A Mixture Estimator of Marginal Likelihood

Zhongfang He*

First Version: May 9, 2024

Abstract

This paper proposes 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 and nests the popular importance sampling and Gelfand-Dey methods as special cases. By varying a mixing parameter between zero and one, 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 encouraging 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 is 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. The new method is built on a simple identity 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)) are special cases of the new method when the mixing parameter equals one and zero respectively. By varying the mixing parameter between zero and one, the new method can easily produce a sequence of log marginal likelihood estimates that are asymptotically consistent. By taking the average of the sequence, one can obtain a more accurate estimate of marginal likelihood.

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 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 equity premium, among which one version allows stochastic volatility while the other allows both stochastic volatility and time-varying regression coefficients. We find that the numerical standard errors of the log marginal likelihood estimates by the new method are substantially smaller than those by the importance sampling and Gelfand-Dey methods.

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 $q(\boldsymbol{\theta}, w) = \left(p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right)^w q(\boldsymbol{\theta})^{1-w}$ 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):

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

where $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{m}$ are i.i.d. draws from the auxiliary distribution $q(\boldsymbol{\theta})$ and $\{\tilde{\boldsymbol{\theta}}^{(j)}\}_{j=1}^{\tilde{m}}$ are draws from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$.

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]

[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 $\exp((w-1)f(\boldsymbol{\theta}))$.

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 averaging over the sequence of marginal likelihood estimates, a final marginal likelihood estimate can be obtained that is likely more accurate. 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. The final LML estimate is $\hat{L} = \frac{1}{n}\sum_{i=1}^n \hat{L}_{w_i}$. The numerical standard error of $\hat{L}$ can be computed by applying suitable versions of the central limit theorem and the delta method. Lemma 2.1 provides the formula for the numerical standard error of $\hat{L}$.

Lemma 2.1. *Let $\hat{\boldsymbol{\Sigma}}_g$ be an estimate of the covariance matrix $\boldsymbol{\Sigma}_g$ of $\{\boldsymbol{g}_j\}_{j=1}^m$ where $\boldsymbol{g}_j = [g_{1,j}\ ...g_{n,j}]'$, $g_{i,j} = exp(w_i f(\boldsymbol{\theta}^{(j)}))$ for $i = 1, ..., n$ and $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^m$ are i.i.d. draws from the auxiliary distribution $q(\boldsymbol{\theta})$. Denote $\hat{v}_g = \frac{1}{m}\boldsymbol{l}_g'\hat{\boldsymbol{\Sigma}}_g\boldsymbol{l}_g$ where $\boldsymbol{l}_g$ is a n-by-1 vector stacking $\{1/(n\bar{g}_i)\}_{i=1}^n$ and $\bar{g}_i = \frac{1}{m}\sum_{j=1}^m g_{i,j}$ for $i = 1, ..., n$. Similarly, let $\hat{\boldsymbol{\Sigma}}_h$ be an estimate of the long-run covariance matrix $\boldsymbol{\Sigma}_h$ of $\{\boldsymbol{h}_j\}_{j=1}^{\tilde{m}}$ where $\boldsymbol{h}_j = [h_{1,j}\ ...h_{n,j}]'$, $h_{i,j} = exp((w_i - 1)f(\tilde{\boldsymbol{\theta}}^{(j)}))$ for $i = 1, ..., n$ and $\{\tilde{\boldsymbol{\theta}}^{(j)}\}_{j=1}^{\tilde{m}}$ are draws from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. Denote $\hat{v}_h = \frac{1}{\tilde{m}}\boldsymbol{l}_h'\hat{\boldsymbol{\Sigma}}_h\boldsymbol{l}_h$ where $\boldsymbol{l}_h$ is a n-by-1 vector stacking $\{1/(n\bar{h}_i)\}_{i=1}^n$ and $\bar{h}_i = \frac{1}{\tilde{m}}\sum_{j=1}^{\tilde{m}} h_{i,j}$ for $i = 1, ..., n$. Then the numerical standard error of $\hat{L}$ is $\sqrt{\hat{v}_g + \hat{v}_h}$.*

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

Remark 1. An estimate of $\boldsymbol{\Sigma}_g$ can be obtained by the sample covariance matrix of $\{\boldsymbol{g}_j\}_{j=1}^m$. For the long-run covariance matrix $\boldsymbol{\Sigma}_h$, there exist many estimation methods (e.g. see the textbook Hayashi (2000)). In this paper, we use the Newey-West method (Newey and West (1987)) to estimate $\boldsymbol{\Sigma}_h$.[2]

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

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 the importance sampling and Gelfand-Dey methods under a common choice of $q(\boldsymbol{\theta})$. Assume that all the elements in $\boldsymbol{\theta}$ take value on the whole real line.[3] 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

the number of posterior draws.

[3]If any element in $\boldsymbol{\theta}$ is restricted, it should 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.

approach of Chan (2023) to calibrate $q(\boldsymbol{\theta})$ and divide $\boldsymbol{\theta}$ into two groups: the latent variables $\boldsymbol{z} = \{\boldsymbol{z}_t\}_{t=1}^T$, 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 poster 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:

$$p(\boldsymbol{y}) = \frac{\int \alpha(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})q(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \alpha(\boldsymbol{\theta})q(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}} \tag{4}$$

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 $\alpha(\boldsymbol{\theta}) = ((p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}))^w q(\boldsymbol{\theta})^{1-w})^{-1}$, 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 one can readily compute a sequence of marginal likelihood estimates by Equation (2) and combining such a sequence of marginal likelihood estimates holds the promise to improve its accuracy.

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 models applied to real economic and financial 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 study the equity premium. The third example expands the model in the second example to allow both stochastic volatility and time-varying regression coefficients. All the three models are popular econometric tools for economists and are routinely used in empirical studies of economic data.

It should be mentioned here that, in all the examples when implementing the mixture estimator of marginal likelihood, the number of draws from the auxiliary distribution $q(\boldsymbol{\theta})$ is the same as the number of posterior draws for convenience. The grid of the mixing

parameter w is $\{0, 0.01, ..., 0.99, 1\}$ with a total of 101 points.

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 recent survey on probit-based models for recession prediction. Data on the binary response is the NBER-based U.S. recession indicator. Table B1 in Appendix B 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.

As discussed in Section 2.1, the auxiliary distribution $q(\boldsymbol{\theta})$ is constructed as a Gaussian distribution with the mean and covariance matrix equal the posterior mean and covariance matrix of $\boldsymbol{\theta}$. We compute the LML as the average of a sequence of Monte Carlo estimates by Equation (3) based on the grid of w. The numerical standard error of the LML estimate is calculated according to Lemma 2.1. For comparison, we also estimate the LML and its numerical standard error under the two special cases $w = 1$ and $w = 0$ that correspond to the importance sampling and Gelfand-Dey methods respectively.

Panel A of Table 1 compares the LML estimates of the probit model by the mixture,

importance sampling and Gelfand-Dey methods. For this example, the LML estimates by the 3 methods are fairly close. The numerical standard error of the mixture estimate of LML is 26% smaller than that of the importance sampling estimate and is 39% smaller than that of the Gelfand-Dey estimate. Panel A of Figure 1 shows the sequence of LML estimates by Equation (3) based on the grid of w. It can be seen that, for this example, the LML estimate steadily increases with w. By averaging over the sequence of LML estimates, the final mixture estimate of LML lies between the importance sampling and Gelfand-Dey estimates.

3.2 Linear Regression with Stochastic Volatility

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:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \exp\left(\frac{z_t}{2}\right)\epsilon_t \tag{5}$$
$$z_t = (1-\rho)\mu + \rho z_{t-1} + s\eta_t$$

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 $\boldsymbol{\delta}$ is a $(J+3)$-by-1 vector $[\boldsymbol{\beta}'\ \mu\ \log\left(\frac{1+\rho}{1-\rho}\right)\ \log(s^2)]'$.[4] The priors of the transformed parameters are adjusted by using the change-of-variables formula.

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

The SV regression model is applied to predict the equity premium. The dependent variable is the value-weighted quarterly return of the S&P500 index minus the corresponding risk free rate. Following Welch and Goyal (2008), we select 12 economic predictors that include stock characteristics, interest rates and other macroeconomic indicators and are listed in Table C1 of Appendix C. In estimation, we also include an intercept and a lag of the equity premium. The sample runs from Q1 1947 to Q4 2020 with a total of 296 quarterly observations. The prediction horizon is 1 quarter. 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 regression model, the model parameters $\boldsymbol{\theta}$ include the latent variable $\boldsymbol{z} = \{z_t\}_{t=1}^T$ 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, importance sampling and Gelfand-Dey methods are reported in Panel B of Table 1.

For this more complicated model, the difference in the LML estimates by the importance sampling method (297.65) and the Gelfand-Dey method (297.52) is larger than in the case of the probit model of Section 3.1. 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. By averaging over the sequence of LML estimates, the numerical standard error of the LML estimate by the mixture method is 54% smaller than that by the importance sampling method and is 76%

smaller than that by the Gelfand-Dey method.

3.3 Linear Regression with Stochastic Volatility and TVP

Greater flexibility can be obtained by allowing the regression coefficients in the SV regression model of Equation (5) to vary over time, which has been found useful to capture changes in economic regimes as well as improve forecasts (e.g. Clark (2011), D'Agostino et al. (2013), Clark and Ravazzolo (2015) and Bitto and Fruhwirth-Schnatter (2019) among many others). A commonly used specification of the SV regression model with time-varying regression coefficients (TVP-SV regression) has the following form:

$$\begin{aligned} y_t &= \boldsymbol{x}_t' \boldsymbol{\beta_t} + \exp\left(\frac{z_t}{2}\right) \epsilon_t \\ z_t &= (1-\rho)\mu + \rho z_{t-1} + s\eta_t \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{s}_\beta \odot \boldsymbol{\xi}_t \end{aligned} \tag{6}$$

for $t = 1, ..., T$. The notations are the same as the SV regression model of Equation (5) in Section 3.2 except for the new feature that the regression coefficients now follow a vector random walk process, where the initial value of the time-varying regression coefficient $\boldsymbol{\beta}_t$ is $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \boldsymbol{s}_\beta \odot \boldsymbol{\xi}_1$, $\boldsymbol{\xi}_t \sim N(\mathbf{0}, \mathrm{I}_J)$ is a J-by-1 vector of i.i.d. innovations of $\boldsymbol{\beta}_t$ for $t = 1, ..., T$ and $\boldsymbol{s}_\beta$ is a J-by-1 vector of model parameters. The notation $\odot$ is the element-wise product of two vectors or matrices of the same size.

The TVP-SV regression model is applied to the same dataset of equity premium in Section 3.2 and is estimated by a Metropolis-within-Gibbs sampler. The priors of the overlapping model parameters are the same as in the SV regression model of Section 3.2. For the new parameters $\{\boldsymbol{\beta}_0, \boldsymbol{s}_\beta\}$, their priors are $N(\mathbf{0}, 10\mathrm{I}_J)$. Similar to the SV regression model, the SV process is estimated based on the algorithm of Kastner and Fruhwirth-Schnatter (2014). The parameters $\{\boldsymbol{\beta}_t\}_{t=1}^T$ are sampled by the simulation smoother of Durbin and Koopman (2002). The interweaving strategy of Yu and Meng (2011) is applied

to boost the sampling quality of the parameters $\{\boldsymbol{\beta}_0, \boldsymbol{s}_\beta\}$. A total of 10,000 posterior draws are collected for analysis after 5,000 burn-ins.

Denote the fixed model parameters $\boldsymbol{\delta} = [\boldsymbol{\beta}_0' \ \log(\boldsymbol{s}_\beta^2)' \ \mu \ \log\left(\frac{1+\rho}{1-\rho}\right) \ \log(s^2)]'$. The latent variables in the TVP-SV regression model include both the log volatility $\boldsymbol{z} = \{z_t\}_{t=1}^T$ and the TVP $\{\boldsymbol{\beta}_t\}_{t=1}^T$. To reduce the number of parameters for estimating the marginal likelihood, we use the partially marginalized likelihood $p(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{z})$ that integrates out the TVP $\{\boldsymbol{\beta}_t\}_{t=1}^T$. The partially marginalized likelihood $p(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{z})$ can be computed analytically with a reasonable computation cost by applying the Kalman filter, whose details can be found in Appendix D. Therefore, for estimating the marginal likelihood, the model parameters $\boldsymbol{\theta}$ include only the fixed model parameters $\boldsymbol{\delta}$ and the latent variable $\boldsymbol{z}$. The auxiliary distribution $q(\boldsymbol{\theta}) = q(\boldsymbol{\delta})q(\boldsymbol{z}|\boldsymbol{\delta})$ is constructed the same way as for the SV regression model of Section 3.2.

Panel C of Table 1 shows the resulting LML estimates and their numerical standard errors by the mixture, importance sampling and Gelfand-Dey methods. The LML estimate by the mixture method (268.32) is between the estimates by the importance sampling method (267.60) and the Gelfand-Dey method (269.68). Panel C of Figure 1 shows the sequence of LML estimates by Equation (3) based on the grid of w. Similar to the SV 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 appear to 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 numerical standard error of the LML estimate by the mixture method is 59% smaller than that by the importance sampling method and is 41% smaller than that by the Gelfand-Dey method.

4 Conclusion

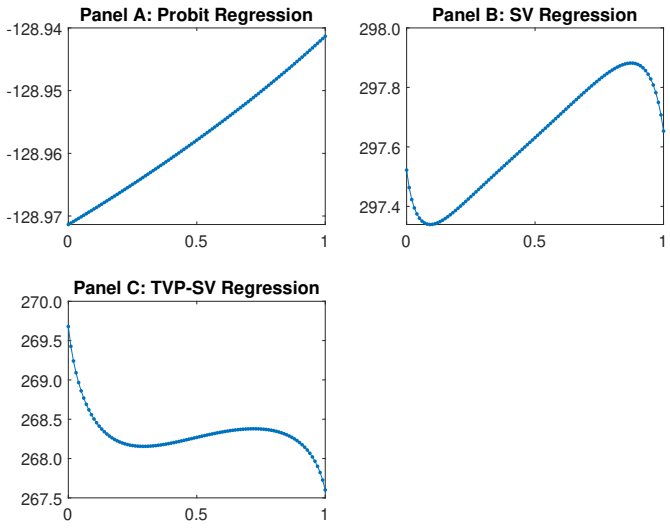
We introduce a new method to compute the marginal likelihood that utilizes a geometric mixture of the unnormalized posterior and an auxiliary distribution of model parameters. The new method nests the popular importance sampling and Gelfand-Dey methods as special cases. By varying the value of the mixing parameter in the geometric mixture, one can easily compute a sequence of marginal likelihood estimates whose average leads to a more accurate estimate of marginal likelihood 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	-128.96	(0.0029)
Importance Sampling	-128.94	(0.0039)
Gelfand-Dey	-128.97	(0.0048)
Panel B: SV Regression		
Mixture	297.62	(0.11)
Importance Sampling	297.65	(0.23)
Gelfand-Dey	297.52	(0.44)
Panel C: TVP-SV Regression		
Mixture	268.32	(0.13)
Importance Sampling	267.60	(0.32)
Gelfand-Dey	269.68	(0.22)

Note: This table compares the log marginal likelihood estimates by the mixture, importance sampling and Gelfand-Dey methods for the three examples of probit regression, linear regression with stochastic volatility (SV regression) and linear regression with stochastic volatility and time-varying regression coefficients (TVP-SV 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 (SV regression) and linear regression with stochastic volatility and time-varying regression coefficients (TVP-SV regression). The number of draws is 10,000 in all cases.

A Proof of Lemma 2.1

Proof. Since $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{m}$ are i.i.d. draws from the auxiliary distribution $q(\boldsymbol{\theta})$, applying the central limit theorem for i.i.d. random variables can show that the scaled average $\sqrt{m}(\bar{\boldsymbol{g}} - E(g_j))$, where $\bar{\boldsymbol{g}} = [\bar{g}_1 \ ... \ \bar{g}_n]'$ and $E(g_j)$ is the expectation of g_j with respect to $q(\boldsymbol{\theta})$ for $j = 1, ..., m$, follows a Gaussian distribution asymptotically. The covariance matrix of the asymptotic Gaussian distribution is $\boldsymbol{\Sigma}_g$. Define $L_g = \frac{1}{n}\mathbf{1}'\log(\bar{\boldsymbol{g}})$, where $\mathbf{1}$ is a n-by-1 vector of ones. Applying the delta method can show that the variance of L_g is $v_g = \frac{1}{m}\boldsymbol{l}_g'\boldsymbol{\Sigma}_g\boldsymbol{l}_g$, which can be estimated by $\hat{v}_g$.

Similarly, assume that the conditions for the Markov chain central limit theorem are satisfied, the scaled average $\sqrt{\tilde{m}}(\bar{\boldsymbol{h}} - E(h_j))$, where $\bar{\boldsymbol{h}} = [\bar{h}_1 \ ... \ \bar{h}_n]'$ and $E(h_j)$ is the expectation of h_j with respect to $p(\boldsymbol{\theta}|\boldsymbol{y})$ for $j = 1, ..., \tilde{m}$, follows a Gaussian distribution asymptotically where the covariance matrix of the asymptotic Gaussian distribution is the long-run covariance matrix $\boldsymbol{\Sigma}_h$. Define $L_h = \frac{1}{n}\mathbf{1}'\log(\bar{\boldsymbol{h}})$. Applying the delta method can show that the variance of L_h is $v_h = \frac{1}{\tilde{m}}\boldsymbol{l}_h'\boldsymbol{\Sigma}_h\boldsymbol{l}_h$, which can be estimated by $\hat{v}_h$.

The average $\hat{L}$ can be written as $L_g - L_h$. Since draws from the auxiliary distribution $q(\boldsymbol{\theta})$ and the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ are independent of each other, the items L_g and L_h are independent of each other as well. Thus, the variance of $\hat{L}$ is the sum of the variances of L_g and L_h. It follows that the standard deviation of $\hat{L}$ is $\sqrt{\hat{v}_g + \hat{v}_h}$.

□

B Regressors for Predicting Regressions

The regressors of the probit regression model in Section 3.1 are listed in Table B1.

Table B1: 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.

C Regressors for Predicting Equity Premium

The regressors of the regression models in Section 3.2 and 3.3 are listed in Table C1.

D Computing the Partially Marginalized Likelihood

This appendix explains how to compute the partially marginalized likelihood $p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\delta})$ of the TVP-SV regression model.

Conditional on $\boldsymbol{z}$ and $\boldsymbol{\delta}$, the TVP-SV regression model of Equation (6) can be treated as a linear Gaussian state space system. The Kalman filter can be applied to compute the filtering distribution $p(\boldsymbol{\beta}_t|\boldsymbol{y}^t,\boldsymbol{z},\boldsymbol{\delta}) = N(\boldsymbol{m}_t,\boldsymbol{M}_t)$ where $\boldsymbol{y}^t = [y_1 \ ... \ y_t]'$ is the history up to time t and the parameters $\boldsymbol{m}_t$ and $\boldsymbol{M}_t$ follow the recursive law of motion:

$$\boldsymbol{M}_t^{-1} = (\boldsymbol{M}_{t-1} + \boldsymbol{W})^{-1} + \boldsymbol{x}_t\boldsymbol{x}_t'\exp(z_t)$$

$$\boldsymbol{M}_t^{-1}\boldsymbol{m}_t = (\boldsymbol{M}_{t-1} + \boldsymbol{W})^{-1}\boldsymbol{m}_{t-1} + \boldsymbol{x}_t y_t\exp(z_t)$$

with $\boldsymbol{m}_0 = \boldsymbol{\beta}_0$ and $\boldsymbol{M}_0 = \boldsymbol{0}$. $\boldsymbol{W}$ is a diagonal matrix with the diagonal elements $\boldsymbol{s}_\beta^2$.

Factorize $p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\delta})$ as $\prod_{t=1}^T p(y_t|\boldsymbol{z},\boldsymbol{\delta},\boldsymbol{y}^{t-1})$ where $\boldsymbol{y}^0$ is the empty set. By applying the law of total expectation and the law of total variance, one can derive $y_t|\boldsymbol{z},\boldsymbol{\delta},\boldsymbol{y}^{t-1} \sim N(\mu_{\boldsymbol{y},t},\sigma_{\boldsymbol{y},t}^2)$ where $\mu_{\boldsymbol{y},t} = \boldsymbol{x}_t'\boldsymbol{m}_{t-1}$ and $\sigma_{\boldsymbol{y},t}^2 = \boldsymbol{x}_t'(\boldsymbol{M}_{t-1}+\boldsymbol{W})\boldsymbol{x}_t + \exp(z_t)$. Thus the density $p(y_t|\boldsymbol{z},\boldsymbol{\delta},\boldsymbol{y}^{t-1}) = (2\pi\sigma_{\boldsymbol{y},t}^2)^{-1/2}\exp\left(-\frac{(y_t-\mu_{\boldsymbol{y},t})^2}{2\sigma_{\boldsymbol{y},t}^2}\right)$.

Table C1: List of Predictors for Equity Premium

Name	Description
Dividend price ratio	Log dividends minus log price
Dividend payout ratio	Log dividends minus log earnings
Stock variance	Log of sum of squared daily returns on the S&P500 index
Book-market ratio	Ratio of book to market value for the Dow Jones Industrial Average index
Equity expansion	Ratio of 12-month moving sums of net issues by NYSE listed stocks divided by the total end-of-year market capitalization of NYSE stocks
Treasury bill rate	Quarterly change of 3-month secondary market Treasury bill rate
Long term yield	Quarterly change of long-term government bond yield from Ibbotson's *Stocks, Bonds, Bills and Inflation Yearbook*
Term spread	Long term yield minus treasury bill rate
Default yield spread	Difference between BAA and AAA-rated corporate bond yields
Default return spread	Difference between long-term corporate and government bond returns
Inflation rate	Consumer price index (all urban consumers)
Investment-capital ratio	Ratio of aggregate (private non-residential fixed) investment to aggregate capital

Note: The data is publicly available from Amit Goyal's website https://sites.google.com/view/agoyal145/?redirpath=/. Detailed descriptions of the variables can be found in Welch and Goyal (2008).

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.

Bitto, A. and S. Fruhwirth-Schnatter (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics 210*, 75–97.

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.

Clark, T. (2011). Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business and Economic Statistics 29*, 327–341.

Clark, T. and F. Ravazzolo (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics 30*, 551–575.

D'Agostino, A., L. Gambetti, and D. Giannone (2013). Macroeconomic forecasting and structural change. *Journal of Applied Econometrics 28*, 82–101.

Durbin, J. and S. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89*, 603–615.

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

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.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 21*(4), 1455–1508.

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.