filtered variables, in GMM, makes such results quite difficult to derive. Some more straightforward extensions of the main results are also possible: the mixture sieve can be extended to accommodate conditional heteroskedasticiy as in Norets (2010), the indepence of the marginals could also be relaxed usign the results of De Jonge & Van Zanten (2010). On the empirical side, estimating the distribution of the shocks in a fully-specified DSGE or Asset Pricing model could be of interest given the importance of non-gaussianity for welfare and the risk premium.

References

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., & Sørensen, 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. (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. & 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.

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.

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

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

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

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*.

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.

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., 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.

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 Jonge, R. & Van Zanten, J. H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Annals of Statistics*, 38(6), 3300–3320.

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.

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.

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.

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.

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). *CRAN R Package Vignette*, (pp. 1–36).

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., & Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8(S1), S85—-S118.

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. & Richard, S. F. (1987). The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models. *Econometrica*, 55(3), 587–613.

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.

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.

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).

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.

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). varepsilon-entropy and varepsilon-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2), 3–86.

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.

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.

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.

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.

Nelder, J. & Mead, R. (1965). A simplex method for function minimization. *The computer journal*.

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.

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.

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.

Peralta-Alva, A. & Santos, M. S. (2014). : (pp. 517–556).

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.

Rio, E. (2000). *Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants*, volume 31 of *Mathématiques et Applications*. Springer Berlin Heidelberg.

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.

Smith, A. (1993). Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8(S1), S63–S84.

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.

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.

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 A Background Material

A.1 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 are associated with 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 (y_t, x_t) has no finite moment.

ii. Boundedness: The CF is bounded $|\psi(\tau)| \leq 1$ for any distribution. As a result, the objective function $\hat{Q}_n^S$ 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:

$$y_t = \mathbb{1}_{x_t'\theta + e_t \geq 0}$$

is discrete but its conditional characteristic function is continuous in both f and θ:

$$\mathbb{E}\left(e^{i\tau_y y_t}|x_t\right) = 1 - F(x_t'\theta) + F(x_t'\theta)e^{i\tau_y},$$

where F is the CDF of $e_t \sim f$. As a result, the joint CF is also continuous:

$$\mathbb{E}\left(e^{i\tau(y_t,x_t)}\right) = \mathbb{E}\left(e^{i\tau_x x_t}[1 - F(x_t'\theta) + F(x_t'\theta)e^{i\tau_y}]\right).$$

In this example, the population CDF is, however, not continuous. As a result, a population objective Q based on the CF is continuous but the one based on a CDF is not.

A.2 Computing the Sample Objective Function $\hat{Q}_n^S$

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 $\tilde{\omega}$ and apply the transformation:

$$\omega_1 = \frac{1}{1 + \sum_{\ell=1}^{k-1} e^{\tilde{\omega}_\ell}}, \quad \omega_j = \frac{e^{\tilde{\omega}_{j-1}}}{1 + \sum_{\ell=1}^{k-1} e^{\tilde{\omega}_\ell}} \text{ for } j = 2, \ldots, k.$$

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 $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_k$ and compute:

$$\sigma_j = \underline{\sigma}_k + e^{\tilde{\sigma}_j}.$$

The resulting σ_j are greater or equal than the lower bound $\underline{\sigma}_k \geq 0$. To impose unit variance, restrict $\tilde{\sigma}_1 = 0$ 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 $\hat{Q}_n^S$. For an integration grid $\tau_1, \ldots, \tau_m$ with weights $\pi_1, \ldots, \pi_m$ compute the vectors:

$$\hat{\psi}_n = (\hat{\psi}_n(\tau_1), \ldots, \hat{\psi}_n(\tau_m))', \quad \hat{\psi}_n^S = (\hat{\psi}_n^S(\tau_1), \ldots, \hat{\psi}_n^S(\tau_m))'$$

and the objective:

$$\hat{Q}_n^S(\beta) = (\hat{\psi}_n - \hat{\psi}_n^S)' \text{diag}(\pi_1, \ldots, \pi_m)(\hat{\psi}_n - \hat{\psi}_n^S).$$

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 $(\hat{\psi}_n - \hat{\psi}_n^S)' B' \text{diag}(\pi_1, \ldots, \pi_m)(\hat{\psi}_n - \hat{\psi}_n^S)'$; a detailed overview on computing the objective function with a linear operator B, using quadrature, is given in the appendix of Carrasco & Kotchoni (2016).

A.3 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[41]:

$$1 - e^{-\frac{\mu_{k(n)}^2}{8\underline{\sigma}_{k(n)}}} \leq \|f_1 - f_2\|_{TV} \leq \sqrt{2}\left(1 - e^{-\frac{\mu_{k(n)}^2}{8\underline{\sigma}_{k(n)}}}\right)^{1/2}.$$

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 implies that

[41] 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: $H(f,g)^2 = 1 - \sqrt{\frac{2\sigma_f\sigma_g}{\sigma_f^2+\sigma_g^2}}e^{-\frac{1}{4}\frac{(\mu_f-\mu_g)^2}{(\sigma_f^2+\sigma_g^2)}}$.

$\|f_{1,k(n)} - f_{2,k(n)}\|_{weak} \to 0$ 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:

$$\begin{aligned}\|f_{1,k(n)} - f_{2,k(n)}\|_\infty &= \sup_{e\in\mathbb{R}} \frac{1}{\underline{\sigma}_{k(n)}} \left|\phi\left(\frac{e}{\underline{\sigma}_{k(n)}}\right) - \phi\left(\frac{e-\mu_{k(n)}}{\underline{\sigma}_{k(n)}}\right)\right| \\ &= \sup_{\tilde{e}\in\mathbb{R}} \frac{|\mu_{k(n)}|}{\underline{\sigma}_{k(n)}^2} \left|\phi'\left(\frac{\tilde{e}}{\underline{\sigma}_{k(n)}}\right)\right| = \frac{|\mu_{k(n)}|}{\underline{\sigma}_{k(n)}^2}\|\phi'\|_\infty\end{aligned}$$

For any $\varepsilon > 0$, pick $\mu_k = \pm\varepsilon\underline{\sigma}_{k(n)}^2/\|\phi'\|_\infty$ 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 $\underline{\sigma}_{k(n)} \to 0$:

$$\frac{\|f_{1,k(n)} - f_{2,k(n)}\|_\infty}{\|f_{1,k(n)} - f_{2,k(n)}\|_{weak}} \geq \underline{\sigma}_{k(n)}^{-2}\varepsilon\|\phi'\|_\infty$$

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 1, 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:

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

A.4 Identification in the Stochastic Volatility Model

This section provides an identification result for the SV model in the first empirical application:

$$\begin{aligned} y_t &= \mu_y + \rho_y y_{t-1} + \sigma_t e_{t,1}, \quad e_{t,1} \overset{iid}{\sim} f \\ \sigma_t^2 &= \mu_\sigma + \rho_\sigma \sigma_{t-1}^2 + \kappa_\sigma e_{t,2} \end{aligned}$$

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.[42] Suppose the CF $\hat{\psi}_n$ includes y_t and two lagged

[42] This assumption makes the derivations easier in terms of notation.

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 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:

$$\mathbb{E}(\hat{\psi}_n^s(\tau, \beta_1)) = \mathbb{E}(\hat{\psi}_n^s(\tau, \beta_2)), \quad \forall \tau \in \mathbb{R}^3. \tag{A.12}$$

Using the notation $\tau = (\tau_1, \tau_2, \tau_3)$ this implies that for any integers $\ell_1, \ell_2, \ell_3 \geq 0$:

$$\begin{aligned} i^{\ell_1+\ell_2+\ell_3}\mathbb{E}_{\beta_1}(y_t^{\ell_1} y_{t-1}^{\ell_2} y_{t-2}^{\ell_3}) &= \frac{d^{\ell_1+\ell_2+\ell_3}\mathbb{E}(\hat{\psi}_n^s(\tau, \beta_1))}{d\tau_1^{\ell_1} d\tau_2^{\ell_2} d\tau_3^{\ell_3}}\Big|_{\tau=0} \\ &= \frac{d^{\ell_1+\ell_2+\ell_3}\mathbb{E}(\hat{\psi}_n^s(\tau, \beta_2))}{d\tau_1^{\ell_1} d\tau_2^{\ell_2} d\tau_3^{\ell_3}}\Big|_{\tau=0} = i^{\ell_1+\ell_2+\ell_3}\mathbb{E}_{\beta_2}(y_t^{\ell_1} y_{t-1}^{\ell_2} y_{t-2}^{\ell_3}) \end{aligned}$$

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):[43] $\rho_{\sigma,1}\kappa_{\sigma,1}^2/(1-\rho_{\sigma,1}^2) = \rho_{\sigma,2}\kappa_{\sigma,2}^2/(1-\rho_{\sigma,2}^2)$ 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 (A.12) implies $\theta_1 = \theta_2$, the parametric component is identified. Since θ is identified, all the moments of σ_t are known. After recentering, this implies that for all $\ell_1 \geq 3$ if $\mathbb{E}_\theta(\sigma_t^{\ell_1}) \neq 0$:

$$\mathbb{E}_{f_1}(e_{t,1}^{\ell_1}) = \mathbb{E}_{f_1}(e_{t,2}^{\ell_1}). \tag{A.13}$$

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 (A.13) holds. Since the moment generating function is analytic and the first two moments are fixed, (A.13) 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.

A.5 Additional Results on Asymptotic Normality

The following provides two additional results on the root-n asymptotic normality of $\hat{\theta}_n$. A positive result is given in Proposition A1 and a negative result is given in Remark A3. The

[43] 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.

results apply to DGPs of the form:[44]

$$y_t = g_{obs}(y_{t-1}, \theta, u_t)$$
$$u_t = g_{latent}(u_{t-1}, \theta, e_t)$$

where g_{obs}, g_{latent} are smooth in θ. In this class of models, the data depends on f only through e_t. Examples 1 and 2, in the main text and Appendix B respectively, satisfy this restriction but dynamic programming models typically don't. The smoothness restriction holds in Example 1 but not Example 2.

Proposition A1 (Sufficient Conditions for Asymptotic Normality of $\hat{\theta}_n$)**.** *If $\mathbb{E}_{\theta_0,f}(\mathbf{y}_t^s)$ and $\mathbb{V}_{\theta_0,f}(\mathbf{y}_t^s)$ do not depend on f then $\hat{\theta}_n$ is root-n asymptotically normal if:*

$$\mathbb{E}_{\theta_0,f_0}\left(\frac{d\mathbf{y}_t^s}{d\theta'}\left[\begin{pmatrix} 1 & \mathbf{y}_t^{s\prime} \end{pmatrix} \otimes I_{d_{\mathbf{y}}}\right]\right)$$

has rank greater or equal than d_θ when $t \to \infty$.

Proposition A1 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 θ 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:

$$\frac{dy_t^s}{d\rho_y} = \sum_{j=1}^{+\infty} j\rho_y^{j-1}\sigma_{t-j}e_{t-j,1}$$
$$\frac{dy_t^s}{d\rho_\sigma} = 0.5\sum_{j=0}^{\infty} \rho_y^j \frac{d\sigma_{t-j}^2}{d\rho_\sigma} e_{t-j,1}/\sigma_{t-j} \quad \textit{where} \quad \frac{d\sigma_{t-j}^2}{d\rho_\sigma} = \sum_{\ell=1}^{t-j} \ell\rho_\sigma^{\ell-1}(\mu_\sigma + \kappa_\sigma e_{\ell,2}).$$

[44]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, $\mathbb{E}\left(\frac{d\mathbf{y}_t^s}{d\theta'}\mathbf{y}_t^s\right)$ *must have rank greater than 2 for Proposition A1 to apply. Now, compute a first set of expectations:*

$$\mathbb{E}(\frac{dy_t^s}{d\rho_y}y_t^s) = \sum_{j=1}^{t} j\rho_y^{2j-1}\mathbb{E}(\sigma_{t-j}^2)$$

$$\mathbb{E}(\frac{dy_t^s}{d\rho_y}y_{t-1}^s) = \sum_{j=0}^{t-1}(j+1)\rho_y^{2j}\mathbb{E}(\sigma_{t-j-1}^2)$$

$$\mathbb{E}(\frac{dy_t^s}{d\rho_y}y_{t-2}^s) = \sum_{j=0}^{t-2}(j+2)\rho_y^{2j+1}\mathbb{E}(\sigma_{t-j-2}^2)$$

$$\mathbb{E}(\frac{dy_{t-1}^s}{d\rho_y}y_t^s) = \sum_{j=1}^{t-1} j\rho_y^{2j}\mathbb{E}(\sigma_{t-j-1})$$

$$\mathbb{E}(\frac{dy_{t-2}^s}{d\rho_y}y_t^s) = \sum_{j=1}^{t-2} j\rho_y^{2j+1}\mathbb{E}(\sigma_{t-j-2}).$$

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:*

$$\mathbb{E}(\frac{dy_t^s}{d\rho_\sigma}y_t^s) = \sum_{j=0}^{t}\rho_y^j\mathbb{E}(\frac{d\sigma_{t-j}^2}{d\rho_\sigma}) = \sum_{j=0}^{t}\rho_y^j\sum_{\ell=1}^{t-j}\ell\rho_\sigma^{2\ell-1}\mu_\sigma$$

$$\mathbb{E}(\frac{dy_t^s}{d\rho_\sigma}y_{t-1}^s) = \sum_{j=1}^{t}\rho_y^{j+1}\mathbb{E}(\frac{d\sigma_{t-j}^2}{d\rho_\sigma}) = \sum_{j=1}^{t}\rho_y^{j+1}\sum_{\ell=1}^{t-j}\ell\rho_\sigma^{2\ell-1}\mu_\sigma$$

$$\mathbb{E}(\frac{dy_t^s}{d\rho_\sigma}y_{t-2}^s) = \sum_{j=2}^{t}\rho_y^{j+2}\mathbb{E}(\frac{d\sigma_{t-j}^2}{d\rho_\sigma}) = \sum_{j=1}^{t}\rho_y^{j+1}\sum_{\ell=1}^{t-j}\ell\rho_\sigma^{2\ell-1}\mu_\sigma.$$

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 A3 ($\hat{\theta}_n$ is generally not an adaptive estimator of θ_0)**.** *For the estimator* $\hat{\theta}_n$ *to be adaptive*[45] *an orthogonality condition is required, namely:*

$$\frac{d^2Q(\beta_0)}{d\theta df}[f - f_0] = 0, \text{ for all } f \in \mathcal{F}_{osn}.$$

[45]If the estimator is adaptive then $\hat{\theta}_n$ is root-n asymptotically normal and its asymptotic variance does not depend on $\hat{f}_n$, 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:

$$\lim_{n\to\infty}\int Real\left(\frac{d\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{d\theta}\overline{\frac{d\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{df}[f-f_0]}\pi(\tau)d\tau\right)=0.$$

Given the restrictions on the DGP and using the notation in the proof of Proposition A1, it implies:

$$\lim_{t\to\infty}\int Real\left(i\tau'\frac{dg_t(\theta_0,e_1)}{d\theta}e^{i\tau'[g_t(\theta_0,e_1)-g_t(\theta_0,e_2)]}f_0(e_1)\Delta f(e_2)\pi(\tau)d\tau de_1de_2\right)=0.$$

After some simplification, the orthogonality condition can be re-written as:

$$\lim_{t\to\infty}\int \tau'\frac{dg_t(\theta_0,e_1)}{d\theta}\sin\left(\tau'[g_t(\theta_0,e_1)-g_t(\theta,e_2)]\right)f_0(e_1)\Delta f(e_2)\pi(\tau)d\tau de_1de_2=0.$$

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 $\frac{dg_t(\theta_0,e_1)}{d\theta}\sin\left(\tau'[g_t(\theta_0,e_1)-g_t(\theta,e_2)]\right)f_0(e_1)\Delta f(e_2)$ over e_1 and e_2 is zero. This is the case if $g_t(\theta_0,e_1)$ is separable in e_1 and f_0,f are symmetric densities which is quite restrictive.

Proof of Proposition A1. Chen & Pouzo (2015), pages 1031-1033 and their Remark A.1, implies that $\hat{\theta}_n$ is root-n asymptotically normal if:

$$\lim_{n\to\infty}\inf_{v\in\overline{V},v_\theta\neq 0}\frac{1}{\|v_\theta\|_1^2}\int\left|\frac{d\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{d\theta}v_\theta+\frac{d\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{df}[v_f]\right|^2\pi(\tau)d\tau>0.$$

By definition of $\overline{V}$ the vector $v=(v_\theta,v_f)$ has the form $v_\theta\in\mathbb{R}^{d_\theta}$ 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 $(\theta_j,f_j)\in\mathcal{B}_{osn}$ for some θ_j. To prove the result, we can proceed by contradiction suppose that for some non-zero v_θ and a v_f:

$$\int\left|\frac{d\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{d\theta}v_\theta+\frac{d\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{df}[v_f]\right|^2\pi(\tau)d\tau=0. \tag{A.14}$$

This implies that $\frac{d\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{d\theta}v_\theta+\frac{d\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{df}[v_f]=0$ for all τ (π almost surely). This implies that the following holds:

$$\frac{d\mathbb{E}(\hat{\psi}_n^s(0,\beta_0))}{d\theta}v_\theta+\frac{d\mathbb{E}(\hat{\psi}_n^s(0,\beta_0))}{df}[v_f]=0 \tag{A.15}$$

$$\frac{d^2\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{d\theta d\tau}\Big|_{\tau=0}v_\theta+\frac{d^2\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{df d\tau}[v_f]\Big|_{\tau=0}=0 \tag{A.16}$$

$$\frac{d^3\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{d\theta d\tau d\tau_\ell}\Big|_{\tau=0}v_\theta+\frac{d^3\mathbb{E}(\hat{\psi}_n^s(\tau,\beta_0))}{df d\tau d\tau_\ell}\Big|_{\tau=0}[v_f]=0 \tag{A.17}$$

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 $\mathbf{y}_t^s = g_t(\theta, e_t^s, \ldots, e_1^s)$ (the dependence on x is removed for simplicity). The first order derivatives can be written as:

$$\frac{d\mathbb{E}(\hat{\psi}_t^s(\tau, \beta_0))}{d\theta} = \int i\tau' \frac{dg_t(\theta_0, e)}{d\theta} e^{i\tau' g_t(\theta_0, e)} f_0(e) de$$
$$\frac{d\mathbb{E}(\hat{\psi}_t^s(\tau, \beta_0))}{df}[v_f] = \sum_{j=0}^{\infty} a_j \int e^{i\tau' g_t(\theta_0, e)} \Delta f_j(e) de$$

For $\tau = 0$ this yields $\frac{d\mathbb{E}(\hat{\psi}_t^s(0, \beta_0))}{d\theta} = 0$ and $\frac{d\mathbb{E}(\hat{\psi}_t^s(0, \beta_0))}{df}[v_f] = 0$, so equality (A.15) holds automatically. Taking derivatives and setting $\tau = 0$ again implies:

$$\frac{d^2\mathbb{E}(\hat{\psi}_t^s(\tau, \beta_0))}{d\theta d\tau}\Big|_{\tau=0} = i \int \frac{dg_t(\theta_0, e)}{d\theta'} f_0(e) de$$
$$\frac{d^2\mathbb{E}(\hat{\psi}_t^s(\tau, \beta_0))}{df d\tau}[v_f]\Big|_{\tau=0} = i \sum_{j=0}^{\infty} a_j \int g_t(\theta_0, e) \Delta f_j(e) de$$

If $\mathbb{E}(\mathbf{y}_t^s)$ does not depend on f then $\int g_t(\theta_0, e)\Delta f_j(e) de = 0$ for all j and $\frac{d^2\mathbb{E}(\hat{\psi}_t^s(\tau, \beta_0))}{df d\tau}[v_f]\Big|_{\tau=0} = 0$ holds automatically. This implies that condition (A.16) becomes:

$$\mathbb{E}\left(\frac{d\mathbf{y}_t^s}{d\theta}\right) v_\theta = 0. \tag{A.18}$$

If $\mathbb{E}\left(\frac{d\mathbf{y}_t^s}{d\theta}\right)$ has rank greater or equal than d_θ then condition (A.18) holds only if $v_\theta = 0$; this is a contradiction. If the rank is less than d_θ, then taking derivatives with respect to τ again yields $\frac{d^3\mathbb{E}(\hat{\psi}_n^s(0, \beta_0))}{df d\tau d\tau'}\Big|_{\tau=0}[v_f] = -\sum_{j=0}^{\infty} a_j \int g_t(\theta, e) g_t(\theta, e)' \Delta f_j(e) de = 0$ assuming $\mathbb{E}(\mathbf{y}_t^s \mathbf{y}_t^{s\prime})$ does not depend on f. Computing the other derivatives implies that condition (A.17) becomes $-v_\theta' \int \frac{dg(\theta_0)}{d\theta'} g(\theta_0, e) f_0(e) de$ i.e.:

$$v_\theta' \mathbb{E}\left(\frac{d\mathbf{y}_t^s}{d\theta'} \mathbf{y}_{t,\ell}^s\right) = 0 \text{ for all } \ell = 1, \ldots, d_{\mathbf{y}}. \tag{A.19}$$

Then, stacking conditions (A.18)-(A.19) together implies:

$$v_\theta' \mathbb{E}\left(\frac{d\mathbf{y}_t^s}{d\theta'} \left[\begin{pmatrix} 1 & \mathbf{y}_t^{s\prime} \end{pmatrix} \otimes I_{d_{\mathbf{y}}}\right]\right) = 0. \tag{A.20}$$

If the matrix has rank greater or equal to d_θ then it implies $v_\theta = 0$ which is a contradiction. Hence (A.14) holds only if $v_\theta = 0$ which proves the result. □

Appendix B 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.

B.1 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 $\hat{\psi}_n$. 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 $(y_t, \ldots, y_1)$ provide additional information about the unobserved u_t. It is used to estimate stochastic volatility models in Sections 4 and 5. 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 $\hat{\eta}_n^{aux}$ (the estimated GARCH parameters). The estimates $\hat{\eta}_n^{aux}$ are computed from the full sample $(y_1, \ldots, y_n, x_1, \ldots, x_n)$ and the auxiliary variables $z_t^{aux}, z_t^{s,aux}$ are computed using the full and simulated samples:[46]

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

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

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

The following assumption provides sufficient conditions on the estimates $\hat{\eta}_n^{aux}$ and the filtering process $g_{t,aux}$ for the asymptotic properties in Section 3 to also hold with auxiliary variables.

Assumption B6 (Auxiliary Variables)**.** *The estimates $\hat{\eta}_n^{aux}$ are such that:*

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

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

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

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

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

$$\|g_{t,aux}(y_t,\ldots,y_1,x_t,\ldots,x_1,\eta_1^{aux}) - g_{t,aux}(y_t,\ldots,y_1,x_t,\ldots,x_1,\eta_2^{aux})\|$$
$$\leq C^{aux}(y_t,\ldots,y_1,x_t,\ldots,x_1) \times \|\eta_1^{aux} - \eta_2^{aux}\|$$

with $\mathbb{E}(C^{aux}(y_t,\ldots,y_1,x_t,\ldots,x_1)^2) \leq \bar{C}^{aux} < \infty$ and $\mathbb{E}(C^{aux}(y_t^s,\ldots,y_1^s,x_t,\ldots,x_1)^2) \leq \bar{C}^{aux} < \infty$. The average of the Lipschitz constants

$$C_n^{aux} = \frac{1}{n}\sum_{t=1}^{n} C^{aux}(y_t,\ldots,y_1,x_t,\ldots,x_1)$$

is uniformly stochastically bounded, i.e. $O_p(1)$, for both sample and simulated data.

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

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

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:

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

The summability condition *iv.* is key in preserving the Hölder continuity and bias accumulation results of Section 3 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 $\hat{\eta}_n^{aux}$ 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 $\phi(\hat{\beta}_n) - \phi(\beta_0)$. It could be shown by assuming an expansion of the form $\hat{\eta}_n^{aux} = \frac{1}{n}\sum_{t=1}^{n}\eta^{aux}(y_t,x_t) + o_p(1/\sqrt{n})$ and expanding $\hat{\psi}_n$, $\hat{\psi}_n^s$ 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:[47]

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

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 $\eta^{aux}, \tilde{\eta}^{aux}$, by recursion:

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

$\overline{\eta}^{aux}$ 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:

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

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 $y_t, \tilde{y}_t$ and a set of auxiliary parameters η^{aux}:

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

By a recursive argument, the inequality above becomes:

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

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.

[47] 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 following corollary shows that the results of Section 3 also hold when addition auxiliary variables to the CF.

Corollary B2 (Asymptotic Properties using Auxiliary Variables)**.** *Suppose the assumptions for Theorems 1, 2 and 3 hold as well as Assumption B6, then the results of Theorems 1, 2 and 3 hold with auxiliary variables. The rate of convergence is unchanged.*

The proof of Corollary B2 is very similar to the proofs of the main results. Rather than repeating the full proofs, Appendix E.5 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 $\hat{\psi}_n(\cdot, \hat{\eta}_n^{aux}) - \hat{\psi}_n^S(\cdot, \beta_0, \hat{\eta}_n^{aux})$ in the sandwich formula for the standard errors. The unknown β_0 is replaced by $\hat{\beta}_n$ in practice.

B.2 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 illustration of that setting.

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

$$y_{j,t} = (x'_{j,t}\theta_1 + u_{j,t})\mathbb{1}_{x'_{j,t}\theta_1 + u_{j,t} \geq 0}$$
$$u_{j,t} = \rho u_{j,t-1} + e_{j,t}$$

where $|\rho| < 1$, $e_{j,t} \overset{iid}{\sim} f$, $\mathbb{E}(e_{j,t}) = 0$. The parameters to be estimated are $\theta = (\theta_1, \rho)$ 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 θ 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.[48] They find that non-Gaussian income shocks help explain transitions between low and higher earnings 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 B7 (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:*

$$\begin{aligned} y_{j,t} &= g_{obs}(x_{j,t}, \beta, u_{j,t}) \\ u_{j,t} &= g_{latent}(u_{j,t-1}, \beta, e_{j,t}) \end{aligned} \tag{B.21}$$

or

$$y_{j,t} = g_{obs}(y_{j,t-1}, x_{j,t}, \beta, e_{j,t}) \tag{B.22}$$

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 B7 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 (B.21), 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

[48] 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.

cross-sectional dimension: $m/n \to c > 0$ as $n \to \infty$. For the DGP in equation (B.22), 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 $\hat{\psi}_n, \hat{\psi}^s_n$ 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 B7 (Impact of the Initial Condition)**.** *Suppose that Assumption B7 holds. If the DGP is given by (B.21) 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 $\mathbf{u}^s_{j,t}$ is geometrically ergodic in t and the integrals*

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

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

$$Q_n(\beta_0) = \int \left| \mathbb{E}\left(\hat{\psi}_n(\tau) - \hat{\psi}^S_n(\tau, \beta_0) \right) \right|^2 \pi(\tau) d\tau = O\left(\bar{\rho}^m_u\right).$$

The effect of the initial condition is exponentially decreasing in m for DGP (B.21). If the DGP is given by (B.22) and the data is simulated with $y^s_{j,1} = y_{j,1}$ fixed then there is no initial condition effect:

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

Simulating a long history $u^s_{j,T}, \ldots, u^s_{j,-m}$ implies that the impact of the initial condition $u^s_{j,m} = u_{-m}$ on the full simulated sample $y^s_{j,1}, \ldots, y^s_{j,T}$ 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 B3 (Asymptotic Properties for Short Panels)**.** *Suppose that Assumption B7 and Lemma B7 hold. For the DGP (B.21) in Assumption B7, 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:*

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

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

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

Remark B4. *For the DGP (B.22), the simulated history is finite and fixed so that the approximation bias is not inflated by the dynamics:*

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

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 $\lim_{n\to\infty}\log[n]/m < -\log[\bar{\rho}_u]$. 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 $\|\theta_1\|$ 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 $|\rho| \leq \bar{\rho} < 1$ and $\mathbb{E}(e_t^2)$ is bounded above. The last condition is a restriction on the density f. Since $|\rho| \leq \bar{\rho} < 1$, 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).*

Appendix C Additional Monte-Carlo Results

C.1 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:

$$\begin{aligned} y_{j,t} &= (\theta_1 + x_{j,t}'\theta_2 + u_{j,t})\mathbb{1}_{\theta_1 + x_{j,t}'\theta_2 + u_{j,t}\geq 0} \\ u_{j,t} &= \rho u_{j,t-1} + e_{j,t} \end{aligned}$$

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 B, 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: $(\rho, \theta_1, \theta_2) = (0.8, -1.25, 1)$ 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:

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

Table C7: Dynamic Tobit: SMM vs. Sieve-SMM Estimates

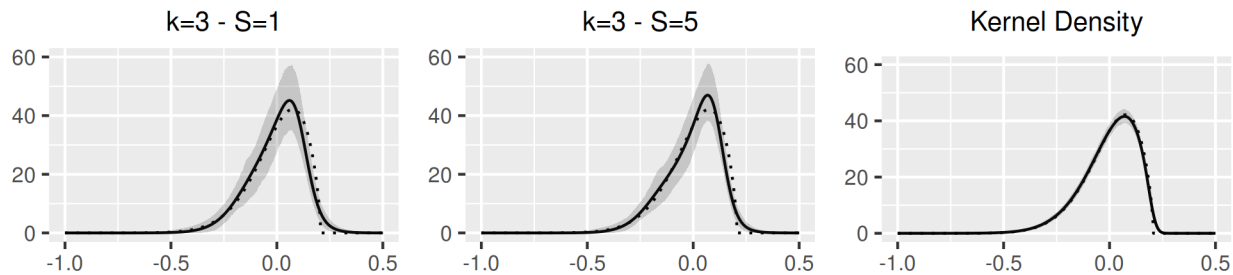
Parameter		$S = 1$		$S = 5$		
		SMM	Sieve-SMM	SMM	Sieve-SMM	True
ρ	Mean	0.796	0.801	0.796	0.796	0.80
	Std. Deviation	(0.042)	(0.039)	(0.031)	(0.031)	-
θ_1	Mean	-1.259	-1.230	-1.250	-1.233	-1.25
	Std. Deviation	(0.234)	(0.200)	(0.178)	(0.169)	-
θ_2	Mean	1.002	1.002	1.000	0.997	1.00
	Std. Deviation	(0.059)	(0.052)	(0.045)	(0.043)	-

Table C7 compares the parametric and Sieve-SMM estimates. The numbers are comparable except for θ_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 θ_1, the average estimate is higher than -1.1. The other estimates were found to have negligible bias.[49]

Figure C6 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 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.

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

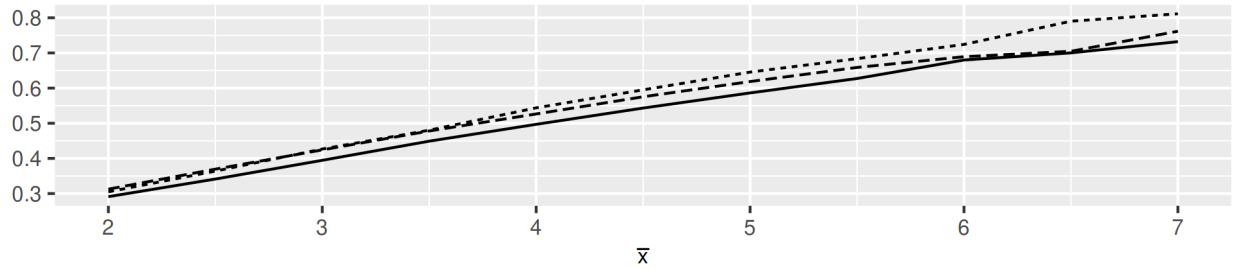
Figure C6: Dynamic Tobit: Sieve-SMM vs. Kernel Density Estimates



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

Figure C7 illustrates the differences between the parametric and Sieve-SMM for a counterfactual that involves the full density f. It shows the estimates of the probability of re-entering the market $\mathbb{P}(y_{j,5} > 0 | y_{j,4} = 0, x_5 = \cdots = x_1 = \bar{x})$ using the true value (θ_0, f_0), the SMM estimates $\hat{\theta}_n^{SMM}$ with Gaussian shocks and the Sieve-SMM estimates $(\hat{\theta}_n, \hat{f}_n)$. 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 $\bar{x}$ is large, as shown in figure C7. For this particular counterfactual, the Sieve-SMM estimates are much closer to the true value for larger values of $\bar{x}$.

Figure C7: Dynamic Tobit: SMM vs. Sieve-SMM Estimates of the Counterfactual



Note: Estimated counterfactual: $\mathbb{P}(y_{j,5} > 0 | y_{j,4} = 0, x_5 = \cdots = x_1 = \bar{x})$ *- 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 B.

Appendix D Additional Empirical Results

D.0.1 First Empirical Application: Welfare Implications

This section discusses the welfare implication of the estimates computed in Section 5. The approach considered here is based on the simple calculation approach of Lucas (1991, 2003).[50] 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,[51] the welfare cost of business cycle fluctuations is now computed under Gaussian and mixture SV dynamics. Table D8 compares the two welfare costs for different levels of risk aversion with the baseline iid Gaussian case of Lucas.[52] 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%.[53] 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 cost that is nearly \$200,

[50] 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).

[51] Expectations are taken over 1,000 Monte-Carlo samples and an horizon of 5,000 months (420 years).

[52] 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.

[53] Additional results for AR(1) processes and linearized DSGE models are given in Reis (2009).

Table D8: 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

or 25%, higher than the parametric SMM welfare estimates. This difference is quite large highlighting the non-negligible role of asymmetric shocks on welfare.

D.1 Second Empirical Application: Additional Content

Table D9 compares the first four moments in the data to those implied by the estimates.[54] 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 United Kingdom European Union membership referendum on June 23rd 2016 which is the largest variation in the sample,[55] 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. The Gaussian and tails mixture is thus the preferred specifications for this dataset.

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

[55] 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.

Table D9: 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.*

Appendix E 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:

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

Since B is bounded linear there exists a $M_B > 0$ such that for any two CFs:

$$\int \left| B\hat{\psi}_n(\tau) - B\hat{\psi}_n^S(\tau,\beta) \right|^2 \pi(\tau) d\tau \leq M_B^2 \int \left| \hat{\psi}_n(\tau) - \hat{\psi}_n^S(\tau,\beta) \right|^2 \pi(\tau) d\tau.$$

As a result, the rate of convergence for the objective function with the weighting B is the same as the rate of convergence without the operator B.[56]

[56]For results on estimating the optimal B see Carrasco & Florens (2000); Carrasco et al. (2007a). Using their method would lead to $M_{\hat{B}} \to \infty$ as $n \to \infty$ resulting in a slower rate of convergence for $\hat{\beta}_n$. On would have to study $\hat{B}_n \hat{\psi}_n^S$ directly, which may not be uniformly bounded and, as a result, does not fall in the class of moments considered in this paper. This is left to future research.

E.1 Properties of the Mixture Sieve

Lemma E8 (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 $\underline{e} < \overline{e}$ such that f_S is weakly decreasing on $(-\infty, \underline{e}]$ and weakly increasing on $[\overline{e}, \infty)$.

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: $\mu_j \in [-\bar{\mu}_k, \bar{\mu}_k]$.

vi. Growth Rate of Bounds: $\bar{\mu}_k = O\left(\log[k]^{1/b}\right)$.

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 $\|\cdot\|_\infty$.

Proof of Lemma 2. :
The difference between e_t^s and $\tilde{e}_t^s$ can be split into two terms:

$$\sum_{j=1}^{k(n)} \left(\mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \omega_l, \sum_{l=0}^{j} \omega_l]} - \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \tilde{\omega}_l, \sum_{l=0}^{j} \tilde{\omega}_l]}\right) \left(\mu_j + \sigma_j Z_{t,j}^s\right) \tag{E.23}$$

$$\sum_{j=1}^{k(n)} \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \tilde{\omega}_l, \sum_{l=0}^{j} \tilde{\omega}_l]} \left(\mu_j - \tilde{\mu}_j + [\sigma_j - \tilde{\sigma}_j] Z_{t,j}^s\right). \tag{E.24}$$

To bound the term (E.23) in expectation, combine the fact that $|\mu_j| \leq \bar{\mu}_{k(n)}, |\sigma_j| \leq \bar{\sigma}$ and ν_t^s and $Z_{t,j}^s$ are independent so that:

$$\left[\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2\leq\delta}\left|\sum_{j=1}^{k(n)}\left(\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\omega_l,\sum_{l=0}^{j}\omega_l]}-\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\tilde{\omega}_l,\sum_{l=0}^{j}\tilde{\omega}_l]}\right)\left(\mu_j+\sigma_j Z_{t,j}^s\right)\right|^2\right)\right]^{1/2}$$
$$\leq \sum_{j=1}^{k(n)}\left[\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2\leq\delta}\left|\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\omega_l,\sum_{l=0}^{j}\omega_l]}-\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\tilde{\omega}_l,\sum_{l=0}^{j}\tilde{\omega}_l]}\right|^2\right)\right]^{1/2}$$
$$\times\left(\bar{\mu}_{k(n)}+\bar{\sigma}\mathbb{E}\left(|Z_{t,j}^s|^2\right)^{1/2}\right).$$

The last term is bounded above by $\bar{\mu}+\bar{\sigma}C_Z$. Next, note that $\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\omega_l,\sum_{l=0}^{j}\omega_l]}-\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\tilde{\omega}_l,\sum_{l=0}^{j}\tilde{\omega}_l]}\in\{0,1\}$ so that:

$$\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2\leq\delta}\left|\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\omega_l,\sum_{l=0}^{j}\omega_l]}-\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\tilde{\omega}_l,\sum_{l=0}^{j}\tilde{\omega}_l]}\right|^2\right)$$
$$=\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2\leq\delta}\left|\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\omega_l,\sum_{l=0}^{j}\omega_l]}-\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\tilde{\omega}_l,\sum_{l=0}^{j}\tilde{\omega}_l]}\right|\right).$$

Also, for any j: $|\sum_{l=0}^{j}\tilde{\omega}_l-\sum_{l=0}^{j}\omega_l|\leq\sum_{l=0}^{j}|\tilde{\omega}_l-\sum_{l=0}^{j}\omega_l|\leq\left(\sum_{l=0}^{j}|\tilde{\omega}_l-\omega_l|^2\right)^{1/2}\leq\|\tilde{\omega}-\omega\|_2\leq\delta$. Following a similar approach to Chen et al. (2003):

$$\left[\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2\leq\delta}\left|\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\omega_l,\sum_{l=0}^{j}\omega_l]}-\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\tilde{\omega}_l,\sum_{l=0}^{j}\tilde{\omega}_l]}\right|\right)\right]^{1/2}$$
$$\leq\left[\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2\leq\delta}\left|\mathbb{1}_{\nu_t^s\in[(\sum_{l=0}^{j-1}\tilde{\omega}_l)-\delta,(\sum_{l=0}^{j}\tilde{\omega}_l)+\delta]}-\mathbb{1}_{\nu_t^s\in[\sum_{l=0}^{j-1}\tilde{\omega}_l,\sum_{l=0}^{j}\tilde{\omega}_l]}\right|\right)\right]^{1/2}$$
$$=\left[\left([(\sum_{l=0}^{j}\tilde{\omega}_l)+\delta]-[(\sum_{l=0}^{j-1}\tilde{\omega}_l)-\delta]-[(\sum_{l=0}^{j}\tilde{\omega}_l)-(\sum_{l=0}^{j-1}\tilde{\omega}_l)]\right)\right]^{1/2}=\sqrt{2\delta}.$$

Overall the term (E.23) is bounded above by $\sqrt{2}(1+C_Z)\left(\bar{\mu}_{k(n)}+\bar{\sigma}+k(n)\right)\sqrt{\delta}$. The term

(E.24) can be bounded above by using the simple fact that $0 \leq \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \tilde{\omega}_l, \sum_{l=0}^{j} \tilde{\omega}_l]} \leq 1$ and:

$$\begin{aligned}
&\left[\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2 \leq \delta} \left| \sum_{j=1}^{k(n)} \mathbb{1}_{\nu_t^s \in [\sum_{l=0}^{j-1} \tilde{\omega}_l, \sum_{l=0}^{j} \tilde{\omega}_l]} \left(\mu_j - \tilde{\mu}_j + [\sigma_j - \tilde{\sigma}_j] Z_{t,j}^s\right) \right|^2 \right)\right]^{1/2} \\
&\leq \sum_{j=1}^{k(n)} \left[\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2 \leq \delta} \left|(\mu_j - \tilde{\mu}_j) + [\sigma_j - \tilde{\sigma}_j] Z_{t,j}^s\right|^2 \right)\right]^{1/2} \\
&\leq \sum_{j=1}^{k(n)} \sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2 \leq \delta} \left(|\mu_j - \tilde{\mu}_j| + |\sigma_j - \tilde{\sigma}_j| C_Z\right) \\
&\leq (1 + C_Z) \sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2 \leq \delta} \left(\sum_{j=1}^{k(n)} |\mu_j - \tilde{\mu}_j|^2 + |\sigma_j - \tilde{\sigma}_j|^2\right)^{1/2} \leq (1 + C_Z)\delta.
\end{aligned}$$

Without loss of generality assume that $\delta \leq 1$ so that:

$$\left[\mathbb{E}\left(\sup_{\|(\omega,\mu,\sigma)-(\tilde{\omega},\tilde{\mu},\tilde{\sigma})\|_2 \leq \delta} \left|e_t^s - \tilde{e}_t^s\right|^2 \right)\right]^{1/2} \leq 2\sqrt{2}(1 + C_Z)\left(1 + \bar{\mu}_{k(n)} + \bar{\sigma} + k(n)\right)\delta^{1/2}.$$

which concludes the proof. □

Lemma E9 (Properties of the Tails Distributions)**.** *Let $\bar{\xi} \geq \xi_1, \xi_2 \geq \underline{\xi} > 0$. 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:

$$\left[\mathbb{E}\left(\sup_{|\xi_1 - \tilde{\xi}_1| \leq \delta} |e_{t,1}^s(\xi_1) - e_{t,1}^s(\tilde{\xi}_1)|^2\right)\right]^{1/2} \leq C\delta, \quad \left[\mathbb{E}\left(\sup_{|\xi_2 - \tilde{\xi}_2| \leq \delta} |e_{t,2}^s(\xi_2) - e_{t,2}^s(\tilde{\xi}_2)|^2\right)\right]^{1/2} \leq C\delta$$

Where the constant C only depends on $\underline{\xi}$ and $\bar{\xi}$.

Proof of Lemma E9. :
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: $(2+\xi_1)\frac{e_1^{1+\xi_1}}{(1-e_1^{2+\xi_1})^2}$. It 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 $\xi_1 \in [\underline{\xi}, \bar{\xi}]$ 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 $|\xi_1 - \tilde{\xi}_1| \leq \delta$, using the mean value theorem, for each ν_1 there exists an intermediate value $\check{\xi}_1 \in [\xi_1, \tilde{\xi}_1]$ such that:

$$\left(\frac{1}{\nu_1} - 1\right)^{\frac{1}{2+\xi_1}} - \left(\frac{1}{\nu_1} - 1\right)^{\frac{1}{2+\check{\xi}_1}} = \frac{1}{2+\check{\xi}_1} \log(\frac{1}{\nu_1} - 1) \left(\frac{1}{\nu_1} - 1\right)^{\frac{1}{2+\check{\xi}_1}} (\xi_1 - \tilde{\xi}_1).$$

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:

$$\left[\mathbb{E}\left(\sup_{|\xi_1 - \tilde{\xi}_1| \leq \delta} |e_{t,1}^s(\xi_1) - e_{t,1}^s(\tilde{\xi}_1)|^2\right)\right]^{1/2} \leq \frac{\delta}{2+\underline{\xi}} \sqrt{\int_2^\infty \log(\nu) \nu^{\frac{2}{2+\underline{\xi}} - 2} d\nu}.$$

□

Proof of Lemma 1. The proof proceeds by recursion. Denote $\Pi_{k(n)} f_j \in \mathcal{F}_{k(n)}$ the mixture approximation of f_j from Lemma E8. For $d_e = 1$, Lemma E8 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 recursively:

$$d_{t+1} = d_t f_{d_e+1} + \Pi_{k(n)} f_1 \times \cdots \times \Pi_{k(n)} f_{d_e} \left(f_{d_e+1} - \Pi_{k(n)} f_{d_e+1} \right).$$

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:

$$\begin{aligned} \|d_{t+1}\|_\infty &\leq \|d_t\|_\infty \|f_{d_e+1}\|_\infty + \|\Pi_{k(n)} f_1 \times \cdots \times \Pi_{k(n)} f_{d_e}\|_\infty \|f_{d_e+1} - \Pi_{k(n)} f_{d_e+1}\|_\infty \\ &\leq \|d_t\|_\infty \left(\|f_{d_e+1}\|_\infty + \|f_1 \times \cdots \times f_{d_e}\|_\infty \|f_{d_e+1} - \Pi_{k(n)} f_{d_e+1}\|_\infty \right) = O\left(\frac{\log[k(n)]^{r/b}}{k(n)^r} \right). \end{aligned}$$

□

Definition E3 (Pseudo-Norm $\|\cdot\|_m$ on $\mathcal{B}_{k(n)}$)**.** *Let $\beta_1, \beta_2 \in \mathcal{B}_{k(n)}$ where $\beta_l = (\theta_l, f_l), l = 1, 2$ with $f_j = f_{1,j} \times \cdots \times f_{d_e,j}$; each $f_{l,j}$ is as in Definition 1. The pseudo-norm $\|\cdot\|_m$ is the ℓ^2 norm on $(\theta, \omega, \mu, \sigma, \xi)$, the associated distance is:*

$$\|\beta_1 - \beta_2\|_m = \|(\theta_1, \omega_1, \mu_1, \sigma_1, \xi_1) - (\theta_2, \omega_2, \mu_2, \sigma_2, \xi_2)\|_2$$

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 $\theta, \omega, \mu, \sigma, \xi$.

Remark E5. *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:

$$\begin{aligned} \|f_1^1 f_1^2 - f_2^1 f_2^2\|_{TV} &\leq \|f_1^2 - f_2^2\|_{TV} + \|f_1^2 - f_2^1\|_{TV} \\ &\leq C_{TV} \frac{\|f_1^2 - f_2^2\|_m + \|f_1^2 - f_2^1\|_m}{\underline{\sigma}_{k(n)}} \leq C_{TV,2} \frac{\|f_1 - f_2\|_m}{\underline{\sigma}_{k(n)}}. \end{aligned}$$

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.

E.2 Consistency

Assumption 2′ (Data Generating Process - L^2-Smoothness)**.** *y_t^s is simulated according to the dynamic model (1)-(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 $0 \leq \bar{C}_1 < 1$:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta} \|g_{obs}(y_t^s(\beta_1), x_t, \beta_1, u_t^s(\beta_1)) - g_{obs}(y_t^s(\beta_2), x_t, \beta_1, u_t^s(\beta_1))\|^2 \Big| y_t^s(\beta_1), y_t^s(\beta_2)\right)\right]^{1/2} \leq \bar{C}_1 \|y_t^s(\beta_1) - y_t^s(\beta_2)\|$$

$y(ii)'$. For some $0 \leq \bar{C}_2 < \infty$:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta} \|g_{obs}(y_t^s(\beta_1), x_t, \beta_1, u_t^s(\beta_1)) - g_{obs}(y_t^s(\beta_1), x_t, \beta_2, u_t^s(\beta_1))\|^2\right)\right]^{1/2} \leq \bar{C}_2 \delta^\gamma$$

$y(iii)'$. For some $0 \leq \bar{C}_3 < \infty$:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta} \|g_{obs}(y_t^s(\beta_1), x_t, \beta_1, u_t^s(\beta_1)) - g_{obs}(y_t^s(\beta_1), x_t, \beta_1, u_t^s(\beta_2))\|^2 \Big| u_t^s(\beta_1), u_t^s(\beta_2)\right)\right]^{1/2} \leq \bar{C}_3 \|u_t^s(\beta_1) - u_t^s(\beta_2)\|^\gamma$$

$u(i)'$. For some $0 \leq \bar{C}_4 < 1$

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta} \|g_{latent}(u_{t-1}^s(\beta_1), \beta, e_t^s(\beta_1)) - g_{latent}(u_{t-1}^s(\beta_2), \beta, e_t^s(\beta_1))\|^2\right)\right]^{1/2} \leq \bar{C}_4 \|u_{t-1}^s(\beta_1) - u_{t-1}^s(\beta_2)\|$$

$u(ii)'$. For some $0 \leq \bar{C}_5 < \infty$:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta} \|g_{latent}(u_{t-1}^s(\beta_1), \beta_1, e_t^s(\beta_1)) - g_{latent}(u_{t-1}^s(\beta_1), \beta_2, e_t^s(\beta_1))\|^2\right)\right]^{1/2} \leq \bar{C}_5 \delta^\gamma$$

$u(iii)'$. For some $0 \leq \bar{C}_5 < \infty$:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_{\mathcal{B}}\leq\delta} \|g_{latent}(u_{t-1}^s(\beta_1), \beta_1, e_t^s(\beta_1)) - g_{latent}(u_{t-1}^s(\beta_1), \beta_1, e_t^s(\beta_2))\|^2 \Big| e_t^s(\beta_1), e_t^s(\beta_2)\right)\right]^{1/2} \leq \bar{C}_6 \|e_t^s(\beta_1) - e_t^s(\beta_2)\|$$

for $\|\beta_1 - \beta_2\|_{\mathcal{B}} = \|\theta_1 - \theta_2\| + \|f_1 - f_2\|_\infty$ *or* $\|\theta_1 - \theta_2\| + \|f_1 - f_2\|_{TV}$.

Proof of Lemma 3: 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 $\sup_\tau \|\tau\|_\infty \pi(\tau)^{\frac{1}{4}} \leq C_\pi < \infty$ and $\pi(\tau)^{\frac{1}{2}} \propto \pi(\tau/\sqrt{2})$ 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 focuses on 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 $\underline{\sigma}_{k(n)} \leq 1$.[57] Applying this inequality recursively, and using the fact that y_0^s, u_0^s are the same regardless of β, yields:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1 - \beta_2\|_m} \|y_t(\beta_1) - y_t(\beta_2)\|^2\right)\right]^{1/2}$$
$$\leq \frac{\overline{C}_2}{1 - \overline{C}_1} \frac{\delta^\gamma}{\underline{\sigma}_{k(n)}^{2\gamma}} + \overline{C}_3 \sum_{l=0}^{t-1} \overline{C}_1^l \left[\mathbb{E}\left(\sup_{\|\beta_1 - \beta_2\|_m} \|u_{t-l}(\beta_1) - u_{t-l}(\beta_2)\|^2\right)\right]^{\gamma/2}.$$

[57] Recall that by assumption $\underline{\sigma}_{k(n)} = O(\frac{log[k(n)]^{2/b}}{k(n)})$ goes to zero.

Using Lemmas 2 and E9 and the same approach as above:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m}\|u_t(\beta_1)-u_t(\beta_2)\|^2\right)\right]^{1/2} \leq \overline{C}_4\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m}\|u_{t-1}(\beta_1)-u_{t-1}(\beta_2)\|^2\right)\right]^{1/2}$$
$$+\overline{C}_5\frac{\delta^\gamma}{\underline{\sigma}^{2\gamma}_{k(n)}}+\overline{C}_6 C\left(k(n)+\bar{\mu}_{k(n)}+\bar{\sigma}\right)\delta^{\gamma/2}.$$

Again, applying this inequality recursively yields:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m}\|u_t(\beta_1)-u_t(\beta_2)\|^2\right)\right]^{1/2} \leq \frac{\overline{C}_5}{1-\overline{C}_4}\frac{\delta^\gamma}{\underline{\sigma}^{2\gamma}_{k(n)}}+\frac{\overline{C}_6}{1-\overline{C}_4}C\left(k(n)+\bar{\mu}_{k(n)}+\bar{\sigma}\right)\delta^{\gamma/2}.$$

Putting everything together:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m}\|y_t(\beta_1)-y_t(\beta_2)\|^2\right)\right]^{1/2}$$
$$\leq \frac{\overline{C}_2}{1-\overline{C}_1}\frac{\delta^\gamma}{\underline{\sigma}^{2\gamma}_{k(n)}}+\frac{\overline{C}_3}{1-\overline{C}_1}\left(\frac{\overline{C}_5}{1-\overline{C}_4}\frac{\delta^\gamma}{\underline{\sigma}^{2\gamma}_{k(n)}}+\frac{\overline{C}_6}{1-\overline{C}_4}C\left(k(n)+\bar{\mu}_{k(n)}+\bar{\sigma}\right)\delta^{\gamma/2}\right)^\gamma.$$

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 E10 (Covering Numbers)**.** *Under the L^2-smoothness of the DGP (as in Lemma 3), the bracketing number satisfies for $x \in (0,1)$ and some $\overline{C}$:*

$$N_{[]}(x,\Psi_{k(n)}(\tau),\|\cdot\|_{L^2})$$
$$\leq (3[k(n)+2]+d_\theta)\left(2\max(\bar{\mu}_{k(n)},\underline{\sigma})\overline{C}^{2/\gamma^2}\frac{\left(k(n)+\overline{\mu}_{k(n)}+\overline{\sigma}\right)^{2/\gamma}+\underline{\sigma}^4_{k(n)}}{x^{2/\gamma^2}}+1\right)^{3[k(n)+2]+d_\theta}.$$

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 $\tilde{C}$:

$$\log\left(N_{[]}(x,\Psi_{k(n)}(\tau),\|\cdot\|_{L^2}\right) \leq \tilde{C}k(n)\log[k(n)])|\log\delta|.$$

Using the above, for some $\tilde{C}_2 < \infty$:

$$\int_0^1 \log^2\left(N_{[]}(x,\Psi_k(n),\|\cdot\|_{L^2}\right)dx \leq \tilde{C}_2 k(n)^2\log[k(n)]^2.$$

Proof of Lemma E10: Since $\mathcal{B}_{k(n)}$ is contained in a ball of radius $\max(\overline{\mu}_{k(n)}, \overline{\sigma}, \|\theta\|_\infty)$ in $\mathbb{R}^{3[k(n)+2]+d_\theta}$ 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).[58] As a result, the covering number $N(x, \mathcal{B}_{k(n)}, \|\cdot\|_m)$ satisfies:

$$N(x, \mathcal{B}_{k(n)}, \|\cdot\|_m) \leq 2\left(3[k(n)+2]+d_\theta\right)\left(\frac{2\max(\bar{\mu}_{k(n)}, \bar{\sigma})}{x}+1\right)^{3[k(n)+2]+d_\theta}.$$

The rest follows from Lemma 3 and Appendix F. □

Proof of Theorem 1: If the assumptions of Corollary F3 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 F8 and F9. Furthermore, by Lemmas 3 and E10, Assumptions 1 with 2 (or 2′) imply Assumption F11 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 4. This implies that all the assumptions for Corollary F3 so that the estimator is consistent if $\sqrt{C_n/n} = o(1)$ which concludes the proof. □

E.3 Rate of Convergence

Proof of Lemma 4: First, using the assumption that B is a bounded linear operator:

$$\begin{aligned}
&Q_n(\Pi_{k(n)}\beta_0) \\
&\leq M_B^2 \int \left|\mathbb{E}\left(\hat{\psi}_n(\tau) - \hat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0)\right)\right|^2 \pi(\tau) d\tau \\
&\leq 3M_B^2 \left(\int \left|\mathbb{E}\left(\hat{\psi}_n(\tau) - \hat{\psi}_n^S(\tau, \beta_0)\right)\right|^2 \pi(\tau) d\tau + \int \left|\mathbb{E}\left(\hat{\psi}_n^S(\tau, \beta_0) - \hat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0)\right)\right|^2 \pi(\tau) d\tau\right)
\end{aligned}$$

Each term can be bounded above individually. Re-write the first term in terms of distribution:

$$\left|\mathbb{E}\left(\hat{\psi}_n(\tau) - \hat{\psi}_n^S(\tau, \beta_0)\right)\right| = \left|\frac{1}{n}\sum_{t=1}^{n} \int e^{i\tau'(\mathbf{y}_t, \mathbf{x_t})}[f_t^*(\mathbf{y}_t, \mathbf{x_t}) - f_t(\mathbf{y}_t, \mathbf{x_t})] d\mathbf{y}_t d\mathbf{x}_t\right|$$

[58] See also Fenton & Gallant (1996) for an application of this result for the sieve estimation of a density and Coppejans (2001) for a CDF.

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:

$$\left|\frac{1}{n}\sum_{t=1}^{n}\int e^{i\tau'(\mathbf{y}_t,\mathbf{x_t})}[f_t^*(\mathbf{y}_t,\mathbf{x_t}) - f_t(\mathbf{y}_t,\mathbf{x_t})]d\mathbf{y}_t d\mathbf{x}_t\right| \leq \frac{1}{n}\sum_{t=1}^{n}\int \left|f_t^*(\mathbf{y}_t,\mathbf{x_t}) - f_t(\mathbf{y}_t,\mathbf{x_t})\right| d\mathbf{y}_t d\mathbf{x}_t$$
$$= \frac{2}{n}\sum_{t=1}^{n}\|f_t^* - f_t\|_{TV} \leq \frac{2C_\rho}{n}\sum_{t=1}^{n}\rho^t \leq \frac{2C_\rho}{(1-\rho)n}$$

for some $\rho \in (0,1)$ and $C_\rho > 0$. This yields a first bound:

$$\int \left|\mathbb{E}\left(\hat{\psi}_n(\tau) - \hat{\psi}_n^S(\tau,\beta_0)\right)\right|^2 \pi(\tau)d\tau \leq \frac{4C_\rho^2}{(1-\rho)^2}\frac{1}{n^2} = O\left(\frac{1}{n^2}\right).$$

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 3 allows to find bounds in terms of $\|\cdot\|_{\mathcal{B}}$ and $\|\cdot\|_{TV}$ for which Lemma 1 gives the approximation rates.

To bound the second term, re-write the simulated data as:

$$y_t^s = g_{obs,t}(\mathbf{x}_{t:1}, \beta, \mathbf{e}_{t:1}^s), \quad u_t^s = g_{latent,t}(\beta, \mathbf{e}_{t:1}^s)$$

with $\beta = (\theta, f)$, $e_t^s \sim f$ and $\mathbf{x}_{t:1} = (x_t, \ldots, x_1), \mathbf{e}_{t:1}^s = (e_t^s, \ldots, e_1^s)$. Under Assumption 2 or 2′, using the same sequence of shocks (e_t^s):

$$\mathbb{E}\left(\left\|g_{obs,t}(\mathbf{x}_{t:1}, \beta_0, \mathbf{e}_{t:1}^s) - g_{obs,t}(\mathbf{x}_{t:1}, \Pi_{k(n)}\beta_0, \mathbf{e}_{t:1}^s)\right\|\right) \leq \overline{C}\|\Pi_{k(n)}f_0 - f_0\|_{\mathcal{B}}^\gamma.$$

This is similar to the proof of Lemma 3, first re-write the difference as:

$$\mathbb{E}\Big(\Big\|g_{obs}(g_{obs,t-1}(\mathbf{x}_{t-1:1}, \beta_0, \mathbf{e}_{t-1:1}^s), x_t, \beta_0, g_{latent}(g_{latent,t-1}(\beta_0, \mathbf{e}_{t-1:1}^s), \beta_0, e_t^s))$$
$$- g_{obs}(g_{obs,t-1}(\mathbf{x}_{t-1:1}, \Pi_{k(n)}\beta_0, \mathbf{e}_{t-1:1}^s), x_t, \Pi_{k(n)}\beta_0, g_{latent}(g_{latent,t-1}(\Pi_{k(n)}\beta_0, \mathbf{e}_{t-1:1}^s), \Pi_{k(n)}\beta_0, e_t^s\Big\|\Big).$$

Using Assumptions 2-2′, the following recursive relationship holds:

$$\begin{aligned}
&\mathbb{E}\Big(\Big\| g_{obs}(g_{obs,t-1}(\mathbf{x}_{t-1:1},\beta_0,\mathbf{e}^s_{t-1:1}),x_t,\beta_0,g_{latent}(g_{latent,t-1}(\beta_0,\mathbf{e}^s_{t-1:1}),\beta_0,e^s_t)) \\
&\quad - g_{obs}(g_{obs,t-1}(\mathbf{x}_{t-1:1},\Pi_{k(n)}\beta_0,\mathbf{e}^s_{t-1:1}),x_t,\Pi_{k(n)}\beta_0, \\
&\qquad g_{latent}(g_{latent,t-1}(\Pi_{k(n)}\beta_0,\mathbf{e}^s_{t-1:1}),\Pi_{k(n)}\beta_0,e^s_t))\Big\|\Big) \\
&\le \Big[\mathbb{E}\Big(\Big\| g_{obs}(g_{obs,t-1}(\mathbf{x}_{t-1:1},\beta_0,\mathbf{e}^s_{t-1:1}),x_t,\beta_0,g_{latent}(g_{latent,t-1}(\beta_0,\mathbf{e}^s_{t-1:1}),\beta_0,e^s_t)) \\
&\quad - g_{obs}(g_{obs,t-1}(\mathbf{x}_{t-1:1},\Pi_{k(n)}\beta_0,\mathbf{e}^s_{t-1:1}),x_t,\Pi_{k(n)}\beta_0, \\
&\qquad g_{latent}(g_{latent,t-1}(\Pi_{k(n)}\beta_0,\mathbf{e}^s_{t-1:1}),\Pi_{k(n)}\beta_0,e^s_t))\Big\|^2\Big)\Big]^{1/2} \\
&\le \overline{C}_1\Big[\mathbb{E}\Big(\Big\| g_{obs,t-1}(\mathbf{x}_{t-1:1},\beta_0,\mathbf{e}^s_{t-1:1}) - g_{obs,t-1}(x_{t-1},\ldots,x_1,\Pi_{k(n)}\beta_0,\mathbf{e}^s_{t-1:1})\Big\|^2\Big)\Big]^{1/2} \\
&+ \overline{C_2}\|\beta_0-\Pi_{k(n)}\beta_0\|^\gamma_{\mathcal{B}} + \overline{C}_3\Big[\mathbb{E}\Big(\Big\| g_{latent,t}(\beta_0,\mathbf{e}^s_{t:1}) - g_{latent,t}(\Pi_{k(n)}\beta_0,\mathbf{e}^s_{t:1})\Big\|^2\Big)\Big]^{\gamma/2}.
\end{aligned}$$

The last term also has a recursive structure:

$$\begin{aligned}
&\Big[\mathbb{E}\Big(\Big\| g_{latent,t}(\beta_0,\mathbf{e}^s_{t:1}) - g_{latent,t}(\Pi_{k(n)}\beta_0,\mathbf{e}^s_{t:1})\Big\|^2\Big)\Big]^{1/2} \\
&\le \overline{C}_4\left[\mathbb{E}\Big(\Big\| g_{latent,t-1}(\beta_0,\mathbf{e}^s_{t-1:1}) - g_{latent,t-1}(\Pi_{k(n)}\beta_0,\mathbf{e}^s_{t-1:1})\Big\|^2\Big)\right]^{1/2} + \overline{C}_5\|\beta_0-\Pi_{k(n)}\beta_0\|^\gamma_{\mathcal{B}}.
\end{aligned}$$

Together these inequalities imply:

$$\begin{aligned}
&\mathbb{E}\Big(\Big\| g_{obs}(g_{obs,t-1}(x_{t-1},\ldots,x_1,\beta_0,\mathbf{e}^s_{t-1:1}),x_t,\beta_0,g_{latent}(g_{latent,t-1}(\beta_0,\mathbf{e}^s_{t-1:1}),\beta_0,e^s_t)) \\
&- g_{obs}(g_{obs,t-1}(x_{t-1},\ldots,x_1,\Pi_{k(n)}\beta_0,\mathbf{e}^s_{t-1:1}),x_t,\Pi_{k(n)}\beta_0, \\
&g_{latent}(g_{latent,t-1}(\Pi_{k(n)}\beta_0,\mathbf{e}^s_{t-1:1}),\Pi_{k(n)}\beta_0,e^s_t\Big\|\Big) \\
&\le \frac{1}{1-\overline{C}_1}\left(\overline{C}_2\|\beta_0-\Pi_{k(n)}\beta_0\|^\gamma_{\mathcal{B}} + \overline{C}_3\frac{\overline{C}_5^\gamma}{(1-\overline{C_4})^\gamma}\|\beta_0-\Pi_{k(n)}\beta_0\|^{\gamma^2}_{\mathcal{B}}\right).
\end{aligned}$$

Recall that $\|\tau\|_\infty\sqrt{\pi(\tau)}$ is bounded above and $\pi(\tau)^{1/4}$ is integrable so that:

$$\begin{aligned}
&\int\left|\mathbb{E}\left(e^{i\tau'(\mathbf{y}_t(\beta_0,\mathbf{x}_{t:1}))} - e^{i\tau'(\mathbf{y}_t(\Pi_{k(n)}\beta_0,\mathbf{x}_{t:1}))}\right)\right|^2\pi(\tau)d\tau \\
&\le \left(\overline{C}_2\|\beta_0-\Pi_{k(n)}\beta_0\|^\gamma_{\mathcal{B}} + \overline{C}_3\frac{\overline{C}_5^\gamma}{(1-\overline{C_4})^\gamma}\|\beta_0-\Pi_{k(n)}\beta_0\|^{\gamma^2}_{\mathcal{B}}\right)\frac{\sup_\tau[\|\tau\|_\infty\sqrt{\pi(\tau)}]\int\pi(\tau)^{1/4}d\tau}{1-\overline{C}_1}.
\end{aligned}$$

To conclude the proof, the difference due to e^s_t needs to be bounded. In order to do so, it suffice to bound the following integral:

$$\int e^{i\tau'(\mathbf{y}_t(y_0,u_0,\mathbf{x}_{t:1},\beta_0,\mathbf{e}^s_{t:1}),\mathbf{x}_t)}\left(\prod_{j=1}^t f_0(e^s_j) - \prod_{j=1}^t \Pi_{k(n)}f_0(e^s_j)\right)f_{\mathbf{x}}(\mathbf{x}_{t:1})d\mathbf{e}^s_{t:1}d\mathbf{x}_{t:1}.$$

A direct bound on this integral yields a term of order of $t\|f_0 - \Pi_{k(n)} f_0\|_{TV}$ which increases with t, which is too fast 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 $\beta = (\theta, f) \in \mathcal{B}$, $e_t^s, \ldots, e_1^s \sim f$ and $\tilde{y}_t^s$ such that $\tilde{y}_{t-m}^s = 0, \tilde{u}_{t-m}^s = 0$ and then $\tilde{y}_j^s = g_{obs}(\tilde{y}_{j-1}^s, x_j, \beta, \tilde{u}_j^s), \tilde{u}_j^s = g_{latent}(\tilde{u}_{j-1}^s, \beta, e_j^s)$ for $t - m + 1 \leq j \leq t$. Each observation t is approximated by its own time-series. For observation $t - m$, by construction:

$$\mathbb{E}\left(\left\|y_{t-m}^s - \tilde{y}_{t-m}^s\right\|\right) = \mathbb{E}\left(\left\|y_{t-m}^s\right\|\right) \leq \left[\mathbb{E}\left(\left\|y_{t-m}^s\right\|^2\right)\right]^{1/2}$$
$$\mathbb{E}\left(\left\|u_{t-m}^s - \tilde{u}_{t-m}^s\right\|\right) = \mathbb{E}\left(\left\|u_{t-m}^s\right\|\right) \leq \left[\mathbb{E}\left(\left\|u_{t-m}^s\right\|^2\right)\right]^{1/2}$$

Then, for any $t \geq \tilde{t} \geq t - m$:

$$\mathbb{E}\left(\left\|u_{\tilde{t}}^s - \tilde{u}_{\tilde{t}}^s\right\|\right) \leq \overline{C}_4 \left[\mathbb{E}\left(\left\|u_{\tilde{t}-1}^s - \tilde{u}_{\tilde{t}-1}^s\right\|^2\right)\right]^{1/2}$$
$$\mathbb{E}\left(\left\|y_{\tilde{t}}^s - \tilde{y}_{\tilde{t}}^s\right\|\right) \leq \overline{C}_3 \overline{C}_4^{\gamma} \left[\mathbb{E}\left(\left\|u_{\tilde{t}-1}^s - \tilde{u}_{\tilde{t}-1}^s\right\|^2\right)\right]^{\gamma/2} + \overline{C}_1 \left[\mathbb{E}\left(\left\|y_{\tilde{t}-1}^s - \tilde{y}_{\tilde{t}-1}^s\right\|^2\right)\right]^{1/2}.$$

The previous two results and a recursion arguments leads to the following inequality:

$$\mathbb{E}\left(\left\|u_t^s - \tilde{u}_t^s\right\|\right) \leq \overline{C}_4^m \left[\mathbb{E}\left(\left\|u_{t-m}^s\right\|^2\right)\right]^{1/2} \tag{E.25}$$
$$\mathbb{E}\left(\left\|y_t^s - \tilde{y}_t^s\right\|\right) \leq \overline{C}_3 \overline{C}_4^{\gamma m} \left[\mathbb{E}\left(\left\|u_{t-m}^s\right\|^2\right)\right]^{\gamma/2} + \overline{C}_1^m \left[\mathbb{E}\left(\left\|y_{t-m}^s\right\|^2\right)\right]^{1/2}. \tag{E.26}$$

For $\beta = \beta_0, \Pi_{k(n)}\beta_0$ since the expectations are finite and bounded by assumption, $\mathbb{E}\left(\left\|y_t^s - \tilde{y}_t^s\right\|\right) \leq \overline{C} \max(\overline{C}_1, \overline{C}_4)^{\gamma m}$ 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, $y_t^s = \tilde{y}_t^s$, so that the bound still holds. The

integral can be split and bounded:

$$\begin{aligned}
&\Big| \int e^{i\tau'(\mathbf{y}_t(y_0,u_0,\mathbf{x}_{t:1},\beta_0,\mathbf{e}^s_{t:1}),\mathbf{x}_t)} \left(\prod_{j=1}^t f_0(e^s_j) - \prod_{j=1}^t \Pi_{k(n)} f_0(e^s_j) \right) f_{\mathbf{x}}(\mathbf{x}_{t:1}) d\mathbf{e}^s_{t:1} d\mathbf{x}_{t:1} \Big| \\
&\leq \left| \mathbb{E} \left([\hat{\psi}^S_n(\tau,\beta_0) - \hat{\psi}^S_n(\tau,\Pi_{k(n)}\beta_0)] - [\tilde{\psi}^S_n(\tau,\beta_0) - \tilde{\psi}^S_n(\tau,\Pi_{k(n)}\beta_0)] \right) \right| \\
&+ \int \Big| \left(\prod_{j=t-m+1}^t f_0(e^s_j) - \prod_{j=t-m+1}^t \Pi_{k(n)} f_0(e^s_j) \right) d\mathbf{e}^s_{t:t-m+1} \Big| \\
&\leq 4\overline{C} \max(\overline{C}_1, \overline{C}_4)^{\gamma m} + 2m \|\Pi_{k(n)} f_0 - f_0\|_{TV}.
\end{aligned}$$

The last inequality is due to the cosine and sine functions being uniformly Lipschitz continuous and equations (E.25)-(E.26). 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 $\|\beta_0 - \Pi_{k(n)}\beta_0\|^{\gamma^2}_{\mathcal{B}} = O(\frac{\log[k(n)]^{4\gamma^2 r/b}}{k(n)^{2\gamma^2 r}})$. 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: This theorem is a corollary of Theorem F5 with a mixture sieve. Lemma 4 gives an explicit derivation of $\sqrt{Q_n(\Pi_{k(n)}\beta_0)}$ in this setting. □

E.4 Asymptotic Normality

Remark E6. *Note that for each τ the matrix $B\frac{d\mathbb{E}(\hat{\psi}^S_n(\tau,\Pi_{k(n)}\beta_0))}{d(\theta,\omega,\mu,\sigma)}' \overline{B\frac{d\mathbb{E}(\hat{\psi}^S_n(\tau,\Pi_{k(n)}\beta_0))}{d(\theta,\omega,\mu,\sigma)}}$ is singular - the requirement is that the average, over τ, of this matrix is invertible. Lemma 5 states that $\hat{\beta}_n$ 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 5: Using the simple inequality $1/2|a|^2 \leq |a-b|^2 + |b|^2$ for any $a, b \in \mathbb{R}$:

$$
\begin{aligned}
&0 \leq 1/2 \int \left| B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta} [\hat{\beta}_n - \Pi_{k(n)}\beta_0] \right|^2 \pi(\tau) d\tau \\
&\leq \int \left| B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \beta_0))}{d\beta} [\hat{\beta}_n - \beta_0] \right|^2 \pi(\tau) d\tau \\
&+ \int \left| B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \beta_0))}{d\beta} [\hat{\beta}_n - \beta_0] - B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta} [\hat{\beta}_n - \Pi_{k(n)}\beta_0] \right|^2 \pi(\tau) d\tau \\
&\leq \int \left| B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \beta_0))}{d\beta} [\hat{\beta}_n - \beta_0] \right|^2 \pi(\tau) d\tau + \int \left| B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta} [\Pi_{k(n)}\beta_0 - \beta_0] \right|^2 \pi(\tau) d\tau \\
&+ \int \left| B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \beta_0))}{d\beta} [\hat{\beta}_n - \beta_0] - B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \beta_0))}{d\beta} [\hat{\beta}_n - \Pi_{k(n)}\beta_0] \right|^2 \pi(\tau) d\tau.
\end{aligned}
$$

By assumption the term on the left is $O_p(\delta_n^2)$, by condition ii. the middle term is $O_p(\delta_n^2)$ and condition i. implies that the term on the right is also $O_p(\delta_n^2)$. It follows that:

$$
\int \left| B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta} [\hat{\beta}_n - \Pi_{k(n)}\beta_0] \right|^2 \pi(\tau) d\tau = O_p(\delta_n^2). \tag{E.27}
$$

Now note that both $\hat{\beta}_n$ and $\Pi_{k(n)}\beta_0$ belong to the finite dimensional space $\mathcal{B}_{k(n)}$ parameterized by $(\theta, \omega, \mu, \sigma)$. To save space, $\hat{\beta}_n$ will be represented by $\hat{\varphi}_n = (\hat{\theta}_n, \hat{\omega}_n, \hat{\mu}_n, \hat{\sigma}_n)$ and $\Pi_{k(n)}\beta_0$ by $\varphi_{k(n)} = (\theta_{k(n)}, \omega_{k(n)}, \mu_{k(n)}, \sigma_{k(n)})$. Using this notation, equation (E.27) becomes:

$$
\begin{aligned}
&\int \left| B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d\beta} [\hat{\beta}_n - \Pi_{k(n)}\beta_0] \right|^2 \pi(\tau) d\tau \\
&= \int \left| B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d(\theta, \omega, \mu, \sigma)} [\hat{\varphi}_n - \varphi_{k(n)}] \right|^2 \pi(\tau) d\tau \\
&= trace \left([\hat{\varphi}_n - \varphi_{k(n)}]' \int B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d(\theta, \omega, \mu, \sigma)}' \overline{B \frac{d\mathbb{E}(\hat{\psi}_n^S(\tau, \Pi_{k(n)}\beta_0))}{d(\theta, \omega, \mu, \sigma)}} \pi(\tau) d\tau [\hat{\varphi}_n - \varphi_{k(n)}] \right) \\
&\geq \underline{\lambda}_n \|\hat{\varphi}_n - \varphi_{k(n)}\|^2 = \underline{\lambda}_n \|\hat{\beta}_n - \Pi_{k(n)}\beta_0\|_m^2.
\end{aligned}
$$

It follows that $0 \leq \underline{\lambda}_n \|\hat{\beta}_n - \Pi_{k(n)}\beta_0\|_m^2 \leq O_p(\delta_n^2)$ so that the rate of convergence in mixture norm is:

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

□

Lemma E11 (Stochastic Equicontinuity)**.** *Let $M_n = \log\log(n+1)$ and $\delta_{mn} = \delta_n/\sqrt{\underline{\lambda}_n}$. Let $\Delta_n^S(\tau, \beta) = \hat{\psi}_n^S(\tau, \beta) - \mathbb{E}(\hat{\psi}_n^S(\tau, \beta))$. Suppose that the assumptions of Lemma 5 and*

Assumption F11 hold then for any $\eta > 0$, uniformly over $\beta \in \mathcal{B}_{k(n)}$:

$$\left[\mathbb{E}\left(\sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m \leq M_n\delta_{mn}} \left|\Delta_n^S(\tau,\beta) - \Delta_n^S(\tau,\Pi_{k(n)}\beta_0)\right|^2 \pi(\tau)^{\frac{2}{2+\eta}}\right)\right]^{1/2} \leq C\frac{(M_n\delta_{mn})^{\frac{\gamma^2}{2}}}{\sqrt{n}} I_{m,n}$$

Where $I_{m,n}$ is defined as:

$$I_{m,n} = \int_0^1 \left(x^{-\vartheta/2}\sqrt{\log N([xM_n\delta_{mn}]^{\frac{2}{\gamma^2}}, \mathcal{B}_{k(n)}, \|\cdot\|_m)} + \log^2 N([xM_n\delta_{mn}]^{\frac{2}{\gamma^2}}, \mathcal{B}_{k(n)}, \|\cdot\|_m)\right) dx$$

For the mixture sieve the integral is a $O(k(n)\log[k(n)] + k(n)|\log(M_n\delta_{mn})|)$ so that:

$$\left[\mathbb{E}\left(\int \sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m \leq M_n\delta_{mn}} \left|\Delta_n^S(\tau,\beta) - \Delta_n^S(\tau,\Pi_{k(n)}\beta_0)\right|^2 \pi(\tau)d\tau\right)\right]^{1/2}$$
$$= O\left((M_n\delta_{mn})^{\frac{\gamma^2}{2}} \max(\log[k(n)]^2, |\log[M_n\delta_{mn}]|^2)\frac{k(n)^2}{\sqrt{n}}\right)$$

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:

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

Also, suppose that $\beta \to \int \mathbb{E}\left|\hat{\psi}_t^s(\tau,\beta_0) - \hat{\psi}_t^s(\tau,\beta)\right|^2 \pi(\tau)d\tau$ is continuous at $\beta = \beta_0$ under the norm $\|\cdot\|_{\mathcal{B}}$, uniformly in $t \geq 1$. Then, the second stochastic equicontinuity result is:

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

Proof of Lemma E11. This proof relies on the results in Lemma 3 together with Lemma G16. First, Lemma 3 implies that, after simplifying the bounds, for some $C > 0$:

$$\left[\mathbb{E}\left(\sup_{\|\beta_1-\beta_2\|_m \leq \delta, \|\beta_j-\Pi_{k(n)}\beta_0\|_m \leq M_n\delta_{m,n}, j=1,2} \left|\hat{\psi}_t^s(\tau,\beta_1) - \hat{\psi}_t^s(\tau,\beta_2)\right|^2\right)\right]^{1/2} \frac{\sqrt{\pi(\tau)}}{(M_n\delta_{m,n})^{\gamma^2/2}}$$
$$\leq Ck(n)^{2\gamma^2}\left(\frac{\delta}{M_n\delta_{m,n}}\right)^{\gamma^2/2}.$$

Next, apply the inequality of Lemma G15 to generate the bound:

$$\left[\mathbb{E}\left(\sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m \leq M_n\delta_{m,n}} \left|\Delta_n^S(\tau,\beta) - \Delta_n^S(\tau,\Pi_{k(n)}\beta_0)\right|^2\right)\right]^{1/2} \sqrt{\pi(\tau)} \leq \overline{C}\frac{(M_n\delta_{m,n})^{\gamma^2/2}}{\sqrt{n}} J_{m,n}$$

for some $\overline{C} > 0, \vartheta \in (0,1)$ and

$$J_{m,n} = \int_0^1 \left(x^{-\vartheta/2} \sqrt{\log N(\left[\frac{xM_n\delta_{mn}}{k(n)^{2\gamma^2}}\right]^{\frac{2}{\gamma^2}}, \mathcal{B}_{k(n)}, \|\cdot\|_m)} + \log^2 N(\left[\frac{xM_n\delta_{mn}}{k(n)^{2\gamma^2}}\right]^{\frac{2}{\gamma^2}}, \mathcal{B}_{k(n)}, \|\cdot\|_m) \right) dx.$$

Since $\int \sqrt{\pi(\tau)}d\tau < \infty$, the term on the left-hand side of the inequality can be squared and multiplied by $\sqrt{\pi(\tau)}$. Then, taking the integral:

$$\left[\mathbb{E}\left(\int \sup_{\|\beta-\Pi_{k(n)}\beta_0\|_m \leq M_n\delta_{m,n}} \left|\Delta_n^S(\tau,\beta) - \Delta_n^S(\tau,\Pi_{k(n)}\beta_0)\right|^2 \pi(\tau)d\tau\right)\right]^{1/2} \leq \overline{C}_\pi \frac{(M_n\delta_{m,n})^{\gamma^2/2}}{\sqrt{n}} J_{m,n}$$

where $\overline{C}_\pi = \overline{C}\int\sqrt{\pi(\tau)}d\tau$. Note that $J_{m,n} = O(k(n)^2 \max(\log[k(n)]^2, \log[M_n\delta_{m,n}]^2))$.

To prove the final statement, notation will be shortened using $\Delta\hat{\psi}_t^s(\tau,\beta) = \hat{\psi}_t^s(\tau,\beta_0) - \hat{\psi}_t^s(\tau,\beta)$. Note that, by applying Davydov (1968)'s inequality:

$$\begin{aligned}
&n\mathbb{E}\left|\Delta\hat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\hat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]\right|^2 \\
&\leq \frac{1}{n}\sum_{t=1}^n \mathbb{E}\left|\Delta\hat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\hat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0)]\right|^2 \\
&+ \frac{24}{n}\sum_{m=1}^n (n-m)\alpha(m)^{1/3} \max_{1\leq t\leq n}\left(\mathbb{E}\left|\Delta\hat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\hat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0)]\right|^6\right)^{2/3} \\
&\leq \left(1 + 24\sum_{m\geq 1}\alpha(m)^{1/3}\right) \max_{1\leq t\leq n}\left(\mathbb{E}\left|\Delta\hat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\hat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0)]\right|^6\right)^{2/3} \\
&\leq 4^{8/3}\left(1 + 24\sum_{m\geq 1}\alpha(m)^{1/3}\right) \max_{1\leq t\leq n}\left(\mathbb{E}\left|\Delta\hat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\hat{\psi}_t^s(\tau,\Pi_{k(n)}\beta_0)]\right|^2\right)^{2/3}.
\end{aligned}$$

The last inequality is due to $|\Delta\hat{\psi}_t^s(\tau,\beta)| \leq 2$. By the continuity assumption the last term is a $o(1)$ when $\|\beta_0 - \Pi_{k(n)}\|_{\mathcal{B}} \to 0$. As a result:

$$\int \mathbb{E}\left|\Delta\hat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0) - \mathbb{E}[\Delta\hat{\psi}_n^S(\tau,\Pi_{k(n)}\beta_0)]\right|^2 \pi(\tau)d\tau = o(1/n).$$

To conclude the proof, apply a triangular inequality and the results above:

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

□

Remark E7. *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 on the term* $k(n)^2\max(\log[k(n)]^2, \log[M_n\delta_{m,n}]^2)$ *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 F6 with two differences: the rate of convergence and the stochastic equicontinuity results in mixture norm. Lemmas 5 and E11 provide these results for the mixture sieve. Hence, given these results, Theorem 3 is a corollary of Theorem F6. □

E.5 Extension 1: Using Auxiliary Variables

Proof of Corollary B2: Since the proof of Corollary B2 is very similar to the main proofs, only the differences in the steps are highlighted.

i. **Consistency**: The objective function with auxiliary variables is:

$$Q_n(\beta) = \int \left| \mathbb{E}\left(\hat{\psi}_n(\tau, \hat{\eta}_n^{aux}) - \hat{\psi}_n^s(\tau, \hat{\eta}_n^{aux}, \beta \right) \right|^2 \pi(\tau) d\tau.$$

To derive its rate of convergence consider:

$$\begin{aligned}
&\int \left| \hat{\psi}_n(\tau, \hat{\eta}_n^{aux}) - \mathbb{E}\left(\hat{\psi}_n(\tau, \hat{\eta}_n^{aux}) \right) \right|^2 \pi(\tau) d\tau \\
&\leq 9 \int \left| \hat{\psi}_n(\tau, \eta^{aux}) - \mathbb{E}\left(\hat{\psi}_n(\tau, \eta^{aux}) \right) \right|^2 \pi(\tau) d\tau \\
&+ 9 \int \left| \hat{\psi}_n(\tau, \hat{\eta}_n^{aux}) - \hat{\psi}_n(\tau, \eta^{aux}) \right|^2 \pi(\tau) d\tau \\
&+ 9 \int \left| \mathbb{E}\left(\hat{\psi}_n(\tau, \hat{\eta}_n^{aux}) - \hat{\psi}_n(\tau, \eta^{aux}) \right) \right|^2 \pi(\tau) d\tau.
\end{aligned}$$

The first term is $O_p(1/n)$. By the Lipschitz condition, the second term satisfies:

$$\begin{aligned}
\int \left| \hat{\psi}_n(\tau, \hat{\eta}_n^{aux}) - \hat{\psi}_n(\tau, \eta^{aux}) \right|^2 \pi(\tau) d\tau &\leq \|\hat{\eta}_n^{aux} - \eta^{aux}\|^2 |C_n^{aux}|^2 \int \|\tau\|_\infty \pi(\tau) d\tau \\
&= O_p(1/n) O_p(1).
\end{aligned}$$

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:

$$\begin{aligned}
\int \left| \hat{\psi}_n(\tau, \hat{\eta}_n^{aux}) - \hat{\psi}_n(\tau, \eta^{aux}) \right|^2 \pi(\tau) d\tau &\leq \mathbb{E}\|\hat{\eta}_n^{aux} - \eta^{aux}\|^2 \mathbb{E}|C_n^{aux}|^2 \int \|\tau\|_\infty \pi(\tau) d\tau \\
&= O_p(1/n^2) O_p(1).
\end{aligned}$$

Altogether, these inequalities imply:

$$\int \left| \hat{\psi}_n(\tau, \hat{\eta}_n^{aux}) - \mathbb{E}\left(\hat{\psi}_n(\tau, \hat{\eta}_n^{aux}) \right) \right|^2 \pi(\tau) d\tau = O_p(1/n^2).$$

The L^2-smoothness result still holds given the summability condition:

$$\begin{aligned}
&\left[\mathbb{E}\Big(\sup_{\|\beta_1 - \beta_2\|_{\mathcal{B}} \leq \delta, \eta \in E} \|g_{aux}(\mathbf{y}_{t:1}^s(\beta_1), \mathbf{x}_{t:1}; \eta) - g_{aux}(\mathbf{y}_{t:1}^s(\beta_2), \mathbf{x}_{t:1}; \eta)\|^2 \Big) \right]^{1/2} \\
&\leq \sum_{j=1}^{t} \rho_j \left[\mathbb{E}\left(\sup_{\|\beta_1 - \beta_2\|_{\mathcal{B}} \leq \delta, \eta \in E} \|y_j^s(\beta_1) - y_j^s(\beta_2)\|^2 \right) \right]^{1/2} \\
&\leq \left(\sum_{j=1}^{+\infty} \rho_j \right) \sup_{t \geq 1} \left[\mathbb{E}\left(\sup_{\|\beta_1 - \beta_2\|_{\mathcal{B}} \leq \delta, \eta \in E} \|y_t^s(\beta_1) - y_t^s(\beta_2)\|^2 \right) \right]^{1/2} \\
&\leq \bar{C} \left(\sum_{j=1}^{+\infty} \rho_j \right) \max\left(\frac{\delta^{\gamma^2}}{\underline{\sigma}_{k(n)}^{2\gamma^2}}, [k(n) + \overline{\mu}_{k(n)} + \overline{\sigma}]^\gamma \delta^{\gamma^2/2} \right)
\end{aligned}$$

where $\mathbf{x}_{t:1} = (x_t, \ldots, x_1)$ and $\mathbf{y}_{t:1}(\beta) = (y_t(\beta), \ldots, y_1(\beta))$. The last inequality is a consequence of Lemma 3.

$$\begin{aligned}
&\int \left|\hat{\psi}_n^s(\tau, \hat{\eta}_n^{aux}) - \mathbb{E}\left(\hat{\psi}_n^s(\tau, \hat{\eta}_n^{aux})\right)\right|^2 \pi(\tau) d\tau \\
&\leq 9 \int \left|\hat{\psi}_n^s(\tau, \eta^{aux}) - \mathbb{E}\left(\hat{\psi}_n^s(\tau, \eta^{aux})\right)\right|^2 \pi(\tau) d\tau \\
&+ 9 \int \left|\hat{\psi}_n^s(\tau, \hat{\eta}_n^{aux}) - \hat{\psi}_n^s(\tau, \eta^{aux})\right|^2 \pi(\tau) d\tau \\
&+ 9 \int \left|\mathbb{E}\left(\hat{\psi}_n^s(\tau, \hat{\eta}_n^{aux}) - \hat{\psi}_n^s(\tau, \eta^{aux})\right)\right|^2 \pi(\tau) d\tau.
\end{aligned}$$

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 $O_p(\delta_n^2)$ 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 $\tilde{y}_t^s$ where part of the history is changed $\tilde{y}_{t-j}^s = \tilde{u}_{t-j}^s = 0$ for $j \geq m$. For a well chosen sequence m, the difference between y_t^s and $\tilde{y}_t^s$ declines exponentially in m. Here $\tilde{z}_t^s$ only depends on a finite number of shocks since $\tilde{y}_{t-m}^s = \cdots = \tilde{y}_1^s = 0$. The difference between z_t^s and $\tilde{z}_t^s$ becomes:

$$\mathbb{E}\left(\|z_t^s - \tilde{z}_t^s\|\right) \leq \sum_{j=1}^{t} \rho_j \mathbb{E}\left(\|y_j^s - \tilde{y}_j^s\|\right) \leq \left(\sum_{j=1}^{+\infty} \rho_j\right) \bar{C} \max(\bar{C}_1, \bar{C}_4)^{\gamma m}$$

where the last inequality comes from Lemma 4. 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 $\hat{\psi}_n(\tau)$ and $\hat{\psi}_n^s(\tau,\beta_0)$ are replaced with $\hat{\psi}_n(\tau,\hat{\eta}_n^{aux})$ and $\hat{\psi}_n^s(\tau,\hat{\eta}_n^{aux},\beta_0)$. Eventually, the expansion implies:

$$\frac{\sqrt{n}}{\sigma_n^*}\left(\phi(\hat{\beta}_n)-\phi(\beta_0)\right)$$
$$=\sqrt{n}Real\left(\int \psi_\beta(\tau,u_n^*,\eta^{aux})\overline{\left(\hat{\psi}_n(\tau,\hat{\eta}_n^{aux})-\hat{\psi}_n^s(\tau,\hat{\eta}_n^{aux},\beta_0)\right)}\pi(\tau)d\tau\right)+o_p(1)$$

where the term on the right is asymptotically normal by assumption.

□

E.6 Extension 2: Using Short Panels

Proof of Lemma B7. The second part of the lemma is implied by Remark B4.

For the first part of Lemma B7, using the notation for the proof of Proposition F4: 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 $\|f_{\mathbf{u}}-f_{\mathbf{u}}^*\|_{TV}\le C_u\bar{\rho}_u^m$ for some $C_u>0$ and $\bar{\rho}_u\in(0,1)$.

$$\sqrt{Q_n(\beta_0)}$$
$$\le M_B\left(\int\left|\mathbb{E}\left(\hat{\psi}_n(\tau)-\hat{\psi}_n^s(\tau,\beta_0)\right)\right|^2\pi(\tau)d\tau\right)^{1/2}$$
$$=M_B\left(\int\left|\frac{1}{n}\sum_{j=1}^n\int e^{i\tau'(\mathbf{y}_{j,t}^s,\mathbf{x}_{j,t})}\left(f(\mathbf{y}_{j,t}^s,\mathbf{x}_{j,t})-f^*(\mathbf{y}_{j,t}^s,\mathbf{x}_{j,t})\right)d\mathbf{y}_{j,t}^s d\mathbf{x}_{j,t}\right|^2\pi(\tau)d\tau\right)^{1/2}$$
$$=M_B\left(\int\left|\frac{1}{n}\sum_{j=1}^n\int e^{i\tau'(\mathbf{y}_{j,t}^s,\mathbf{x}_{j,t})}f^*(\mathbf{y}_{j,t}^s,\mathbf{x}_{j,t}|\mathbf{u}_{j,t}^s)\left(f(\mathbf{u}_{j,t}^s)-f^*(\mathbf{u}_{j,t}^s)\right)d\mathbf{y}_{j,t}^s d\mathbf{x}_{j,t}d\mathbf{u}_{j,t}^s\right|^2\pi(\tau)d\tau\right)^{1/2}$$
$$\le M_B\int f^*(\mathbf{y}_{j,t}^s,\mathbf{x}_{j,t}|\mathbf{u}_{j,t}^s)\left|f(\mathbf{u}_{j,t}^s)-f^*(\mathbf{u}_{j,t}^s)\right|d\mathbf{y}_{j,t}^s d\mathbf{x}_{j,t}d\mathbf{u}_{j,t}^s.$$

Applying the Cauchy-Schwarz inequality implies:

$$\int f^*(\mathbf{y}_{j,t}^s,\mathbf{x}_{j,t}|\mathbf{u}_{j,t}^s)\left|f(\mathbf{u}_{j,t}^s)-f^*(\mathbf{u}_{j,t}^s)\right|d\mathbf{y}_{j,t}^s d\mathbf{x}_{j,t}d\mathbf{u}_{j,t}^s$$
$$\le\left(\int f^*(\mathbf{y}_{j,t}^s,\mathbf{x}_{j,t}|\mathbf{u}_{j,t}^s)^2\left|f(\mathbf{u}_{j,t}^s)-f^*(\mathbf{u}_{j,t}^s)\right|d\mathbf{y}_{j,t}^s d\mathbf{x}_{j,t}d\mathbf{u}_{j,t}^s\right)^{1/2}\left(\int\left|f(\mathbf{u}_{j,t}^s)-f^*(\mathbf{u}_{j,t}^s)\right|d\mathbf{u}_{j,t}^s\right)^{1/2}.$$

By assumption the first term is finite and bounded while the second term is a $O(\bar{\rho}_u^{m/2})$. Taking squares on both sides on the inequality concludes the proof. □

Proof of Corollary B3: As discussed in Section B asymptotic are conducted over the cross-sectional dimension n for the moments:

$$\hat{\psi}_j(\tau) = \frac{1}{T}\sum_{t=1}^{T} e^{i\tau'(\mathbf{y}_{j,t},\mathbf{x}_{j,t})}, \quad \hat{\psi}_j^s(\tau) = \frac{1}{T}\sum_{t=1}^{T} e^{i\tau'(\mathbf{y}_{j,t}^s,\mathbf{x}_{j,t})}$$

which are iid under the stated assumptions. The bias can accumulate dynamically for DGP (B.21), 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 B7 shows how the assumed DGPs handle the initial condition problem in the panel setting. Note that:

$$n\bar{\rho}_u^m = e^{\log[n]+m\log[\bar{\rho}_u]} = e^{m(\log[n]/m+\log[\bar{\rho}_u])} \to 0$$

as $m, n \to \infty$ if $\lim_{m,n\to\infty} \log[n]/m < -\log[\bar{\rho}_u] > 0$. Given, this result and the dynamic bias accumulation the results for the iid case apply with an inflation bias term for DGP (B.21). □