A semiparametric approach to marginal likelihood computation for multimodal posterior distributions*

Ryohei Oishi†

September 2026

Abstract

I propose a semiparametric modified harmonic mean estimator of the marginal likelihood that flexibly captures multimodal posterior distributions while remaining computationally feasible. Building on the approach introduced by Sims, Waggoner and Zha (2008), this estimator uses a multidimensional elliptical distribution as the weighting density but nonparametrically approximates a one-dimensional distribution within the ellipsoid. It calculates the normalizing constant of the weighting density by combining rejection sampling from this nonparametric distribution with direct sampling from the ellipsoid. Monte Carlo simulations confirm that the new semiparametric estimator approximates complex distributions more accurately than its parametric counterpart. Furthermore, I show that estimators based on Sims et al. (2008) outperform both sequential Monte Carlo and Geweke (1999) estimators. I also demonstrate that the simplified, simulation-free variants occasionally used in the literature suffer from upward bias.

Keywords: Marginal likelihood, Sequential Monte Carlo, State-space model, Kernel density estimation, Semiparametric estimation

*I am grateful to O. Deniz Akyildiz, Stephen Hansen, Samuel Livingstone, Morten O. Ravn, Yuxuan Ren, Vincent Sterk, and seminar participants at University College London for their comments and discussions.

†University College London, Gower Street, London, WC1E 6BT, UK (e-mail: ryohei.oishi.22@ucl.ac.uk)

1 Introduction

Recent advances in Bayesian statistics and computational methods have enabled researchers to estimate increasingly complex quantitative models. These models often yield nonstandard posterior distributions characterized by multimodality and fat tails. For example, multimodal posterior distributions are reported in standard dynamic stochastic general equilibrium (DSGE) models (Chib and Ramamurthy, 2010; Herbst and Schorfheide, 2014; Cai, Del Negro, Herbst, Matlin, Sarfati and Schorfheide, 2021) and DSGE models with indeterminacy (Bianchi and Nicolò, 2021; Ettmeier and Kriwoluzky, 2024), in addition to Markov-switching vector autoregression (VAR) and DSGE models.

For complex models with nonstandard posterior distributions, estimating the marginal likelihood is computationally challenging due to the tradeoff between flexibility and computational costs.[1] A flexible nonparametric approach requires substantial computational resources and quickly becomes infeasible and less accurate due to the curse of dimensionality. In contrast, a computationally inexpensive parametric approach lacks sufficient flexibility for multimodal distributions.

This tradeoff applies to the widely used harmonic mean estimators (Newton and Raftery, 1994; Gelfand and Dey, 1994) of the marginal likelihood, where the reciprocal of the marginal likelihood is calculated as a weighted harmonic mean of the likelihood evaluated at posterior samples.[2] The parametric weighting density, such as the truncated Gaussian distribution of Geweke (1999), is commonly used in the applied macroeconomic literature thanks to its high accuracy under common econometric setups (Chib and Jeliazkov, 2001; Chan and Grant, 2015; Oishi, 2026); however, its performance is questioned when the posterior distribution is multimodal and highly non-Gaussian (Sims et al., 2008). In contrast, a fully nonparametric weighting density, such as kernel density estimates from posterior samples (McEwen, Wallis, Price and Mancini, 2021), is often

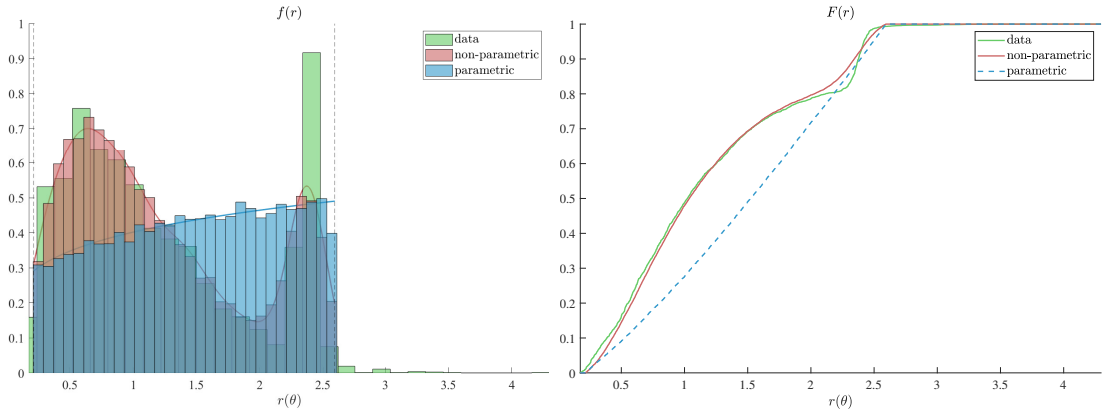
[1] The marginal likelihood is defined as the probability of the data given a model and is also known as the marginal data density, integrated likelihood, and normalizing constant (of the posterior distribution). This is an important quantity in Bayesian statistics, as it is used for model selection, model averaging, and hyperparameter estimation, among other applications.

[2] See Ardia, Baştürk, Hoogerheide and Van Dijk (2012), Friel and Wyse (2012), and Llorente, Martino, Delgado and Lopez-Santiago (2023) for recent surveys.

computationally prohibitive, particularly in high-dimensional settings.

To address these limitations and combine the advantages of parametric and nonparametric approaches, this paper proposes a semiparametric modified harmonic mean estimator for models with multimodal posterior distributions. This estimator retains the general elliptical weighting density introduced by Sims, Waggoner and Zha (2008, SWZ) to overcome the limitations of Geweke's estimator for multimodal posterior distributions. However, the key difference is that it approximates a *one-dimensional* density within the elliptical weighting density nonparametrically, while Sims et al. (2008) do so parametrically with a monotonic function. This semiparametric approach balances computational costs and flexibility by limiting nonparametric approximation to the one-dimensional density, which inherits multimodality from the underlying posterior distribution, rather than approximating the entire posterior distribution nonparametrically.

Figure 1 illustrates the benefit of the nonparametric approach. The figure shows the histogram and cumulative distribution of a random variable r whose density must be approximated for the SWZ estimator, based on simulated data from a multimodal state-space model. The nonparametric approach approximates the left side of the distribution almost perfectly, though it exhibits minor deviations in the right tail; in contrast, the parametric estimator entirely fails to capture the multimodality of the distribution.



(a) Probability density function (b) Cumulative distribution function

Figure 1: Approximated PDF and CDF of a one-dimensional density function in the weighting density of Sims et al. (2008) from a single simulated dataset

Note: This figure illustrates the approximated PDF and CDF of a one-dimensional density function $f(r)$ in the weighting density of Sims et al. (2008) in (5) using simulated data from the multimodal state-space model in Section 4. "parametric" uses the parametric specification in Sims et al. (2008), while "nonparametric" uses the kernel density estimation.

Monte Carlo simulations confirm that the semiparametric approach approximates the multimodal distribution better than its parametric counterpart. Furthermore, both the parametric and semiparametric SWZ estimators outperform the Geweke (1999) estimator and the sequential Monte Carlo estimator. To the best of my knowledge, this is the first evidence in the literature that SWZ estimators, designed to handle non-Gaussian posterior distributions, outperform alternative methods when the posterior is indeed multimodal. I also show that simplified variants occasionally adopted in the literature that bypass simulation when computing the normalizing constant suffer from upward bias. This is because these variants either fix the normalizing constant at a preset value or evaluate it using posterior draws rather than samples from the weighting density, which tends to yield larger estimates of the constant.[3]

The remainder of the paper is organized as follows. Section 2 reviews the modified harmonic mean estimator of the marginal likelihood. Section 3 explains the original SWZ estimator, proposes its semiparametric extension, and details the computational algorithm. Section 4 evaluates the properties of the semiparametric estimator through Monte Carlo simulations, using a state-space model that exhibits a multimodal posterior distribution. Section 5 concludes.

2 Review: The modified harmonic mean estimators of the marginal likelihood

This section defines the marginal likelihood and the modified harmonic mean estimator. Readers familiar with marginal likelihood estimation can skip to the next section.

The marginal likelihood is the probability of observing the data under a given model and plays a crucial role in Bayesian model selection and averaging. Formally, suppose

[3]This simplification likely occurs because the computational algorithm for the SWZ estimator has not, to my knowledge, been fully documented in the literature. For instance, the description in Sims et al. (2008) omits the procedure for sampling from a one-dimensional density within the weighting function, while Herbst and Schorfheide (2016) assumes that we can sample from the weighting density without explaining the implementation. Oishi (2026) recently provides a complete algorithm.

we have a model M parametrized by $\theta \in \Theta_M \subseteq \mathbb{R}^k$, and we observe data y. Then, the marginal likelihood is defined as

$$p(y|M) = \int_{\Theta_M} p(y|\theta, M)p(\theta|M)d\theta, \tag{1}$$

where $p(y|\theta, M)$ is the likelihood and $p(\theta|M)$ is the prior distribution. For notational simplicity, I suppress conditioning on M unless otherwise unclear. Except in a limited number of cases, (1) cannot be solved analytically; hence, we need numerical methods to approximate it.

Equation (1) implies that a naive Monte Carlo estimator could, at least in principle, estimate the marginal likelihood. Specifically, we can draw samples from the prior distribution, compute the likelihoods, and average them. However, this naive estimator is rarely used in practice due to its computational inefficiency, especially in high-dimensional cases. Because the prior is much more dispersed than the likelihood function, most samples contribute negligibly to the integral. The limitations of the naive Monte Carlo estimator led to the development of other estimators, such as harmonic-mean-type estimators (Newton and Raftery, 1994; Gelfand and Dey, 1994).

The modified harmonic mean estimator (Gelfand and Dey, 1994) relies on the following identity and approximation:

$$p(y)^{-1} = E_{p(\theta|y)}\left[\frac{w(\theta)}{p(y|\theta)p(\theta)}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{w(\theta^{(i)})}{p(y|\theta^{(i)})p(\theta^{(i)})}, \tag{2}$$

where $E_{p(\theta|y)}[\cdot]$ denotes the expectation with respect to the posterior distribution, $w(\theta)$ is a weighting density that satisfies $\int_\Theta w(\theta)d\theta = 1$, $\theta^{(i)}$ is a sample from $p(\theta|y)$, N is the number of samples, and $p(y|\theta^{(i)})$ is the likelihood evaluated at the ith posterior sample.

The choice of weighting density is crucial, as it directly affects both computational efficiency and estimation accuracy. To see this, we can rewrite (2) in importance-sampling form:

$$E_{p(\theta|y)}\left[\frac{w(\theta)}{p(y|\theta)p(\theta)}\right] = p(y)^{-1}\int \frac{w(\theta)}{p(\theta|y)}p(\theta|y)d\theta,$$

where the target and proposal distributions are the weighting density and the posterior in the integral, respectively. Importance sampling is more computationally efficient and accurate when the target and proposal distributions exhibit larger overlap. Furthermore, the estimator becomes the exact marginal likelihood (up to a small computational error) when the weighting density equals the posterior distribution.[4]

One of the most widely used weighting densities in economics is the truncated normal distribution proposed by Geweke (1999). Formally, Geweke's estimator uses the following weighting density:

$$w_{Geweke}(\theta) = q^{-1}(2\pi)^{-k/2}|\bar{\Omega}|^{-1/2}\exp\left(-\frac{1}{2}(\theta-\bar{\theta})'\bar{\Omega}^{-1}(\theta-\bar{\theta})\right)\mathbb{1}(\theta\in\bar{\Theta}), \tag{3}$$

where $\bar{\Theta} = \{\theta : (\theta-\bar{\theta})'\bar{\Omega}^{-1}(\theta-\bar{\theta}) \leq F^{-1}_{\chi^2_k}(q)\}$, $\bar{\theta}$ and $\bar{\Omega}$ are the posterior mean and covariance matrix of θ estimated from the posterior samples, k is the dimension of θ, $F^{-1}_{\chi^2_k}(q)$ is the inverse CDF of a χ^2 distribution with k degrees of freedom, and $q \in (0,1)$ is a tuning parameter that determines the area of truncation, which is typically set to $q = 0.9$ or $q = 0.5$. This estimator is particularly convenient because it requires only MCMC output, such as posterior draws and unnormalized posterior densities, without further simulation. Moreover, when the posterior distribution is well approximated by a normal density, Geweke's estimator becomes highly efficient and accurate. However, Sims et al. (2008) argue that Geweke's estimator can be inaccurate when the posterior distribution is far from Gaussian and propose a new weighting density.

3 The Sims–Waggoner–Zha estimator and its semi-parametric extension

This section first reviews the original SWZ estimator, which was developed to handle non-standard posterior distributions, such as multimodal and fat-tailed distributions. Then,

[4] If the absolute continuity condition of importance sampling is not satisfied in simulation, meaning that the empirical posterior support is a subset of the support of the weighting density, the modified harmonic mean estimator is upward-biased in practice. See Lenk (2009) and Oishi (2026) for details.

I introduce the semiparametric extension and provide the computational algorithm.

3.1 The original Sims–Waggoner–Zha estimator

Sims et al. (2008) proposed a new weighting density to handle multimodal posterior distributions more effectively by extending Geweke's estimator in three respects. First, the weighting density takes the more general form of an elliptical density function $g(\theta)$. Second, it centers the distribution around a posterior mode. Third, it introduces an additional truncation criterion based on the unnormalized posterior density.

Formally, they propose the following weighting density:

$$w_{SWZ}(\theta) = q_L^{-1} g(\theta) \mathbb{1}(\theta \in \widehat{\Theta}) \tag{4}$$

where

$$g(\theta) = \frac{\Gamma(k/2)}{2\pi^{k/2}|\widehat{S}|} \frac{f(r(\theta))}{r(\theta)^{k-1}}, \tag{5}$$

$r(\theta) = \sqrt{(\theta - \hat{\theta})'\widehat{\Omega}^{-1}(\theta - \hat{\theta})}$, $\hat{\theta}$ is a posterior mode, $\widehat{\Omega} = \frac{1}{N}\sum_{i=1}^{N}(\theta^{(i)} - \hat{\theta})(\theta^{(i)} - \hat{\theta})'$, $\widehat{S}$ is the lower triangular Cholesky factor of $\widehat{\Omega}$, $f(\cdot)$ is any one-dimensional probability density defined on the positive real numbers, q_L is the normalizing constant ensuring that $\int_\Theta w_{SWZ}(\theta)d\theta = 1$, and $\widehat{\Theta}$ determines the truncation in the parameter space.[5] Their truncation criterion is based on two conditions: $\widehat{\Theta} = \{\theta : p(y|\theta)p(\theta) > L_{1-q}, r(\theta) \in [a, b]\}$ whereL_{1-q} is the empirical $(1-q)$ quantile of the unnormalized posterior densities $\{p(y|\theta^{(i)})p(\theta^{(i)})\}_{i=1}^{N}$.[6] Sims et al. (2008) recommend $q = 0.9$. a, b, and ν are chosen to truncate approximately 10% of the tails of the posterior distribution, with their formal definitions provided below. Importantly, estimating the normalizing constant q_L requires additional simulations outside of the MCMC posterior sampling.

Since the elliptical distribution in (5) only requires $f(r)$ to be a one-dimensional

[5] Some papers define the denominator of $g(\theta)$ as $(2\pi)^{k/2}$, not $2\pi^{k/2}$, but this definition distorts the evaluation of q_L if the direct sampling is employed, as in Sims et al. (2008) and the following algorithm.

[6] In the original article by Sims et al. (2008), they define $\hat{\Theta} = \{\theta : p(y|\theta)p(\theta) > L_{1-q}\}$, meaning that they do not condition on $r(\theta) \in [a, b]$. Here, I follow Herbst and Schorfheide (2016) and impose the condition on $r(\theta)$.

probability density function defined on the positive real numbers, Sims et al. (2008) set it parametrically:

$$f^p(r) = \begin{cases} \frac{\nu r^{\nu-1}}{b^\nu - a^\nu} & \text{if } r \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where a, b and ν are set as follows. Let c_1, c_{10}, and c_{90} be the 1st, 10th, and 90th percentiles of the posterior draws $\{r^{(i)}\}_{i=1}^N$ computed as $r^{(i)} = \sqrt{(\theta^{(i)} - \hat{\theta})' \widehat{\Omega}^{-1} (\theta^{(i)} - \hat{\theta})}$. Then, set $\nu = \frac{\log(1/9)}{\log(c_{10}/c_{90})}, a = c_1$ and $b = \frac{c_{90}}{0.9^{\frac{1}{\nu}}}$ so that $r \in [a, b]$ contains approximately 90% of the posterior samples.

3.2 A semiparametric approach

Since the parametric form of $f(r)$ in Sims et al. (2008) is a monotonic function, it is not sufficiently flexible to approximate multimodal distributions. Therefore, I propose to approximate the empirical distribution of $r(\theta)$ nonparametrically with $f^{np}(r)$ to improve the fit. This novel semiparametric estimator better captures multimodal distributions without significantly increasing computational cost because it limits the nonparametric approximation to a one-dimensional density function, rather than applying it to the entire posterior distribution. Indeed, Figure 1 in Section 1 shows that the nonparametric estimator proposed in this paper fits the multimodal distribution better than the parametric form in (6).

In addition to the computational efficiency gain, improving the fit mitigates the tradeoff between q and the resulting estimate of the normalizing constant $\hat{q}_L$ in practice. Sims et al. (2008) argue that one should increase q when $\hat{q}_L$, which they refer to as the overlap measure, is small, and as a rule of thumb, they recommend $\hat{q}_L \geq 1.0 \times 10^{-5}$. However, larger q also increases the chance of computational instability by introducing extremely small values of the unnormalized posterior density in the denominator of the modified harmonic mean estimator (2). This tradeoff is particularly severe when the model is complex because $\hat{q}_L$ tends to be lower. For example, Sims et al. (2008) report that $\hat{q}_L$ is as low as 1.6×10^{-5} in their Markov-Switching VAR model.

However, the nonparametric approximation of $f(r)$ poses two unique challenges in the current setup. First, since the original SWZ estimator truncates the distribution of $r(\theta)$ such that $r \in (a, b)$, we need to approximate a truncated distribution nonparametrically. Second, we need a numerical method to sample from the truncated nonparametric distribution to obtain $\hat{q}_L$.

To address the first challenge, I propose the following two-step approach. The first step is to estimate $f(r)$ nonparametrically without taking the truncation into account and obtain $\tilde{f}^{np}(r)$. The second step is to compute the truncation area numerically and obtain the truncated nonparametric function $f^{np}(r)$. Specifically, we first evaluate $Q_a = \int_{-\infty}^{a} \tilde{f}^{np}(r)dr$ and $Q_b = \int_{b}^{\infty} \tilde{f}^{np}(r)dr$ by numerical integration, where, in practice, the infinite integration limits are replaced with sufficiently large numerical bounds. Then, we define the nonparametric approximation of $f(r)$ as follows:

$$f^{np}(r) = \begin{cases} \frac{1}{1-Q_a-Q_b}\tilde{f}^{np}(r) & \text{if } r \in [a, b] \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

For $f^{np}(r)$, I use kernel density estimation with a Gaussian kernel.[7] Note, however, that one can choose different kernels depending on the shape of the distribution of $r(\theta)$ in practice. Since the empirical distribution of $r(\theta)$ is observable, we can first plot a histogram and then specify an appropriate kernel. For the bandwidth selection, I use two different methods. The first is Silverman's Gaussian rule of thumb (Silverman, 1986), and the second is the plug-in method developed by Botev, Grotowski and Kroese (2010). While I apply these methods to automate bandwidth selection with small computational cost for the simulation exercise, one can use other methods, such as cross-validation, in practice.

To address the second challenge, I propose using rejection sampling (e.g., Robert and Casella, 2004) to sample from $f^{np}(r)$. Since $f^{np}(r)$ is a one-dimensional probability density, the computational cost of rejection sampling remains small.[8] The next subsection

[7] See Li and Racine (2007); Ullah and Pagan (1999) for kernel density estimation.

[8] The rejection sampling method draws candidates from a proposal distribution and accepts them

presents the algorithm.

3.3 Algorithm

The key difficulty in implementing the SWZ-type estimator is the sampling procedure from $g(\theta)$ to evaluate q_L. In the algorithm below, I follow Sims et al. (2008) and sample directly from $g(\theta)$.[9] The algorithm closely follows Oishi (2026) but replaces the inversion sampler for $f^p(r)$ with the rejection sampler for $f^{np}(r)$.[10]

1. Compute a posterior mode $\hat{\theta}$ and a covariance matrix $\widehat{\Omega}$ centered at the posterior mode. Choose the threshold value q and other parameters, such as a, b and ν. Compute $f^{np}(r)$ in (7).

2. Simulate N^{sim} i.i.d. samples $\{\theta^{(j)}\}_{j=1}^{N^{sim}}$ from $g(\theta)$. Steps 2-1 and 2-2 are rejection sampling from $f^{np}(r)$. Step 2-3 is direct sampling from $g(\theta)$:

   2-1. Sample $r^{(j)}$ from $f^{np}(r)$ by using the rejection sampling. First, draw $r^* \sim \pi$, where π is a proposal distribution.

   2-2. Compute $f^{np}(r^*)$ and accept $r^{(j)} = r^*$ with probability $\frac{f^{np}(r^*)}{M \cdot \pi(r^*)}$, i.e., accept if $u < \frac{f^{np}(r^*)}{M \cdot \pi(r^*)}$ for $u \sim \text{Uniform}[0,1]$, where M is chosen such that $M = \sup_r \frac{f^{np}(r)}{\pi(r)}$. If rejected, go back to the previous step and repeat.

   2-3. Sample $\theta^{(j)}$ by

$$\theta^{(j)} = \frac{r^{(j)}}{||x^{(j)}||} \widehat{S} x^{(j)} + \hat{\theta}$$

   where $||x||$ is the Euclidean norm, $\widehat{S}$ is the lower triangular Cholesky factor of $\widehat{\Omega}$, and $x^{(j)}$ is a random sample from the k-dimensional standard multivariate normal

based on a probability proportional to the target density. See, for instance, Robert and Casella (2004) for more details.

[9] Instead of direct sampling, one could use importance sampling in theory; however, I found that direct sampling is more stable and accurate.

[10] In practice, it is computationally more stable to take the natural log whenever it is possible, even though this algorithm is written without doing so to keep the notation concise. For instance, $\log(p(y))$ is typically reported. Furthermore, when computing the natural log of the weighting density, using the following formula will help avoid computational instability: $\log\left(|\widehat{S}|\right) = \frac{1}{2}\log\left(|\widehat{\Omega}|\right) = \sum_{s=1}^{k} \log\left(\widehat{S}_{s,s}\right)$ where $\widehat{S}_{s,s}$ denotes the sth diagonal element of $\widehat{S}$. The first equality follows because $|A| = |A'|$ for a square matrix A and $\Omega = SS'$. The second equality follows because S is lower triangular.

distribution.

3. Estimate the normalizing constant q_L by $\hat{q}_L = \frac{1}{N^{sim}} \sum_{j=1}^{N^{sim}} \mathbb{1}(\theta^{(j)} \in \widehat{\Theta})$. Increase q if $\hat{q}_L$ is too small.[11]

4. For all $i = 1, 2, ..., N$, compute $w_{SWZ}^{(i)} = w_{SWZ}(\theta^{(i)})$ in (4) with $q_L = \hat{q}_L$ and $f(r) = f^{np}(r)$.

5. Approximate the marginal likelihood $p(y)$ as $p(y) = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{w_{SWZ}^{(i)}}{p(y|\theta^{(i)})p(\theta^{(i)})} \right]^{-1}$.

In this paper, I use Uniform$[a, b]$ for the proposal π. Hence, we have $\pi(r^*) = (b-a)^{-1}$ for all r^* and $M = (b-a)f^*(1+\epsilon)$, where f^* is the maximum of $f^{np}(r)$, which can be evaluated, for example, by a one-dimensional grid search, and ϵ is a small positive number.

4 Monte Carlo simulation using a state-space model

In this section, I use Monte Carlo simulations to evaluate the accuracy of the semi-parametric SWZ estimator and its variants, as well as other popular estimators, such as Geweke's estimator and the sequential Monte Carlo estimator. This simulation exercise uses the state-space model of Herbst and Schorfheide (2016), whose posterior can be multimodal because of global non-identification. This setup is particularly suited for evaluating the SWZ-type estimators, which were specifically developed to handle non-Gaussian and multimodal posterior distributions. Moreover, the specification resembles a modern business cycle model.

4.1 Model

Consider the following state-space model:

$$y_t = [1, 1]s_t, \quad s_t = \begin{bmatrix} \phi_1 & 0 \\ \phi_3 & \phi_2 \end{bmatrix} s_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \tag{8}$$

[11] As a rule of thumb, Sims et al. (2008) recommend $\hat{q}_L \geq 1.0 \times 10^{-5}$.

where $\phi_1 = \theta_1^2, \phi_2 = 1 - \theta_1^2, \phi_3 - \phi_2 = -\theta_1\theta_2$, and the parameters of interest are θ_1 and θ_2. This model can be interpreted as a small-scale business cycle model, where the first state variable is an exogenous driver of a business cycle (such as technology), the second state variable is an endogenous variable (such as capital stock), and the observed variable is a linear combination of these two variables (such as output).

In the model (8), θ_1 and θ_2 are not identified, and the posterior distribution of θ_1 and θ_2 tends to be multimodal.[12] This lack of identification becomes clear when (8) is represented as ARMA models as follows:

$$(1 - \theta_1^2 L)(1 - (1 - \theta_1^2)L)y_t = (1 - \theta_1\theta_2 L)\varepsilon_t, \tag{9}$$

where L is the lag-operator such that $Lx_t = x_{t-1}$.[13] First, θ_2 appears only multiplicatively with θ_1; therefore, if $\theta_1 = 0$, it cannot be identified. Second, plugging $\theta_1^2 = 1 - \tilde{\theta}_1^2$ and $\theta_1\theta_2 = \tilde{\theta}_1\tilde{\theta}_2$ into (9) provides

$$(1 - \tilde{\theta}_1^2 L)(1 - (1 - \tilde{\theta}_1^2)L)y_t = (1 - \tilde{\theta}_1\tilde{\theta}_2 L)\varepsilon_t,$$

which is equivalent to the original model (9). Therefore, $\tilde{\theta}_1 = \sqrt{1 - \theta_1^2}$ and $\tilde{\theta}_2 = \frac{\theta_1\theta_2}{\tilde{\theta}_1}$ yield an observationally equivalent model, which tends to induce a multimodal posterior distribution.

4.2 Simulation setup

For each simulation, I generate $T = 200$ data points following the state-space model (8), with true parameters $\theta_1 = 0.25$ and $\theta_2 = 0.85$. Their prior distributions are uniform on $[0, 1]$. I assume that the initial state is at the unconditional mean, i.e., $s_0 = 0$.

Since the analytical expression for the marginal likelihood is unavailable for (8) under the uniform prior, I use the naive Monte Carlo estimator as a benchmark, which directly

[12] One example of the posterior distribution from this simulation can be found in Supplementary Appendix B.

[13] See Supplementary Appendix A for the derivation.

approximates (1). Specifically, I draw $50,000$ samples from the prior and compute the average likelihood. This Monte Carlo estimator works exceptionally well in the current setup because there are only two parameters, and their prior distributions are uniform over $[0,1]$. I confirm that the Monte Carlo estimator is stable after $10,000$ draws.

To sample from the posterior distribution, I use a sequential Monte Carlo (SMC) sampler, where the likelihood is evaluated by the Kalman filter. The SMC sampler approximates the posterior distribution by iteratively perturbing the likelihood and gradually updating the prior:

$$\pi_n(\theta) = \frac{p(Y|\theta)^{\phi_n} p(\theta)}{\int p(Y|\theta)^{\phi_n} p(\theta) d\theta},$$

where $n = 1, ..., N_\phi$, N_ϕ is the total number of stages, and ϕ_n is an increasing sequence with $\phi_1 = 0$ and $\phi_{N_\phi} = 1$. Then, $\pi_n(\theta)$ corresponds to the posterior distribution when $n = N_\phi$.[14]

4.3 Other marginal likelihood estimators for comparison

In addition to the original parametric and the novel semiparametric SWZ estimator, I compute the marginal likelihood using three different estimators. The first is Geweke's estimator (3).

The second is a simpler variant of the SWZ estimator occasionally used in the literature, which bypasses the simulation for $\hat{q}_L$. Specifically, instead of running the simulation in Step 2 of the aforementioned algorithm, one simply sets:

$$\hat{q}_L = 1 - q \tag{10}$$

[14] I use the algorithm provided in Herbst and Schorfheide (2016, Chap. 5). I generate 5120 particles, and the number of stages is 50. I use a non-linear tempering schedule, as described in Herbst and Schorfheide (2016), with a tempering parameter of $\lambda = 2$. I resample to avoid weight-degeneracy issues when the effective sample size drops below half the number of particles. I repeat the Metropolis-Hastings step 5 times during the mutation and use a Gaussian proposal distribution. I set the target acceptance ratio to 0.25 and adaptively adjust the scaling parameter to achieve this target.

or computes q_L by using the posterior samples $\theta^{(i)}$, not the samples from $g(\theta)$:

$$\hat{q}_L = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{1}(p(y|\theta^{(i)})p(\theta^{(i)}) > L) \times \mathbb{1}(r^{(i)} \in [a,b])\right). \tag{11}$$

These simplifications are especially attractive when evaluating the model likelihood is computationally expensive. Theoretically, however, these simplifications invalidate the assumption that $w_{SWZ}(\theta)$ is a probability density because $\int_\Theta w_{SWZ}(\theta)d\theta = 1$ no longer holds, which distorts the estimate.

The third is the marginal likelihood estimate provided by the SMC sampler as a by-product of the sampling procedure, which is calculated as follows:

$$\hat{p}_{SMC}(y) = \prod_{n=1}^{N_\phi}\left(\sum_{i=1}^{N}\tilde{w}_n^i W_{n-1}^i\right), \tag{12}$$

where $\tilde{w}_n^i$ is the unnormalized weight at the correction step and W_n^i is the normalized weight at the selection step (Herbst and Schorfheide, 2014, 2016).

4.4 Simulation results

Table 1 summarizes the simulation results using the multimodal state-space model (8). I focus on $q = 0.9$, as recommended by Sims et al. (2008), though the results hold when $q = 0.5$. There are three main findings.

First, the semiparametric estimator approximates the multimodal distribution better than the parametric SWZ estimator. Table 1 shows that the semiparametric estimators using either the bandwidth of Silverman (1986) or Botev, Grotowski and Kroese (2010, BGK) achieve a larger overlap measure $\hat{q}_L$ than the parametric SWZ estimator. Specifically, the semiparametric estimator with the BGK bandwidth achieves a mean $\hat{q}_L$ of 0.38 and a minimum of 0.038 (the Silverman bandwidth yields similar results). In contrast, the parametric SWZ estimator yields 0.26 and 0.019, respectively.

Furthermore, the Kolmogorov–Smirnov test implies that the nonparametric approximation ($f^{np}(r)$ in (7)) is more accurate than the original parametric approximation ($f^{p}(r)$

in (6)). The Kolmogorov–Smirnov test statistic evaluates whether the samples from $f^p(r)$ and $f^{np}(r)$ are indistinguishable from the empirical distribution of $r(\theta)$. The sixth row of Table 1 shows that the average of the test statistic is smaller for the semiparametric estimators than for the SWZ estimator, regardless of the bandwidth selection rule. Moreover, the number of simulations that fail to reject the null of the Kolmogorov–Smirnov test at the 1% significance level is 160 and 51 out of 160 for the semiparametric estimator with BGK and Silverman bandwidths, respectively, while the null is rejected in every simulation for the parametric estimator. This result can also be graphically confirmed by Figure 1, where the nonparametric estimator (7) approximates the empirical PDF and CDF of $r(\theta)$ better than the parametric version (6).

Second, both semiparametric and parametric SWZ estimates are more accurate than SMC and Geweke's estimates. The first row of Table 1 shows that the mean errors of the SWZ estimators are almost zero, while those for Geweke's estimator and SMC are 0.42 and 0.13, respectively. The SWZ estimators also achieve a smaller standard deviation of the errors than the other estimators, resulting in a much smaller root mean squared error. I am not aware of prior simulation evidence showing that the SWZ estimators outperform other popular methods for multimodal distributions, as Sims et al. (2008) do not benchmark their estimator against alternatives. Although these differences are "barely worth mentioning" on the Kass and Raftery (1995) scale in Table 2 for this simple state-space model, they could be significantly larger in the complex models studied in the applied macroeconomic literature.

Third, the simplified SWZ estimators overestimate the marginal likelihood. The mean errors of Simple1 and Simple2 are 1.46 and 1.42, respectively, which are classified as "Positive" according to the criteria of Kass and Raftery (1995) in Table 2. When correctly evaluated through simulation, $\hat{q}_L$ tends to be lower than the value implied by the simplified calculation, as it is challenging to capture the multimodal distribution with $g(\theta)$. Consequently, the simplified versions overestimate $\hat{q}_L$, leading to positive estimation errors.

q		SP: Silverman	SP: BGK	SWZ	Simple1	Simple2	Geweke	SMC
0.9	ME	-2.4e-03	-2.4e-03	-2.5e-03	1.46	1.42	0.42	0.13
	Std	0.03	0.03	0.03	0.76	0.78	0.30	0.02
	RMSE	0.03	0.03	0.03	1.65	1.62	0.51	0.13
	mean: q_L	0.37	0.38	0.26	0.9	0.86	-	-
	min: q_L	0.036	0.038	0.019	0.9	0.799	-	-
	mean: KS	0.026	0.009	0.259	-	-	-	-
	$\#(p_{KS} > 0.01)$	51	160	0	-	-	-	-
0.5	ME	-2.3e-03	-2.3e-03	-5.5e-03	1.63	1.58	0.22	0.13
	Std	0.04	0.04	0.04	0.98	1.00	0.29	0.02
	RMSE	0.04	0.04	0.04	1.90	1.87	0.37	0.13
	mean: q_L	0.20	0.20	0.13	0.5	0.48	-	-
	min: q_L	0.011	0.016	0.005	0.5	0.438	-	-
	mean: KS	0.026	0.009	0.259	-	-	-	-
	$\#(p_{KS} > 0.01)$	50	159	0	-	-	-	-

Table 1: Summary of the Monte Carlo simulation using the multimodal state-space model

Note: This table presents the mean error (ME, estimate - Monte Carlo approximation), standard deviation of the error (Std), root mean squared error (RMSE), along with the mean of $\hat{q}_L$ and its minimum value for 160 simulations. The row of mean: KS shows the average Kolmogorov-Smirnov test statistic comparing the estimated $f(r)$ with its empirical counterpart. The bottom row indicates how many times the p-value of the Kolmogorov-Smirnov test statistic exceeded 0.01 across 160 simulations. The truncation parameter q applies to the Geweke and SWZ estimators. The columns "SP: Silverman" and "SP: BGK" are the semiparametric estimators using Gaussian kernel density estimation for $f(r)$ and their bandwidths are chosen by the method of Silverman (1986) and Botev et al. (2010), respectively; "Parametric" is the original SWZ estimator; "Simple 1" and "Simple 2" refer to the simplified SWZ estimators, computed using equations (10) and (11) for $\hat{q}_L$. "Geweke" and "SMC" are the estimates by Geweke's estimator (3) and the sequential Monte Carlo sampler (12). These values are based on 160 simulations.

$\mathbf{2\ ln(B_{1,2})}$	**Grades of Evidence**
0 to 2	Barely worth mentioning
2 to 6	Positive
6 to 10	Strong
> 10	Very strong

Table 2: Kass and Raftery (1995) scale

Note: This table summarizes the interpretation of evidence strength proposed by Kass and Raftery (1995), based on differences in marginal likelihoods when comparing two models.

5 Conclusion

This paper extends the SWZ estimator of the marginal likelihood to handle multimodal posterior distributions. By incorporating nonparametric kernel density estimation, this approach improves the approximation accuracy and mitigates the tradeoff between truncation and computational stability. Monte Carlo simulations confirm that the semipara-

metric SWZ estimator provides a more accurate approximation and yields a larger overlap measure, $\hat{q}_L$, than the parametric version. Furthermore, the study demonstrates that both the parametric and semiparametric SWZ estimators outperform Geweke's estimator and SMC when the posterior distribution is multimodal.

References

Ardia, D., Baştürk, N., Hoogerheide, L. and Van Dijk, H. K. (2012), 'A comparative study of monte carlo methods for efficient evaluation of marginal likelihood', *Computational Statistics & Data Analysis* **56**(11), 3398–3414.

Bianchi, F. and Nicolò, G. (2021), 'A generalized approach to indeterminacy in linear rational expectations models', *Quantitative Economics* **12**(3), 843–868.

Botev, Z. I., Grotowski, J. F. and Kroese, D. P. (2010), 'Kernel density estimation via diffusion', *The Annals of Statistics* **38**(5), 2916 – 2957.

Cai, M., Del Negro, M., Herbst, E., Matlin, E., Sarfati, R. and Schorfheide, F. (2021), 'Online estimation of dsge models', *The Econometrics Journal* **24**(1), C33–C58.

Chan, J. C. and Grant, A. L. (2015), 'Pitfalls of estimating the marginal likelihood using the modified harmonic mean', *Economics Letters* **131**, 29–33.

Chib, S. and Jeliazkov, I. (2001), 'Marginal likelihood from the metropolis–hastings output', *Journal of the American statistical association* **96**(453), 270–281.

Chib, S. and Ramamurthy, S. (2010), 'Tailored randomized block mcmc methods with application to dsge models', *Journal of Econometrics* **155**(1), 19–38.

Ettmeier, S. and Kriwoluzky, A. (2024), 'Active or passive? revisiting the role of fiscal policy during high inflation', *European Economic Review* **170**, 104874.

Friel, N. and Wyse, J. (2012), 'Estimating the evidence–a review', *Statistica Neerlandica* **66**(3), 288–308.

Gelfand, A. E. and Dey, D. K. (1994), 'Bayesian model choice: asymptotics and exact calculations', *Journal of the Royal Statistical Society: Series B (Methodological)* **56**(3), 501–514.

Geweke, J. (1999), 'Using simulation methods for bayesian econometric models: inference, development, and communication', *Econometric reviews* **18**(1), 1–73.

Herbst, E. and Schorfheide, F. (2014), 'Sequential monte carlo sampling for dsge models', *Journal of Applied Econometrics* **29**(7), 1073–1098.

Herbst, E. and Schorfheide, F. (2016), *Bayesian estimation of DSGE models*, Princeton University Press.

Kass, R. E. and Raftery, A. E. (1995), 'Bayes factors', *Journal of the american statistical association* **90**(430), 773–795.

Lenk, P. (2009), 'Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods', *Journal of Computational and Graphical Statistics* **18**(4), 941–960.

Li, Q. and Racine, J. S. (2007), *Nonparametric econometrics: theory and practice*, Princeton University Press.

Llorente, F., Martino, L., Delgado, D. and Lopez-Santiago, J. (2023), 'Marginal likelihood computation for model selection and hypothesis testing: an extensive review', *SIAM review* **65**(1), 3–58.

McEwen, J. D., Wallis, C. G., Price, M. A. and Mancini, A. S. (2021), 'Machine learning assisted bayesian model comparison: learnt harmonic mean estimator', *arXiv preprint arXiv:2111.12720* .

Newton, M. A. and Raftery, A. E. (1994), 'Approximate bayesian inference with the weighted likelihood bootstrap', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **56**(1), 3–26.

Oishi, R. (2026), 'On simulation pseudo-bias and truncation in the modified harmonic mean estimator', *Available at SSRN 6502721* .

Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, Springer Texts in Statistics, 2 edn, Springer, New York, NY.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis / B.W. Silverman.*, Monographs on statistics and applied probability (Series) ; 26, Chapman and Hall, London.

Sims, C. A., Waggoner, D. F. and Zha, T. (2008), 'Methods for inference in large multiple-equation markov-switching models', *Journal of econometrics* **146**(2), 255–274.

Ullah, A. and Pagan, A. (1999), *Nonparametric econometrics*, Cambridge university press Cambridge.

Supplementary Appendix (Not for publication)

A Derivation of the ARMA representation of the state-space model

In Section 4, we consider the following state-space model of Herbst and Schorfheide (2016):

$$y_t = [1,1]s_t, \quad s_t = \begin{bmatrix} \phi_1 & 0 \\ \phi_3 & \phi_2 \end{bmatrix} s_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \tag{8}$$

where $\phi_1 = \theta_1^2, \phi_2 = 1 - \theta_1^2, \phi_3 - \phi_2 = -\theta_1\theta_2$, and the parameters of interest are θ_1 and θ_2. This model can be represented as an ARMA model:

$$(1 - \theta_1^2 L)(1 - (1 - \theta_1^2)L)y_t = (1 - \theta_1\theta_2 L)\varepsilon_t, \tag{9}$$

where L is the lag-operator such that $Lx_t = x_{t-1}$.

The derivation of (9) proceeds as follows. First, write the state-space model as

$$y_t = H(I - \Phi L)^{-1} G\varepsilon_t, \tag{A.1}$$

where $H = [1,1]$, $G = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, and $\Phi = \begin{bmatrix} \phi_1 & 0 \\ \phi_3 & \phi_2 \end{bmatrix}$. Then, since we have

$$(I - \Phi L)^{-1} = \frac{1}{(1 - \phi_1 L)(1 - \phi_2 L)} \begin{bmatrix} 1 - \phi_2 L & 0 \\ \phi_3 L & 1 - \phi_1 L \end{bmatrix},$$

we can write (A.1) as

$$(1-\phi_1 L)(1-\phi_2 L)y_t = [1,1]\begin{bmatrix} 1-\phi_2 L & 0 \\ \phi_3 L & 1-\phi_1 L \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\varepsilon_t$$
$$= (1+(\phi_3-\phi_2)L)\varepsilon_t,$$

which yields the desired ARMA representation in (9).

B A posterior distribution from a single simulated dataset

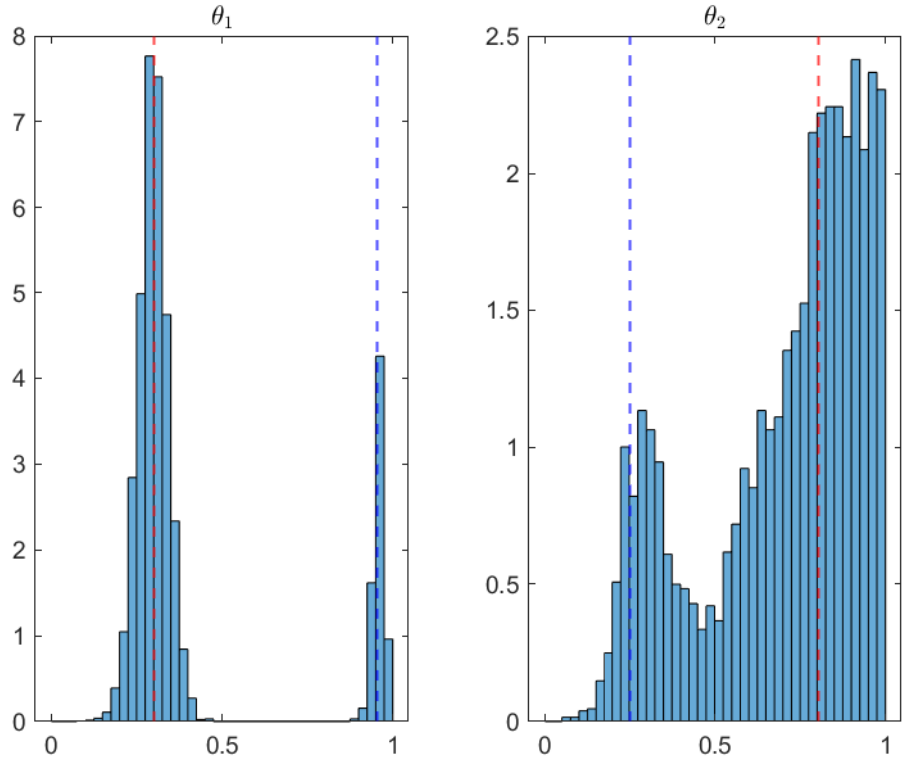


Figure B.1: Posterior distribution in a simulated data from (8)

This figure displays an example of the posterior distributions for θ_1 and θ_2 derived from the state-space model (8), as studied in the Monte Carlo simulations in Section 4. The red lines show the true values of the parameters, while the blue lines show the observationally equivalent parameters, calculated as $\tilde{\theta}_1 = \sqrt{1-\theta_1^2}$ and $\tilde{\theta}_2 = \frac{\theta_1\theta_2}{\tilde{\theta}_1}$.