Marginal Likelihood Estimation with the Cross-Entropy Method

Joshua Chan and Eric Eisenstat

Australian National University, Spiru Haret University

2012

Marginal Likelihood Estimation with the Cross-Entropy Method

Joshua C.C. Chan[1] Eric Eisenstat[2]

[1]Centre for Applied Macroeconomic Analysis,
Research School of Economics,
Australian National University, Australia
[2]Faculty of Marketing and International Business,
Spiru Haret University, Bucharest, Romania

April 2012

Abstract

We consider an adaptive importance sampling approach to estimating the marginal likelihood, a quantity that is fundamental in Bayesian model comparison and Bayesian model averaging. This approach is motivated by the difficulty of obtaining an accurate estimate through existing algorithms that use Markov chain Monte Carlo (MCMC) draws, where the draws are typically costly to obtain and highly correlated in high-dimensional settings. In contrast, we use the cross-entropy (CE) method, a versatile adaptive Monte Carlo algorithm originally developed for rare-event simulation. The main advantage of the importance sampling approach is that random samples can be obtained from some convenient density with little additional costs. As we are generating *independent* draws instead of *correlated* MCMC draws, the increase in simulation effort is much smaller should one wish to reduce the numerical standard error of the estimator. Moreover, the importance density derived via the CE method is grounded in information theory, and therefore, is in a well-defined sense optimal. We demonstrate the utility of the proposed approach by two empirical applications involving women's labor market participation and U.S. macroeconomic time series. In both applications the proposed CE method compares favorably to existing estimators.

Keywords: importance sampling, model selection, probit, logit, time-varying parameter vector autoregressive model, dynamic factor model

JEL classification codes: C11, C15, C32, C52

Acknowledgment: J. C. C. Chan's research is supported by the Australian Research Council (Discovery Grant DP0985177 and DP0987170).

1 Introduction

This paper proposes a simple yet effective adaptive importance sampling approach to estimating marginal likelihoods that is readily implementable in a vast variety of econometric applications. The calculation of marginal likelihood, which is obtained by integrating the likelihood function with respect to the prior distribution, has attracted considerable interest in the Bayesian literature due to its importance in Bayesian model comparison and Bayesian model averaging. In fact, there is a vast literature on estimating the normalizing constant of a given density by Markov chain Monte Carlo (MCMC) methods; see Gelfand and Dey (1994), Newton and Raftery (1994), Chib (1995), Gelman and Meng (1998), Chib and Jeliazkov (2001), Fruhwirth-Schnatter and Wagner (2008) and Friel and Pettitt (2008), among many others. Although substantial progress has been made from the statistics side in the last two decades, estimation of marginal likelihood, particularly in high-dimensional settings, remains a difficult problem. It continues to be especially elusive in applied economic work due to the practical and computational burdens characterizing existing methods.

More specifically, not only do existing approaches often require nontrivial programming efforts, most involve using MCMC draws to compute certain Monte Carlo averages, which are then used to derive an estimate of the normalizing constant. One major drawback of this approach is that MCMC draws are typically costly to obtain. To compound the problem in high-dimensional settings, the draws are also often highly correlated. This is especially problematic, for example, when the researcher wishes to compare several similar models with a small dataset, where the marginal likelihoods under each model are expected to be similar, and therefore accurate estimates are required. For instance, to reduce the numerical standard error of an estimator with *independent* samples by a tenfold, one needs to increase the simulation effort by a factor of 100. With correlated MCMC draws, however, the increase in sample size could be much larger. In a typical situation in high-dimensional settings where the effective sample size is, say, 0.02 (i.e., inefficiency factor equals 50), one needs to increase the simulation effort by 5000 to achieve the same reduction in the numerical standard error. In addition, since posterior draws are usually costly to obtain, the computational effort required to discern the competing models may be formidable.

In view of these drawbacks, we consider instead an adaptive importance sampling approach to estimating the marginal likelihood, where random samples can be obtained from some convenient density with little additional costs. In addition, since we are generating *independent* draws, the increase in simulation effort is much smaller should one wish to reduce the numerical standard error of the estimator. The importance sampling approach is perhaps one of the earliest attempts to tackle the problem of marginal likelihood estimation; in fact, an early application of the importance sampling idea can be traced back to Geweke (1989). Of course, the trade-off in applying direct importance sampling is the complexity related to choosing an *importance density* from which to sample, and such a choice is not obvious in general. Typically, for importance sampling to be successful, the importance density should at least satisfy the following three properties: (i) it is heavier-tailed than the target density, such that the resulting estimator has finite variance, (ii) it is practically simple to draw samples from, and (iii) it is easy to evaluate at any particular point in its support. Consequently, an *ad hoc* choice of the importance density is likely to render the method impractical; yet, the choice of a "good" importance density is not always trivial.

Recently, Hoogerheide et al. (2007) present an innovative technique for estimation by fitting adaptively a mixture of Student-t distributions to the posterior density. Importance sampling is then used to obtain various quantities of interest, using the fitted mixture as an importance density. Ardia et al. (2010) later show that this approach allows an efficient and reliable estimation of marginal likelihood that performs well against existing methods. We build on this line of research by proposing a methodological procedure for deriving an importance density that is in a well-defined sense optimal. Moreover, the proposed approach is generally applicable to a vast class of econometric models, including various high-dimensional latent data models (see Section 5). More specifically, we use the cross-entropy (CE) method, which is a versatile adaptive Monte Carlo algorithm originally developed for rare-event simulation by Rubinstein (1997). Since its inception, it has been applied to a diverse range of estimation problems, such as network reliability estimation in telecommunications (Hui et al., 2005), efficient simulation of buffer overflow probabilities in queuing networks (de Boer et al., 2004), adaptive independence sampler design for MCMC methods (Keith et al., 2008), estimation of large portfolio loss probabilities in credit risk models (Chan and Kroese, 2010), and other rare-event probability estimation problems involving light- and heavy-tailed random variables (Kroese and Rubinstein, 2004; Asmussen et al., 2005). A recent review of the CE method and its applications can be found in Kroese (2010), and a book-length treatment is given in Rubinstein and Kroese (2004).

The CE method involves the following fundamental insights. First, for marginal likelihood estimation there exists an importance density that gives a zero-variance estimator.[1] In fact, if we use the posterior density as the importance density, then the associated importance sampling estimator has zero variance. Clearly, while in principle the posterior density gives the best estimator possible, it cannot be used in practice as its normalizing constant is the very unknown quantity we seek to estimate. On the other hand, a computationally tractable density that is "close" to the posterior density should be a good candidate, in the sense that its associated estimator has a small variance. The CE method seeks to locate within a given parametric family the importance density that is the "closest" to the zero-variance importance density, using the *Kullback-Leibler divergence*, or the *cross-entropy distance* as a measure of closeness between the two densities. As long as the parametric family embodies the properties of convenient density evaluation and sample generation, the resulting density will typically yield an efficient and straightforward importance sampling algorithm.

From a statistics perspective, therefore, estimating the marginal likelihood by the CE method proposed here is attractive because it inherits the well-known properties of importance sampling estimation. Moreover, employing an information-theoretic criterion for optimality is both conceptually and practically appealing to economists. Indeed, information theory presents a familiar approach to conducting econometric inference with minimal *a priori* knowledge of the data-generating process (Zellner, 1991; Maasoumi, 1993; Golan et al., 1996; Golan, 2008); it has likewise been applied (dating back to Arrow, 1970; Marschak, 1971) in the theoretical modeling of risk and uncertainty. On the practical side, it turns out that deriving the importance density by minimizing its cross-entropy distance to the posterior amounts to an exercise identical to likelihood maximization. Finally, we note that although the proposed approach also requires MCMC draws for obtaining the optimal importance density, the number of draws needed is typically small (a few thousands or less).

[1] More generally, zero-variance importance distributions exist for estimation problems of the form $\mathbb{E}H(\mathbf{X})$, where H is a positive function.

In consequence, one obtains a marginal likelihood estimator that is relatively simple to implement and does not require auxiliary *reduced-run* simulations in a wide variety of applications. In addition, the proposed approach leads to a straightforward procedure to perform prior sensitivity analysis without the need to have multiple MCMC runs. This is important as the value of the marginal likelihood is often sensitive to the prior specification. It is therefore sensible to investigate if the conclusions based on the marginal likelihood criterion are robust under different, yet reasonable, priors.

Despite their marked importance, model comparison and prior sensitivity analysis are often omitted in applied econometric work due to the programming and computational complexities characterizing the existing methods for computing marginal likelihoods. The CE method aims to overcome this by providing a more accessible methodology to applied economists without sacrificing effectiveness: as demonstrated by the empirical examples of Section 5, the CE method generally compares favorably with the most popular existing methods — and quite often provides substantial gains — in terms of computational efficiency.

The rest of this article is organized as follows. We first review the problem of estimating the marginal likelihood in Section 2, which is of importance in Bayesian econometrics and statistics. We then discuss two practical adaptive importance sampling approaches to tackle the problem in Section 3: the variance minimization (VM) and cross-entropy (CE) methods, with particular focus on the latter. We then discuss in Section 4 a straightforward procedure to perform a prior sensitivity analysis via the proposed approach without having multiple MCMC runs. In Section 5 we present two empirical examples to demonstrate the utility and effectiveness of the proposed approach. The first example involves women's labor market participation. There we compare three different binary response models in order to find the one that best fits the data. The second example considers three popular vector autoregressive (VAR) models to analyze the interdependence and structural stability of four U.S. macroeconomic time series: GDP growth, inflation, unemployment rate and interest rate.

2 Marginal Likelihood Estimation

To set the stage, consider the problem of comparing a collection of models $\{M_1, \ldots, M_K\}$. Each model M_k is formally defined by a likelihood function $p(\mathbf{y} \,|\, \boldsymbol{\theta}_k, M_k)$ and a prior on the model-specific parameter vector $\boldsymbol{\theta}_k$ denoted as $p(\boldsymbol{\theta}_k \,|\, M_k)$. A popular criterion to compare between models M_i and M_j is the *posterior odds ratio* between the two models:

$$\mathrm{PO}_{ij} \equiv \frac{p(M_i \,|\, \mathbf{y})}{p(M_j \,|\, \mathbf{y})} = \frac{p(M_i)}{p(M_j)} \times \frac{p(\mathbf{y} \,|\, M_i)}{p(\mathbf{y} \,|\, M_j)},$$

where $p(M_k \,|\, \mathbf{y})$ is the *posterior model probability* of model M_k and

$$p(\mathbf{y} \,|\, M_k) = \int p(\mathbf{y} \,|\, \boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k \,|\, M_k) \mathrm{d}\boldsymbol{\theta}_k$$

is the *marginal likelihood* under model M_k. The ratio $p(\mathbf{y} \,|\, M_i)/p(\mathbf{y} \,|\, M_j)$ is often referred to as the *Bayes factor* in favor of model M_i against model M_j. If both models are equally probable *a priori*, i.e., $p(M_i) = p(M_j)$, the posterior odds ratio between the two models is then equal to the Bayes factor. In addition, if we assume that the set of models under consideration is exhaustive,

i.e., $\sum_{k=1}^{K} p(M_k \,|\, \mathbf{y}) = 1$, we can use the posterior odds ratios to obtain the posterior model probabilities. Specifically, it is easy to check that

$$p(M_i \,|\, \mathbf{y}) = \left(1 + \sum_{k \neq i}^{K} \mathrm{PO}_{ki} \right)^{-1}.$$

For a more detailed discussion of the marginal likelihood and its role in Bayesian model comparison and model averaging, see Gelman et al. (2003) and Koop (2003).

For moderately high-dimensional problems, analytic calculation of the marginal likelihood is almost never possible. But with the advent of the Markov chain Monte Carlo methods, substantial progress has been made to estimate this quantity using simulation methods. For notational convenience, we suppress the model index from here onwards and write the marginal likelihood as $p(\mathbf{y})$. There are many different approaches to estimate $p(\mathbf{y})$ via MCMC methods, and here we mention two popular ones. Gelfand and Dey (1994) first realize that for any probability density function f with support contained in the support of the posterior density, one has

$$\begin{aligned} \mathbb{E}\left[\frac{f(\boldsymbol{\theta})}{p(\boldsymbol{\theta}) p(\mathbf{y} \,|\, \boldsymbol{\theta})} \,\middle|\, \mathbf{y} \right] &= \int \frac{f(\boldsymbol{\theta})}{p(\boldsymbol{\theta}) p(\mathbf{y} \,|\, \boldsymbol{\theta})} p(\boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta} \\ &= \int \frac{f(\boldsymbol{\theta})}{p(\boldsymbol{\theta}) p(\mathbf{y} \,|\, \boldsymbol{\theta})} \frac{p(\boldsymbol{\theta}) p(\mathbf{y} \,|\, \boldsymbol{\theta})}{p(\mathbf{y})} \mathrm{d}\boldsymbol{\theta} = p(\mathbf{y})^{-1}. \end{aligned} \tag{1}$$

Therefore, they propose the following estimator for $p(\mathbf{y})$:

$$\widehat{p}_{\mathrm{GD}}(\mathbf{y}) = \left\{ \frac{1}{L} \sum_{l=1}^{L} \frac{f(\boldsymbol{\theta}_l)}{p(\mathbf{y} \,|\, \boldsymbol{\theta}_l) p(\boldsymbol{\theta}_l)} \right\}^{-1}, \tag{2}$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ are posterior draws. Even though the equality in (1) holds for any density function f, the estimator $\widehat{p}_{\mathrm{GD}}(\mathbf{y})$ is biased, i.e., $\mathbb{E}[\widehat{p}_{\mathrm{GD}}(\mathbf{y})] \neq p(\mathbf{y})$ in general. Moreover, its accuracy depends crucially on the choice of the tuning function f. In view of this, Geweke (1999) proposes choosing f to be a normal approximation to the posterior density with a tail truncation determined by asymptotic arguments. We note in passing that the CE method described in this paper may be easily adapted to provide an alternative to the tuning function suggested in Geweke (1999). However, the resulting estimator $\widehat{p}_{\mathrm{GD}}(\mathbf{y})$ would nevertheless be based on correlated MCMC draws from the posterior distribution, rather than truly independent draws from the importance density, thereby still embodying a disadvantage relative to importance sampling algorithm we propose.

Another approach to computing the marginal likelihood, which is due to Chib (1995), does not require a tuning function. Indeed, the principal motivation for this approach is that "good" tuning functions are often difficult to find. To proceed, note that the marginal likelihood can be written as

$$p(\mathbf{y}) = \frac{p(\mathbf{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \,|\, \mathbf{y})}.$$

Hence, a natural estimator for $p(\mathbf{y})$ is the quantity (written in logarithmic scale)

$$\log \widehat{p}(\mathbf{y}) = \log p(\mathbf{y} \,|\, \boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta}^*) - \log \widehat{p}(\boldsymbol{\theta}^* \,|\, \mathbf{y})$$

where $\boldsymbol{\theta}^*$ is any point in the support of the posterior distribution. In practice, it is often chosen to be some "high density" point such as the posterior mean or mode. The only unknown

quantity is the posterior ordinate $p(\boldsymbol{\theta}^* \,|\, \mathbf{y})$, which may be estimated by Monte Carlo methods. In particular, if posterior draws are obtained via the Gibbs sampler — i.e., when *all* the full conditional distributions are known — then $p(\boldsymbol{\theta}^* \,|\, \mathbf{y})$ can be estimated by sampling draws from a series of suitably designed Gibbs samplers, the so-called *reduced runs*. This approach, however, requires complete knowledge of all conditional distributions in the Gibbs scheme, which greatly restricts its applicability in practice. To overcome this limitation, Chib and Jeliazkov (2001) extend the basic approach to include cases when the full conditional distributions are intractable and Metropolis-Hastings steps are required. However, implementation of these extensions are considerably more involved, especially for computing numerical standard errors of marginal likelihood estimates. In addition, like the estimator of Gelfand and Dey (1994), the Chib's estimator is also biased. There are many other approaches to marginal likelihood estimation, and we refer interested readers to the articles Han and Carlin (2001) and Friel and Pettitt (2008) for a more comprehensive review.

3 The Proposed Cross-Entropy Estimator

3.1 Importance Sampling and the Cross-Entropy Method

The basic idea of importance sampling — a fundamental Monte Carlo approach to estimating the expected value of an arbitrary function of random variables — is to bias the sampling distribution in such a way that more "important values" are generated in the simulation. The simulation output is then weighted to correct for the use of the biased distribution to give an unbiased estimator. Specifically, to estimate the marginal likelihood $p(\mathbf{y}) = \int p(\mathbf{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$, we consider the following importance sampling estimator:

$$\widehat{p}_{\mathrm{IS}}(\mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} \frac{p(\mathbf{y} \,|\, \boldsymbol{\theta}_n) p(\boldsymbol{\theta}_n)}{g(\boldsymbol{\theta}_n)}, \tag{3}$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ are independent draws obtained from the importance density $g(\cdot)$ that dominates $p(\mathbf{y} \,|\, \cdot) p(\cdot)$, i.e., $g(\mathbf{x}) = 0 \Rightarrow p(\mathbf{y} \,|\, \mathbf{x}) p(\mathbf{x}) = 0$. Although the estimator in (3) is unbiased and consistent for any such g, the performance of the estimator, particularly its variance, depends critically on the choice of the importance density. In what follows, we apply the cross-entropy (CE) method, which establishes a well-defined criterion for the *optimal* choice of g, and generates a straightforward procedure to derive the optimal importance density. The CE method derives its name from the cross-entropy or Kullback-Leibler divergence — a fundamental concept in modern information theory. Textbook treatments of the CE method can be found in Rubinstein and Kroese (2004) and Kroese et al. (2011, ch. 13); a recent survey of entropy and information-theoretic methods in econometrics is given in Golan (2008).

As previously discussed, the principal motivation of the CE approach is the fact that there exists an importance density that gives a zero variance estimator. That is, if we use the posterior density as the importance density, i.e., $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \,|\, \mathbf{y}) = p(\mathbf{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) / p(\mathbf{y})$, then the associated importance sampling estimator (3) has zero variance:

$$\widehat{p}_{\mathrm{IS}}(\mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} \frac{p(\mathbf{y} \,|\, \boldsymbol{\theta}_n) p(\boldsymbol{\theta}_n)}{g(\boldsymbol{\theta}_n)} = \frac{1}{N} \sum_{n=1}^{N} \frac{p(\boldsymbol{\theta}_n) p(\mathbf{y} \,|\, \boldsymbol{\theta}_n)}{p(\boldsymbol{\theta}_n) p(\mathbf{y} \,|\, \boldsymbol{\theta}_n) / p(\mathbf{y})} = p(\mathbf{y}),$$

and we need to produce only $N = 1$ sample. For later reference we denote this zero-variance importance density as g^*. Although in principle g^* gives the best possible estimator for $p(\mathbf{y})$, it cannot be used in practice because the normalization constant of the density depends on the marginal likelihood, which is exactly the quantity one wishes to estimate. However, this suggests a practical approach to obtain an optimal importance density. Intuitively, if one chooses g "close enough" to g^* so that both behave similarly, the resulting importance sampling estimator should have reasonable accuracy. Hence, our goal is to locate a convenient density that is in a well-defined sense "close" to g^*. To this end, consider a parametric family $\mathcal{F} = \{f(\boldsymbol{\theta}; \mathbf{v})\}$ indexed by some parameter vector $\mathbf{v}$ within which we locate the importance density. We will shortly discuss various considerations in choosing such a family; for the moment we assume that $\mathcal{F}$ is given. Now, we wish to find the density $f(\boldsymbol{\theta}; \mathbf{v}^*) \in \mathcal{F}$ such that it is the "closest" to g^*. We consider two popular directed divergence measures of densities, which correspond to the variance minimization (VM) and cross-entropy methods (CE). Let h_1 and h_2 be two probability density functions. The *Pearson* χ^2 *measure* between h_1 and h_2 is defined as:

$$\mathcal{D}_2(h_1, h_2) = \frac{1}{2}\left(\int \frac{h_1(\mathbf{x})^2}{h_2(\mathbf{x})}\mathrm{d}\mathbf{x} - 1\right).$$

Since every density in $\mathcal{F}$ can be represented as $f(\cdot; \mathbf{v})$ for some $\mathbf{v}$, the problem of obtaining the optimal importance density now reduces to the following parametric minimization problem:

$$\mathbf{v}^*_{\mathrm{vm}} = \operatorname*{argmin}_{\mathbf{v}} \mathcal{D}_2(g^*, f(\cdot; \mathbf{v})).$$

Note that the value that minimizes any affine transformation of $\mathcal{D}_2(g^*, f(\cdot; \mathbf{v}))$ also minimizes $\mathcal{D}_2(g^*, f(\cdot; \mathbf{v}))$ itself. Hence, we also have

$$\mathbf{v}^*_{\mathrm{vm}} = \operatorname*{argmin}_{\mathbf{v}} p(\mathbf{y})^2[2\mathcal{D}_2(g^*, f(\cdot; \mathbf{v})) + 1] = \operatorname*{argmin}_{\mathbf{v}} \int \frac{p(\mathbf{y} \mid \boldsymbol{\theta})^2 p(\boldsymbol{\theta})^2}{f(\boldsymbol{\theta}; \mathbf{v})}\mathrm{d}\boldsymbol{\theta}. \tag{4}$$

From this it can be shown that minimizing the Pearson χ^2 measure is the same as minimizing the variance of the associated importance sampling estimator. In other words, $f(\boldsymbol{\theta}; \mathbf{v}_{\mathrm{vm}})$ gives the minimum variance estimator within the parametric class $\mathcal{F}$. For this reason, this procedure is termed the variance minimization method. For a more detailed discussion of the Pearson χ^2 measure and its applications, we refer the readers to Botev and Kroese (2011).

In practice, the optimization problem in (4) is often difficult to solve analytically. Instead, we consider its stochastic counterpart:

$$\widehat{\mathbf{v}}_{\mathrm{vm}} = \operatorname*{argmin}_{\mathbf{v}} \frac{1}{L}\sum_{l=1}^{L} \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_l)p(\boldsymbol{\theta}_l)}{f(\boldsymbol{\theta}_l; \mathbf{v})}, \tag{5}$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ are draws obtained from the posterior density $p(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$. Note that the optimization problem in (5) is equivalent to solving the following equation

$$\frac{1}{L}\sum_{l=1}^{L} \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_l)p(\boldsymbol{\theta}_l)}{f(\boldsymbol{\theta}_l; \mathbf{v})}\nabla_{\mathbf{v}} \log f(\boldsymbol{\theta}_l; \mathbf{v}) = 0, \tag{6}$$

where $\nabla_{\mathbf{v}}$ is the gradient operator, which can be solved by various root-finding algorithms. The VM method is summarized as follows:

Algorithm 1. *VM Algorithm for Marginal Likelihood Estimation*

1. *Obtain a sample* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ *from the posterior density* $g^*(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})$ *and find the solution to* (5) *or equivalently* (6)*, which is denoted as* $\widehat{\mathbf{v}}^*_{\text{vm}}$.

2. *Generate a random sample* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ *from the density* $f(\cdot; \widehat{\mathbf{v}}^*_{\text{vm}})$ *and estimate* $p(\mathbf{y})$ *via importance sampling, as in (3).*

One potential problem with the VM method is that in high-dimensional settings, the optimization problem in (5) or (6) might be difficult to solve. This motivates the consideration of a more convenient measure of "distance" between densities that leads to straightforward optimization routines in obtaining the optimal importance density. Let h_1 and h_2 be two probability density functions. The *Kullback-Leibler divergence*, or *cross-entropy distance* between h_1 and h_2 is defined as follows:

$$\mathcal{D}_1(h_1, h_2) = \int h_1(\mathbf{x}) \log \frac{h_1(\mathbf{x})}{h_2(\mathbf{x})} \mathrm{d}\mathbf{x}.$$

As before, given the cross-entropy distance and a parametric family $\mathcal{F}$, we then locate the density $f(\cdot; \mathbf{v}) \in \mathcal{F}$ such that $\mathcal{D}_1(g^*, f(\cdot; \mathbf{v}))$ is minimized. Further, note that

$$\mathcal{D}_1(g^*, f(\cdot; \mathbf{v})) = \int g^*(\boldsymbol{\theta}) \log g^*(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - p(\mathbf{y})^{-1} \int p(\mathbf{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \log f(\boldsymbol{\theta}; \mathbf{v}) \mathrm{d}\boldsymbol{\theta},$$

where the first term on the right-hand side does not depend on $\mathbf{v}$. Therefore, solving the CE minimization problem is equivalent to finding

$$\mathbf{v}^*_{\text{ce}} = \operatorname*{argmax}_{\mathbf{v}} \int p(\mathbf{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \log f(\boldsymbol{\theta}; \mathbf{v}) \mathrm{d}\boldsymbol{\theta}. \tag{7}$$

Once again, we consider the stochastic counterpart of (7):

$$\widehat{\mathbf{v}}^*_{\text{ce}} = \operatorname*{argmax}_{\mathbf{v}} \frac{1}{L} \sum_{l=1}^{L} \log f(\boldsymbol{\theta}_l; \mathbf{v}), \tag{8}$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ are draws from the posterior density. In other words, $\widehat{\mathbf{v}}^*_{\text{ce}}$ is exactly the maximum likelihood estimate for $\mathbf{v}$ if we view $f(\boldsymbol{\theta}; \mathbf{v})$ as the likelihood function with parameter vector $\mathbf{v}$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ an observed sample. Since finding the maximum likelihood estimator (MLE) is a very well understood problem, solving (8) is typically easy. We summarize the CE method as below:

Algorithm 2. *CE Algorithm for Marginal Likelihood Estimation*

1. *Obtain a sample* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ *from the posterior density* $g^*(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})$ *and find the solution to* (8)*, which is denoted as* $\widehat{\mathbf{v}}^*_{\text{ce}}$.

2. *Generate a random sample* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ *from the density* $f(\cdot; \widehat{\mathbf{v}}^*_{\text{ce}})$ *and estimate* $p(\mathbf{y})$ *via importance sampling, as in (3).*

3.2 Choice of the Parametric Family $\mathcal{F}$

We now discuss various considerations in choosing the parametric family $\mathcal{F}$. First of all, it is obvious that it should be easy to generate random samples from any members of $\mathcal{F}$; otherwise it would defeat the purpose of the proposed approach. This requirement, however, can be easily met as efficient random variable generation is a very well-understood research area, and there is a large variety of densities at our disposal from which samples can be readily obtained. Second, the computational burden of evaluating the density $f(\boldsymbol{\theta}; \mathbf{v})$ at any point $\boldsymbol{\theta}$ should be sufficiently low, as this evaluation must be performed for each importance draw. To that end, most multivariate distributions that satisfy the first criterion of easy random sample generation are likewise straightforward to evaluate.

Additionally, the optimization problem (8) should be easily solved either analytically or numerically. Once again, this requirement can be easily met as solving (8) amounts to finding the MLE if one views $f(\cdot; \mathbf{v}) \in \mathcal{F}$ as the likelihood for $\mathbf{v}$. In fact, for the applications in the next section, the optimization problem (8) can either be solved analytically, or reduced to a one-dimensional maximization problem. In addition, by choosing $f(\boldsymbol{\theta}; \mathbf{v})$ as a product of densities, e.g., $f(\boldsymbol{\theta}; \mathbf{v}) = f(\boldsymbol{\theta}_1; \mathbf{v}_1) \times \cdots \times f(\boldsymbol{\theta}_B; \mathbf{v}_B)$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B)$ and $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_B)$, one can reduce the possibly high-dimensional maximization problem (8) into B low-dimensional problems, which can then be readily solved (albeit at some expense of "closeness" in the resulting importance density to the posterior). It is likewise worth mentioning that the *exponential family* provides a particularly appealing class of distributions in choosing $\mathcal{F}$, since analytical solutions to (8) can often be found explicitly for such distributions (e.g., Rubinstein and Kroese, 2004, p. 70).

A final, and slightly less obvious to handle consideration is that the importance density must be heavier-tailed than the posterior density in order for $\widehat{p}_{\mathrm{IS}}(\mathbf{y})$ to have finite variance. Suppose that the posterior/likelihood is sufficiently peaked and well-behaved such that $\int p(\mathbf{y} \,|\, \boldsymbol{\theta})^2 p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty$, then the crude Monte Carlo estimator, i.e., the importance sampling estimator associated with the importance density $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, has finite variance. Hence, if we choose the family $\mathcal{F}$ such that $p(\boldsymbol{\theta}) \in \mathcal{F}$, then by construction the estimator associated with $f(\cdot; \mathbf{v}^*_{\mathrm{vm}})$ has a smaller variance than the crude Monte Carlo estimator, and consequently its variance is also finite; see Proposition 2 in the next section for the exact statement, and Appendix B for the proof. As for the CE estimator (with importance density $f(\cdot; \mathbf{v}^*_{\mathrm{ce}})$), although the question of whether it always provides variance reduction (compared to the crude Monte Carlo estimator) remains open in general, in a wide variety of applications the CE and VM estimators have very similar performance (see, e.g., de Boer et al., 2004; Chan et al., 2011). Furthermore, there are numerous heuristics to monitor the stability of the importance sampling estimator. For example, if the variance is indeed finite, we should observe that an increase in the size of the importance sample by a factor of ϕ decreases the numerical standard error by a factor of $\sqrt{\phi}$. Since sampling from any member of $\mathcal{F}$ is by construction inexpensive, this monitoring presents little additional difficulty. Consequently, if one finds that additional sampling fails to decrease the numerical standard error of the marginal likelihood estimate, the parametric family should be appropriately expanded/modified. We emphasize, however, that in our experience this complication is rarely present, and in such exceptional cases when it does occur, the adjustments to $\mathcal{F}$ necessary to achieve the finite variance property are often obvious.

3.3 Properties of the Proposed Estimators

In this section we document various desirable properties of the VM and CE estimators. Recall that both are importance sampling estimators of the form given in (3), where the importance density $g(\boldsymbol{\theta})$ is $f(\cdot; \mathbf{v}^*_{\text{vm}})$ for the VM estimator and $f(\cdot; \mathbf{v}^*_{\text{ce}})$ for the CE estimator. In particular, one can easily show that both the VM and CE estimators are unbiased, in contrast to the estimators of Gelfand and Dey (1994) and Chib (1995), which are biased. Throughout this section we assume:

Assumption 1. For any given data $\mathbf{y}$, the likelihood function is bounded in the parameter vector $\boldsymbol{\theta}$, i.e., there exists $\widehat{\boldsymbol{\theta}}$ such that $p(\mathbf{y} \,|\, \boldsymbol{\theta}) \leqslant p(\mathbf{y} \,|\, \widehat{\boldsymbol{\theta}})$ for all $\boldsymbol{\theta}$ in the parameter space.

Assumption 1 is a mild regularity condition that, for example, is necessary for the existence of the maximum likelihood estimator. We need this assumption to show that the marginal likelihood $p(\mathbf{y})$ is in fact finite.

Proposition 1. Suppose $p(\boldsymbol{\theta})$ is a proper prior and Assumption 1 holds. Further assume that both $f(\cdot; \mathbf{v}^*_{\text{vm}})$ and $f(\cdot; \mathbf{v}^*_{\text{ce}})$ dominate $p(\mathbf{y} \,|\, \cdot)p(\cdot)$, i.e., $f(\mathbf{x}; \mathbf{v}^*_{\text{vm}}) = 0 \Rightarrow p(\mathbf{y} \,|\, \mathbf{x})p(\mathbf{x}) = 0$. Then the VM and CE estimators are consistent and unbiased.

The proof is given in Appendix B. In addition, if the likelihood is sufficiently well-behaved, one can further show that the variance of the VM estimator is finite, and therefore the central limit theorem applies to give an asymptotic distribution for the estimator.

Proposition 2. Suppose $\int p(\mathbf{y} \,|\, \boldsymbol{\theta})^2 p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty$ and Assumption 1 holds. If the parametric family $\mathcal{F}$ contains the prior density $p(\boldsymbol{\theta})$, then the variance of the VM estimator is finite. Furthermore, the VM estimator has an asymptotic normal distribution.

The reader is referred to Appendix B for a detailed proof.

4 Prior Sensitivity Analysis

A common criticism of the Bayes factor as a criterion in comparing non-nested models is that the value of marginal likelihood is often sensitive to the prior specification of the model-specific parameters. Unless the researcher has a strong theoretical justification of the priors entertained, it is often a good practice to perform a prior sensitivity analysis to investigate if the conclusions made based on the Bayes factor criterion are robust under different, yet reasonable, priors. We describe in this section how one can perform a prior sensitivity analysis via the proposed approach without the need to have multiple MCMC runs to estimate the marginal likelihood under each prior specification.

When two models under comparison (e.g. M_i and M_j) have the same parameter vector $\boldsymbol{\theta}$ (i.e. $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j = \boldsymbol{\theta}$), then the Bayes factor may be computed in a rather straightforward fashion that does not involve explicit marginal likelihood computation, as long as draws from one of the

posterior distributions are available. To see this, note that

$$\begin{aligned} p(\mathbf{y} \mid M_i) &= \int \frac{p(\mathbf{y} \mid \boldsymbol{\theta}, M_i) p(\boldsymbol{\theta} \mid M_i)}{p(\boldsymbol{\theta} \mid \mathbf{y}, M_j)} p(\boldsymbol{\theta} \mid \mathbf{y}, M_j) \mathrm{d}\boldsymbol{\theta} \\ &= \int \frac{p(\mathbf{y} \mid \boldsymbol{\theta}, M_i) p(\boldsymbol{\theta} \mid M_i)}{p(\mathbf{y} \mid \boldsymbol{\theta}, M_j) p(\boldsymbol{\theta} \mid M_j) / p(\mathbf{y} \mid M_j)} p(\boldsymbol{\theta} \mid \mathbf{y}, M_j) \mathrm{d}\boldsymbol{\theta} \\ &= p(\mathbf{y} \mid M_j) \int \frac{p(\mathbf{y} \mid \boldsymbol{\theta}, M_i) p(\boldsymbol{\theta} \mid M_i)}{p(\mathbf{y} \mid \boldsymbol{\theta}, M_j) p(\boldsymbol{\theta} \mid M_j)} p(\boldsymbol{\theta} \mid \mathbf{y}, M_j) \mathrm{d}\boldsymbol{\theta}. \end{aligned}$$

Therefore, we have

$$\mathrm{BF}_{ij} = \int r_{ij}(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}, M_j) \mathrm{d}\boldsymbol{\theta},$$

where

$$r_{ij}(\boldsymbol{\theta}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}, M_i) p(\boldsymbol{\theta} \mid M_i)}{p(\mathbf{y} \mid \boldsymbol{\theta}, M_j) p(\boldsymbol{\theta} \mid M_j)}.$$

Hence, given the sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ from the posterior density $p(\boldsymbol{\theta} \mid \mathbf{y}, M_j)$, an importance sampling estimate of BF_{ij} (the Bayes factor in favor of M_i) is

$$\widehat{\mathrm{BF}}_{ij} = \frac{1}{L} \sum_{l=1}^{L} r_{ij}\left(\boldsymbol{\theta}_l\right). \tag{9}$$

In fact, Kim et al. (1998) utilize this importance sampling idea to perform exact inference when $\boldsymbol{\theta}$ is sampled from an approximated model. In their case, the "true" model is a nonlinear state space model for which it is difficult to construct an efficient MCMC sampling scheme. Their solution is to approximate this nonlinear model using a mixture of linear models, where each of these models can be estimated efficiently using the Gibbs sampler. Given the draws obtained from the approximated model, they then use importance reweighing to compute posterior moments of parameters under the true model.

For our purpose, suppose that we have a marginal likelihood estimate under a certain prior distribution, and we would like to obtain an estimate under a different prior. To this end, consider models M_i and M_j where the two models differ only in their prior distributions, i.e., $p(\mathbf{y} \mid \boldsymbol{\theta}, M_i) = p(\mathbf{y} \mid \boldsymbol{\theta}, M_j)$. Now, $r_{ij}(\boldsymbol{\theta})$ reduces to simply the ratio of the two priors. Consequently, if one has an estimate of the marginal likelihood under M_j and posterior draws from $p(\mathbf{y} \mid \boldsymbol{\theta}, M_j)$, one can obtain an estimate for the marginal likelihood under M_i via (9). However, it is readily observed that for the importance sampling estimator (9) to work well, it is crucial that the variability of the ratio $r_{ij}(\boldsymbol{\theta})$ is small under the posterior distribution $p(\mathbf{y} \mid \boldsymbol{\theta}, M_j)$; otherwise, estimates of BF_{ij} might be driven by a select few realizations of $\boldsymbol{\theta}$, and hence, be rendered rather unstable. In other words, this approach can only reasonably accommodate subtle to moderate differences in prior specifications. For radical differences, they will need to be compared by computing marginal likelihoods $p(\mathbf{y} \mid M_i)$ and $p(\mathbf{y} \mid M_j)$, separately. Hence, the accurate computation of marginal likelihoods is of central concern in model comparison exercises and related interests.

To that end, we note that the proposed importance sampling approach of Section 3 likewise provides a straightforward way to perform a prior sensitivity analysis without the need to have multiple MCMC runs. In addition, unlike the approach in (9), it is robust across rather different prior specifications, as long as the posterior distribution does not change drastically under the

different priors. To motivate the procedure, recall that the CE method consists of two steps: first, we obtain MCMC draws (under a certain prior) to locate the optimal importance density. Second, we generate independent draws from the importance density to give an estimate as in (3). Now suppose that we have posterior draws from $p(\boldsymbol{\theta} \,|\, \mathbf{y}, M_j)$, and wish to estimate the marginal likelihood $p(\mathbf{y} \,|\, M_i)$ with the corresponding prior $p(\boldsymbol{\theta} \,|\, M_i)$. As shown previously, the zero-variance importance density for estimating $p(\mathbf{y} \,|\, M_i)$ is the posterior density $p(\boldsymbol{\theta} \,|\, \mathbf{y}, M_i)$. However, in many situations, even though the priors are very different under the two models M_i and M_j, the corresponding posterior distributions are similar (e.g., when the sample size is moderately large). In those cases, draws from $p(\boldsymbol{\theta} \,|\, \mathbf{y}, M_j)$ are a good representation of draws from $p(\boldsymbol{\theta} \,|\, \mathbf{y}, M_i)$. Hence, we can simply use the MCMC draws from the former to obtain $\widehat{\mathbf{v}}^*_{\text{ce}}$ as in (8), then deliver the importance sampling estimator (3) with the prior $p(\boldsymbol{\theta} \,|\, M_i)$. In what follows, we demonstrate that this approach works well across quite different prior specifications, and further show that the numerical standard error of the importance sampling estimator hardly changes.

5 Empirical Applications

In this section, we present two empirical examples to illustrate the proposed importance sampling approach for estimating the marginal likelihood. In each of these examples, the implementation is straightforward, and the estimate is accurate even for relatively small sample sizes. In the first example we consider three binary response models for women's labor market participation with logit, probit and t links. Since the link function is often chosen out of convenience rather than being based on modeling considerations, a formal model comparison exercise seems appropriate to determine which model best fits the data. The second empirical example is concerned with the US post-war macroeconomic time-series data involving GDP growth, inflation, unemployment and interest rate. We fit three popular vector autoregressive (VAR) models and formally compare them.

5.1 Binary Response Models for Women's Labor Market Participation

In the first application we analyze a dataset from Mroz (1987) that deals with the labor market participation decision of 200 married women using three binary response models with logit, probit and t links. Eight covariates are used to explain the binary indicator of labor market participation — a constant, non-wife income, number of years of education, years of experience, experience squared (divided by 100), age, number of children less than six years of age in the household, and number of children older than six years of age in the household. The likelihood for each of the binary response models has the form

$$p(\mathbf{y} \,|\, \boldsymbol{\beta}) = \prod_{i=1}^{200} p_i^{y_i} (1 - p_i)^{1 - y_i}, \tag{10}$$

where $\boldsymbol{\beta}$ is a $k \times 1$ parameter vector, $p_i = F(\mathbf{x}_i' \boldsymbol{\beta})$ is the probability that the i^{th} subject will participate in the labor market, $F(\cdot)$ is a cumulative distribution function (cdf), $y_i \in \{0, 1\}$ is the binary outcome variable, and $\mathbf{x}_i$ is the covariate vector of the i^{th} subject in the sample.

For the binary logit model, the cdf F is assumed to be logistic: $F(z) = (1 + e^{-z})^{-1}$; for the binary probit model, $F(z) = \Phi(z)$, where Φ is the cdf of the standard normal distribution. Lastly, F is assumed to be the cdf of a standard t distribution with degree of freedom $\nu = 10$ in the t model.

5.1.1 The Priors and Marginal Likelihood Estimation

For each of the three models, we assume a multivariate normal prior for $\boldsymbol{\beta}$: $\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{\beta}_0, \mathbf{V}_0)$ with $\boldsymbol{\beta}_0 = \mathbf{0}$. Since the three models imply different variances ($\pi^2/3$ for the logit model, 1 for the probit model, and $\nu/(\nu-2)$ for the t model), we scale the covariance matrix $\mathbf{V}_0$ accordingly to ensure the same prior is assumed across models. Specifically, we set $\mathbf{V}_0 = \tau\mathbf{I}$ for the logit model, $\mathbf{V}_0 = 3/\pi^2 \times \tau\mathbf{I}$ for the probit model, and $\mathbf{V}_0 = 3\nu/(\pi^2(\nu-2)) \times \tau\mathbf{I}$ for the t model. The form of the prior (multivariate normal) is chosen for computational convenience — it is a conjugate prior for both the probit and the t-link models — and is standard in the literature. To investigate the effect of the hyperparameters of the prior, we include a prior sensitivity analysis over a set of reasonably non-informative priors (with $\tau = 5, 10, 100$) in the subsequent model comparison exercise.

We locate the optimal importance density within the parametric family $\mathcal{F} = \{f_{\mathsf{N}}(\boldsymbol{\beta}; \mathbf{b}, \mathbf{B})\}$ indexed by $(\mathbf{b}, \mathbf{B})$, where $f_{\mathsf{N}}(\cdot\,; \boldsymbol{\mu}, \Sigma)$ is the multivariate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$ be the posterior draws from a given model. To obtain the CE reference parameters $(\widehat{\mathbf{b}}, \widehat{\mathbf{B}})$ for the optimal importance density, we need to solve the maximization problem (8). It is easy to check that the solution is simply

$$\widehat{\mathbf{b}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\beta}_i, \quad \widehat{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{\beta}_i - \widehat{\mathbf{b}})(\boldsymbol{\beta}_i - \widehat{\mathbf{b}})'. \tag{11}$$

In this case, therefore, the somewhat "obvious" choice of importance density, i.e. multivariate normal parameterized by the posterior mean and posterior variance, is justified as the *optimal* choice among all possible multivariate normal distributions. That is, setting the mean and variance at the respective posterior counterparts minimizes its cross-entropy distance to the posterior density. As it turns out, it also leads to a well-performing importance sampling algorithm that generates efficient estimates of the marginal likelihood.

To demonstrate the latter, for each of the three models we obtain a sample of size $L = 5000$ from the posterior density under the prior variance $\tau = 10$ after a burn-in period of 500 draws, and compute $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{B}}$ as in (11). Given the optimal importance density $f_{\mathsf{N}}(\boldsymbol{\beta}; \widehat{\mathbf{b}}, \widehat{\mathbf{B}})$, we estimate the marginal likelihoods of the three models, each with three different prior variances ($\tau = 5, 10, 100$) by the importance sampling estimator in (3). For the main importance sampling run, we set the sample size to be $N = 5000$.

5.1.2 Empirical Results

The importance sampling estimates for the log marginal likelihoods of the binary logit, probit and t models, as well as their numerical standard errors, are given in Table 1. Since the dataset is relatively small (sample size 200), the estimated marginal likelihoods for the three models

are similar. Therefore, it is important in this case for the numerical standard errors for the estimates to be small. Under the set of priors considered, the probit model seems to be favored by the data. In fact, if one assumes that the three models are equally likely *a priori*, then the posterior probabilities for the logit, probit and t models are respectively about 25%, 45% and 30% under all three prior specifications. Hence, for this particular dataset, the data shows moderate support for the probit model, while the logit and t models are more or less equally likely.

Table 1: Log marginal likelihood (numerical standard error) for the logit, probit and t models for various prior specifications.

	logit	probit	t-link
$\tau = 5$	$-138.68(0.005)$	$-138.13(0.004)$	$-138.85(0.004)$
$\tau = 10$	$-141.24(0.004)$	$-140.67(0.003)$	$-141.08(0.003)$
$\tau = 100$	$-150.23(0.005)$	$-149.64(0.004)$	$-150.06(0.005)$

We also report the model parameter estimates for the probit model for $\tau = 10$ in Table 2, together with the posterior standard deviations and the probabilities that the parameters are positive. All parameter estimates have signs that are consistent with economic theory. For instance, non-wife income, age and the number of children in the household have a negative impact on the probability that an average woman will participate in the labor market. On the other hand, education and experience both have a positive impact, though the impact of experience is diminishing (the impact is quadratic in years of experience). However, note the large posterior standard errors for the coefficients associated with the variables experience squared and children older than six.

Table 2: Parameter posterior means, standard deviations, and probabilities being positive for the probit model.

Covariate	$\mathbb{E}(\cdot \mid \mathbf{y})$	$\sqrt{\mathrm{Var}(\cdot \mid \mathbf{y})}$	$\mathbb{P}(\cdot > 0 \mid \mathbf{y})$
Constant	0.282	0.832	0.631
Non-wife income	-0.012	0.009	0.085
Education	0.113	0.046	0.995
Experience	0.100	0.041	0.992
Experience squared	-0.074	0.147	0.292
Age	-0.049	0.014	0.001
Children younger than six	-0.811	0.222	0
Children older than six	-0.005	0.080	0.475

5.1.3 Comparison with Other Methods

We now compare the empirical performance of the proposed CE estimator with two other popular methods for estimating the marginal likelihood. The first is the estimator of Gelfand and Dey (1994) with the tuning function suggested in Geweke (1999) — more precisely, a multivariate normal density truncated to its 95% highest density region — and we call this the GD-G method. The second estimator is the the method of Chib (1995) (for Gibbs output)

and Chib and Jeliazkov (2001) (for Metropolis-Hastings output), and we refer to this the CJ method. We use each of the three methods to estimate the marginal likelihoods of the three binary response models, and compare the methods in terms of computation time and numerical accuracy. Implementation details are given as follows. For the CE method, we first use $L = 2500$ MCMC draws to compute the optimal importance sampling density for each of the three models, and then sample $N = 50000$ draws from the density obtained for the main importance sampling run. For both the GD-G and CJ methods, we use $L = 50000$ MCMC draws to estimate the marginal likelihood. But instead of one long chain, we run 10 parallel chains each of length $L = 5000$, so that the numerical standard errors of the estimates can be readily computed.[2]

It is worth noting the different factors affecting the computation times of the three methods. Since both the GD-G and CJ methods require a lot of MCMC draws, when those draws are more costly to obtain (e.g., more costly in the t-link model than in the probit model), both methods tend to be slower. When reduced runs are required, as in the logit model, the CJ method tends to be slower (the other two methods do not require reduced runs). Finally, when the evaluation of the (integrated) likelihood function is more costly (e.g., more costly in the t-link model than in the probit model), the GD-G and CE methods — both require numerous evaluations of the likelihood function — tend to be slower.

Table 3: Log marginal likelihood estimate, numerical standard error, and computation time (in second) for the three methods. Variance reduction is approximately the amount of time needed to have the same level of accuracy as the CE method.

	Probit			
	estimate	NSE	time (s)	var. reduction
GD-G	-140.69	0.0025	33	15.5
CJ	-140.68	0.0081	27	133.1
CE	-140.67	0.0011	11	1.0
	Logit			
	estimate	NSE	time (s)	var. reduction
GD-G	-141.26	0.0028	36	8.7
CJ	-141.24	0.0050	66	50.9
CE	-141.24	0.0018	10	1.0
	t-link			
	estimate	NSE	time (s)	var. reduction
GD-G	-141.16	0.0062	54	13.9
CJ	-140.83	0.0179	37	79.2
CE	-141.08	0.0024	26	1.0

We report in Table 3 the marginal likelihood estimate (in log), its numerical standard error and the computation time (in second) for each of the three methods. For all the binary response models, the proposed CE method is both the quickest and provides the most accurate estimates. To compare the methods with the same denominator, we also provide the "variance reduction" of the CE method compared to the other two methods. For example, to achieve the same level of accuracy as the CE method for estimating the marginal likelihood of the probit model, the

[2]More precisely, the numerical standard error is calculated as the sample standard deviation of the 10 estimates divided by $\sqrt{10}$.

GD-G method would take about 15.5 $(33/11 \times (0.0025/0.0011)^2)$ times longer than the CE method. It is obvious from the results that the CE method provides substantial improvements over the other two competing methods.

5.2 Vector Autoregressive Models for U.S. Macroeconomic Time Series

In the second empirical application we analyze a dataset obtained from the US Federal Reserve Bank at St. Louis website[3] that consists of 248 quarterly observations from 1948Q1 to the 2009Q4 on $n = 4$ U.S. macroeconomic series: output (GDP) growth, interest rate, unemployment rate, and inflation.

One fundamental question is: which time-series specification considered previously in the literature best models the evolution and interdependence among the time series? To make a first step towards answering this question, we perform a formal Bayesian model comparison exercise to compare three popular time-series specifications. The first model we consider is a basic vector autoregressive (VAR) model. Pioneered by Sims (1980), it has traditionally been employed to capture the interdependence among the time series. Specifically, for $t = 1, \ldots, T$, we consider the first-order VAR model

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Omega}), \tag{12}$$

where $\mathbf{y}_t$ is a vector containing measurements on the aforementioned $n = 4$ macroeconomic variables (output growth, unemployment, income, and inflation), $\boldsymbol{\mu}$ is an $n \times 1$ vector of intercepts, $\boldsymbol{\Gamma}$ is an $n \times n$ matrix of parameters that governs the interdependence among the time series, and $\boldsymbol{\Omega}$ is an unknown $n \times n$ positive definite covariance matrix. The analysis is performed conditionally on the initial data point and consequently our sample consists of $T = 247$ observations.

For the purpose of estimation, (12) is rewritten in the form of a seemingly unrelated regression (SUR) model as follows:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \tag{13}$$

where $\mathbf{X}_t = \mathbf{I} \otimes (1, \mathbf{y}'_{t-1})$, $\boldsymbol{\beta}$ is a $q \times 1$ $(q = n^2 + n)$ vector containing the corresponding parameters from $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$, ordered equation by equation, i.e., $\boldsymbol{\beta} = \mathrm{vec}\,((\boldsymbol{\mu}, \boldsymbol{\Gamma})')$, and the $\mathrm{vec}(\cdot)$ operator stacks the columns of a matrix into a vector. Stacking the observations in (13) over t, we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathsf{N}(\mathbf{0}, \mathbf{I} \otimes \boldsymbol{\Omega}),$$

where

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_T \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_T \end{pmatrix}.$$

Although the traditional VAR model is a popular specification for macroeconomic series, it is not flexible enough to accommodate potential structural instabilities in the times series. In view of this, in our second specification we extend the basic VAR model (12) by allowing the parameter vector $\boldsymbol{\beta}$ to evolve over time in order to study the potential structural change of the time series.

[3]The website is: `http://research.stlouisfed.org/fred2/`.

In particular, we consider the following first-order time-varying parameter vector autoregressive (TVP-VAR) model (e.g., Canova, 1992; Carter and Kohn, 1994; Koop and Korobilis, 2010):

$$\mathbf{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\Gamma}_t \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Omega}), \tag{14}$$

where the parameters $\boldsymbol{\mu}_t$ and $\boldsymbol{\Gamma}_t$ now have a subscript t to denote the time period. The evolution of the parameter vector $\boldsymbol{\beta}_t = \text{vec}\left((\boldsymbol{\mu}_t, \boldsymbol{\Gamma}_t)'\right)$ for $t = 2, \ldots, T$ is governed by the evolution equation

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\nu}_t, \tag{15}$$

with $\boldsymbol{\beta}_1 \sim \mathsf{N}(\mathbf{0}, \mathbf{D})$, where $\boldsymbol{\nu}_t \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$ is an unknown $q \times q$ diagonal matrix. The transition equation (15) allows the elements of $\boldsymbol{\beta}_t$ to evolve gradually over time by penalizing large changes between successive values, thus *a priori* favoring the simple VAR model with time-invariant parameters.

Another approach to extend the traditional VAR model (12) is the following dynamic factor VAR (DF-VAR) model (e.g., Bernanke et al., 2005; Koop and Korobilis, 2010):

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma} \mathbf{y}_{t-1} + \mathbf{A} f_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Omega}), \tag{16}$$

where $\boldsymbol{\Omega} = \text{diag}(\omega_1^2, \ldots, \omega_n^2)$ is a diagonal matrix, $\mathbf{A}$ is an $n \times 1$ vector of factor loadings, and f_t is an unobserved factor that captures economy-wide macroeconomic volatility, which is assumed to evolve as

$$f_t = \gamma f_{t-1} + \eta_t, \quad \eta_t \sim \mathsf{N}(0, \sigma_f^2), \tag{17}$$

for $t = 2, \ldots, T$, and the initial factor f_1 is distributed according to the stationary distribution $f_1 \sim \mathsf{N}(0, \sigma_f^2/(1-\gamma^2))$ with $|\gamma| < 1$. Since neither the factor loadings $\mathbf{A}$ nor the factorf_t is observed, the scale and sign of $\mathbf{A}$ and f_t are not individually identified because $\mathbf{A} f_t = (c\mathbf{A})(f_t/c)$ for any $c \neq 0$. One popular identification assumption is to fixed the first element of $\mathbf{A}$ at 1, i.e., $\mathbf{A} = (1, \mathbf{a}')'$ and $\mathbf{a}$ is an $(n-1) \times 1$ vector of free parameters. With this assumption, both the sign and scale identification problems are resolved.

5.2.1 The Priors and Marginal Likelihood Estimation

In this subsection we describe the priors used in the three models and provide the implementation details of the proposed importance sampling approach. Since the value of marginal likelihood in general is sensitive to the choice of priors, we maintain the same priors across models where appropriate. In addition, we consider a set of reasonable yet relatively non-informative priors in a prior sensitivity analysis to see how the marginal likelihood estimates are affected by the choice of priors.

For the VAR model, the parameters are $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ with independent prior distributions: $\boldsymbol{\Omega} \sim \mathsf{IW}(\nu_0, \mathbf{S}_0)$ and $\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{\beta}_0, \mathbf{V}_{\boldsymbol{\beta}})$, where $\mathsf{IW}(\cdot, \cdot)$ is the inverse-Wishart distribution, $\nu_0 = n + 3, \boldsymbol{\beta}_0 = \mathbf{0}$ and $\mathbf{V}_{\boldsymbol{\beta}} = 10 \times \mathbf{I}$. These are conjugate priors and are standard in the literature. For $\mathbf{S}_0$ we consider two cases: $\mathbf{S}_0 = \mathbf{I}$ and $\mathbf{S}_0 = 0.1 \times \mathbf{I}$. Standard results from the SUR model can be applied to sample from the posterior density, e.g., see Koop (2003). To estimate the marginal likelihood of the model, we consider the parametric family $\mathcal{F} = \{f_{\mathsf{N}}(\boldsymbol{\beta}; \mathbf{b}, \mathbf{B}) f_{\mathsf{IW}}(\boldsymbol{\Omega}; \nu, \mathbf{S})\}$, where $f_{\mathsf{IW}}(\cdot; \nu, \mathbf{S})$ is the density of $\mathsf{IW}(\nu, \mathbf{S})$. Given the posterior output $\{\boldsymbol{\beta}_l, \boldsymbol{\Omega}_l\}_{l=1}^L$, the optimal CE

reference parameters $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{B}}$ can be obtained as in (11). As for $\widehat{\nu}$ and $\widehat{\mathbf{S}}$, first note that

$$\widehat{\mathbf{S}}^{-1} = \frac{1}{\nu L} \sum_{l=1}^{L} \mathbf{\Omega}_l^{-1}.$$

Now by substituting $\mathbf{S} = \widehat{\mathbf{S}}^{-1}$ into the density $f_{\mathsf{IW}}(\cdot; \nu, \mathbf{S})$, $\widehat{\nu}$ can be obtained by any one-dimensional root-finding algorithm (e.g., Newton-Raphson method). Alternatively, simply fixing $\widehat{\nu} = T$ also gives reasonable results. The former approach is the one we use for the empirical example.

For the TVP-VAR model, we assume that $\mathbf{\Omega} \sim \mathsf{IW}(\nu_0, \mathbf{S}_0)$ and each diagonal element of $\mathbf{\Sigma}$ follows, independently, an inverse-gamma distribution: $\sigma_i^2 \sim \mathsf{IG}(\nu_{0i}/2, S_{0i}/2)$, for $i = 1, \ldots, q$, with $\mathbf{D} = 10 \times \mathbf{I}$, $\nu_0 = n + 3$, and $\nu_{0i} = 6$, $S_{0i} = 0.01$ for $i = 1, \ldots, q$. For $\mathbf{S}_0$, we consider two cases: $\mathbf{S}_0 = \mathbf{I}$ and $\mathbf{S}_0 = 0.1 \times \mathbf{I}$. Posterior inference is based on a recently proposed algorithm in Chan and Jeliazkov (2009), which we briefly describe in the appendix. More importantly for our purpose, Chan and Jeliazkov (2009) also give an efficient way to evaluate the integrated likelihood, defined as the likelihood function marginal of the state vector $\boldsymbol{\beta}$. Hence, to compute the marginal likelihood of the TVP-VAR model, we can work with the integrated likelihood $f(\mathbf{y} \,|\, \mathbf{\Omega}, \mathbf{\Sigma})$ rather than the complete-data likelihood $f(\mathbf{y} \,|\, \boldsymbol{\beta}, \mathbf{\Omega}, \mathbf{\Sigma})$. This approach substantially reduces the variance of the importance sampling estimator as we analytically "integrate out" $\boldsymbol{\beta}$ that consists of 247 $\boldsymbol{\beta}_t$'s, each of which is a 20-dimensional vector. Generally, in situations where the integrated likelihood $f(\mathbf{y} \,|\, \boldsymbol{\theta})$ cannot be efficiently evaluated, one has to work with the complete-data likelihood $f(\mathbf{y} \,|\, \mathbf{z}, \boldsymbol{\theta})$, where $\mathbf{z}$ is the latent variable vector. The proposed approach proceeds as before, with the optimal importance density $f(\boldsymbol{\theta}, \mathbf{z}; \widehat{\mathbf{v}}_{\text{ce}})$ constructed from the posterior draws. In this case, however, the variance of the associated estimator is no smaller than the case where integrated likelihood $f(\mathbf{y} \,|\, \boldsymbol{\theta})$ can be efficiently evaluated. Now, to obtain the optimal importance density for the TVP-VAR model, we consider the parametric family

$$\mathcal{F} = \{f_{\mathsf{IW}}(\mathbf{\Omega}; \nu, \mathbf{S}) \prod_{i=1}^{q} f_{\mathsf{IG}}(\sigma_i^2; m_i, s_i)\}.$$

The optimal CE parameters $\{\widehat{\nu}, \widehat{\mathbf{S}}\}$ are obtained as in the previous case. Moreover, since the inverse-gamma density is a special case of the inverse-Wishart density, $(\widehat{m}_i, \widehat{s}_i), i = 1, \ldots, q$ can be obtained similarly.

Lastly, the priors for the parameters in the DF-VAR model are given as follows: $\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{\beta}_0, \mathbf{V}_{\boldsymbol{\beta}})$, $\mathbf{a} \sim \mathsf{N}(\mathbf{a}_0, \mathbf{V}_{\mathbf{a}})$, and γ is distributed uniformly on the interval $(-1, 1)$, with $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{V}_{\boldsymbol{\beta}} = 10 \times \mathbf{I}$ and $\mathbf{a}_0 = \mathbf{0}$. For $\mathbf{V}_{\mathbf{a}}$, we consider two cases $\mathbf{V}_{\mathbf{a}} = \mathbf{I}$ and $\mathbf{V}_{\mathbf{a}} = 5 \times \mathbf{I}$. In addition, each diagonal element of $\mathbf{\Omega} = \text{diag}(\omega_1^2, \ldots, \omega_q^2)$ follows, independently, an inverse-gamma distribution: $\omega_i^2 \sim \mathsf{IG}(\nu_{0i}/2, S_{0i}/2)$, for $i = 1, \ldots, q$, and finally, $\sigma^2 \sim \mathsf{IG}(\nu_f, S_f)$, where $\nu_f = \nu_{0i} = 6$, $S_f = S_{0i} = 1$. Posterior inference is based on an efficient MCMC sampler discussed in Chan and Jeliazkov (2009). Specifically, instead of sequentially drawing from the full conditionals, we implement the following collapsed Markov sampler that substantially improves the mixing properties and reduces the autocorrelation of the Markov chain:

Algorithm 3. *Collapsed MCMC Sampling of* $\boldsymbol{\beta}$ *and* $\mathbf{A}$

1. $[\boldsymbol{\beta} \,|\, \mathbf{y}, \mathbf{A}, \mathbf{\Omega}, \gamma, \sigma^2]$, *which does not depend on* $\mathbf{f} = (f_1, \ldots, f_T)'$

2. $[\mathbf{a}, \mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\Omega}, \gamma, \sigma^2]$, *which is done using*

   (a) $[\mathbf{a} \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\Omega}, \gamma, \sigma^2]$, *which does not depend on* $\mathbf{f}$

   (b) $[\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\Omega}, \gamma, \sigma^2]$

3. $[\boldsymbol{\Omega} \,|\, \mathbf{y}, \boldsymbol{\beta}, \mathbf{A}, \mathbf{f}]$

4. $[\gamma \,|\, \mathbf{f}, \sigma^2]$

5. $[\sigma^2 \,|\, \mathbf{f}, \gamma]$

Now, to locate the optimal importance density, we consider the following parametric family

$$\mathcal{F} = \left\{ f_{\mathsf{N}}(\boldsymbol{\theta}; \mathbf{b}_{\boldsymbol{\theta}}, \mathbf{B}_{\boldsymbol{\theta}}) f_{\mathsf{N}}(\gamma; b_\gamma, B_\gamma) f_{\mathsf{IG}}(\sigma^2; m, s) \} \prod_{i=1}^{q} f_{\mathsf{IG}}(\omega_i^2; m_i, s_i) \right\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{a}')'$. Since the densities are either normal or inverse-gamma, the optimal CE reference parameters can be obtained exactly in the same way as before.

5.2.2 Empirical Results

Estimation results of the VAR, TVP-VAR and DF-VAR models for the U.S. macroeconomic data are each based on a sample of $L = 10000$ MCMC draws after a burn-in of 500 draws, which are obtained by the MCMC samplers described in Section 5.2.1. We first present in Figure 1 the main parameter of interest in the TVP-VAR model, $\boldsymbol{\beta}$, the time-varying intercepts and VAR coefficients. The figure reveals that except for those coefficients in the unemployment equation, all the others exhibit considerable instability over the sampled periods, particularly the early part of the sample and the late 70s/early 80s. This suggests that the TVP-VAR model, which allows for structural changes in the time series, seems to be a more suitable model for the U.S. data compared to the restrictive VAR model, which assumes constant coefficients over time. In fact, results from a formal exercise of Bayesian model comparison shows that this is indeed the case (see below). Another point that is worth noting is that the impact of all the lagged macroeconomic variables on GDP growth appears to have drifted towards zero in the latter part of the sample. This supports the contention of a "Great Moderation" in U.S. output growth volatility that economic growth in the U.S. has become more stable over time due to lower dependence on past outcomes.

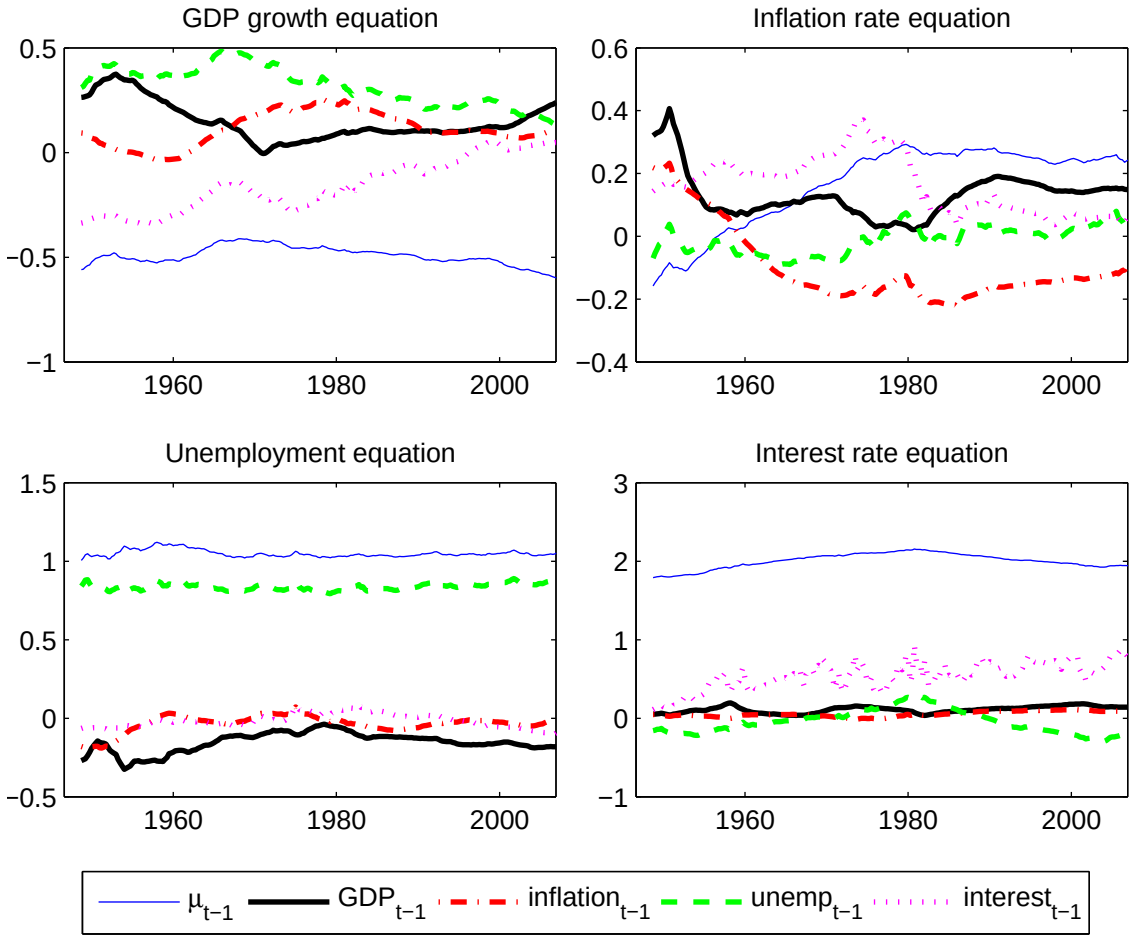

Figure 1: Evolution of the parameters $\boldsymbol{\beta}_t$ in the TVP-VAR model.

We also estimate the DF-VAR model using the same U.S. post-war macroeconomic data, and the estimated latent dynamic factor is reported in Figure 2, as well as the timing of officially announced U.S. recessions (shaded regions). We first note that from the figure it is apparent that the latent factor captures the timing of these recessions rather well. The ability of the model to capture these features is remarkable because what followed the officially announced end of the 1990–1991 recession was a period of jobless growth when unemployment actually increased and remained elevated for several years. Recovery from the 2001 recession was stymied by the immediate aftermath of the 9/11 attacks, which affected consumer behavior, investor sentiment and demand in key industries. In addition, the dynamic factor approach also gives an indication of the relative severity and evolution of each recession. In fact, it suggests that the most recent recession following the subprime mortgage crisis and the collapse of several U.S. largest financial institutions is the most severe one in the post-war period. See Jeliazkov and Liu (2010) for a formal analysis on the ranking of U.S. post-war recessions.

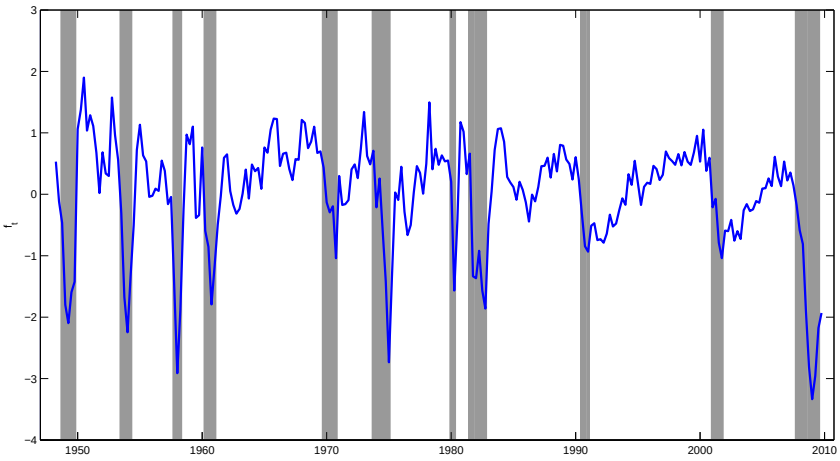

Figure 2: Evolution of the latent dynamic factor f_t in the DF-VAR model, together with the timing of officially announced U.S. recessions (shaded regions).

To formally compare the three time-series specifications, we compute the marginal likelihoods for the models via the proposed importance sampling approach. More specifically, for each model, we first use the posterior draws to estimate the optimal CE reference parameters for the importance density. Then we obtain an estimate for the marginal likelihood via importance sampling (which involves only normal and inverse-Wishart distributions) with a sample size of $N = 10000$. In particular, no reduced MCMC runs nor auxiliary simulations are needed. Moreover, should the researcher desire a more accurate estimate, additional draws can be obtained from the importance density with little computational costs. A prior sensitivity analysis is also done following the procedure described in Section 4. More particularly, for each model only one Markov sampler is run, and the MCMC draws are used to estimate the marginal likelihoods under both prior specifications. The estimated log marginal likelihoods are verified by the method of Chib and Jeliazkov (2001). The estimates, together with the numerical standard errors, are reported in Table 4. The result suggests that the TVP-VAR model is the best model for the U.S. post-war data among the three, under a variety of prior specifications. However, the relative ranking of the VAR and DF-VAR models changes as we use different priors, highlighting the importance of a prior sensitivity analysis in marginal likelihood estimation.

Table 4: Log marginal likelihood (numerical standard error) for the three time-series models.

VAR		TVP-VAR		DF-VAR	
$\mathbf{S}_0 = \mathbf{I}$	$\mathbf{S}_0 = 0.1 \times \mathbf{I}$	$\mathbf{S}_0 = \mathbf{I}$	$\mathbf{S}_0 = 0.1 \times \mathbf{I}$	$\mathbf{V_a} = 5 \times \mathbf{I}$	$\mathbf{V_a} = \mathbf{I}$
$-905.4(0.005)$	$-925.6(0.005)$	$-899.8(0.24)$	$-896.0(0.61)$	$-906.0(0.46)$	$-903.9(0.28)$

5.2.3 Comparison with Other Methods

We now compare the empirical performance of the three marginal likelihood estimators: the CE estimator, the estimator of Gelfand and Dey (1994) with the tuning function suggested in

Geweke (1999) (GD-G method) and the method of Chib (1995) and Chib and Jeliazkov (2001) (CJ method). We use each of the three methods to estimate the marginal likelihoods of the three VAR models, and compare the methods in terms of computation time and numerical accuracy.

Table 5: Log marginal likelihood estimate, numerical standard error, and computation time (in second) for the three methods. Variance reduction is approximately the amount of time needed to have the same level of accuracy as the CE method.

	VAR			
	estimate	NSE	time (s)	var. reduction
GD-G	-905.46	0.0030	60	3.8
CJ	-905.37	0.0003	61	0.03
CE	-905.37	0.0015	64	1.0
	TVP-VAR			
	estimate	NSE	time (s)	var. reduction
GD-G	-898.9	0.09	4099	2.8
CJ	-899.4	0.11	3132	3.2
CE	-899.8	0.07	2386	1.0
	DF-VAR			
	estimate	NSE	time (s)	var. reduction
GD-G	-906.0	0.03	640	0.4
CJ	-904.8	0.09	740	3.9
CE	-903.9	0.09	190	1.0

First, for computing the marginal likelihood of the VAR model via the CE method, we use $L = 2500$ MCMC draws to obtain the optimal importance sampling density, and then sample $N = 100000$ draws from the density obtained for the main importance sampling run. For both the GD-G and CJ methods, we use $L = 50000$ MCMC draws to estimate the marginal likelihood. Again, instead of one long chain, we run 10 parallel chains each of length $L = 5000$, such that the numerical standard error of the estimate can be computed easily. We report in Table 5 the estimate (in log), its numerical standard error and the computation time (in second) for each of the three methods, as well as the approximate amount of time needed to have the same level of accuracy as the CE method. For this relatively low-dimensional model, all three methods work very well, and the CJ method gives the most accurate estimate per unit of computation time.

We then perform the same exercise for the TVP-VAR model. The CE method uses $L = 5000$ MCMC draws for obtaining the optimal importance sampling density and $N = 100000$ draws for the main importance sampling run; both the GD-G and CJ methods use $L = 100000$ MCMC draws. Estimation of the marginal likelihood is much more time-consuming for this high-dimensional model: the GD-G, CJ and CE methods take respectively 68, 52 and 40 minutes. For the TVP-VAR model, the CE method is the fastest and most accurate. For instance, the CJ method would need about 129 minutes to achieve the same level of accuracy as the CE method (which takes only 40 minutes).

Finally, for the DF-VAR model, the CE method uses $L = 5000$ MCMC draws and $N = 100000$ draws from the importance density; both the GD-G and CJ methods use $L = 100000$ MCMC

draws. Although the CE method is the fastest, the GD-G method seems to provide the most accurate estimate given the same computation time. However, it is also worth noting that the GD-G estimate is slightly different from those obtained via the CJ and CE methods (taking into account of the small numerical standard errors), and this might indicate that the bias of the GD-G estimate in this example is substantial.

6 Concluding Remarks

In this article we introduce the CE method to tackle an important problem in Bayesian econometrics and statistics, namely, the estimation of marginal likelihood. Although the problem is well-studied and many approaches already exist, the proposed method has the merit of being both conceptually and computationally simple. In particular, it is essentially an importance sampling approach, with the choice of importance density guided by a methodology founded on formal optimization by minimizing the *cross-entropy distance* to the posterior density (i.e. the intractable zero-variance importance density). Therefore, the importance density is chosen to systematically minimize the variance of the resulting estimate, yet random samples can be obtained from it conveniently with negligible computational costs. Furthermore, since the draws are independent by construction, the simulation effort is much less compared to other approaches should one wish to reduce the numerical standard error of the estimator. In the two empirical applications studied, the CE method compares favorably to existing estimators.

Generally, this approach relies on the researcher designating a *family* of densities on which the CE-based optimization method is implemented to operate. For many applications, as demonstrated through two empirical examples in the text, this choice of distribution family is relatively straightforward. Typically, it will be sufficient to match the chosen family with the class of prior distributions employed by the model. Regardless of the family chosen, however, the eventual optimization over this collection, as prescribed by the CE approach, amounts to an exercise directly analogous to likelihood maximization.

This suggests, therefore, that when a particular choice of distribution family is inadequate, straightforward extensions are readily available. For example, it may be worthwhile to consider expanding a family of multivariate normal distributions to one consisting of a *mixture* of multivariate normal distributions. Such a generalization presents little additional computational difficulties – multivariate normal mixture likelihoods, while no longer analytically tractable, are easily maximized by Expectation-Maximization based techniques. To that end, an interesting and important extension of the algorithm presented in this paper would be one incorporating the automatic selection of the parametric family. We leave the latter open for future work.

Aside from the fact that following the CE approach leads to a simple, yet methodological construction of importance densities, the empirical examples discussed in Section 5 further illustrate the overall utility of implementing importance sampling in this way to estimate marginal likelihoods. In both cases, the proposed algorithm is not only fast and and easy to implement, but it also gives very accurate marginal likelihood estimates with a relatively small sample size.

Appendix A: Efficient Simulation of the Gaussian State Space Models

In this appendix we present an efficient Markov sampler proposed in Chan and Jeliazkov (2009) for simulation and integrated likelihood estimation in state space models. We consider the following linear Gaussian state space model:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{G}_t\boldsymbol{\eta}_t + \boldsymbol{\varepsilon}_t, \tag{18}$$

$$\boldsymbol{\eta}_t = \mathbf{Z}_t\boldsymbol{\gamma} + \mathbf{F}_t\boldsymbol{\eta}_{t-1} + \boldsymbol{\nu}_t, \tag{19}$$

for $t = 1, \ldots, T$, where $\mathbf{y}_t$ is an $n \times 1$ vector of observations, $\boldsymbol{\eta}_t$ is a $q \times 1$ latent state vector, (19) is initialized with $\boldsymbol{\eta}_1 \sim \mathsf{N}(\mathbf{Z}_1\boldsymbol{\gamma}, \mathbf{D})$, and

$$\begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \boldsymbol{\nu}_t \end{pmatrix} \sim \mathsf{N}\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{22} \end{pmatrix}\right).$$

Equation (18) is often referred to as the *measurement* or *observation* equation, while (19) is called the *transition* or *evolution* equation. Define $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)$, and let $\boldsymbol{\theta}$ represent the parameters in the state space model (i.e. $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\{\mathbf{G}_t\}$, $\{\mathbf{F}_t\}$, and the unique elements of $\boldsymbol{\Omega}_{11}$, $\boldsymbol{\Omega}_{22}$, and $\mathbf{D}$). The covariates $\{\mathbf{X}_t\}$ and $\{\mathbf{Z}_t\}$ are taken as given and will be suppressed in the conditioning sets below.

From (18)–(19) it is easily seen that the joint sampling density $f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$ is Gaussian. In fact, stacking (18)–(19) over the T time periods, we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\eta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_T \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} \mathbf{G}_1 & & \\ & \ddots & \\ & & \mathbf{G}_T \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_T \end{pmatrix},$$

with $\boldsymbol{\varepsilon} \sim \mathsf{N}(0, \mathbf{I} \otimes \boldsymbol{\Omega}_{11})$. A change of variable from $\boldsymbol{\varepsilon}$ to $\mathbf{y}$ implies that

$$f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = f_{\mathsf{N}}(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\eta}, \mathbf{I} \otimes \boldsymbol{\Omega}_{11}). \tag{20}$$

For the prior distribution of $\boldsymbol{\eta}$, we note that the directed conditional structure for $f(\boldsymbol{\eta}_t \mid \boldsymbol{\theta}, \boldsymbol{\eta}_{t-1})$ in (19) implies that the joint density for $\boldsymbol{\eta}$ is also Gaussian. To see this, define

$$\mathbf{H} = \begin{pmatrix} \mathbf{I} & & & & \\ -\mathbf{F}_2 & \mathbf{I} & & & \\ & -\mathbf{F}_3 & \mathbf{I} & & \\ & & \ddots & \ddots & \\ & & & -\mathbf{F}_T & \mathbf{I} \end{pmatrix} \quad \text{and} \quad \mathbf{S} = \begin{pmatrix} \mathbf{D} & & & & \\ & \boldsymbol{\Omega}_{22} & & & \\ & & \boldsymbol{\Omega}_{22} & & \\ & & & \ddots & \\ & & & & \boldsymbol{\Omega}_{22} \end{pmatrix},$$

so that (19) can be written as $\mathbf{H}\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\nu}$, where

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_T \end{pmatrix} \quad \text{and} \quad \boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\nu}_1 \\ \vdots \\ \boldsymbol{\nu}_T \end{pmatrix} \sim \mathsf{N}(\mathbf{0}, \mathbf{S}).$$

Therefore, by a simple change of variable from $\boldsymbol{\nu}$ to $\boldsymbol{\eta}$, we have

$$\boldsymbol{\eta} \,|\, \boldsymbol{\theta} \sim \mathsf{N}\left(\widetilde{\boldsymbol{\eta}}, \mathbf{K}^{-1}\right), \tag{21}$$

where $\widetilde{\boldsymbol{\eta}} = \mathbf{H}^{-1}\mathbf{Z}\boldsymbol{\gamma}$ and the $Tq \times Tq$ precision matrix $\mathbf{K}$ is given by $\mathbf{K} = \mathbf{H}'\mathbf{S}^{-1}\mathbf{H}$, i.e.,

$$\mathbf{K} = \begin{pmatrix} \mathbf{F}_2'\boldsymbol{\Omega}_{22}^{-1}\mathbf{F}_2 + \mathbf{D}^{-1} & -\mathbf{F}_2'\boldsymbol{\Omega}_{22}^{-1} & & & \\ -\boldsymbol{\Omega}_{22}^{-1}\mathbf{F}_2 & \mathbf{F}_3'\boldsymbol{\Omega}_{22}^{-1}\mathbf{F}_3 + \boldsymbol{\Omega}_{22}^{-1} & -\mathbf{F}_3'\boldsymbol{\Omega}_{22}^{-1} & & \\ & \ddots & \ddots & \ddots & \\ & & -\boldsymbol{\Omega}_{22}^{-1}\mathbf{F}_{T-1} & \mathbf{F}_T'\boldsymbol{\Omega}_{22}^{-1}\mathbf{F}_T + \mathbf{S}_{22}^{-1} & -\mathbf{F}_T'\boldsymbol{\Omega}_{22}^{-1} \\ & & & -\boldsymbol{\Omega}_{22}^{-1}\mathbf{F}_T & \boldsymbol{\Omega}_{22}^{-1} \end{pmatrix}. \tag{22}$$

It is important to realize that the precision matrix $\mathbf{K}$ is block-banded and contains a small number of non-zero elements in a narrow band around the main diagonal. This observation is important as it affords substantial saving in storage space and computation costs.

Since the densities $f(\mathbf{y} \,|\, \boldsymbol{\theta}, \boldsymbol{\eta})$ and $f(\boldsymbol{\eta} \,|\, \boldsymbol{\theta})$ in (18) and (19) are both Gaussian, the standard update for Gaussian linear regression (see, e.g. Koop, 2003; Gelman et al., 2003) implies that the conditional posterior density $f(\boldsymbol{\eta} \,|\, \mathbf{y}, \boldsymbol{\theta}) \propto f(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}) f(\boldsymbol{\eta} \,|\, \boldsymbol{\theta})$ is also Gaussian:

$$\boldsymbol{\eta} \,|\, \mathbf{y}, \boldsymbol{\theta} \sim \mathsf{N}(\widehat{\boldsymbol{\eta}}, \mathbf{P}^{-1}), \tag{23}$$

where the precision $\mathbf{P}$ and the mean $\widehat{\boldsymbol{\eta}}$ are given by

$$\mathbf{P} = \mathbf{K} + \mathbf{G}'\left(\mathbf{I} \otimes \boldsymbol{\Omega}_{11}^{-1}\right)\mathbf{G}, \tag{24}$$

$$\widehat{\boldsymbol{\eta}} = \mathbf{P}^{-1}\left(\mathbf{K}\widetilde{\boldsymbol{\eta}} + \mathbf{G}'\left(\mathbf{I} \otimes \boldsymbol{\Omega}_{11}^{-1}\right)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right). \tag{25}$$

Since $\mathbf{G}'\left(\mathbf{I} \otimes \boldsymbol{\Omega}_{11}^{-1}\right)\mathbf{G}$ is banded, it follows that $\mathbf{P}$ is also banded. This allows sampling $\boldsymbol{\eta} \,|\, \mathbf{y}, \boldsymbol{\theta}$ without the need to carry out an inversion to obtain $\mathbf{P}^{-1}$ and $\widehat{\boldsymbol{\eta}}$ in (25). Specifically, the mean $\widehat{\boldsymbol{\eta}}$ can be found in two steps. First, we compute the (banded) Cholesky factor $\mathbf{C}$ of $\mathbf{P}$ such that $\mathbf{C}'\mathbf{C} = \mathbf{P}$. Second, we solve

$$\mathbf{C}'\mathbf{C}\widehat{\boldsymbol{\eta}} = \mathbf{K}\widetilde{\boldsymbol{\eta}} + \mathbf{G}'(\mathbf{I} \otimes \boldsymbol{\Omega}_{11}^{-1})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{26}$$

for $\mathbf{C}\widehat{\boldsymbol{\eta}}$ by forward-substitution and then using the result to solve for $\widehat{\boldsymbol{\eta}}$ by back-substitution. The two steps are done in $\mathcal{O}(Tq^2)$ operations. Similarly, to obtain a random draw from $\mathsf{N}(\widehat{\boldsymbol{\eta}}, \mathbf{P}^{-1})$ efficiently, sample $\mathbf{u} \sim \mathsf{N}(\mathbf{0}, \mathbf{I})$, and solve $\mathbf{C}\mathbf{x} = \mathbf{u}$ for $\mathbf{x}$ by back-substitution. It follows that $\mathbf{x} \sim \mathsf{N}(\mathbf{0}, \mathbf{P}^{-1})$. Adding the mean $\widehat{\boldsymbol{\eta}}$ to $\mathbf{x}$, one obtains a draw $\boldsymbol{\eta} \sim \mathsf{N}(\widehat{\boldsymbol{\eta}}, \mathbf{P}^{-1})$. We summarize the above procedures in the following algorithm.

Algorithm 4. *Efficient State Smoothing and Simulation*

1. *Compute* $\mathbf{P}$ *in* (24) *and obtain its Cholesky factor* $\mathbf{C}$ *such that* $\mathbf{P} = \mathbf{C}'\mathbf{C}$.
2. `Smoothing:` *Solve* (26) *by forward- and back-substitution to obtain* $\widehat{\boldsymbol{\eta}}$.
3. `Simulation:` *Sample* $\mathbf{u} \sim \mathsf{N}(\mathbf{0}, \mathbf{I})$, *and solve* $\mathbf{C}\mathbf{x} = \mathbf{u}$ *for* $\mathbf{x}$ *by back-substitution and take* $\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}} + \mathbf{x}$, *so that* $\boldsymbol{\eta} \sim \mathsf{N}(\widehat{\boldsymbol{\eta}}, \mathbf{P}^{-1})$.

Lastly, we describe an efficient method to evaluate the integrated likelihood $f(\mathbf{y} \mid \boldsymbol{\theta})$, defined as

$$f(\mathbf{y} \mid \boldsymbol{\theta}) = \int f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) f(\boldsymbol{\eta} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\eta}, \tag{27}$$

where $f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$ is the likelihood and $f(\boldsymbol{\eta} \mid \boldsymbol{\theta})$ represents the prior. This quantity is often used in likelihood evaluation for classical and Bayesian problems involving optimization and model comparison. Evaluating the integrated likelihood via the multivariate integration in (27) is computationally intensive and often impractical. However, the derivation of the full conditional density $f(\boldsymbol{\eta}|\mathbf{y}, \boldsymbol{\theta})$ discussed above provides a simple way to evaluate $f(\mathbf{y} \mid \boldsymbol{\theta})$ without invoking (27). Specifically, it follows from the Bayes' theorem that integrated likelihood as

$$f(\mathbf{y} \mid \boldsymbol{\theta}) = \frac{f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) f(\boldsymbol{\eta} \mid \boldsymbol{\theta})}{f(\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\theta})}, \tag{28}$$

where $f(\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\theta})$ denotes the full conditional posterior density of $\boldsymbol{\eta}$ given $(\mathbf{y}, \boldsymbol{\theta})$. Due to the banded nature of the prior and posterior precision matrices ($\mathbf{K}$ and $\mathbf{P}$, respectively), evaluation of the integrated likelihood $f(\mathbf{y} \mid \boldsymbol{\theta})$ at the point $\boldsymbol{\theta}$ can be done efficiently simply by evaluating the right-hand-side of (28) at a single point $(\boldsymbol{\theta}, \boldsymbol{\eta})$. Note that the choice is arbitrary, in particular, the choice $\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}$ eliminates the need to compute the exponential part of the density function in the denominator density.

Appendix B: Proofs of Propositions

Proof of Proposition 1. We prove the claims for the VM estimator; the proof for the CE estimator follows similarly. We first show that the marginal likelihood $p(\mathbf{y})$ is finite for any given $\mathbf{y}$. Using Assumption 1 and the fact that the prior is proper, we have

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leqslant p(\mathbf{y} \mid \widehat{\boldsymbol{\theta}}) \int p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = p(\mathbf{y} \mid \widehat{\boldsymbol{\theta}}) < \infty.$$

Now write the VM estimator as the the average of N iid copies of the random variable Z_n:

$$\widehat{p}_{\mathrm{vm}}(\mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} Z_n,$$

where $Z_n = p(\mathbf{y} \mid \boldsymbol{\theta}_n) p(\boldsymbol{\theta}_n) / f(\boldsymbol{\theta}_n; \mathbf{v}^*_{\mathrm{vm}})$ and $\boldsymbol{\theta}_n \sim f(\boldsymbol{\theta}_n; \mathbf{v}^*_{\mathrm{vm}})$. It is easy to see that

$$\mathbb{E} Z_n = \int \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_n) p(\boldsymbol{\theta}_n)}{f(\boldsymbol{\theta}_n; \mathbf{v}^*_{\mathrm{vm}})} f(\boldsymbol{\theta}_n; \mathbf{v}^*_{\mathrm{vm}}) \mathrm{d}\boldsymbol{\theta}_n = \int p(\mathbf{y} \mid \boldsymbol{\theta}_n) p(\boldsymbol{\theta}_n) \mathrm{d}\boldsymbol{\theta}_n = p(\mathbf{y}) < \infty.$$

Therefore, by the (weak) law of large number, $\widehat{p}_{\mathrm{vm}}(\mathbf{y})$ converges in probability to $p(\mathbf{y})$ as $N \to \infty$. Next, note that

$$\mathbb{E}\left[\widehat{p}_{\mathrm{vm}}(\mathbf{y})\right] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} Z_n\right] = \frac{1}{N} \sum_{n=1}^{N} p(\mathbf{y}) = p(\mathbf{y}). \tag{29}$$

That is, the VM estimator is also unbiased. □

If the parametric family $\mathcal{F}$ contains the prior density $p(\cdot)$, then the variance of the VM estimator is finite. Furthermore, the VM estimator has an asymptotic normal distribution.

Proof of Proposition 2. Let $Z = p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})/f(\boldsymbol{\theta}; \mathbf{v}^*_{\mathrm{vm}})$, where $\boldsymbol{\theta} \sim f(\boldsymbol{\theta}; \mathbf{v}^*_{\mathrm{vm}})$. Then the VM estimator $\widehat{p}_{\mathrm{vm}}(\mathbf{y})$ is simply the average of N independent copies of Z. We will first show that the second moment of Z, $\mathbb{E}Z^2$, is finite. By the definition of $\mathbf{v}^*_{\mathrm{vm}}$ in (4) and the assumption that the prior density $p(\boldsymbol{\theta})$ is in the parametric family $\mathcal{F}$, it follows that

$$\mathbb{E}Z^2 = \int \frac{p(\mathbf{y} \mid \boldsymbol{\theta})^2 p(\boldsymbol{\theta})^2}{f(\boldsymbol{\theta}; \mathbf{v}^*_{\mathrm{vm}})} \mathrm{d}\boldsymbol{\theta} \leqslant \int \frac{p(\mathbf{y} \mid \boldsymbol{\theta})^2 p(\boldsymbol{\theta})^2}{p(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} = \int p(\mathbf{y} \mid \boldsymbol{\theta})^2 p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty.$$

Now, we have

$$\mathrm{Var}(\widehat{p}_{\mathrm{vm}}(\mathbf{y})) = \frac{\mathrm{Var}(Z)}{N} \leqslant \frac{\mathbb{E}Z^2}{N} < \infty.$$

Since $\mathbb{E}Z = p(\mathbf{y})$ and $\sigma^2 \equiv \mathrm{Var}(Z) < \infty$, by the central limit theorem, we finally have

$$\sqrt{N}(\widehat{p}_{\mathrm{vm}}(\mathbf{y}) - p(\mathbf{y})) \xrightarrow{\mathrm{d}} \mathsf{N}(0, \sigma^2),$$

as $N \to \infty$. □

References

D. Ardia, N. Basturk, L. Hoogerheide, and H. K. van Dijk. A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics and Data Analysis*, 2010. In press.

K. J. Arrow. *Essays in the Theory of Risk Bearing*. North-Holland, Amsterdam, 1970.

S. Asmussen, R. Y. Rubinstein, and D. P. Kroese. Heavy tails, importance sampling and cross-entropy. *Stochastic Models*, 21:57–76, 2005.

B. Bernanke, J. Boivin, and P. S. Eliasz. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, 120(1):387–422, 2005.

Z. I. Botev and D. P. Kroese. The generalized cross-entropy method, with applications to probability density estimation. *Methodology and Computing in Applied Probability*, 13:1–27, 2011.

F. Canova. Modelling and forecasting exchange rates with a Bayesian time-varying coefficient model. *Journal of Economic Dynamics and Control*, 17:233–262, 1992.

C. K. Carter and R. Kohn. On Gibbs sampling for state space models. *Biometrika*, 81:541–553, 1994.

J. C. C. Chan and I. Jeliazkov. Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation*, 1:101–120, 2009.

J. C. C. Chan and D. P. Kroese. Efficient estimation of large portfolio loss probabilities in t-copula models. *European Journal of Operational Research*, 205:361–367, 2010.

J. C. C. Chan, P. W. Glynn, and D. P. Kroese. A comparison of cross-entropy and variance minimization strategies. *Journal of Applied Probability*, 48A:183–194, 2011.

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

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

P. T. de Boer, D. P. Kroese, and R. Y. Rubinstein. A fast cross-entropy method for estimating buffer overflows in queueing networks. *Management Science*, 50:883–895, 2004.

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

S. Fruhwirth-Schnatter and Helga Wagner. Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling. *Computational Statistics and Data Analysis*, 52(10):4608 – 4624, 2008.

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

A. Gelman and X. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, New York, second edition, 2003.

J. Geweke. Exact predictive densities for linear models with ARCH disturbances. *Journal of Econometrics*, 40(1):63 – 86, 1989.

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

A. Golan. Information and entropy econometrics – a review and synthesis. *Foundations and Trends in Econometrics*, 2(1-2):1–145, 2008.

A. Golan, G. G. Judge, and D. Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Wiley, New York, 1996.

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

L. F. Hoogerheide, J. F. Kaashoek, and H. K. van Dijk. On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks. *Journal of Econometrics*, 139(1):154–180, 2007.

K-P. Hui, N. Bean, M. Kraetzl, and D. P. Kroese. The cross-entropy method for network reliability estimation. *Annals of Operations Research*, 134:101–118, 2005.

I. Jeliazkov and R. Liu. A model-based ranking of U.S. recessions. *Economics Bulletin*, 30(3): 2289–2296, 2010.

J. M. Keith, D. P. Kroese, and G. Y. Sofronov. Adaptive independence samplers. *Statistics and Computing*, 18:409–420, 2008.

S. Kim, N. Shephard, and S. Chib. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, 65(3):361–393, 1998.

G. Koop. *Bayesian Econometrics*. Wiley & Sons, New York, 2003.

G. Koop and D. Korobilis. Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, 3(4):267–358, 2010.

D. P. Kroese. The cross-entropy method. In James J. Cochran, Louis A. Cox, Pinar Keskinocak, Jeffrey P. Kharoufeh, and J. Cole Smith, editors, *Wiley Encyclopedia of Operations Research and Management Science*. Wiley & Sons, New York, 2010.

D. P. Kroese and R. Y. Rubinstein. The transform likelihood ratio method for rare event simulation with heavy tails. *Queueing Systems*, 46:317–351, 2004.

D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo Methods*. John Wiley & Sons, New York, 2011.

E. Maasoumi. A compendium to information theory in economics and econometrics. *Econometric Reviews*, 12:137–181, 1993.

J. Marschak. Economics of information systems. *Journal of the American Statistical Association*, 66:192–219, 1971.

T. A. Mroz. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55:765–799, 1987.

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

R. Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99:89–112, 1997.

R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization Monte-Carlo Simulation, and Machine Learning*. Springer-Verlag, New York, 2004.

C. A. Sims. Macroeconomics and reality. *Econometrica*, 48:1–48, 1980.

A. Zellner. Bayesian methods and entropy in economics and econometrics. In W. T. Grandy and L. H. Schick, editors, *Maximum Entropy and Bayesian Methods*, volume 43 of *Fundamental Theories of Physics*, pages 17–31. Springer Netherlands, 1991.