A Mixture Estimator of Marginal Likelihood

Zhongfang He*

First Version: May 9, 2024

Abstract

This paper develops a new method of computing the marginal likelihood for model comparison in the Bayesian framework. The new method utilizes a simply identity based on a geometric mixture of the unnormalized posterior and an auxiliary distribution of model parameters that nests the popular importance sampling and Gelfand-Dey methods as special cases. By varying a mixing parameter in the geometric mixture, one can easily compute a sequence of marginal likelihood estimates that can be combined to obtain a more accurate estimate. The new method is illustrated in three examples studying economic data and shows good results.

Keywords: Model Comparison, Marginal Likelihood, Model Evidence
JEL Codes: C11, C22, C25, E37

*Email: hezhongfang2004@yahoo.com. Royal Bank of Canada, 155 Wellington St W, Toronto, ON, Canada, M5V 3H6. The views in this paper are solely the author's responsibility and are not related to the company the author works in.

1 Introduction

In the Bayesian framework, the marginal likelihood, i.e. the integral of the model likelihood function with respect to the prior distribution of model parameters including any latent variable, is the central quantify for model specification analysis (e.g. Aitkin (1991), Carlin and Chib (1995) and Kass and Raftery (1995)). However, computing the marginal likelihood in practice is often difficult when the number of model parameters is large (e.g. in models with latent variables). The naive Monte Carlo estimate of marginal likelihood by directly simulating from the prior can be vastly inefficient and unworkable in practice. For improved efficiency, various methods have been proposed in the literature including the importance sampling method (Hammersley and Handscomb (1964), Kloek and Dijk (1978), Geweke (1989)), the harmonic mean method (Newton and Raftery (1994), Gelfand and Dey (1994)), the Chib method (Chib (1995), Chib and Jeliazkov (2001)), the bridge sampling method (Meng and Wong (1996), Meng and Schilling (2002)) and the path sampling method (Gelman and Meng (1998), Friel and Pettitt (2008)), among many others. See Han and Carlin (2001) and Ardia et al. (2012) for reviews of marginal likelihood computation methods.

In this paper, I add to the literature by proposing a new method of computing the marginal likelihood that is both simple and effective. The new method is built on a simple identity for the marginal likelihood that arises from a geometric mixture of the unnormalized posterior and an auxiliary distribution of model parameters indexed by a mixing parameter. It turns out that the importance sampling method and the Gelfand-Dey method (Gelfand and Dey (1994)), which are arguably the most popular methods of computing marginal likelihood in practice, can be derived as special cases of the identity when the mixing parameter equals one and zero respectively. By varying the mixing parameter between zero and one, one can easily produce a sequence of log marginal likelihood estimates

that are asymptotically consistent. By taking a properly weighted average of the sequence, one can obtain a more accurate estimate of marginal likelihood than any individual estimate that uses a specific value of the mixing parameter (e.g. the importance sampling and the Gelfand-Dey methods).

Implementing the new method requires draws of model parameters from the posterior distribution as well as from the auxiliary distribution in the geometric mixture that is constructed to approximate the posterior distribution and is easy to sample from. No additional draws are needed. Thus, the computation cost is low. In all the three empirical examples examined in this paper, estimating the log marginal likelihood and its numerical standard error by the new method takes only a few seconds.

The new method is demonstrated in three examples: a probit model applied to predict economic recessions and two versions of the linear regression model to study the inflation rate, among which one version specifies stochastic volatility and normal regression residual while the other allows both stochastic volatility and fat-tailed regression residual. We find that the numerical standard errors of the log marginal likelihood estimates by the new method are smaller than competing methods that include the importance sampling and Gelfand-Dey methods as well as the optimal bridge sampling method of Meng and Wong (1996) and Meng and Schilling (2002) and the bridge sampling method with the optimal geometric bridge function.

In the remainder of the paper, Section 2 describes the proposed approach in detail. Empirical examples are provided in Section 3. Section 4 concludes. Additional details are provided in the appendices.

2 The Approach

Given the data $\boldsymbol{y}$, the marginal likelihood of a model is $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, where $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the model likelihood function and $p(\boldsymbol{\theta})$ is the prior of model parameters $\boldsymbol{\theta}$ that is assumed to be proper. To compute $p(\boldsymbol{y})$, we introduce an auxiliary distribution $q(\boldsymbol{\theta})$ and construct a geometric mixture $q(\boldsymbol{\theta}, w) = (p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}))^w \, q(\boldsymbol{\theta})^{1-w}$ with a mixing parameter $w \in [0, 1]$. We note that the mixture $q(\boldsymbol{\theta}, w)$ can be written in two equivalent ways:

$$q(\boldsymbol{\theta}, w) = \exp(wf(\boldsymbol{\theta}))q(\boldsymbol{\theta})$$

where $f(\boldsymbol{\theta}) = \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)$, as well as

$$q(\boldsymbol{\theta}, w) = \exp((w-1)f(\boldsymbol{\theta}))p(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{y})$$

where $p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$ is the posterior. Therefore, we can obtain an identity by computing the integral $\int q(\boldsymbol{\theta}, w)d\boldsymbol{\theta}$:

$$\int \exp(wf(\boldsymbol{\theta}))q(\boldsymbol{\theta})d\boldsymbol{\theta} = \int \exp((w-1)f(\boldsymbol{\theta}))p(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{\theta} \tag{1}$$

Re-arranging the items in the identity of Equation (1) leads to an estimator of the marginal likelihood:

$$p(\boldsymbol{y}) = \frac{\int \exp(wf(\boldsymbol{\theta}))q(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \exp((w-1)f(\boldsymbol{\theta}))p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}} \tag{2}$$

A Monte Carlo estimate of the log marginal likelihood (LML) is readily available based on Equation (2):

$$\hat{L}_w \equiv \log(\hat{p}(\boldsymbol{y})) = \log\left(\frac{1}{m}\sum_{j=1}^{m}\exp\left(wf(\boldsymbol{\theta}^{(j)})\right)\right) - \log\left(\frac{1}{m}\sum_{j=1}^{m}\exp\left((w-1)f(\tilde{\boldsymbol{\theta}}^{(j)})\right)\right) \tag{3}$$

where $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^m$ are i.i.d. draws from the auxiliary distribution $q(\boldsymbol{\theta})$ and $\{\tilde{\boldsymbol{\theta}}^{(j)}\}_{j=1}^m$ are draws from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ by a Markov chain Monte Carlo sampler. Note that the number of draws from $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is set to be the same in this paper for convenience.

It is clear that, when the mixing parameter $w = 1$, the estimator of Equation (2) reduces to the importance sampling estimator where the auxiliary distribution $q(\boldsymbol{\theta})$ is the importance sampler. On the other hand, the mixing parameter $w = 0$ will lead to the Gelfand-Dey estimator of Gelfand and Dey (1994), where $q(\boldsymbol{\theta})$ works as the tuning function. Thus, the estimator of Equation (2) nests the importance sampling and Gelfand-Dey estimators as two special cases and combines draws from the auxiliary distribution $q(\boldsymbol{\theta})$ and the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ to compute the marginal likelihood. Relative to the importance sampling and Gelfand-Dey estimators, the mixing parameter w in the estimator of Equation (2) is able to dampen the variances of the integrands and hence helps improve the accuracy of the Monte Carlo estimate when $0 < w < 1$.[1]

It is particularly interesting that the mixing parameter w is not identified in the estimator of Equation (2). That is, any value of $w \in [0, 1]$ will lead to the same estimate of $p(\boldsymbol{y})$. Thus, one can easily compute a sequence of marginal likelihood estimates by specifying a grid for the mixing parameter w. By properly combining such a sequence of marginal likelihood estimates, a final estimate can be obtained that is more accurate than the individual estimates. We dub such an approach as a mixture estimator of marginal likelihood.

Specifically, denote as $\{w_i\}_{i=1}^n$ a grid of values for the mixing parameter w. Let $\hat{L}_{w_i}$ be the Monte Carlo estimate of LML computed via Equation (3) that corresponds to w_i and $\hat{\boldsymbol{L}} = [\hat{L}_{w_1} \ ... \ \hat{L}_{w_n}]'$ be a n-by-1 vector collecting the individual LML estimates. Denote the asymptotic covariance matrix of $\hat{\boldsymbol{L}}$ as $\boldsymbol{\Sigma}$, whose formula is given in Lemma 2.1. Theorem 2.2 establishes that the weighted average $\hat{\boldsymbol{L}}'\boldsymbol{r}^*$, where $\boldsymbol{r}^* = (\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{1}$ and $\mathbf{1}$ is a n-by-1 vector of ones, is an asymptotically consistent estimator of the LML and has a smaller asymptotic variance than any linear combination of the individual estimates in $\hat{\boldsymbol{L}}$

[1]The estimator of Equation (2) is actually valid without requiring $w \in [0, 1]$. However, the restriction $w \in [0, 1]$ avoids the undesirable situations where w further increases the variances of the integrands $\exp(wf(\boldsymbol{\theta}))$ and $\exp((w-1)f(\boldsymbol{\theta}))$.

that converges in probability to the LML, which includes any individual estimate in $\hat{\boldsymbol{L}}$. Replacing $\boldsymbol{\Sigma}$ by its estimate $\hat{\boldsymbol{\Sigma}}$, the (feasible) mixture estimator is:

$$\hat{L}_M = \hat{\boldsymbol{L}}' \hat{\boldsymbol{r}}^* \tag{4}$$

where $\hat{\boldsymbol{r}}^* = (\mathbf{1}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1})^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}$. Estimating the covariance matrix $\boldsymbol{\Sigma}$ is described in Remark 2 below. The numerical standard error of $\hat{L}_M$ can be estimated based on Corollary 2.1.1 as $(m\mathbf{1}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1})^{-1/2}$.[2]

Lemma 2.1. *Let $\boldsymbol{g}(\boldsymbol{\theta})$ be a n-by-1 vector collecting $\{exp(w_i f(\boldsymbol{\theta}))\}_{i=1}^n$ and $\boldsymbol{h}(\boldsymbol{\theta})$ be a n-by-1 vector collecting $\{exp((w_i - 1)f(\boldsymbol{\theta}))\}_{i=1}^n$. Denote $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ as the mean and covariance matrix of $\boldsymbol{g}(\boldsymbol{\theta})$ with respect to the auxiliary distribution $q(\boldsymbol{\theta})$. Similarly, denote $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ as the mean and long-run covariance matrix of $\boldsymbol{h}(\boldsymbol{\theta})$ with respect to the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. Let $\boldsymbol{A}_g$ and $\boldsymbol{A}_h$ be two n-by-n diagonal matrices with the diagonals $1/\boldsymbol{\mu}_g$ and $1/\boldsymbol{\mu}_h$ respectively and $\boldsymbol{\Sigma} = \boldsymbol{A}_g\boldsymbol{\Sigma}_g\boldsymbol{A}_g + \boldsymbol{A}_h\boldsymbol{\Sigma}_h\boldsymbol{A}_h$. We have $\sqrt{m}(\hat{\boldsymbol{L}} - log(p(y))\boldsymbol{1}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$.*

Proof. The proof can be found in Appendix A. □

Corollary 2.1.1. *$\sqrt{m}(\hat{\boldsymbol{L}}'\boldsymbol{r} - log(p(y))) \xrightarrow{d} N(0, \boldsymbol{r}'\boldsymbol{\Sigma}\boldsymbol{r})$, where $\boldsymbol{r}$ is a n-by-1 vector satisfying $\boldsymbol{1}'\boldsymbol{r} = 1$. In particular, $\sqrt{m}(\hat{\boldsymbol{L}}'\boldsymbol{r}^* - log(p(y))) \xrightarrow{d} N(0, (\boldsymbol{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{1})^{-1})$.*

Proof. From Lemma 2.1, we have $\sqrt{m}(\hat{\boldsymbol{L}} - \log(p(y))\mathbf{1}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$. Since $\boldsymbol{r}'\mathbf{1} = 1$, it follows that $\sqrt{m}(\hat{\boldsymbol{L}}'\boldsymbol{r} - \log(p(y))) \xrightarrow{d} N(0, \boldsymbol{r}'\boldsymbol{\Sigma}\boldsymbol{r})$. Note that $\boldsymbol{r}^* = (\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{1}$ satisfies $\mathbf{1}'\boldsymbol{r}^* = 1$. Inserting $\boldsymbol{r}^*$ gives $\sqrt{m}(\hat{\boldsymbol{L}}'\boldsymbol{r}^* - \log(p(y))) \xrightarrow{d} N(0, (\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1})$ □

[2] One might instead treat the grid $\{w_i\}_{i=1}^n$ as random draws of w from a uniform distribution and wonder if an "optimal" distribution of w can be derived based on certain statistical criteria. While this is an interesting angle to explore in future works, in the present paper we prefer a deterministic grid for w as it avoids adding further stochastic variation to the marginal likelihood estimate due to sampling w and simplifies its use in practice.

Theorem 2.2. *The estimator* $\hat{\boldsymbol{L}}'\boldsymbol{r}^*$ *converges in probability to the log marginal likelihood* $log(p(y))$ *and has the lowest asymptotic variance among all the linear combinations* $\hat{\boldsymbol{L}}'\boldsymbol{r}$*, where* $\boldsymbol{r}$ *is a n-by-1 vector of real numbers, that converge in probability to* $log(p(y))$*.*

Proof. The proof can be found in Appendix B. □

Remark 1. Any individual estimate $\hat{L}_{w_i}$ in $\hat{\boldsymbol{L}}$ can be treated as a linear combination $\hat{\boldsymbol{L}}'\boldsymbol{r}$ where the element in $\boldsymbol{r}$ equals 1 at the position corresponding to $\hat{L}_{w_i}$ and equals 0 otherwise. Therefore, Theorem 2.2 can be applied to show that the asymptotic variance of the weighted average $\hat{\boldsymbol{L}}'\boldsymbol{r}^*$ is smaller than that of any individual estimate $\hat{L}_{w_i}$ in $\hat{\boldsymbol{L}}$.

Remark 2. An estimate of $\boldsymbol{\Sigma}_g$ can be obtained by the sample covariance matrix of $\{\boldsymbol{g}(\boldsymbol{\theta}_j)\}_{j=1}^m$ where $\{\boldsymbol{\theta}_j\}_{j=1}^m$ are i.i.d. draws from the auxiliary distribution $q(\boldsymbol{\theta})$. For the long-run covariance matrix $\boldsymbol{\Sigma}_h$, there exist many estimation methods (e.g. see the textbook Hayashi (2000)) that utilize $\{\boldsymbol{h}(\tilde{\boldsymbol{\theta}}_j)\}_{j=1}^m$ where $\{\tilde{\boldsymbol{\theta}}_j\}_{j=1}^m$ are posterior draws. In this paper, we use the Newey-West method (Newey and West (1987)) to estimate $\boldsymbol{\Sigma}_h$.[3] Replacing the population means $\boldsymbol{\mu}_g$ and $\boldsymbol{\mu}_h$ by their sample counterparts, an estimate of the matrices $\boldsymbol{A}_g$, $\boldsymbol{A}_h$ and thus the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{A}_g\boldsymbol{\Sigma}_g\boldsymbol{A}_g + \boldsymbol{A}_h\boldsymbol{\Sigma}_h\boldsymbol{A}_h$ can be readily computed.

Remark 3. In cases where the matrix $\hat{\boldsymbol{\Sigma}}$ is near singular, it can be difficult to compute its inverse for implementing the mixture estimator. A practical workaround is adding a matrix $\epsilon\mathbf{I}_n$ to $\hat{\boldsymbol{\Sigma}}$, where ϵ is a small positive number (e.g. 1e-10) and $\mathbf{I}_n$ is a n-by-n matrix. Thus the (practical) mixture estimator is $\hat{\boldsymbol{L}}'(\hat{\boldsymbol{\Sigma}} + \epsilon\mathbf{I}_n)^{-1}\mathbf{1}\left(\mathbf{1}'(\hat{\boldsymbol{\Sigma}} + \epsilon\mathbf{I}_n)^{-1}\mathbf{1}\right)^{-1}$. The numerical standard error of the LML estimate should be adjusted accordingly as $m^{-1/2}\left(\mathbf{1}'(\hat{\boldsymbol{\Sigma}} + \epsilon\mathbf{I}_n)^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}} + \epsilon\mathbf{I}_n)^{-1}\mathbf{1}\right)^{1/2}\left(\mathbf{1}'(\hat{\boldsymbol{\Sigma}} + \epsilon\mathbf{I}_n)^{-1}\mathbf{1}\right)^{-1}$.

[3]The number of lags when implementing the Newey-West method is $floor\left(4(m/100)^{2/9}\right)$ where m is the number of posterior draws.

2.1 Choice of the Auxiliary Distribution

Regarding the auxiliary distribution $q(\boldsymbol{\theta})$, the estimator of marginal likelihood in Equation (2) only requires that the density function of $q(\boldsymbol{\theta})$ can be easily computed and it is easy to sample from $q(\boldsymbol{\theta})$. So it seems that any probability distribution satisfying these two conditions can be used. However, in reality, an inappropriate choice of $q(\boldsymbol{\theta})$ could lead to nonsensical estimate of marginal likelihood with extremely large sampling variance.

Since the importance sampling and Gelfand-Dey methods are special cases of the mixture estimator, the requirement for a good importance sampler in the importance sampling method or a good tuning function in the Gelfand-Dey method provides a clue for selecting $q(\boldsymbol{\theta})$. Inspecting the definition of $f(\boldsymbol{\theta})$, a Monte Carlo estimate of $p(\boldsymbol{y})$ by Equation (2) would have zero sampling variance if the auxiliary distribution $q(\boldsymbol{\theta})$ equals the posterior $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. In general when this ideal is unachievable, one should aim to find a $q(\boldsymbol{\theta})$ such that it is close to $p(\boldsymbol{\theta}|\boldsymbol{y})$ to minimize the variance of the Monte Carlo estimate. Such a target is the same as in the case of finding a good importance sampler or tuning function for the importance sampling and Gelfand-Dey methods.

In this paper, we use a simple approach to calibrate $q(\boldsymbol{\theta})$ and focus on comparing the performance of the mixture estimator with competing methods under a common choice of $q(\boldsymbol{\theta})$. Assume that all the elements in $\boldsymbol{\theta}$ take value on the whole real line.[4] When the size of $\boldsymbol{\theta}$ is modest, we calibrate $q(\boldsymbol{\theta})$ as a Gaussian distribution with the mean and covariance matrix equal the posterior mean and covariance matrix of $\boldsymbol{\theta}$.

For models with latent variables, the size of $\boldsymbol{\theta}$ is large. We adopt the cross entropy approach of Chan (2023) to calibrate $q(\boldsymbol{\theta})$ that minimizes the Kullback-Leibler divergence between $q(\boldsymbol{\theta})$ and $p(\theta|y)$. We divide $\boldsymbol{\theta}$ into two groups: the latent variables $\boldsymbol{z} = \{\boldsymbol{z}_t\}_{t=1}^T$,

[4]If any element in $\boldsymbol{\theta}$ is restricted, it can be re-parameterized such that the support of the re-parameterized element is the whole real line. The prior of this element should be adjusted accordingly by the change-of-variables formula.

where T is the data sample size, and the other fixed model parameters $\boldsymbol{\delta}$. The auxiliary distribution $q(\boldsymbol{\theta})$ is factorized accordingly as $q(\boldsymbol{\delta})q(\boldsymbol{z}|\boldsymbol{\delta})$. The part $q(\boldsymbol{\delta})$ is calibrated as a Gaussian distribution whose mean and covariance matrix equal the posterior mean and covariance matrix of $\boldsymbol{\delta}$. For the other part $q(\boldsymbol{z}|\boldsymbol{\delta})$, we construct it as $q(\boldsymbol{z}_1)\prod_{t=2}^{T} q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$, where $q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ for $t > 1$ is a Gaussian distribution of $\boldsymbol{z}_t$ with the mean $\boldsymbol{a}_t + \text{diag}(\boldsymbol{b}_t)\boldsymbol{z}_{t-1} + \boldsymbol{C}_t\boldsymbol{\delta}$ and the covariance matrix $\boldsymbol{D}_t$ while $q(\boldsymbol{z}_1)$ is a Gaussian distribution with the mean $\boldsymbol{a}_1 + \boldsymbol{C}_1\boldsymbol{\delta}$ and the covariance matrix $\boldsymbol{D}_1$. The parameters $\{\boldsymbol{a}_t, \boldsymbol{C}_t, \boldsymbol{D}_t\}_{t=1}^{T}$ and $\{\boldsymbol{b}_t\}_{t=2}^{T}$ are vectors and matrices of conformable sizes. Note that we choose a diagonal matrix $\text{diag}(\boldsymbol{b}_t)$ for the autoregressive coefficients to simplify the computation. The free parameters $\{\boldsymbol{a}_t, \boldsymbol{b}_t, \boldsymbol{C}_t\}_{t=2}^{T}$ are calibrated by running regressions of $\boldsymbol{z}_t$ on a conformable vector of ones, $\boldsymbol{z}_{t-1}$ and $\boldsymbol{\delta}$ based on posterior draws of $\boldsymbol{z}_t$, $\boldsymbol{z}_{t-1}$ and $\boldsymbol{\delta}$ for each $t = 2, ..., T$. Similarly, the free parameters $\{\boldsymbol{a}_1, \boldsymbol{C}_1\}$ are calibrated by running a regression of $\boldsymbol{z}_1$ on a conformable vector of ones and $\boldsymbol{\delta}$ based on posterior draws of $\boldsymbol{z}_1$ and $\boldsymbol{\delta}$. The covariance matrices $\{\boldsymbol{D}_t\}_{t=1}^{T}$ are calibrated as the sample covariance matrices of the regression residuals.

2.2 Relation with the Bridge Sampling Method

The marginal likelihood estimator of Equation (2) is reminiscent of the bridge sampling method of Meng and Wong (1996) that, by using the notations in this paper, utilizes the equation:

$$p(\boldsymbol{y}) = \frac{\int \alpha(\boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \alpha(\boldsymbol{\theta}) q(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}} \tag{5}$$

where $\alpha(\boldsymbol{\theta})$ is an auxiliary function of $\boldsymbol{\theta}$ and is often called the bridge function. In fact, by specifying the bridge function $\alpha(\boldsymbol{\theta}) = \left((p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}))^{1-w} q(\boldsymbol{\theta})^{w}\right)^{-1}$, the resulting bridge sampling estimator is identical to the marginal likelihood estimator of Equation (2). In this sense, Equation (2) is a special case of the bridge sampling estimator.

What is unique about the approach of this paper is that the bridge function used in

Equation (2) introduces an additional mixing parameter w, which effectively leads to a space of bridge functions indexed by w. Instead of searching for an optimal bridge function as in Meng and Wong (1996), this paper makes use of the feature that the mixing parameter w is unidentified and thus a sequence of marginal likelihood estimates by Equation (2) can be easily computed by varying the value of w. Properly combining such a sequence of marginal likelihood estimates is able to improve the accuracy over an individual estimate. In the empirical examples, we will show that the proposed approach outperforms the bridge sampling method. Details of the LML estimator by the bridge sampling method as well as its numerical standard error are provided in Appendix C.

The above comparison clearly leads to the question of whether alternative spaces of bridge functions might be better than the geometric mixing one used in this paper. For example, the optimal bridge function of Meng and Wong (1996) appears to suggest that the arithmetic mixing function $\alpha(\boldsymbol{\theta}) = \left(w\left(p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right) + (1-w)q(\boldsymbol{\theta})\right)^{-1}$ is an interesting alternative to investigate. We leave further explorations of this route for future work and focus on the geometric mixing function in the present paper.

3 Empirical Examples

The proposed approach is illustrated in three econometric models applied to real economic data. The first example is a binary probit model applied to predict economic recessions. The second example is a linear regression model with stochastic volatility applied to predict the inflation rate. The third example expands the model in the second example to allow fat tails in the regression residual.

Implementing the mixture estimator needs to specify the grid of the mixing parameter w. It may appear that a dense grid between 0 and 1 would produce more individual LML estimates for computing the average and hence be more efficient. However, a dense grid

of w will result in a large covariance matrix of the individual LML estimates that can be difficult to be estimated accurately. Moreover, the individual LML estimates based on adjacent values of w in a dense grid would be highly correlated and could lead to numerically unstable estimate of the covariance matrix of the individual LML estimates. In this paper, we set the grid of w to be $\{0, 0.02, ..., 0.98, 1\}$ with a total of 51 points.[5] Given that, in all the three empirical exercises, we use 10,000 samples from $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$ respectively for computing the LML, the number of free parameters when estimating $\boldsymbol{\Sigma}$ with such a grid of w is about 13% of the number of samples from $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$.

In each empirical example, we compare the LML estimate by the proposed mixture method with those by the bridge sampling method under 4 bridge functions: the optimal bridge function of Meng and Wong (1996) and Meng and Schilling (2002) and the geometric function with $w = 1$ (the importance sampling method), $w = 0$ (the Gelfand-Dey method) and $w = w^*$ where w^* is the value of the mixing parameter that leads to the LML estimate with the smallest numerical standard error in the grid for applying the mixture estimator (labeled as the *minimum variance* method for convenience).

3.1 Binary Probit Regression

The probit model specifies the model likelihood function $p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p(y_t|\boldsymbol{\theta})$, where $p(y_t|\boldsymbol{\theta}) = \Phi(\boldsymbol{x}_t'\boldsymbol{\theta})^{y_t} (1 - \Phi(\boldsymbol{x}_t'\boldsymbol{\theta}))^{1-y_t}$, $y_t \in \{0, 1\}$ is the binary response, $\boldsymbol{x}_t$ is a J-by-1 vector of regressors, $\boldsymbol{\theta}$ is the regression coefficients and $\Phi(\cdot)$ is the cumulative density function of a standard Gaussian distribution. The prior $N(\mathbf{0}, \sigma^2 \mathrm{I}_J)$ is placed for $\boldsymbol{\theta}$, where σ^2 is the prior variance common to all regression coefficients and I_J is a J-by-J identity matrix. In estimation, the prior variance σ^2 is set to be 100.

The probit model is applied to predict economic recessions. See Nissila (2020) for a

[5] We also experimented with alternative grids of w that divide the interval $[0, 1]$ evenly by 41 and 61 points. The results are very close to the ones using the grid of 51 points.

recent survey on probit-based models for recession prediction. Data on the binary response is the NBER-based U.S. recession indicator. Table D1 in Appendix D lists the 11 employed predictors including macroeconomic variables such as the inflation rate as well as financial variables such as the stock market return and the credit spread. The data sample is quarterly from Q3 1953 to Q2 2021 with a total of 272 observations. Including a constant, there are a total of 12 regressors in the model. The prediction horizon is 4 quarters. Posterior draws of $\boldsymbol{\theta}$ are simulated by the method in Liu and Wu (1999). We discard 2,000 burn-ins and keep the next 10,000 posterior draws for analysis.

Panel A of Table 1 compares the LML estimates of the probit model by the mixture method and 4 alternative methods. For this relatively simple example, the LML estimates by these 5 methods are fairly close. The numerical standard error of the LML estimate by the mixture method is the smallest (0.00253), followed by the optimal bridge sampling method (0.00254) and the minimum variance method (0.00267). The numerical standard errors of the LML estimates by the importance sampling (0.00392) and Gelfand-Dey (0.00476) methods are substantially larger.

Panel A of Figure 1 shows the sequence of LML estimates by Equation (3) corresponding to the grid of w as well as the 95% confidence intervals of the LML estimates (that is, the LML estimates plus and minus 1.96 times their numerical standard errors). It can be seen that, for this example, the LML estimates steadily increase with w. The confidence intervals are narrower towards the middle of the grid of w. The minimum of the numerical standard errors of the LML estimate sequence occurs at $w = 0.64$.

3.2 SV-Normal Linear Regression

The volatility of economic time series is often found to fluctuate over time. To capture this phenomenon, the linear regression model is usually expanded to allow stochastic volatility

(SV) for studying economic time series:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \exp\left(\frac{z_t}{2}\right)\epsilon_t \tag{6}$$

$$z_t = (1-\rho)\mu + \rho z_{t-1} + s\eta_t$$

for $t = 1, ..., T$, where y_t is the scalar regressand, $\boldsymbol{x}_t$ is a J-by-1 vector of regressors, ϵ_t and η_t are i.i.d. standard Gaussian residuals, and z_t is the log volatility with the initial value $z_1 = \mu + \sqrt{\frac{s^2}{1-\rho^2}}\eta_1$. For Bayesian analysis, the priors for the fixed model parameters are $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2 \mathrm{I}_J)$, $\mu \sim N(0, \sigma_\mu^2)$, $\frac{1+\rho}{2} \sim Beta(a_{\rho,0}, b_{\rho,0})$, and $s \sim N(0, \sigma_s^2)$. As discussed in Section 2.1, the fixed model parameters are re-parameterized when needed such that they all take value on the whole real line. Thus, the group of fixed model parameters is a $(J+3)$-by-1 vector $\boldsymbol{\delta} = [\boldsymbol{\beta}'\ \mu\ \log\left(\frac{1+\rho}{1-\rho}\right)\ \log(s^2)]'$.[6] The priors of the transformed parameters are adjusted by using the change-of-variables formula.

The SV-Normal regression model is applied to predict the quarter-to-quarter change of the U.S. inflation rate as measured by the quarterly change of log GDP deflator. A total of 20 exogenous predictors are considered including real activity variables, interest rates and other macroeconomic indicators. A list of the exogenous variables and their descriptions are provided in Table E1. Along with the intercept and an autoregressive lag, the one-quarter-ahead prediction model includes 22 regressors. The data sample is quarterly from Q2 1966 to Q1 2021 with a total of 220 observations. The hyper parameters of the priors are $\sigma^2 = 100$, $\sigma_\mu^2 = 10$, $a_{\rho,0} = 100$, $b_{\rho,0} = 100/19$ and $\sigma_s^2 = 1$. An Metropolis-within-Gibbs sampler is applied to estimate the model where estimation of the SV process is based on the algorithm of Kastner and Fruhwirth-Schnatter (2014). Analysis is based on 10,000 posterior draws after 2,000 burn-ins.

In this SV-Normal regression model, the model parameters $\boldsymbol{\theta}$ include the latent variable

[6] Though the prior of s is Gaussian, its posterior is often bi-modal. The re-parametrization $\log(s^2)$ helps restore a bell-shaped posterior distribution to facilitate constructing the importance sampler.

$\boldsymbol{z} = \{z_t\}_{t=1}^T$ and the fixed model parameters $\boldsymbol{\delta}$. Following the discussions in Section 2.1, the auxiliary distribution $q(\boldsymbol{\theta})$ is decomposed as $q(\boldsymbol{\delta})q(\boldsymbol{z}|\boldsymbol{\delta})$. The part $q(\boldsymbol{\delta})$ is constructed as a Gaussian distribution with the mean and covariance matrix equal the posterior mean and covariance matrix of $\boldsymbol{\delta}$. For the part $q(\boldsymbol{z}|\boldsymbol{\delta})$, a sequence of adaptive Gaussian distributions are fitted to each z_t, $t = 1, ..., T$, based on posterior draws of $\boldsymbol{z}$ and $\boldsymbol{\delta}$. The resulting LML estimates and their numerical standard errors by the mixture method and the alternative methods are reported in Panel B of Table 1.

For this more complicated model, the difference in the LML estimates by the 5 methods is larger than in the case of the probit model of Section 3.1. Among the LML estimates, the one by the mixture method has the smallest numerical standard error (0.031), followed by the optimal bridge sampling method (0.042) and the minimum variance method (0.047). The LML estimates by the importance sampling and Gelfand-Dey methods have substantially larger numerical standard errors (0.598 and 0.086 respectively).

Panel B of Figure 1 shows that the sequence of LML estimates by Equation (3) exhibits an "S" shape when plotted against the grid of w. The LML estimate moves noticeably when the value of w changes. Due to occasional "bad" draws from the auxiliary distribution $q(\theta)$, the 95% confidence intervals of the LML estimates near the end of $w = 1$ (i.e. the importance sampling estimator) are much wider than those near the other end $w = 0$ (i.e. the Gelfand-Dey estimator). By taking an optimally weighted average over the sequence of LML estimates, the mixture estimator is able to significantly reduce the numerical standard error of the LML estimate.

3.3 SV-Student Linear Regression

Greater flexibility can be obtained by specifying a student-t distribution for the regression residuals in the SV-Normal regression model of Equation (6) to capture possible fat tails. That is, $\epsilon_t \sim t(v)$, where $v > 0$ is the degrees of freedom parameter.

The prior for v is $IG(a_{v,0}, b_{v,0})$ where IG denotes the inverse gamma distribution. We re-parameterize $v^* = \log(v)$ such that the support of the new parameter is the whole real line. In estimation, the hyper parameters for v are $a_{v,0} = 7$ and $b_{v,0} = 60$. The priors of the other model parameters are the same as in the Gaussian residual case of Section 3.2.

The model is applied to the same dataset of inflation rate in Section 3.2 and is estimated by a Metropolis-within-Gibbs sampler. In the model estimation, the student-t distribution $\epsilon_t \sim t(v)$ is re-framed as a hierarchical Gaussian distribution $\epsilon_t \sim N(0, d_t)$ and $d_t \sim IG\left(\frac{v}{2}, \frac{v}{2}\right)$ for $t = 1, ..., T$. To improve sampling efficiency, the parameters μ and s in the SV process are sampled by integrating out the linear coefficients $\boldsymbol{\beta}$ and are drawn by a Metropolis-Hastings step with a random walk proposal tuned adaptively by the approach of Garthwaite et al. (2016). Similar to the Gaussian residual case of Section 3.2, the interweaving strategy of Kastner and Fruhwirth-Schnatter (2014), which originates from Yu and Meng (2011), is applied to further boost the sampling efficiency of μ and s. The degrees of freedom parameter v is sampled by a Metropolis-Hastings step with a random walk proposal tuned by the approach of Garthwaite et al. (2016). A total of 10,000 posterior draws are collected for analysis after 2,000 burn-ins.

Denote the fixed model parameters $\boldsymbol{\delta} = [\boldsymbol{\beta}' \ \mu \ \log\left(\frac{1+\rho}{1-\rho}\right) \ \log(s^2) \ \log(v)]'$. The auxiliary distribution $q(\boldsymbol{\theta}) = q(\boldsymbol{\delta})q(\boldsymbol{z}|\boldsymbol{\delta})$ is constructed the same way as for the SV-Normal regression model of Section 3.2. Panel C of Table 1 shows the resulting LML estimates and their numerical standard errors by the mixture method and the other 4 alternative methods. The LML estimates by the mixture, optimal bridge sampling and minimum variance methods are close to each other while the LML estimates by the importance sampling and Gelfand-Dey methods are farther apart. The ranking of the 5 methods in terms of the numerical standard errors of the LML estimates remains the same as in the two other empirical exercises: the numerical standard error of the LML estimate by the mixture estimator is the smallest (0.037), followed by the optimal bridge sampling method (0.045), the minimum

variance method (0.049), and the importance sampling and Gelfand-Dey methods (0.166 and 0.111 respectively).

Panel C of Figure 1 shows the sequence of LML estimates by Equation (3) based on the grid of w and their 95% confidence intervals. Similar to the SV-Normal regression model, the LML sequence shows an "S" shape against the grid of w. All the LML estimates produced by varying the mixing parameter w fall between the importance sampling and Gelfand-Dey estimates (i.e. the two end points in the LML sequence corresponding to $w = 1$ and $w = 0$). The confidence intervals around the middle of the grid of w are narrower than those near the two end points $w = 1$ and $w = 0$. The minimum variance estimator is at the point $w = 0.50$ in this exercise.

4 Conclusion

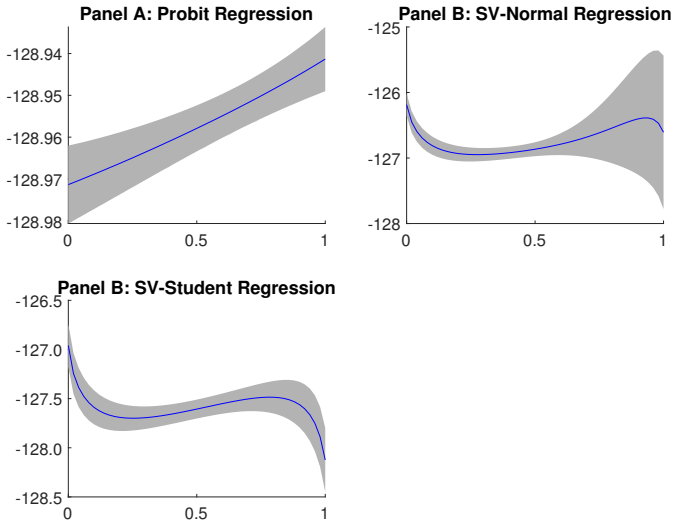
We utilize a simple identity of the marginal likelihood produced by integrating a geometric mixture of the unnormalized posterior and an auxiliary distribution of model parameters that nests the popular importance sampling and Gelfand-Dey methods as special cases. By varying the value of the mixing parameter in the geometric mixture, we develop a new method that combines a sequence of marginal likelihood estimates and leads to a more accurate marginal likelihood estimate as demonstrated in a number of empirical applications to economic data. We hope that the new method can be useful towards the ultimate goal of convenient and routine calculation of marginal likelihoods in applied Bayesian works.

Table 1: Comparing Log Marginal Likelihood Estimates

Panel A: Probit Regression		
Mixture	-128.95	(0.00253)
Optimal Bridge Sampling	-128.95	(0.00254)
Minimum Variance ($w = 0.64$)	-128.95	(0.00267)
Importance Sampling	-128.94	(0.00392)
Gelfand-Dey	-128.97	(0.00476)
Panel B: SV-Normal Regression		
Mixture	-126.76	(0.031)
Optimal Bridge Sampling	-126.83	(0.042)
Minimum Variance ($w = 0.40$)	-126.92	(0.047)
Importance Sampling	-126.61	(0.598)
Gelfand-Dey	-126.19	(0.086)
Panel C: SV-Student Regression		
Mixture	-127.89	(0.037)
Optimal Bridge Sampling	-127.52	(0.045)
Minimum Variance ($w = 0.50$)	-127.61	(0.049)
Importance Sampling	-128.12	(0.166)
Gelfand-Dey	-126.96	(0.111)

Note: This table compares the log marginal likelihood estimates by the mixture, optimal bridge sampling, minimum variance, importance sampling and Gelfand-Dey methods for the three examples of probit regression, linear regression with stochastic volatility and normal residual (SV-Normal regression), and linear regression with stochastic volatility and student-t residual (SV-Student regression). The number of draws is 10,000 in all cases. Numbers in the brackets are the numerical standard errors of the log marginal likelihood estimates.

Figure 1: Sequence of Log Marginal Likelihood Estimates



Note: This figure shows the sequence of log marginal likelihood estimates by Equation (3) against the grid of the mixing parameter w for the three examples of probit regression, linear regression with stochastic volatility and normal residual (SV-Normal regression), and linear regression with stochastic volatility and student-t residual (SV-Student regression). The shaded area is the sequence of the LML estimates plus and minus 1.96 times their numerical standard errors. The number of draws is 10,000 in all cases.

Appendix

A Proof of Lemma 2.1

Proof. Let $\{\boldsymbol{\theta}_j\}_{j=1}^m$ be i.i.d. draws from the auxiliary distribution $q(\boldsymbol{\theta})$ and $\{\boldsymbol{g}(\boldsymbol{\theta}_j)\}_{j=1}^m$ be the corresponding values of $\boldsymbol{g}(\boldsymbol{\theta})$. Applying the central limit theorem for i.i.d. random variables can show that the scaled average $\sqrt{m}(\bar{\boldsymbol{g}}-\boldsymbol{\mu}_g)$, where $\bar{\boldsymbol{g}} = \frac{1}{m}\sum_{j=1}^m \boldsymbol{g}(\boldsymbol{\theta}_j)$ and $\boldsymbol{\mu}_g$ is a n-by-1 vector stacking $\int \exp(w_1 f(\boldsymbol{\theta}))q(\boldsymbol{\theta})d\boldsymbol{\theta}$, ..., $\int \exp(w_n f(\boldsymbol{\theta}))q(\boldsymbol{\theta})d\boldsymbol{\theta}$, follows a zero-mean Gaussian distribution asymptotically. The covariance matrix of the asymptotic Gaussian distribution is $\boldsymbol{\Sigma}_g = \text{var}(\boldsymbol{g}(\boldsymbol{\theta}))$. Applying the delta method can show that $\sqrt{m}(\log(\bar{\boldsymbol{g}}) - \log(\boldsymbol{\mu}_g))$ converges in distribution to $N(\boldsymbol{0}, \boldsymbol{A}_g\boldsymbol{\Sigma}_g\boldsymbol{A}_g)$, where $\boldsymbol{A}_g$ is a n-by-n diagonal matrix with the diagonal elements $1/\boldsymbol{\mu}_g$.

Similarly, let $\{\tilde{\boldsymbol{\theta}}_j\}_{j=1}^m$ be draws from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ by some Markov chain Monte Carlo sampler and $\{\boldsymbol{h}(\tilde{\boldsymbol{\theta}}_j)\}_{j=1}^m$ be the corresponding values of $\boldsymbol{h}(\boldsymbol{\theta})$. Assume that the conditions for the Markov chain central limit theorem are satisfied, the scaled average $\sqrt{m}(\bar{\boldsymbol{h}} - \boldsymbol{\mu}_h)$, where $\bar{\boldsymbol{h}} = \frac{1}{m}\sum_{j=1}^m \boldsymbol{h}(\tilde{\boldsymbol{\theta}}_j)$ and $\boldsymbol{\mu}_h$ is a n-by-1 vector stacking $\int \exp((w_1 - 1)f(\boldsymbol{\theta}))p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$, ..., $\int \exp((w_n - 1)f(\boldsymbol{\theta}))p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$, follows a zero-mean Gaussian distribution asymptotically with the asymptotic covariance matrix $\boldsymbol{\Sigma}_h$ being the long-run covariance matrix of $\boldsymbol{h}(\boldsymbol{\theta})$. Applying the delta method can show that $\sqrt{m}(\log(\bar{\boldsymbol{h}}) - \log(\boldsymbol{\mu}_h))$ converges in distribution to $N(\boldsymbol{0}, \boldsymbol{A}_h\boldsymbol{\Sigma}_h\boldsymbol{A}_h)$, where $\boldsymbol{A}_h$ is a n-by-n diagonal matrix with the diagonal elements $1/\boldsymbol{\mu}_h$.

The sequence $\hat{\boldsymbol{L}}$ can be written as $\log(\bar{\boldsymbol{g}}) - \log(\bar{\boldsymbol{h}})$ based on Equation (3). It is clear from Equation (2) that $\log(\boldsymbol{\mu}_g) - \log(\boldsymbol{\mu}_h) = \log(p(y))\boldsymbol{1}$ where $\boldsymbol{1}$ is a n-by-1 vector of ones. Since draws from the auxiliary distribution $q(\boldsymbol{\theta})$ and the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ are mutually independent, it follows that $\sqrt{m}(\hat{\boldsymbol{L}} - \log(p(y))\boldsymbol{1})$ converges in distribution to $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \boldsymbol{A}_g\boldsymbol{\Sigma}_g\boldsymbol{A}_g + \boldsymbol{A}_h\boldsymbol{\Sigma}_h\boldsymbol{A}_h$.

□

B Proof of Theorem 2.2

Proof. Based on Lemma 2.1, we have that $\sqrt{m}(\hat{\boldsymbol{L}}'\boldsymbol{r} - \log(p(y))\mathbf{1}'\boldsymbol{r})$ converges in distribution to $N(0, \boldsymbol{r}'\boldsymbol{\Sigma}\boldsymbol{r})$. It is clear that the product $\mathbf{1}'\boldsymbol{r}$ should equal 1 for $\hat{\boldsymbol{L}}'\boldsymbol{r}$ to converge in probability to $\log(p(y))$. In particular, $\boldsymbol{r}^* = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}$ satisfies $\mathbf{1}'\boldsymbol{r}^* = 1$. Thus, the estimator $\hat{\boldsymbol{L}}'\boldsymbol{r}^*$ converges in probability to $\log(p(y))$.

Now let $\boldsymbol{d} = \boldsymbol{r} - \boldsymbol{r}^*$. The conditions $\mathbf{1}'\boldsymbol{r} = 1$ and $\mathbf{1}'\boldsymbol{r}^* = 1$ lead to $\mathbf{1}'\boldsymbol{d} = 0$. We write the asymptotic variance for $\hat{\boldsymbol{L}}'\boldsymbol{r}$ as $\boldsymbol{r}'\boldsymbol{\Sigma}\boldsymbol{r} = (\boldsymbol{r}^* + \boldsymbol{d})'\boldsymbol{\Sigma}(\boldsymbol{r}^* + \boldsymbol{d})$, which in turn equals $(\boldsymbol{r}^*)'\boldsymbol{\Sigma}\boldsymbol{r}^* + \boldsymbol{d}'\boldsymbol{\Sigma}\boldsymbol{d} + 2\boldsymbol{d}'\boldsymbol{\Sigma}\boldsymbol{r}^*$. Note that the term $\boldsymbol{d}'\boldsymbol{\Sigma}\boldsymbol{r}^* = \frac{\boldsymbol{d}'\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} = \frac{(\mathbf{1}'\boldsymbol{d})'}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} = 0$. It follows that $\boldsymbol{r}'\boldsymbol{\Sigma}\boldsymbol{r} = (\boldsymbol{r}^*)'\boldsymbol{\Sigma}\boldsymbol{r}^* + \boldsymbol{d}'\boldsymbol{\Sigma}\boldsymbol{d}$. Given that $\boldsymbol{\Sigma}$ is positive definite, we conclude that $\boldsymbol{r}'\boldsymbol{\Sigma}\boldsymbol{r} > (\boldsymbol{r}^*)'\boldsymbol{\Sigma}\boldsymbol{r}^*$ for any $\boldsymbol{d} \neq \mathbf{0}$. Thus, the estimator $\hat{\boldsymbol{L}}'\boldsymbol{r}^*$ has a lower asymptotic variance than any consistent estimator $\hat{\boldsymbol{L}}'\boldsymbol{r}$ where $\boldsymbol{r} \neq \boldsymbol{r}^*$. □

C Details of the Bridge Sampling Method

The bridge sampling method utilized the equation:

$$p(\boldsymbol{y}) = \frac{\int \alpha(\boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \alpha(\boldsymbol{\theta}) q(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}}$$

where $\alpha(\boldsymbol{\theta})$ is the bridge function. The log marginal likelihood can be estimated as:

$$\log(\hat{p}(\boldsymbol{y})) = \log\left(\frac{1}{m}\sum_{j=1}^{m} \alpha(\boldsymbol{\theta}_j) p(\boldsymbol{y}|\boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j)\right) - \log\left(\frac{1}{m}\sum_{j=1}^{m} \alpha(\tilde{\boldsymbol{\theta}}_j) q(\tilde{\boldsymbol{\theta}}_j)\right)$$

where $\{\boldsymbol{\theta}_j\}_{j=1}^m$ are i.i.d. draws from the auxiliary distribution $q(\boldsymbol{\theta})$ and $\{\tilde{\boldsymbol{\theta}}_j\}_{j=1}^m$ are draws from the posterior $p(\boldsymbol{\theta}|y)$ by some Markov chain Monte Carlo sampler. By applying the central limit theorem for i.i.d. random variables and the delta method, it can be shown

that:

$$\sqrt{m}\left(\log\left(\frac{1}{m}\sum_{j=1}^{m}\alpha(\boldsymbol{\theta}_j)p(\boldsymbol{y}|\boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j)\right)-\log(\mu_1)\right)\xrightarrow{d}N(0,\sigma_1^2/\mu_1^2)$$

where $\mu_1=\int\alpha(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})q(\boldsymbol{\theta})d\boldsymbol{\theta}$ and σ_1^2 is the variance of $\alpha(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ with respect to $q(\boldsymbol{\theta})$. Similarly, assume that the conditions for the Markov chain central limit theorem are satisfied. Applying the Markov chain central limit theorem and the delta method gives:

$$\sqrt{m}\left(\log\left(\frac{1}{m}\sum_{j=1}^{m}\alpha(\tilde{\boldsymbol{\theta}}_j)q(\tilde{\boldsymbol{\theta}}_j)\right)-\log(\mu_2)\right)\xrightarrow{d}N(0,\sigma_2^2/\mu_2^2)$$

where $\mu_2=\int\alpha(\boldsymbol{\theta})q(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$ and σ_2^2 is the long-run variance of $\alpha(\boldsymbol{\theta})q(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})$ with respect to $p(\boldsymbol{\theta}|\boldsymbol{y})$. Since draws from $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$ are mutually independent, we can derive:

$$\sqrt{m}\left(\log(\hat{p}(\boldsymbol{y}))-\log(p(\boldsymbol{y}))\right)\xrightarrow{d}N(0,\sigma_1^2/\mu_1^2+\sigma_2^2/\mu_2^2)$$

Thus, the numerical standard error of the LML estimate can be computed as $m^{-1/2}(\hat{\sigma}_1^2/\hat{\mu}_1^2+\hat{\sigma}_2^2/\hat{\mu}_2^2)^{1/2}$, where $\hat{\mu}_1=\frac{1}{m}\sum_{j=1}^{m}\alpha(\boldsymbol{\theta}_j)p(\boldsymbol{y}|\boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j)$, $\hat{\mu}_2=\frac{1}{m}\sum_{j=1}^{m}\alpha(\tilde{\boldsymbol{\theta}}_j)q(\tilde{\boldsymbol{\theta}}_j)$, $\hat{\sigma}_1^2$ is the sample variance of $\{\alpha(\boldsymbol{\theta}_j)p(\boldsymbol{y}|\boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j)\}_{j=1}^{m}$, and $\hat{\sigma}_2^2$ is an estimate of the long-run variance based on $\{\alpha(\tilde{\boldsymbol{\theta}}_j)q(\tilde{\boldsymbol{\theta}}_j)\}_{j=1}^{m}$ (e.g. by the Newey-West method of Newey and West (1987)).

Based on Meng and Wong (1996) and Meng and Schilling (2002), the optimal bridge function is $\alpha(\boldsymbol{\theta})\propto(\phi p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})+q(\boldsymbol{\theta}))^{-1}$ where $\phi=(1-\phi_1)(1+\phi_1)^{-1}p(\boldsymbol{y})^{-1}$ and ϕ_1 is the first order serial correlation of the log likelihood estimates based on $\{\tilde{\boldsymbol{\theta}}_j\}_{j=1}^{m}$. Given the presence of $p(\boldsymbol{y})$ in the optimal bridge function, an iterative procedure is applied to compute the LML. In this paper, the importance sampling estimate of the LML is used as the starting value of the iterations. A total of 10 iterations are applied after which the LML estimate appears to have been stabilized.

We know from Section 2.2 that the LML estimator of Equation (3) corresponds to the bridge function $\alpha(\boldsymbol{\theta})=\left((p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}))^{1-w}\,q(\boldsymbol{\theta})^{w}\right)^{-1}$. Since the importance sampling method and the Gelfand-Dey method are the special cases $w=1$ and $w=0$, the LML estimates and the corresponding numerical standard errors by these two methods can be readily

computed by plugging in the associated bridge functions into the general formula of the bridge sampling estimator described in the preceding paragraphs. Similarly, for a grid of w, we can compute the LML estimate and the corresponding numerical standard error for each element in the grid. The LML estimate with the smallest numerical standard error in the grid is labeled as the minimum variance estimator and is also compared in the empirical exercises.

D Regressors for Predicting Regressions

The regressors of the probit regression model in Section 3.1 are listed in Table D1.

E Regressors for Predicting Inflation Rate

The regressors of the regression models in Section 3.2 and 3.3 are listed in Table E1.

Table D1: List of Predictors for Economic Recessions

Name	Description
Term spread	Difference between 10-year Treasury constant maturity rate and 3-month Treasury bill rate
Long spread	Difference between 20- and 10-year Treasury constant maturity rates
Short rate	Change of 3-month Treasury bill rate
Long rate	Change of 20-year Treasury constant maturity rate
Inflation	Change of inflation rate measured as log change of GDP deflator
Expenditure	Log change of real government consumption expenditures and gross investment
Debt	Log change of consumer credit to households and non-profit organizations
Mortgage	Log change of one-to-four-family residential mortgages
S&P500	Log change of average daily closing price
AAA	Log change of average monthly Moody's Aaa corporate bond yield relative to 10-year Treasury constant maturity rate
BAA	Log change of average monthly Moody's Baa corporate bond yield relative to 10-year Treasury constant maturity rate

Note: Data on the S&P500 index is from Robert Shiller's website http://www.econ.yale.edu/ shiller/data.htm. Data on all other variables are from the FRED database of the U.S. Federal Reserve Bank of St. Louis.

Table E1: List of Predictors for Inflation Rate

Name	Description
GDP	Log change of real GDP
Investment	Log change of real gross private domestic investment
Expenditure	Log change of real government consumption expenditures and gross investment
Imports	Log change of imports of goods and services
Potential GDP	Log change of real potential GDP
Employee	Log change of total non-farm employees
Unemployment	Change of unemployment rate
Wage	Log change of average hourly earnings of production and non-supervisory employees
House start	Log change of new privately-owned housing unites started
House supply	Change of the ratio of houses for sale to houses sold
Public debt	Change of the ratio of public debt to GDP
Consumer debt	Log change of consumer credit to households and non-profit organizations
Mortgage	Log change of one-to-four-family residential mortgages
Energy price	Log change of consumer price index for energy in U.S. city average
Producer price	Log change of producer price index for all commodities
Short rate	Change of 3-month Treasury bill rate
Term spread	Difference between 10-year Treasury constant maturity rate and 3-month Treasury bill rate
S&P500	Log change of average daily closing price
M1	Log change of M1 money stock
M2	Log change of M2 money stock

Note: Data on the S&P500 index is from Robert Shiller's website `http://www.econ.yale.edu/~shiller/data.htm`. Data on all other variables are from the FRED database of the U.S. Federal Reserve Bank of St. Louis.

References

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 53*, 111–142.

Ardia, D., N. Basturk, L. Hoogerheide, and H. V. Dijk (2012). A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics and Data Analysis 56*, 3398–3414.

Carlin, B. and S. Chib (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 57*(3), 473–484.

Chan, J. (2023). Comparing stochastic volatility specifications for large Bayesian VARs. *Journal of Econometrics 235*(2), 1419–1446.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*(432), 1313–1321.

Chib, S. and I. Jeliazkov (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association 96*(453), 270–281.

Friel, N. and A. Pettitt (2008). Marginal likelihood estimation via power posteriors. *Journal Royal Statistical Society, Series B (Statistical Methodology) 70*, 589–607.

Garthwaite, P., Y. Fan, and S. Sisson (2016). Adaptive optimal scaling of Metropolis-Hastings algorithms using the Robbins-Monro process. *Communications in Statistics - Theory and Methods 45*(17), 5098–5111.

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

Gelman, A. and X. Meng (1998). Simulating normalizing contants: from importance sampling to bridge sampling to path sampling. *Statistical Science 13*, 163–185.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–1339.

Hammersley, J. and D. Handscomb (1964). *Monte Carlo Methods*. Methuen, London.

Han, C. and B. Carlin (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association 96*, 1122–1132.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kastner, G. and S. Fruhwirth-Schnatter (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics and Data Analysis 76*, 408–423.

Kloek, T. and H. V. Dijk (1978). Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econometrica 46*, 1–20.

Liu, J. and Y. Wu (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association 94*, 1264–1274.

Meng, X. and S. Schilling (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics 11*, 552–586.

Meng, X. and W. Wong (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica 6*, 831–860.

Newey, W. and K. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*, 703–708.

Newton, M. and A. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 56*, 3–48.

Nissila, W. (2020). Probit based time series models in recession forecasting – a survey with an empirical illustration for Finland. BoF Economic Review, No.7/2020, Bank of Finland, Helsinki.

Yu, Y. and X. Meng (2011). To center or not to center: That is not the question - an ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics 20*(3), 531–570.