Efficient Computation of Marginal Likelihood

Zhongfang He*

First Version: April 3, 2024

Abstract

A marginal likelihood is the likelihood of a model that integrates out all model parameters and is the essential quantity for Bayesian model comparison. Though conceptually simple, computing the marginal likelihood in practice is often difficult when the number of model parameters is large, particularly for models with latent variables. In this paper, I propose a new approach of constructing the importance sampler to implement the importance sampling method for estimating the marginal likelihood based on the principle of minimizing its Monte Carlo sampling variance. The same approach can be used to construct the tuning function for implementing the Gelfand-Dey method of computing the marginal likelihood. The proposed approach is simple yet efficient. Performance of the new approach is illustrated in empirical examples including the probit model and linear regressions with stochastic volatility.

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

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

1 Introduction

The marginal likelihood integrates the likelihood function of a model with respect to the prior distribution of model parameters and lies at the heart of model comparison in the Bayesian framework (e.g. Aitkin (1991), Carlin and Chib (1995) and Kass and Raftery (1995)). Specifically, the marginal likelihood of the data $\boldsymbol{y} = \{\boldsymbol{y}_t\}_{t=1}^T$ for a given model is:

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{1}$$

where $\boldsymbol{\theta}$ is the model parameters, $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$ (assumed to be proper) and $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the model likelihood function. Computing $p(\boldsymbol{y})$ in practice is often difficult. Except in rare cases, the integral in Equation (1) is not analytically available and requires numerical integration that can be very challenging when the dimension of $\boldsymbol{\theta}$ is large. Various methods have been proposed in the literature to compute $p(\boldsymbol{y})$ efficiently. Examples include the importance sampling method (Hammersley and Handscomb (1964), Kloek and Dijk (1978), Geweke (1989)), the harmonic mean method (Newton and Raftery (1994), Gelfand and Dey (1994)), the Chib method (Chib (1995), Chib and Jeliazkov (2001)), the bridge sampling method (Meng and Wong (1996), Meng and Schilling (2002)) and the path sampling method (Gelman and Meng (1998), Friel and Pettitt (2008)), among many others. Although substantial progress has been made over the years, calculating marginal likelihoods for model comparison remains a hurdle in many applied studies, especially for models with latent variables. Many of the existing methods for marginal likelihood computation are conceptually simple but can be difficult to implement in practice for various reasons. This paper contributes to the literature by proposing a new approach of constructing the auxiliary distribution to implement the importance sampling (IS hereafter) method and the modified harmonic mean method of Gelfand and Dey (1994) (GD hereafter). The new approach is both simple and effective.

The basic idea of the proposed approach is that an auxiliary function $g(\boldsymbol{\theta})$ of the model

parameters can be introduced to approximate the log model likelihood function $\log(p(\boldsymbol{y}|\boldsymbol{\theta}))$ such that the pseudo posterior proportional to $p(\boldsymbol{\theta})\exp\left(g(\boldsymbol{\theta})\right)$ has an analytically closed form while being close to the true posterior. By using this pseudo posterior as the auxiliary distribution to implement the IS or GD method, the sampling variation of the Monte Carlo estimate of the marginal likelihood is limited to the approximation error between $\exp(g(\boldsymbol{\theta}))$ and $p(\boldsymbol{y}|\boldsymbol{\theta})$ and can be greatly reduced. Relative to the approaches that construct an auxiliary distribution to directly approximate the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ (e.g. the cross entropy approach of Chan and Eisenstat (2015) and Chan (2023)), the proposed approach is more flexible to allow complex dependence between elements of $\boldsymbol{\theta}$ in the constructed auxiliary distribution (e.g. the dependence between latent variables and fixed model parameters in latent variable models) and hence can be more accurate for approximating the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$.

The computational cost of the proposed approach is low as the free parameters of the auxiliary function $g(\boldsymbol{\theta})$ can be determined by a sequence of low-dimensional linear regressions that are guided by directly minimizing the Monte Carlo variance of the marginal likelihood estimate. Both the importance sampler of the IS method and the tuning function of the GD method can be constructed by the proposed approach. Encouraging results are found when the new approach is applied to a number of empirical examples including the probit model and the linear regression models with stochastic volatility.

In the remainder of the paper, Section 2 describes the proposed approach in detail. The approaches for comparison are described in Section 3. Empirical examples are provided in Section 4. Section 5 concludes. Additional details are provided in the appendices.

2 The Approach

The proposed approach is motivated by searching for a suitable importance sampler to implement the IS method for computing the marginal likelihood. Hence the discussion will first focus on the IS method in Section 2.1 and 2.2. Later in Section 2.3, I will discuss the use of the same approach to construct the tuning function for the GD method.

Without loss of generality, I assume that the model parameters take values on the whole real line for convenience.[1] In this paper, I will focus on the Gaussian distribution as the importance sampler.

2.1 The Simple Case

I begin with the simple case where the prior $p(\boldsymbol{\theta})$ is Gaussian and will discuss the more general case in Section 2.2. Let $\boldsymbol{\theta}$ denote a K-by-1 vector of the model parameters. The IS method introduces an auxiliary distribution $q(\boldsymbol{\theta})$ to compute the marginal likelihood via the following equation:

$$p(\boldsymbol{y}) = \int \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2}$$

A Monte Carlo estimate of the marginal likelihood is:

$$\hat{p}(\boldsymbol{y}) = \frac{1}{m} \sum_{i=1}^{m} \frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})p(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})} \tag{3}$$

where $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{m}$ are i.i.d. draws from the importance sampler $q(\boldsymbol{\theta})$. When $q(\boldsymbol{\theta})$ is close to the posterior $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, the IS estimator of Equation (3) enjoys small sampling variance and hence greater accuracy. An obvious choice of $q(\boldsymbol{\theta})$, when $\boldsymbol{\theta}$ is of modest dimension, is a Gaussian distribution whose mean and covariance matrix equal the posterior mean and covariance matrix of $\boldsymbol{\theta}$. However, when the dimension of $\boldsymbol{\theta}$ increases, constructing a suitable $q(\boldsymbol{\theta})$ in a practical way becomes more challenging.

[1] If a parameter takes values on a subset of the real line (e.g. $\mathbb{R}_+$ or $(0, 1)$), one can always re-parameterize the parameter such that its support becomes $\mathbb{R}$ and adjust its prior accordingly.

In this paper, we choose the importance sampler $q(\boldsymbol{\theta})$ as a Gaussian distribution and calibrate its free parameters based on the principle of minimizing the variance of the IS estimator:

$$\begin{aligned} V(\hat{p}(\boldsymbol{y})) &= \frac{1}{m}\int\left(\frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}-p(\boldsymbol{y})\right)^2 q(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \frac{p(\boldsymbol{y})^2}{m}\int\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{q(\boldsymbol{\theta})}+\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})}-2\right)p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \end{aligned} \tag{4}$$

where the second line of the equation is derived by using the Bayes' law $p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$. For the task of minimizing $V(\hat{p}(\boldsymbol{y}))$, one can in theory calibrate $q(\boldsymbol{\theta})$ to directly minimize the Monte Carlo counterpart of the integral in Equation (4) based on draws from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. But such a direct optimization scheme is usually infeasible as the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ appears inside the integrand and is rarely analytically available.

To simplify the problem, let $f(\boldsymbol{\theta}) = \log(p(\boldsymbol{\theta}|\boldsymbol{y})) - \log(q(\boldsymbol{\theta}))$. A second-order Taylor expansion at $f(\boldsymbol{\theta}) = 0$ gives $\exp(f(\boldsymbol{\theta})) + \exp(-f(\boldsymbol{\theta})) - 2 \approx f(\boldsymbol{\theta})^2$. Inserting this approximation into Equation (4) returns:

$$\begin{aligned} V(\hat{p}(\boldsymbol{y})) &\approx \frac{p(\boldsymbol{y})^2}{m}\int\left(\log(p(\boldsymbol{\theta}|\boldsymbol{y}))-\log(q(\boldsymbol{\theta}))\right)^2 p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \\ &= \frac{p(\boldsymbol{y})^2}{m}\int\left(\log(p(\boldsymbol{\theta}))+\log(p(\boldsymbol{y}|\boldsymbol{\theta}))-\log(p(\boldsymbol{y}))-\log(q(\boldsymbol{\theta}))\right)^2 p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \end{aligned} \tag{5}$$

Thus the task of minimizing $V(\hat{p}(\boldsymbol{y}))$ can be reduced to a least-squares problem based on posterior draws of $\boldsymbol{\theta}$. However, directly minimizing the Monte Carlo counterpart of the integral in Equation (5) is still difficult as one needs to estimate the mean and in particular the covariance matrix for $q(\boldsymbol{\theta})$.

Our solution is to decompose $p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1}, \boldsymbol{\theta})$, where $\boldsymbol{y}^t = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_t\}$ is the history up to point t and $\boldsymbol{y}^0$ is an empty set, and introduce a series of auxiliary functions $g_t(\boldsymbol{\theta}; \boldsymbol{a}_t) = a_{1,t} + \boldsymbol{a}_{2,t}'\boldsymbol{\theta} + a_{3,t}\boldsymbol{\theta}'\boldsymbol{A}_t\boldsymbol{\theta}$ that is quadratic in $\boldsymbol{\theta}$ to approximate $\log(p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1}, \boldsymbol{\theta}))$ for each $t = 1, ..., T$, where $\boldsymbol{a}_t = \{a_{1,t}, \boldsymbol{a}_{2,t}, a_{3,t}\}$ denotes the free parameters in each auxiliary function and $\boldsymbol{A}_t$ is a known K-by-K positive semi-definite matrix.[2] Define the

[2] As guided by a second-order Taylor expansion of $\log(p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1}, \boldsymbol{\theta}))$ at a given value $\bar{\boldsymbol{\theta}}$ (e.g. the posterior

function $g(\boldsymbol{\theta};\boldsymbol{a})$ as:

$$\begin{aligned} g(\boldsymbol{\theta};\boldsymbol{a}) &= \sum_{t=1}^{T} g_t(\boldsymbol{\theta};\boldsymbol{a}_t) \\ &= a_1 + \boldsymbol{a}_2'\boldsymbol{\theta} + \boldsymbol{\theta}'\boldsymbol{A}\boldsymbol{\theta} \end{aligned} \tag{6}$$

where $a_1 = \sum_{t=1}^{T} a_{1,t}$, $\boldsymbol{a}_2 = \sum_{t=1}^{T} \boldsymbol{a}_{2,t}$, $\boldsymbol{A} = \sum_{t=1}^{T} a_{3,t}\boldsymbol{A}_t$ and $\boldsymbol{a} = \{\boldsymbol{a}_t\}_{t=1}^{T}$. Since the prior $p(\boldsymbol{\theta})$ is Gaussian, it can be shown that, under suitable conditions for $\{a_{3,t}\}_{t=1}^{T}$, the product $p(\boldsymbol{\theta})\exp\left(g(\boldsymbol{\theta};\boldsymbol{a})\right)$ is proportional to a Gaussian density function of $\boldsymbol{\theta}$ (see Theorem 2.1 below). Denote the normalizing constant $c(\boldsymbol{a}) = \int p(\boldsymbol{\theta})\exp\left(g(\boldsymbol{\theta};\boldsymbol{a})\right) d\boldsymbol{\theta}$. Clearly $c(\boldsymbol{a})$ is analytically available. We construct the importance sampler as $q(\boldsymbol{\theta};\boldsymbol{a}) = p(\boldsymbol{\theta})\exp\left(g(\boldsymbol{\theta};\boldsymbol{a})\right)/c(\boldsymbol{a})$. Inserting $q(\boldsymbol{\theta};\boldsymbol{a})$ into Equation (5) leads to:

$$V(\hat{p}(\boldsymbol{y})) \approx \frac{p(\boldsymbol{y})^2}{m} \int \left(\log(p(\boldsymbol{y}|\boldsymbol{\theta})) - g(\boldsymbol{\theta};\boldsymbol{a}) + \log(c(\boldsymbol{a})) - \log(p(\boldsymbol{y}))\right)^2 p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \tag{7}$$

For practicality, we ignore the cross product terms and approximate the square function inside the integral in Equation (7) as:

$$\sum_{t=1}^{T} \left(\log(p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1},\boldsymbol{\theta})) - g_t(\boldsymbol{\theta};\boldsymbol{a}_t)\right)^2 + \left(\log(c(\boldsymbol{a})) - \log(p(\boldsymbol{y}))\right)^2$$

Thus, minimizing $V(\hat{p}(\boldsymbol{y}))$ of Equation (7) is simplified to minimize the components:

$$\begin{aligned} &\int \left(\log(p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1},\boldsymbol{\theta})) - g_t(\boldsymbol{\theta};\boldsymbol{a}_t)\right)^2 p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \\ =&\int \left(\log(p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1},\boldsymbol{\theta})) - a_{1,t} - \boldsymbol{a}_{2,t}'\boldsymbol{\theta} - a_{3,t}\boldsymbol{\theta}'\boldsymbol{A}_t\boldsymbol{\theta}\right)^2 p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \end{aligned} \tag{8}$$

for each $t = 1, ..., T$ and $\left(\log(c(\boldsymbol{a})) - \log(p(\boldsymbol{y}))\right)^2$.

To minimize each component of Equation (8), one can simply run a regression of the function $\log(p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1},\boldsymbol{\theta}))$ on a constant, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'\boldsymbol{A}_t\boldsymbol{\theta}$ based on posterior draws of $\boldsymbol{\theta}$ and determine the optimal value of the free parameters $\boldsymbol{a}_t^* = \{a_{1,t}^*, \boldsymbol{a}_{2,t}^*, a_{3,t}^*\}$ for each auxiliary

mean of $\boldsymbol{\theta}$), one can set $\boldsymbol{A}_t \propto \frac{\partial^2 \log(p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1},\bar{\boldsymbol{\theta}}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$ if a positive semi-definite $\boldsymbol{A}_t$ can be obtained. Otherwise, one could simply set $\boldsymbol{A}_t$ as an identify matrix.

function $g_t(\boldsymbol{\theta}; \boldsymbol{a}_t)$, $t = 1, ..., T$. For $g_t(\boldsymbol{\theta}; \boldsymbol{a}_t^*)$ to be useful, we require that $a_{3,t}^* \leq 0$ for all $t = 1, ..., T$ (see Theorem 2.1 and Remark 1 below).[3] On the other hand, the other component $(\log(c(\boldsymbol{a})) - \log(p(\boldsymbol{y})))^2$ is minimized when $c(\boldsymbol{a}) = p(\boldsymbol{y})$. Heuristically, as $\boldsymbol{a}_t^*$ minimizes the difference between $\log(p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1}, \boldsymbol{\theta}))$ and $g_t(\boldsymbol{\theta}; \boldsymbol{a}_t)$, we have $\exp(g(\boldsymbol{\theta}; \boldsymbol{a}^*)) \approx p(\boldsymbol{y}|\boldsymbol{\theta})$, where $\boldsymbol{a}^* = \{\boldsymbol{a}_t^*\}_{t=1}^T$, and thus naturally $c(\boldsymbol{a}^*) \approx p(\boldsymbol{y})$ by its definition. Therefore, no separate optimization is performed for this component. Inserting $\boldsymbol{a}^*$ into $q(\boldsymbol{\theta}; \boldsymbol{a})$ determines the optimized importance sampler $q(\boldsymbol{\theta}; \boldsymbol{a}^*)$.

Further intuition of the proposed approach can be gained by going back to the IS representation of the marginal likelihood in Equation (2). Heuristically, recall that the goal is to find an importance sampler $q(\boldsymbol{\theta}; \boldsymbol{a})$ that is close to the posterior $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. By introducing the auxiliary functions $\{g_t(\boldsymbol{\theta}; \boldsymbol{a}_t)\}_{t=1}^T$, one can re-write the marginal likelihood as:

$$p(\boldsymbol{y}) = \int \left(\frac{p(\boldsymbol{y}|\boldsymbol{\theta})}{\exp(g(\boldsymbol{\theta}; \boldsymbol{a}))} \right) \left(\frac{p(\boldsymbol{\theta})\exp(g(\boldsymbol{\theta}; \boldsymbol{a}))}{q(\boldsymbol{\theta}; \boldsymbol{a})} \right) q(\boldsymbol{\theta}; \boldsymbol{a}) d\boldsymbol{\theta} \tag{9}$$

For the first term $\frac{p(\boldsymbol{y}|\boldsymbol{\theta})}{\exp(g(\boldsymbol{\theta};\boldsymbol{a}))}$ of the integrand, the denominator $\exp(g(\boldsymbol{\theta}; \boldsymbol{a}))$ is made "close" to the numerator $p(\boldsymbol{y}|\boldsymbol{\theta})$ through a series of simple linear regressions as discussed in the preceding paragraph, while the second term $\frac{p(\boldsymbol{\theta})\exp(g(\boldsymbol{\theta};\boldsymbol{a}))}{q(\boldsymbol{\theta};\boldsymbol{a})}$ of the integrand is a constant as per the construction of $q(\boldsymbol{\theta}; \boldsymbol{a})$. Thus the new approach achieves the goal of making the importance sampler $q(\boldsymbol{\theta}; \boldsymbol{a})$ close to $p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$ by using the auxiliary functions $\{g_t(\boldsymbol{\theta}; \boldsymbol{a}_t)\}_{t=1}^T$ as a series of "bridges". Instead of directly jumping from $q(\boldsymbol{\theta}, \boldsymbol{a})$ to the target $p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$, the new approach takes a shortcut via the constructed bridges, which facilitates easy construction of a useful importance sampler with minimal computation cost.

Theorem 2.1. *Suppose $\boldsymbol{\theta}$ is a K-by-1 vector of random variables taking values on the whole real line. If $p(\boldsymbol{\theta})$ is a Gaussian density function $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $g(\boldsymbol{\theta}) = a_1 + \boldsymbol{a}_2'\boldsymbol{\theta} + \boldsymbol{\theta}'\boldsymbol{A}\boldsymbol{\theta}$ is a quadratic function of $\boldsymbol{\theta}$ where a_1 is a scalar, $\boldsymbol{a}_2$ is a K-by-1 vector and $\boldsymbol{A}$ is a K-*

[3] If the regression estimate of $a_{3,t}$ is positive, one can set $a_{3,t}^* = 0$ and re-run the regression by keeping only the constant and $\boldsymbol{\theta}$ as regressors.

by-K negative semi-definite matrix, then the product $p(\boldsymbol{\theta})exp(g(\boldsymbol{\theta}))$ *is proportional to the Gaussian density function* $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *where* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *satisfy* $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_0^{-1} - 2\boldsymbol{A}$ *and* $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{a}_2$.

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

Remark 1. In the simple case of this section, $\boldsymbol{A} = \sum_{t=1}^{T} a_{3,t}\boldsymbol{A}_t$ where each $\boldsymbol{A}_t$ is a positive semi-definite matrix (Equation (6)). Given that $\boldsymbol{\Sigma}_0$ is positive definite by definition, a sufficient condition for the matrix $\boldsymbol{\Sigma}$ satisfying $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_0^{-1} - 2\boldsymbol{A}$ to be positive definite is that $a_{3,t} \leq 0$ for all t .

2.2 The General Case

In the general case, consider a model with latent variables $\boldsymbol{\theta} = \{\boldsymbol{\theta}_t\}_{t=1}^{T}$ and fixed model parameters $\boldsymbol{\delta}$, where the prior $p(\boldsymbol{\theta}, \boldsymbol{\delta}) = p(\boldsymbol{\delta})p(\boldsymbol{\theta}|\boldsymbol{\delta})$ is such that $p(\boldsymbol{\delta})$ could be non-Gaussian while $p(\boldsymbol{\theta}|\boldsymbol{\delta})$ is Gaussian. Let $\boldsymbol{\delta}$ be an L-by-1 vector, $\boldsymbol{\theta}_t$ a K-by-1 vector and $\boldsymbol{\theta} = [\boldsymbol{\theta}_1' \ ... \ \boldsymbol{\theta}_T']'$. Given an importance sampler $q(\boldsymbol{\theta}, \boldsymbol{\delta})$, the marginal likelihood can be written as:

$$p(\boldsymbol{y}) = \int \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\delta})p(\boldsymbol{\theta}, \boldsymbol{\delta})}{q(\boldsymbol{\theta}, \boldsymbol{\delta})} q(\boldsymbol{\theta}, \boldsymbol{\delta}) d\boldsymbol{\theta} d\boldsymbol{\delta} \tag{10}$$

with the Monte Carlo estimate:

$$\hat{p}(\boldsymbol{y}) = \frac{1}{m}\sum_{i=1}^{m} \frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)}, \boldsymbol{\delta}^{(i)})p(\boldsymbol{\theta}^{(i)}, \boldsymbol{\delta}^{(i)})}{q(\boldsymbol{\theta}^{(i)}, \boldsymbol{\delta}^{(i)})} \tag{11}$$

where $\{\boldsymbol{\theta}^{(i)}, \boldsymbol{\delta}^{(i)}\}_{i=1}^{m}$ are independent draws from the importance sampler $q(\boldsymbol{\theta}, \boldsymbol{\delta})$.

Constructing a suitable importance sampler $q(\boldsymbol{\theta}, \boldsymbol{\delta})$ here is trickier than in the simple case of Section 2.1 due to the presence of $\boldsymbol{\delta}$. We focus on the case where the model likelihood function can be decomposed as $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\delta}) = \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1}, \boldsymbol{\theta}, \boldsymbol{\delta}) = \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1}, \boldsymbol{\theta}_t, \boldsymbol{\delta})$. To simplify notations, we will write $p(\boldsymbol{y}_t|\boldsymbol{y}^{t-1}, \boldsymbol{\theta}_t, \boldsymbol{\delta})$ as $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \boldsymbol{\delta})$ hereafter. Section 2.2.1 will discuss the auxiliary function to approximate each $\log(p(\boldsymbol{y_t}|\boldsymbol{\theta}_t, \boldsymbol{\delta}))$, which is then used to construct the importance sampler described in Section 2.2.2.

2.2.1 Constructing the Auxiliary Function

We decompose the importance sampler $q(\boldsymbol{\theta}, \boldsymbol{\delta}) = q(\boldsymbol{\theta}|\boldsymbol{\delta})q(\boldsymbol{\delta})$ and discuss the two parts $q(\boldsymbol{\theta}|\boldsymbol{\delta})$ and $q(\boldsymbol{\delta})$ separately. Given the high dimension of $\boldsymbol{\theta}$, the key challenge of constructing the importance sampler $q(\boldsymbol{\theta}, \boldsymbol{\delta})$ is to determine the part $q(\boldsymbol{\theta}|\boldsymbol{\delta})$. To address this challenge, we aim to first find an auxiliary function $g_t(\boldsymbol{\theta}_t; \boldsymbol{\delta}, \boldsymbol{a}_t)$ that is quadratic in $\boldsymbol{\theta}_t$ to approximate $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \boldsymbol{\delta}))$ for each $t = 1, ..., T$.

A straightforward auxiliary function to approximate $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta_t}, \boldsymbol{\delta}))$, similar to the simple case of Section 2.1, would be $g(\boldsymbol{\theta}_t; \boldsymbol{\delta}, \boldsymbol{a}_t) = a_{1,t} + \boldsymbol{a}'_{2,t}\boldsymbol{\delta} + \boldsymbol{a}'_{3,t}\boldsymbol{\theta}_t + a_{4,t}\boldsymbol{\theta}'_t\boldsymbol{A}_t\boldsymbol{\theta}_t$, where $a_t = \{a_{1,t}, \boldsymbol{a}_{2,t}, \boldsymbol{a}_{3,t}, a_{4,t}\}$ are the free parameters and $\boldsymbol{A}_t$ is a known K-by-K positive semi-definite matrix. However, such a function is often too rigid to capture the complex relation between $\boldsymbol{\theta}_t$ and $\boldsymbol{\delta}$ in $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta_t}, \boldsymbol{\delta}))$. We have found that, by exploiting the specific feature of the model likelihood function under study, one can often construct a more flexible auxiliary function to improve the approximation.

To that end, we note that one can always write the function $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta_t}, \boldsymbol{\delta}))$ in the following form:

$$\log(p(\boldsymbol{y}_t|\boldsymbol{\theta_t}, \boldsymbol{\delta})) = f_{1,t}(\boldsymbol{\delta}) + f_{2,t}(\boldsymbol{\delta})'\boldsymbol{\theta}_t + f_{3,t}(\boldsymbol{\delta})h_t(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t) \tag{12}$$

where $f_{1,t}(\cdot) : \mathbb{R}^L \to \mathbb{R}$, $f_{2,t}(\cdot) : \mathbb{R}^L \to \mathbb{R}^K$ and $f_{3,t}(\cdot) : \mathbb{R}^L \to \{\mathbb{R}_+, 0\}$ are known functions of $\boldsymbol{\delta}$, $\boldsymbol{\xi}_t = f_{4,t}(\boldsymbol{\delta})$ is a known function mapping $\boldsymbol{\delta}$ to an H-by-1 vector (H could be different from K or L), and $h_t(\cdot) : \mathbb{R}^{K+H} \to \mathbb{R}$ is a known function of $\boldsymbol{\theta}_t$ and $\boldsymbol{\xi}_t$. Such a formulation of $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta_t}, \boldsymbol{\delta}))$ is always valid. For example, one can trivially let $f_{1,t}(\boldsymbol{\delta}) = 0$, $f_{2,t}(\boldsymbol{\delta}) = \boldsymbol{0}$, $f_{3,t}(\boldsymbol{\delta}) = 1$, $\boldsymbol{\xi}_t = f_{4,t}(\boldsymbol{\delta}) = \boldsymbol{\delta}$ and $h_t(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t) = \log(p(\boldsymbol{y}_t|\boldsymbol{\theta_t}, \boldsymbol{\delta}))$. However, depending on the specific model under study, other non-trivial forms of the functions $f_{1,t}(\boldsymbol{\delta})$, $f_{2,t}(\boldsymbol{\delta})$, $f_{3,t}(\boldsymbol{\delta})$, $f_{4,t}(\boldsymbol{\delta})$ and $h_t(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)$ can be derived. Examples are illustrated for the empirical models in Section 4. By re-writing $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta_t}, \boldsymbol{\delta}))$ in the form of Equation (12), one only needs to find a function of $\boldsymbol{\theta}_t$ and $\boldsymbol{\xi}_t$ that is quadratic in $\boldsymbol{\theta}_t$ to approximate the component

$h_t(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)$, while the functions $f_{1,t}(\boldsymbol{\delta})$, $f_{2,t}(\boldsymbol{\delta})$ and $f_{3,t}(\boldsymbol{\delta})$ can be added to the approximation of $h_t(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)$ to form an auxiliary function that allows complex relation between $\boldsymbol{\theta}_t$ and $\boldsymbol{\delta}$ to approximates $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta_t}, \boldsymbol{\delta}))$.

Specifically, guided by a second-order Taylor expansion of $h_t(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)$, we introduce the function:

$$\tilde{g}_t(\boldsymbol{\theta}_t; \boldsymbol{\delta}, \boldsymbol{a}_t) = a_{1,t} + \boldsymbol{a}'_{2,t}\boldsymbol{\xi}_t + \boldsymbol{a}'_{3,t}\text{vech}(\boldsymbol{\xi}_t\boldsymbol{\xi}'_t) + \boldsymbol{a}'_{4,t}\text{vec}(\boldsymbol{\xi}_t\boldsymbol{\theta}'_t) + \boldsymbol{a}'_{5,t}\boldsymbol{\theta}_t + a_{6,t}\boldsymbol{\theta}'_t\boldsymbol{A}_t\boldsymbol{\theta}_t \tag{13}$$

to approximate $h_t(\boldsymbol{\theta}_t, \boldsymbol{\xi_t})$, where $\text{vec}(\cdot)$ is the vectorization operator of a matrix that stacks the columns of the matrix from left to right to form a single column, $\text{vech}(\cdot)$ is a half-vectorization operator for a symmetric matrix, $\boldsymbol{a}_t = \{a_{1,t}, \boldsymbol{a}_{2,t}, \boldsymbol{a}_{3,t}, \boldsymbol{a}_{4,t}, \boldsymbol{a}_{5,t}, a_{6,t}\}$ are free parameters to be determined, and $\boldsymbol{A}_t$ is a known K-by-K positive semi-definite matrix.[4] Adding the other terms of $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta_t}, \boldsymbol{\delta}))$ in Equation (12) leads to the following auxiliary function to approximate $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \boldsymbol{\delta}))$ for each $t = 1, ..., T$:

$$\begin{aligned} g_t(\boldsymbol{\theta}_t; \boldsymbol{\delta}, \boldsymbol{a}_t) =& f_{1,t}(\boldsymbol{\delta}) + f_{2,t}(\boldsymbol{\delta})'\boldsymbol{\theta}_t + f_{3,t}(\boldsymbol{\delta})\tilde{g}_t(\boldsymbol{\theta}_t; \boldsymbol{\delta}, \boldsymbol{a}_t) \\ =& g_{0,t}(\boldsymbol{\delta}) + a_{1,t}g_{1,t}(\boldsymbol{\delta}) + \boldsymbol{a}'_{2,t}g_{2,t}(\boldsymbol{\delta}) + \boldsymbol{a}'_{3,t}g_{3,t}(\boldsymbol{\delta}) + \boldsymbol{a}'_{4,t}g_{4,t}(\boldsymbol{\delta})\boldsymbol{\theta}_t \\ &+ \boldsymbol{a}'_{5,t}g_{5,t}(\boldsymbol{\delta})\boldsymbol{\theta}_t + g_{6,t}(\boldsymbol{\delta})'\boldsymbol{\theta}_t + a_{6,t}\boldsymbol{\theta}'_t\boldsymbol{A}_t(\boldsymbol{\delta})\boldsymbol{\theta}_t \end{aligned} \tag{14}$$

where $g_{0,t}(\boldsymbol{\delta}) = f_{1,t}(\boldsymbol{\delta})$, $g_{1,t}(\boldsymbol{\delta}) = f_{3,t}(\boldsymbol{\delta})$, $g_{2,t}(\boldsymbol{\delta}) = f_{3,t}(\boldsymbol{\delta})\boldsymbol{\xi}_t$, $g_{3,t}(\boldsymbol{\delta}) = f_{3,t}(\boldsymbol{\delta})\text{vech}(\boldsymbol{\xi}_t\boldsymbol{\xi}'_t)$, $g_{4,t}(\boldsymbol{\delta}) = f_{3,t}(\boldsymbol{\delta})(\text{I}_K \otimes \boldsymbol{\xi}_t)$, $g_{5,t}(\boldsymbol{\delta}) = f_{3,t}(\boldsymbol{\delta})$, $g_{6,t}(\boldsymbol{\delta}) = f_{2,t}(\boldsymbol{\delta})$ and $A_t(\boldsymbol{\delta}) = f_{3,t}(\boldsymbol{\delta})\boldsymbol{A}_t$. The item $g_{4,t}(\boldsymbol{\delta})$ is obtained by using $\text{vec}(\boldsymbol{\xi}_t\boldsymbol{\theta}'_t) = (\text{I}_K \otimes \boldsymbol{\xi}_t)\boldsymbol{\theta}_t$, where $\otimes$ denotes the Kronecker product of two matrices.

[4]Similar to the simple case of Section 2.1, one can set $\boldsymbol{A}_t \propto \frac{\partial^2 h_t(\bar{\boldsymbol{\theta}}_t, \bar{\boldsymbol{\xi}}_t)}{\partial\boldsymbol{\theta}_t\partial\boldsymbol{\theta}'_t}$, where $\bar{\boldsymbol{\theta}}_t$ and $\bar{\boldsymbol{\xi}}_t$ are given values of $\boldsymbol{\theta}_t$ and $\boldsymbol{\xi}_t$ (e.g. the posterior mean of $\boldsymbol{\theta}_t$ and $\boldsymbol{\xi}_t$), as guided by a second-order Taylor expansion of $h_t(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)$ at $\bar{\boldsymbol{\theta}}_t$ and $\bar{\boldsymbol{\xi}}_t$ if a positive semi-definite $\boldsymbol{A}_t$ can be obtained. Otherwise, one could simply set $\boldsymbol{A}_t$ as an identify matrix.

2.2.2 Constructing the Importance Sampler

Now let $g(\boldsymbol{\theta};\boldsymbol{\delta},\boldsymbol{a}) = \sum_{t=1}^{T} g_t(\boldsymbol{\theta}_t;\boldsymbol{\delta},\boldsymbol{a}_t)$ where $\boldsymbol{a} = \{\boldsymbol{a}_t\}_{t=1}^{T}$ is the union of free parameters. One can write $g(\boldsymbol{\theta};\boldsymbol{\delta},\boldsymbol{a}) = \tilde{a}_1 + \tilde{\boldsymbol{a}}_2'\boldsymbol{\theta} + \boldsymbol{\theta}'\tilde{\boldsymbol{A}}\boldsymbol{\theta}$, where $\tilde{a}_1 = \sum_{t=1}^{T} g_{0,t}(\boldsymbol{\delta}) + a_{1,t}g_{1,t}(\boldsymbol{\delta}) + \boldsymbol{a}_{2,t}'g_{2,t}(\boldsymbol{\delta}) + \boldsymbol{a}_{3,t}'g_{3,t}(\boldsymbol{\delta})$, $\tilde{\boldsymbol{a}}_2$ is a KT-by-1 vector stacking $\{\boldsymbol{a}_{4,t}'g_{4,t}(\boldsymbol{\delta}) + \boldsymbol{a}_{5,t}'g_{5,t}(\boldsymbol{\delta}) + g_{6,t}(\boldsymbol{\delta})'\}_{t=1}^{T}$, and $\tilde{\boldsymbol{A}}$ is a KT-by-KT block-diagonal matrix with the diagonal block $a_{6,t}A_t(\boldsymbol{\delta})$ for $t = 1$, ..., T. Since $p(\boldsymbol{\theta}|\boldsymbol{\delta})$ is Gaussian, one can apply Theorem 2.1 to show that, under suitable conditions of $\{a_{6,t}\}_{t=1}^{T}$, the product $p(\boldsymbol{\theta}|\boldsymbol{\delta})\exp\left(g(\boldsymbol{\theta};\boldsymbol{\delta},\boldsymbol{a})\right)$ is proportional to a Gaussian density function of $\boldsymbol{\theta}$ and thus construct the part $q(\boldsymbol{\theta}|\boldsymbol{\delta})$ of the importance sampler as a Gaussian distribution $q(\boldsymbol{\theta}|\boldsymbol{\delta};\boldsymbol{a}) = p(\boldsymbol{\theta}|\boldsymbol{\delta})\exp\left(g(\boldsymbol{\theta};\boldsymbol{\delta},\boldsymbol{a})\right)/c(\boldsymbol{\delta},\boldsymbol{a})$ where $c(\boldsymbol{\delta},\boldsymbol{a}) = \int p(\boldsymbol{\theta}|\boldsymbol{\delta})\exp\left(g(\boldsymbol{\theta};\boldsymbol{\delta},\boldsymbol{a})\right)d\boldsymbol{\theta}$ is the normalizing constant and is analytically available. On the other hand, assuming that the dimension of the fixed parameters $\boldsymbol{\delta}$ is modest, the other part $q(\boldsymbol{\delta})$ of the importance sampler is specified as a Gaussian distribution for simplicity.

To calibrate the free parameters in the importance sampler, we note that, similar to the derivation of Equation (5), the Monte Carlo variance of the resulting IS estimate of the marginal likelihood can be approximated as:

$$V(\hat{p}(\boldsymbol{y})) \approx \frac{p(\boldsymbol{y})^2}{m}\int \left(\log(p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y})) - \log(q(\boldsymbol{\theta}|\boldsymbol{\delta};\boldsymbol{a})) - \log(q(\boldsymbol{\delta}))\right)^2 p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y})d\boldsymbol{\theta}d\boldsymbol{\delta} \qquad (15)$$

By inserting $q(\boldsymbol{\theta}|\boldsymbol{\delta};\boldsymbol{a}) = p(\boldsymbol{\theta}|\boldsymbol{\delta})\exp\left(g(\boldsymbol{\theta};\boldsymbol{\delta},\boldsymbol{a})\right)/c(\boldsymbol{\delta},\boldsymbol{a})$ and applying the Bayes' law, the squared function in the integrand of Equation (15) can be re-written as:

$$\left(f_1(\boldsymbol{\theta},\boldsymbol{\delta};\boldsymbol{a}) + f_2(\boldsymbol{\delta};\boldsymbol{a})\right)^2$$

where $f_1(\boldsymbol{\theta},\boldsymbol{\delta};\boldsymbol{a}) = \log(p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\delta})) - g(\boldsymbol{\theta};\boldsymbol{\delta},\boldsymbol{a}) = \sum_{t=1}^{T}\log(p(\boldsymbol{y}_t|\boldsymbol{\theta}_t,\boldsymbol{\delta})) - g_t(\boldsymbol{\theta}_t;\boldsymbol{\delta},\boldsymbol{a}_t)$ and $f_2(\boldsymbol{\delta};\boldsymbol{a}) = \log(c(\boldsymbol{\delta};\boldsymbol{a})) + \log(p(\boldsymbol{\delta})) - \log(p(\boldsymbol{y})) - \log(q(\boldsymbol{\delta}))$. By ignoring the cross product terms, minimizing the Monte Carlo variance $V(\hat{p}(\boldsymbol{y}))$ is simplified to minimize the following

components:

$$\begin{aligned}
&\int \left(\log(p(\boldsymbol{y}_t|\boldsymbol{\theta}_t,\boldsymbol{\delta})) - g_t(\boldsymbol{\theta}_t;\boldsymbol{\delta},\boldsymbol{a}_t)\right)^2 p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y})d\boldsymbol{\theta}d\boldsymbol{\delta} \\
=&\int \left(\log(p(\boldsymbol{y}_t|\boldsymbol{\theta}_t,\boldsymbol{\delta})) - g_{0,t}(\boldsymbol{\delta}) - a_{1,t}g_{1,t}(\boldsymbol{\delta}) - \boldsymbol{a}'_{2,t}g_{2,t}(\boldsymbol{\delta}) - \boldsymbol{a}'_{3,t}g_{3,t}(\boldsymbol{\delta})\right. \\
&\left. -\boldsymbol{a}'_{4,t}g_{4,t}(\boldsymbol{\delta})\boldsymbol{\theta}_t - \boldsymbol{a}'_{5,t}g_{5,t}(\boldsymbol{\delta})\boldsymbol{\theta}_t - g_{6,t}(\boldsymbol{\delta})'\boldsymbol{\theta}_t - a_{6,t}\boldsymbol{\theta}'_t A_t(\boldsymbol{\delta})\boldsymbol{\theta}_t)^2 p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y})d\boldsymbol{\theta}d\boldsymbol{\delta}\right.
\end{aligned} \tag{16}$$

for each $t = 1, ..., T$ and:

$$\begin{aligned}
&\int \left(f_2(\boldsymbol{\delta};\boldsymbol{a})\right)^2 p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y})d\boldsymbol{\theta}d\boldsymbol{\delta} \\
=&\int \left(\log(c(\boldsymbol{\delta};\boldsymbol{a})) + \log(p(\boldsymbol{\delta})) - \log(p(\boldsymbol{y})) - \log(q(\boldsymbol{\delta}))\right)^2 p(\boldsymbol{\delta}|\boldsymbol{y})d\boldsymbol{\delta}
\end{aligned} \tag{17}$$

To minimize the component of Equation (16), one can in principle run a regression of $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta}_t,\boldsymbol{\delta})) - g_{0,t}(\boldsymbol{\delta}) - g_{6,t}(\boldsymbol{\delta})'\boldsymbol{\theta}_t$ on the regressors of $g_{1,t}(\boldsymbol{\delta})$, $g_{2,t}(\boldsymbol{\delta})$, $g_{3,t}(\boldsymbol{\delta})$, $g_{4,t}(\boldsymbol{\delta})\boldsymbol{\theta}_t$, $g_{5,t}(\boldsymbol{\delta})\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}'_t A_t(\boldsymbol{\delta})\boldsymbol{\theta}_t$ based on posterior draws of $\boldsymbol{\theta}_t$ and $\boldsymbol{\delta}$ to obtain the optimal value $\boldsymbol{a}_t^* = \{a_{1,t}^*, \boldsymbol{a}_{2,t}^*, \boldsymbol{a}_{3,t}^*, \boldsymbol{a}_{4,t}^*, \boldsymbol{a}_{5,t}^*, a_{6,t}^*\}$ for each $t = 1, ..., T$. However, given the structure of $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta}_t,\boldsymbol{\delta}))$ in Equation (12) and the construction of $g_t(\boldsymbol{\theta}_t;\boldsymbol{\delta},\boldsymbol{a}_t)$ as in Equation (13) and (14), such a minimization scheme is equivalent to a simpler regression of $h(\boldsymbol{\theta}_t,\boldsymbol{\xi}_t)$ on the regressors of a constant, $\boldsymbol{\xi}_t$, $\text{vech}(\boldsymbol{\xi}_t\boldsymbol{\xi}'_t)$, $\text{vec}(\boldsymbol{\xi}_t\boldsymbol{\theta}'_t)$, $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}'_t\boldsymbol{A}_t\boldsymbol{\theta}_t$ based on posterior draws of $\boldsymbol{\theta}_t$ and $\boldsymbol{\delta}$ to determine the optimal value $\boldsymbol{a}_t^*$ for all t. Inserting $\boldsymbol{a}^* = \{\boldsymbol{a}_t^*\}_{t=1}^T$ gives the optimized part $q(\boldsymbol{\theta}|\boldsymbol{\delta};\boldsymbol{a}^*)$ of the importance sampler.

To calibrate the part $q(\boldsymbol{\delta})$ of the importance sampler, one can in principle apply a non-linear least squares procedure to estimate the mean and covariance matrix of $q(\boldsymbol{\delta})$ based on Equation (17) by using $\boldsymbol{a}^*$ and posterior draws of $\boldsymbol{\delta}$, subject to the constraint that the covariance matrix must be positive definite. In this paper, we opt for a simpler heuristic approach. Given that $g_t(\boldsymbol{\theta}_t;\boldsymbol{\delta},\boldsymbol{a}_t^*)$ approximates $\log(p(\boldsymbol{y}_t|\boldsymbol{\theta}_t,\boldsymbol{\delta}))$ for all t via the linear regressions discussed in the preceding paragraph, we have $g(\boldsymbol{\theta};\boldsymbol{\delta},\boldsymbol{a}^*) \approx \log(p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\delta}))$ and thus $c(\boldsymbol{\delta};\boldsymbol{a}^*) \approx p(\boldsymbol{y}|\boldsymbol{\delta})$ by the definition of $c(\boldsymbol{\delta};\boldsymbol{a})$. Therefore, the item of Equation (17) can be usefully approximated by $\int \left(\log(p(\boldsymbol{\delta}|\boldsymbol{y})) - \log(q(\boldsymbol{\delta}))\right)^2 p(\boldsymbol{\delta}|\boldsymbol{y})d\boldsymbol{\delta}$, which is minimized

when $q(\boldsymbol{\delta}) = p(\boldsymbol{\delta}|\boldsymbol{y})$. Since $\boldsymbol{\delta}$ takes values on the whole real line, we approximate the posterior $p(\boldsymbol{\delta}|\boldsymbol{y})$ as a Gaussian distribution whose mean and covariance matrix equal the posterior mean and covariance matrix of $\boldsymbol{\delta}$, which we take as the part $q(\boldsymbol{\delta})$ of the importance sampler.

To directly apply Theorem 2.1 in this general case, one needs to work with a full covariance matrix of the size TK-by-TK, which could be computationally cumbersome. Since in practice the conditional prior $p(\boldsymbol{\theta}|\boldsymbol{\delta})$ often admits a Markovian structure $p(\boldsymbol{\theta}|\boldsymbol{\delta}) = p(\boldsymbol{\theta}_1|\boldsymbol{\delta}) \prod_{t=2}^{T} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\delta})$, the construction of $q(\boldsymbol{\theta}|\boldsymbol{\delta})$ can be simplified by a backward sequential matching procedure as in Theorem 2.2.

Theorem 2.2. *Suppose $\boldsymbol{\theta}_t$ is a K-by-1 vector of random variables taking values on the whole real line and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_t\}_{t=1}^{T}$. If $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1) \prod_{t=2}^{T} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$, where $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = N(\boldsymbol{\mu}_0 + \boldsymbol{\Phi}_0\boldsymbol{\theta}_{t-1}, \boldsymbol{\Sigma}_0)$ for $t > 1$ and $p(\boldsymbol{\theta}_1) = N(\tilde{\boldsymbol{\mu}}_0, \tilde{\boldsymbol{\Sigma}}_0)$, and $g(\boldsymbol{\theta}) = \sum_{t=1}^{T} a_{1,t} + \boldsymbol{a}_{2,t}'\boldsymbol{\theta}_t + \boldsymbol{\theta}_t'\boldsymbol{A}_t\boldsymbol{\theta}_t$ is a quadratic function of $\boldsymbol{\theta}$ where $a_{1,t}$ is a scalar, $\boldsymbol{a}_{2,t}$ is a K-by-1 vector and $\boldsymbol{A}_t$ is a K-by-K negative semi-definite matrix, then the product $p(\boldsymbol{\theta})exp(g(\boldsymbol{\theta}))$ is proportional to the function $q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_1) \prod_{t=2}^{T} q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ where $q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = N(\boldsymbol{\mu}_t + \boldsymbol{\Phi}_t\boldsymbol{\theta}_{t-1}, \boldsymbol{\Sigma}_t)$ for $t > 1$ and $p(\boldsymbol{\theta}_1) = N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. The parameters $\boldsymbol{\mu}_t$, $\boldsymbol{\Phi}_t$ and $\boldsymbol{\Sigma}_t$ follow a recursive law of motion:*

- *For $t = T$, $\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_0^{-1} - 2\boldsymbol{A}_t$, $\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{a}_{2,t}$, $\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Phi}_t = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Phi}_0$.*
- *For $1 < t < T$, $\boldsymbol{\Sigma}_t^{-1} + \boldsymbol{\Phi}_{t+1}'\boldsymbol{\Sigma}_{t+1}^{-1}\boldsymbol{\Phi}_{t+1} = \boldsymbol{\Sigma}_0^{-1} - 2\boldsymbol{A}_t + \boldsymbol{\Phi}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Phi}_0$, $\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t - \boldsymbol{\Phi}_{t+1}'\boldsymbol{\Sigma}_{t+1}^{-1}\boldsymbol{\mu}_{t+1} = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{a}_{2,t} - \boldsymbol{\Phi}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Phi}_t = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Phi}_0$.*
- *For $t = 1$, $\boldsymbol{\Sigma}_t^{-1} + \boldsymbol{\Phi}_{t+1}'\boldsymbol{\Sigma}_{t+1}^{-1}\boldsymbol{\Phi}_{t+1} = \tilde{\boldsymbol{\Sigma}}_0^{-1} - 2\boldsymbol{A}_t + \boldsymbol{\Phi}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Phi}_0$, $\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t - \boldsymbol{\Phi}_{t+1}'\boldsymbol{\Sigma}_{t+1}^{-1}\boldsymbol{\mu}_{t+1} = \tilde{\boldsymbol{\Sigma}}_0^{-1}\tilde{\boldsymbol{\mu}}_0 + \boldsymbol{a}_{2,t} - \boldsymbol{\Phi}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$.*

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

Remark 2. The computation cost in Theorem 2.2 is limited to working with K-by-K matrices, which is much cheaper than in Theorem 2.1 where one would directly work with a

full covariance matrix of the size TK-by-TK.

Remark 3. The relation between the latent variables $\boldsymbol{\theta}$ and the fixed model parameters $\boldsymbol{\delta}$ in the posterior $p(\boldsymbol{\theta}, \boldsymbol{\delta}|\boldsymbol{y})$ can be complicated. It is clear from the preceding discussions and Theorem 2.2 that the part $q(\boldsymbol{\theta}|\boldsymbol{\delta})$ of the importance sampler thus constructed allows complex non-linear dependence between $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ and hence can be more flexible for capturing the relation between $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ in the posterior $p(\boldsymbol{\theta}, \boldsymbol{\delta}|\boldsymbol{y})$ than alternatives that attempt to directly approximate $p(\boldsymbol{\theta}, \boldsymbol{\delta}|\boldsymbol{y})$ (e.g. using a joint Gaussian distribution to approximate $p(\boldsymbol{\theta}, \boldsymbol{\delta}|\boldsymbol{y})$).

2.3 The Gelfand-Dey Method

In essence, the approach described in Section 2.1 and 2.2 calibrates an auxiliary distribution of model parameters that approximates their joint posterior distribution. This approach should be equally useful in other settings where an approximation to a complicated posterior distribution is in need. One such example is the GD method for computing the marginal likelihood.

Assuming that the model parameters are $\{\boldsymbol{\theta}, \boldsymbol{\delta}\}$ as in Section 2.2, the GD method uses the following equation:

$$\frac{1}{p(\boldsymbol{y})} = \int \frac{q(\boldsymbol{\theta}, \boldsymbol{\delta})}{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\delta})p(\boldsymbol{\theta}, \boldsymbol{\delta})} p(\boldsymbol{\theta}, \boldsymbol{\delta}|\boldsymbol{y}) d\boldsymbol{\theta} d\boldsymbol{\delta} \tag{18}$$

where $q(\boldsymbol{\theta}, \boldsymbol{\delta})$ is a tuning function satisfying $\int q(\boldsymbol{\theta}, \boldsymbol{\delta}) d\boldsymbol{\theta} d\boldsymbol{\delta} = 1$. The Monte Carlo estimate of the marginal likelihood is:

$$\hat{p}(\boldsymbol{y}) = \left(\frac{1}{m} \sum_{i=1}^{m} \frac{q(\boldsymbol{\theta}^{(i)}, \boldsymbol{\delta}^{(i)})}{p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)}, \boldsymbol{\delta}^{(i)})p(\boldsymbol{\theta}^{(i)}, \boldsymbol{\delta}^{(i)})} \right)^{-1} \tag{19}$$

where $\{\boldsymbol{\theta}^{(i)}, \boldsymbol{\delta}^{(i)}\}_{i=1}^{m}$ are draws from the posterior. High efficiency is obtained when the tuning function $q(\boldsymbol{\theta}, \boldsymbol{\delta})$ is close to the posterior $p(\boldsymbol{\theta}, \boldsymbol{\delta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta}, \boldsymbol{\delta})p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\delta})$, which is exactly what the proposed approach aims to achieve. In Appendix C, we provide a heuristic proof

to show that an auxiliary distribution constructed as in Section 2.2 can also be a useful tuning function for minimizing the variability of the marginal likelihood estimate by the GD method. The key precondition is that, when calculating the marginal likelihood as in Equation (19), the posterior draws of model parameters should be adequately thinned to minimize their serial correlations. Thinning the posterior draws ensures that the variance of the summands in Equation (19) is a meaningful measure to be minimized in order to reduce the variability of the Monte Carlo estimate. When implementing the GD method in the empirical examples, we thin the posterior draws by keeping 1 in every 10 draws that appears adequate to reduce their serial correlations to very low levels.

3 Approaches for Comparison

This section describes two representatives of existing approaches for constructing importance samplers, which will be used as the benchmarks to evaluate the performance of the proposed approach.

For the simple case described in Section 2.1, a straightforward importance sampler for the model parameters $\boldsymbol{\theta}$ would be a Gaussian distribution whose mean and covariance matrix equal the posterior mean and covariance matrix of $\boldsymbol{\theta}$. This simple approach is often effective in practice when the dimension of $\boldsymbol{\theta}$ is modest. For easier reference, I will refer to this approach as the *direct calibration* approach.

For the more general case of models with latent variables $\boldsymbol{\theta} = \{\boldsymbol{\theta}_t\}_{t=1}^T$ and fixed parameters $\boldsymbol{\delta}$ (see Section 2.2), I consider a recent contribution in the literature by Chan (2023) that proposes a cross entropy approach to construct the importance sampler. Basically, minimizing the Kullback-Leibler divergence (i.e. the cross entropy plus a constant) between the posterior $p(\boldsymbol{\theta}, \boldsymbol{\delta}|\boldsymbol{y})$ and the importance sampler $q(\boldsymbol{\theta}, \boldsymbol{\delta})$ leads to a maximum likelihood estimator of $q(\boldsymbol{\theta}, \boldsymbol{\delta})$ based on posterior draws from $p(\boldsymbol{\theta}, \boldsymbol{\delta}|\boldsymbol{y})$. Chan (2023) constructs an

importance sampler for $\boldsymbol{\theta}$ by specifying an order-1 vector autoregressive (VAR) process for $\boldsymbol{\theta}_t$ with time-dependent parameters to maximize flexibility and shows that the resulting importance sampler is effective in models with stochastic volatility.

To apply the cross entropy approach of Chan (2023) in the setup of this paper, I specify a VAR(1) model $\boldsymbol{\theta}_t = \boldsymbol{a}_t + \boldsymbol{B}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{C}_t\boldsymbol{\delta} + N(\mathbf{0}, \boldsymbol{D}_t)$ to construct the component $q(\boldsymbol{\theta}|\boldsymbol{\delta})$ of the importance sampler, where the parameter $\boldsymbol{a}_t$ is a K-by-1 vector, $\boldsymbol{B}_t$ is a K-by-K matrix, $\boldsymbol{C}_t$ is a K-by-L matrix and $\boldsymbol{D}_t$ is a K-by-K covariance matrix.[5] The parameters of the VAR(1) model are estimated at each t based on posterior draws $\{\boldsymbol{\theta}_t^{(i)}, \boldsymbol{\theta}_{t-1}^{(i)}, \boldsymbol{\delta}^{(i)}\}_{i=1}^m$ by first running ordinary least squares (OLS) regressions for each element of $\boldsymbol{\theta}_t$ and then using the residuals from the OLS regressions to estimate the covariance matrix $\boldsymbol{D}_t$. The other component $q(\boldsymbol{\delta})$ of the importance sampler is a Gaussian distribution whose mean and covariance matrix equal the posterior mean and covariance matrix of $\boldsymbol{\delta}$. It is useful to note here that the dependence between $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ is linear in the importance sampler constructed by the cross entropy approach thus described, while the importance sampler constructed by the proposed approach described in Section 2.2 allows more complex non-linear dependence between $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$.

4 Empirical Examples

The proposed approach is illustrated in 3 models that have been routinely applied in empirical studies of economic data. The first is a binary probit regression model applied to predict economic recessions, which would correspond to the simple case discussed in Section 2.1. The second and third examples concern the linear regression model with stochastic volatility and correspond to the general case of Section 2.2. The second example specifies Gaussian model residual for the regressand while the third example allows fat fails by

[5]The initial value is $\boldsymbol{\theta}_1 = \boldsymbol{a}_1 + \boldsymbol{C}_1\boldsymbol{\delta} + N(\mathbf{0}, \boldsymbol{D}_1)$.

specifying a student-t distribution for the model residual.

4.1 Binary Probit Regression

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

The probit model is studied in the context of predicting economic recessions. See Nissila (2020) for a recent survey on probit-based models for recession prediction. Data on the binary response is the NBER-based U.S. recession indicator. Table D1 in Appendix D lists the 11 employed predictors including macroeconomic variables such as the inflation rate as well as financial variables such as the stock market return and the credit spread. The data sample is quarterly from Q3 1953 to Q2 2021 with a total of 272 observations. Including a constant, there are a total of 12 regressors in the model. The prediction horizon is 4 quarters. Setting the prior variance $\sigma^2 = 100$, posterior draws of $\boldsymbol{\theta}$ are produced by the method in Liu and Wu (1999). A total of 10,000 posterior draws are kept after 2,000 burn-ins.

The probit model under study corresponds to the simple case discussed in Section 2.1. The auxiliary function is $g_t(\boldsymbol{\theta}; \boldsymbol{a}_t) = a_{1,t} + \boldsymbol{a}_{2,t}'\boldsymbol{\theta} + a_{3,t}\boldsymbol{\theta}'\boldsymbol{A}_t\boldsymbol{\theta}$ where $\boldsymbol{A}_t = \boldsymbol{x}_t\boldsymbol{x}_t'$ since $\boldsymbol{x}_t\boldsymbol{x}_t' \propto \frac{\partial^2 \log(p(y_t|\bar{\boldsymbol{\theta}}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$ is a positive semi-definite matrix, where $\bar{\boldsymbol{\theta}}$ is a given value of $\boldsymbol{\theta}$. I run the following regression:

$$\log\left(p(y_t|\boldsymbol{\theta}^{(i)}\right) = a_{1,t} + \boldsymbol{a}_{2,t}'\boldsymbol{\theta}^{(i)} + a_{3,t}\left(\boldsymbol{\theta}^{(i)}\right)'\boldsymbol{A}_t\boldsymbol{\theta}^{(i)} + \text{residual} \tag{20}$$

for each $t = 1, ..., T$ where $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{m}$ are posterior draws of $\boldsymbol{\theta}$. To ensure $a_{3,t} \leq 0$, I drop

the quadratic term and re-run the regression whenever the original regression of Equation (20) returns a positive $a_{3,t}$.[6] Denote the regression estimates as $\boldsymbol{a}_t^* = \{a_{1,t}^*, \boldsymbol{a}_{2,t}^*, a_{3,t}^*\}$ and $\boldsymbol{a}^* = \{\boldsymbol{a}_t^*\}_{t=1}^T$. Let $g(\boldsymbol{\theta}; \boldsymbol{a}^*) = \sum_{t=1}^T g_t(\boldsymbol{\theta}; \boldsymbol{a}_t^*)$ as defined in Equation (6). The auxiliary distribution $q(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})\exp(g(\boldsymbol{\theta}; \boldsymbol{a}^*))$ can be easily constructed by using Theorem 2.1 to approximate the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$.

Table 1 compares the log marginal likelihood (LML) estimates by the IS and GD methods where the auxiliary distribution is constructed by the direct calibration approach and the new one. The new approach produces fairly close LML estimates under the IS and GD methods (-128.9365 versus -128.9370), while those by the direct calibration approach shows larger difference (-128.9413 versus -128.9694). Thus there is less finite sample bias in the LML estimates by the new approach than those by the direct calibration one. Moreover, the numerical standard error of the LML estimate by the new approach is 38% smaller than that by the direct calibration one under the IS method and is 25% smaller under the GD method.

4.2 Regression with Stochastic Volatility and Gaussian Residual

For time series studies, the linear regression model is commonly augmented with stochastic volatility (SV) to capture the phenomenon of volatility clustering often observed in economic time series data. A typical form of the model is:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \exp\left(\frac{\theta_t}{2}\right)\epsilon_t \tag{21}$$

$$\theta_t = (1-\rho)\mu + \rho\theta_{t-1} + s\eta_t$$

for $t = 1, ..., T$, where y_t is the scalar regressand, $\boldsymbol{x}_t$ is a J-by-1 vector of regressors, ϵ_t and η_t are i.i.d. standard Gaussian residuals, and θ_t is the log volatility with the initial value

[6] The R-squared of the regressions across $t = 1, ..., T$ ranges from 50% to over 99% with the average of 94%.

$\theta_1 = \mu + \sqrt{\frac{s^2}{1-\rho^2}}\eta_1$. For Bayesian analysis, the priors for the fixed model parameters are $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2 \mathrm{I}_J)$, $\mu \sim N(0, \sigma_\mu^2)$, $\frac{1+\rho}{2} \sim Beta(a_{\rho,0}, b_{\rho,0})$, and $s \sim N(0, \sigma_s^2)$. The autoregressive parameter ρ is transformed by $\rho^* = \log\left(\frac{1+\rho}{1-\rho}\right)$ such that it takes values on the whole real line. For the parameter s, the re-parametrization is $s^* = \log(s^2)$.[7] The priors of the transformed parameters are adjusted by using the change-of-variables formula.

The regression model is applied to predict the equity premium with the predictors examined in Welch and Goyal (2008). The dependent variable is the value-weighted quarterly return of the S&P500 index minus the corresponding risk free rate. A total of 12 economic predictors are selected based on Welch and Goyal (2008) that include stock characteristics, interest rates and other macroeconomic indicators and are listed in Table E1 of Appendix E. In estimation, I also include an intercept and an AR 1 lag of the equity premium. The sample runs from Q1 1947 to Q4 2020 with a total of 296 quarterly observations. The prediction horizon is 1 quarter. The hyper parameters of the priors are $\sigma^2 = 100$, $\sigma_\mu^2 = 10$, $a_{\rho,0} = 8$, $b_{\rho,0} = 2$ and $\sigma_s^2 = 1$. An Metropolis-within-Gibbs sampler is applied to estimate the model where estimation of the SV process is based on the algorithm of Kastner and Fruhwirth-Schnatter (2014). Analysis is based on 10,000 posterior draws after 2,000 burn-ins.

Let $\boldsymbol{\delta} = \{\boldsymbol{\beta}, \mu, \rho^*, s^*\}$ collect the fixed model parameters and $\boldsymbol{\theta} = \{\theta_t\}_{t=1}^T$ the latent variables. The marginal likelihood can be computed by the IS and GD methods following the discussions in Section 2.2 and 2.3. The auxiliary distribution is $q(\boldsymbol{\theta}, \boldsymbol{\delta}) = q(\boldsymbol{\theta}|\boldsymbol{\delta})q(\boldsymbol{\delta})$ where the component $q(\boldsymbol{\delta})$ is a Gaussian density with its mean and covariance matrix equal the posterior mean and covariance matrix of $\boldsymbol{\delta}$. To construct the component $q(\boldsymbol{\theta}|\boldsymbol{\delta})$, the log likelihood function $\log(p(y_t|\boldsymbol{\theta}, \boldsymbol{\delta})) = -\frac{\log(2\pi)}{2} - \frac{\theta_t}{2} - \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{2}\exp(-\theta_t)$ is matched with the structure of Equation (12), which gives $f_{1,t}(\boldsymbol{\delta}) = -\frac{\log(2\pi)}{2}$, $f_{2,t}(\boldsymbol{\delta}) = -\frac{1}{2}$, $f_{3,t}(\boldsymbol{\delta}) = \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{2}$,

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

$\xi_t = f_{4,t}(\boldsymbol{\delta}) = 0$, and $h_t(\theta_t, \xi_t) = -\exp(-\theta_t)$. Inserting $\xi_t = 0$ into Equation (13) leads to the auxiliary function $\tilde{g}_t(\theta_t; \boldsymbol{\delta}, \boldsymbol{a}_t) = a_{1,t} + a_{5,t}\theta_t + a_{6,t}\theta_t^2$ to approximate $h_t(\theta_t, 0)$, where the parameter $\boldsymbol{A}_t$ in Equation (13) equals 1 since the second-order derivative $\frac{\partial^2 h(\bar{\theta}_t, 0)}{\partial \theta_t^2} \propto 1$ at a given value $\bar{\theta}_t$ of θ_t. It follows by Equation (14) that the auxiliary function to approximate $\log(p(y_t|\theta_t, \boldsymbol{\delta}))$ is:

$$g_t(\theta_t; \boldsymbol{\delta}, \boldsymbol{a}_t) = \frac{a_{1,t}(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2 - \log(2\pi)}{2} + \frac{a_{5,t}(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2 - 1}{2}\theta_t + \frac{a_{6,t}(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{2}\theta_t^2$$

Following the discussion in Section 2.2.2, one can determine the optimal value $\boldsymbol{a}_t^*$ of the free parameters of $g_t(\theta_t; \boldsymbol{\delta}, \boldsymbol{a}_t)$ by running the following regression:

$$h_t\left(\theta_t^{(i)}, 0)\right) = -\exp\left(-\theta_t^{(i)}\right) = a_{1,t} + a_{5,t}\theta_t^{(i)} + a_{6,t}\left(\theta_t^{(i)}\right)^2 + \text{residual}$$

subject to $a_{6,t} \leq 0$ based on posterior draws $\{\theta_t^{(i)}\}_{i=1}^m$ for each $t = 1, ..., T$.[8] With $g_t(\theta_t; \boldsymbol{\delta}, \boldsymbol{a}_t^*)$ at hand, one can compute $g(\boldsymbol{\theta}; \boldsymbol{\delta}, \boldsymbol{a}^*) = \sum_{t=1}^T g_t(\theta_t; \boldsymbol{\delta}, \boldsymbol{a}_t^*)$. On the other hand, the AR(1) process for log volatility θ_t in Equation (21) would lead to the prior $p(\boldsymbol{\theta}|\boldsymbol{\delta}) = p(\theta_1|\boldsymbol{\delta}) \prod_{t=2}^T p(\theta_t|\theta_{t-1}, \boldsymbol{\delta})$ with $p(\theta_t|\theta_{t-1}, \boldsymbol{\delta}) = N\left((1-\rho)\mu + \rho\theta_{t-1}, s^2\right)$ for $t > 1$ and $p(\theta_1|\boldsymbol{\delta}) = N\left(\mu, \frac{s^2}{1-\rho^2}\right)$. Combining $p(\boldsymbol{\theta}|\boldsymbol{\delta})$ and $g(\boldsymbol{\theta}; \boldsymbol{\delta}, \boldsymbol{a}^*)$, one can readily apply Theorem 2.2 to construct $q(\boldsymbol{\theta}|\boldsymbol{\delta})$.

Table 2 compares the log marginal likelihood (LML) estimates by the IS and GD methods where the auxiliary distribution is constructed by the cross entropy approach and the new one. The difference between the IS and GD estimates is similar when the auxiliary distributions are produced by these two approaches and is fairly small (297.88 vs. 297.81 with the cross entropy approach and 298.41 vs. 298.33 with the new approach). However, the numerical standard error of the LML estimate by the new approach is 65% smaller than that by the cross entropy one under the IS method and is 70% smaller under the GD method.

[8] The R-squared of the regressions ranges from 76% to 99% with the average of 95%.

To see the efficiency of the new approach more intuitively, Figure 1 compares the log ratio $\{w^{(i)}\}_{i=1}^m$ that are computed by the cross entropy approach and the new one under both the IS and GD methods, where $w^{(i)} = \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)})p(\boldsymbol{\theta}^{(i)}|\boldsymbol{\delta}^{(i)})p(\boldsymbol{\delta}^{(i)})}{q(\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)})}\right)$ and $\{\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)}\}_{i=1}^m$ are draws from the importance sampler under the IS method and are posterior draws under the GD method. It is clear from Equation (11) and (19) that, by substituting the log ratio $\{w^{(i)}\}_{i=1}^m$, the IS estimate of the marginal likelihood can be re-written as:

$$\hat{p}_{IS}(\boldsymbol{y}) = \frac{1}{m}\sum_{i=1}^{m}\exp\left(w^{(i)}\right)$$

where $\{w^{(i)}\}_{i=1}^m$ are based on $\{\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)}\}_{i=1}^m$ from the importance sampler, and the GD estimate as:

$$\hat{p}_{GD}(\boldsymbol{y}) = \left(\frac{1}{m}\sum_{i=1}^{m}\exp\left(-w^{(i)}\right)\right)^{-1}$$

where $\{w^{(i)}\}_{i=1}^m$ are based on posterior draws $\{\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)}\}_{i=1}^m$. A more efficient importance sampler or tuning function would return draws of the log ratio $\{w^{(i)}\}_{i=1}^m$ that are more concentrated and less volatile. Compared to the summand $\{\exp\left(w^{(i)}\right)\}_{i=1}^m$ or $\{\exp\left(-w^{(i)}\right)\}_{i=1}^m$, the distribution of the log ratio $\{w^{(i)}\}_{i=1}^m$ is usually much closer to a bell shape and is more convenient to analyze. In Figure 1, it can be seen that draws of the log ratio $\{w^{(i)}\}_{i=1}^m$ by the new approach are clearly much more stable than those by the cross entropy approach under both the IS and GD methods. The standard deviation of the log ratio $\{w^{(i)}\}_{i=1}^m$ by the new approach, under either the IS or GD method, is an order of magnitude smaller than that by the cross entropy approach.

A concern of the IS method is that, in extreme cases, the denominator may be too thin-tailed for the Monte Carlo estimate of the marginal likelihood to have a finite variance (Geweke (1989)). The same concern applies to the GD estimator. We address this concern by examining the behavior of the log ratio $\{w^{(i)}\}_{i=1}^m$. If the denominator when applying the IS method is too thin-tailed, we should see extremely large values of the log ratio from time to time, while in the case of the GD method, occasional negative values of extreme

magnitude for the log ratio should be observed. In Figure 1, we find that the log ratio shows neither of these patterns and thus alleviates the infinite variance concern in this empirical study.

4.3 Regression with Stochastic Volatility and Student-t Residual

Fatter tails than the Gaussian distribution are often found in economic time series data. To allow fat tails, the regression model of Equation (21) can be modified to specify a student-t distribution for the model residual $\epsilon_t \sim t(v)$, where $v > 0$ is the degrees of freedom parameter. The prior for v is $IG(a_{v,0}, b_{v,0})$ where IG denotes the inverse gamma distribution. We re-parameterize $v^* = \log(v)$ such that the support of the new parameter is the whole real line. In estimation, the hyper parameters for v are $a_{v,0} = 7$ and $b_{v,0} = 60$. The setup of other model parameters are the same as in the Gaussian residual case of Section 4.2.

The model is applied to the same dataset of equity premium in Section 4.2 and is estimated by a Metropolis-within-Gibbs sampler. In the model estimation, the student-t distribution $\epsilon_t \sim t(v)$ is re-framed as a hierarchical Gaussian distribution $\epsilon_t \sim N(0, d_t)$ and $d_t \sim IG\left(\frac{v}{2}, \frac{v}{2}\right)$ for $t = 1, ..., T$. To improve sampling efficiency, the parameters μ and s in the SV process are sampled by integrating out the linear coefficients $\boldsymbol{\beta}$ and are drawn by a Metropolis-Hastings step with a random walk proposal which is tuned by the approach of Garthwaite et al. (2016). Similar to the Gaussian residual case of Section 4.2, the interweaving strategy of Kastner and Fruhwirth-Schnatter (2014), which originates from Yu and Meng (2011), is applied to further boost the sampling efficiency of μ and s. The degrees of freedom parameter v is sampled by a Metropolis-Hastings step with the proposal tuned by the approach of Garthwaite et al. (2016). A total of 10,000 posterior draws are collected for analysis after 2,000 burn-ins.

Denote $\boldsymbol{\delta} = \{\boldsymbol{\beta}, \mu, \rho^*, s^*, v^*\}$. The construction the part $q(\boldsymbol{\delta})$ of the importance sam-

pler is the same as in the Gaussian residual case of Section 4.2. The steps of constructing the other part $q(\boldsymbol{\theta}|\boldsymbol{\delta})$ are also the same as in Section 4.2 except that now the log likelihood function becomes $\log(p(y_t|\boldsymbol{\theta},\boldsymbol{\delta})) = \log\left(\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v}\Gamma\left(\frac{v}{2}\right)}\right) - \frac{\theta_t}{2} - \frac{1+v}{2}\log\left(1 + \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2 \exp(-\theta_t)}{v}\right)$. Matching the log likelihood function with the structure of Equation (12) produces $f_{1,t}(\boldsymbol{\delta}) = \log\left(\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v}\Gamma\left(\frac{v}{2}\right)}\left(\frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{v}\right)^{-\frac{1+v}{2}}\right)$, $f_{2,t}(\boldsymbol{\delta}) = \frac{v}{2}$, $f_{3,t}(\boldsymbol{\delta}) = \frac{1+v}{2}$, $\xi_t = f_{4,t}(\boldsymbol{\delta}) = -\log\left(\frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{v}\right)$ and $h_t(\theta_t, \xi_t) = -\log\left(1 + \exp(\theta_t + \xi_t)\right)$. The auxiliary function $\tilde{g}_t(\theta_t; \boldsymbol{\delta}, \boldsymbol{a}_t)$ is $a_{1,t} + a_{2,t}\xi_t + a_{3,t}\xi_t^2 + a_{4,t}\xi_t\theta_t + a_{5,t}\theta_t + a_{6,t}\theta_t^2$ per Equation (13), where the parameter $\boldsymbol{A}_t$ in Equation (13) equals 1 since the second-order derivative $\frac{\partial^2 h(\bar{\theta}_t, \bar{\xi}_t)}{\partial \theta_t^2} \propto 1$ at a given value $(\bar{\theta}_t, \bar{\xi}_t)$ of (θ_t, ξ_t). The free parameters $\boldsymbol{a}_t = \{a_{j,t}\}_{j=1}^6$ are determined by running the following regression:

$$h_t\left(\theta_t^{(i)}, \xi_t^{(i)}\right) = a_{1,t} + a_{2,t}\xi_t^{(i)} + a_{3,t}\left(\xi_t^{(i)}\right)^2 + a_{4,t}\xi_t^{(i)}\theta_t^{(i)} + a_{5,t}\theta_t^{(i)} + a_{6,t}\left(\theta_t^{(i)}\right)^2 + \text{residual}$$

subject to $a_{6,t} \leq 0$ based on posterior draws $\{\theta_t^{(i)}, \boldsymbol{\delta}^{(i)}\}_{i=1}^m$ for each $t = 1, ..., T$.[9]

The LML estimates by the cross entropy approach and the new one are provided in the right columns of Table 2. For this more complicated model than the Gaussian residual case, the difference in the LML estimates between the IS and GD methods is smaller when the new approach is used. The new approach is also able to push the numerical standard error of the LML estimate 81% smaller under the IS method and 55% smaller under the GD method than the cross entropy approach. The log ratio $\{w^{(i)}\}_{i=1}^m$ shown in Figure 2 shows a similar pattern as in the Gaussian residual case and is much less volatile when the new approach is used.

By examining the behavior of the log ratio $\{w^{(i)}\}_{i=1}^m$ in Figure 2, we find no evidence of the infinite variance problem in the IS and GD estimates of the marginal likelihood. It is also interesting to note that, while the IS and GD estimates of the LML by the cross entropy approach show a mixed message regarding the relative merit of the student-t specification, the IS and GD estimates of the LML by the new approach unanimously support, albeit

[9]The R-squared of the regressions ranges from 99.5% to almost 100% with the average of 99.96%.

slightly, that the student-t specification actually under-performs the Gaussian one in this example.

5 Conclusion

In summary, this paper has demonstrated a new approach to construct the auxiliary distribution of model parameters to implement the IS and GD methods for computing the marginal likelihood. The new approach is guided by the principle of minimizing the Monte Carlo variance of the marginal likelihood estimate and is simple to implement. The computational cost of the new approach is low as only a series of low dimensional linear regressions based on posterior draws of model parameters are required. Encouraging results are found in the examples examined and should encourage further efforts to apply the idea of this paper for Bayesian model specification analysis.

Table 1: Log Marginal Likelihood Estimates for the Probit Regression Model

	Direct Calibration	New Approach
Importance Sampling	-128.9413	-128.9365
	(0.0039)	(0.0024)
Gelfand-Dey	-128.9694	-128.9370
	(0.0089)	(0.0067)

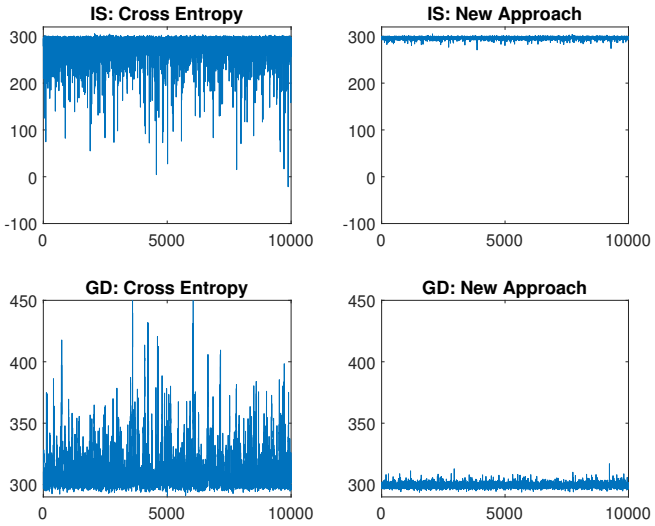
Note: This table compares the log marginal likelihood estimates for the probit regression model of Section 4.1 that are computed by the importance sampling and Gelfand-Dey methods where the auxiliary distribution is constructed by the direct calibration approach and the new one. The number of draws is 10,000 in all cases. Numbers in the brackets are the numerical standard errors of the log marginal likelihood estimates computed by the delta method. For the Gelfand-Dey method, the posterior sample is thinned 1-in-10 and the standard error is calculated by the Newey-West method (Newey and West (1987)) with the number of lags floor $\left(4(m/100)^{2/9}\right)$ where m is the number of draws in the thinned posterior sample.

Table 2: Log Marginal Likelihood Estimates for the Linear Regression Model with SV

	Gaussian Residual		Student-t Residual	
	Cross Entropy	New Approach	Cross Entropy	New Approach
Importance Sampling	297.8849	298.4089	298.1101	297.4860
	(0.2631)	(0.0927)	(0.5407)	(0.1004)
Gelfand-Dey	297.8144	298.3308	297.7320	297.6403
	(0.2903)	(0.0883)	(0.2738)	(0.1241)

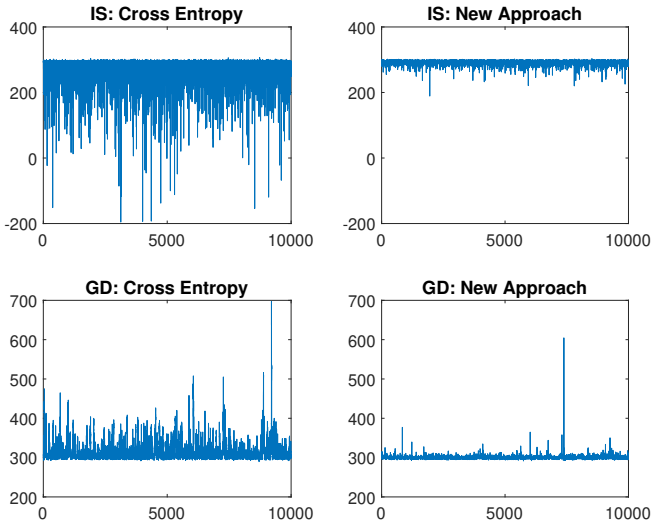
Note: This table compares the log marginal likelihood estimates for the linear regression model with stochastic volatility and Gaussian/student-t residuals (Section 4.2 and 4.3 respectively) that are computed by the importance sampling and Gelfand-Dey methods where the auxiliary distribution is constructed by the direct calibration approach and the new one. The number of draws is 10,000 in all cases. Numbers in the brackets are the numerical standard errors of the log marginal likelihood estimates computed by the delta method. For the Gelfand-Dey method, the posterior sample is thinned 1-in-10 and the standard error is calculated by the Newey-West method (Newey and West (1987)) with the number of lags floor $\left(4(m/100)^{2/9}\right)$ where m is the number of draws in the thinned posterior sample.

Figure 1: Log Ratio for the SV-Gaussian Linear Regression Model



Note: This figure compares the log ratio $\{w^{(i)}\}_{i=1}^m$ for the linear regression model with stochastic volatility and Gaussian residual (Section 4.2) that are computed by the cross entropy approach and the new one under both the importance sampling (IS) and Gelfand-Dey (GD) methods, where $w^{(i)} = \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)})p(\boldsymbol{\theta}^{(i)}|\boldsymbol{\delta}^{(i)})p(\boldsymbol{\delta}^{(i)})}{q(\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)})}\right)$ and $\{\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)}\}_{i=1}^m$ are draws from the importance sampler under the IS method and are posterior draws under the GD method.

Figure 2: Log Ratio for the SV-Student Linear Regression Model



Note: This figure compares the log ratio $\{w^{(i)}\}_{i=1}^m$ for the linear regression model with stochastic volatility and student-t residual (Section 4.3) that are computed by the cross entropy approach and the new one under both the importance sampling (IS) and Gelfand-Dey (GD) methods, where $w^{(i)} = \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)})p(\boldsymbol{\theta}^{(i)}|\boldsymbol{\delta}^{(i)})p(\boldsymbol{\delta}^{(i)})}{q(\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)})}\right)$ and $\{\boldsymbol{\theta}^{(i)},\boldsymbol{\delta}^{(i)}\}_{i=1}^m$ are draws from the importance sampler under the IS method and are posterior draws under the GD method.

A Proof of Theorem 2.1

Proof. Since $p(\boldsymbol{\theta})$ is the Gaussian density for $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, it follows:

$$p(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

Multiplying $p(\boldsymbol{\theta})$ by $\exp(g(\boldsymbol{\theta}))$ and re-arranging items give:

$$p(\boldsymbol{\theta})\exp(g(\boldsymbol{\theta})) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}'(\boldsymbol{\Sigma}_0^{-1} - 2\boldsymbol{A})\boldsymbol{\theta} + \boldsymbol{\theta}'(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{a}_2)\right)$$

which is the kernel of the Gaussian density $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_0^{-1} - 2\boldsymbol{A}$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{a}_2$. □

B Proof of Theorem 2.2

Proof. Given the assumption $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1)\prod_{t=2}^{T} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ where $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = N(\boldsymbol{\mu}_0 + \boldsymbol{\Phi}_0\boldsymbol{\theta}_{t-1}, \boldsymbol{\Sigma}_0)$ for $t > 1$ and $p(\boldsymbol{\theta}_1) = N(\tilde{\boldsymbol{\mu}_0}, \tilde{\boldsymbol{\Sigma}}_0)$, we have:

$$p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}_t'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta}_t + \boldsymbol{\mu}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta}_t + \boldsymbol{\theta}_{t-1}'\boldsymbol{\Phi}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta}_t - \frac{1}{2}\boldsymbol{\theta}_{t-1}'\boldsymbol{\Phi}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Phi}_0\boldsymbol{\theta}_{t-1} - \boldsymbol{\mu}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Phi}_0\boldsymbol{\theta}_{t-1}\right)$$

for $t > 1$ and

$$p(\boldsymbol{\theta}_t) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}_t'\tilde{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{\theta}_t + \tilde{\boldsymbol{\mu}}_0'\tilde{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{\theta}_t\right)$$

for $t = 1$.

Similarly, given that $q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_1)\prod_{t=2}^{T} q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ where $q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = N(\boldsymbol{\mu}_t + \boldsymbol{\Phi}_t\boldsymbol{\theta}_{t-1}, \boldsymbol{\Sigma}_t)$ for $t > 1$ and $q(\boldsymbol{\theta}_1) = N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, the functional forms of $q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ and $q(\boldsymbol{\theta}_1)$ would be the same as those for $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ and $p(\boldsymbol{\theta}_1)$ except for the notational change of replacing $\boldsymbol{\mu}_0$, $\boldsymbol{\Phi}_0$, $\boldsymbol{\Sigma}_0$, $\tilde{\boldsymbol{\mu}}_0$ and $\tilde{\boldsymbol{\Sigma}}_0$ by appropriate $\boldsymbol{\mu}_t$, $\boldsymbol{\Phi}_t$ and $\boldsymbol{\Sigma}_t$.

Let $g_t(\boldsymbol{\theta}) = a_{1,t} + \boldsymbol{a}_{2,t}'\boldsymbol{\theta}_t + \boldsymbol{\theta}_t'\boldsymbol{A}_t\boldsymbol{\theta}_t$. Start with $t = T$. Retain only the items involving $\boldsymbol{\theta}_T$

in $p(\boldsymbol{\theta}_T|\boldsymbol{\theta}_{T-1})\exp(g_T(\boldsymbol{\theta}))$ and match them with the items in $q(\boldsymbol{\theta}_T|\boldsymbol{\theta}_{T-1})$. This would leave:

$$\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_0^{-1} - 2\boldsymbol{A}_T$$

$$\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_T = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{a}_{2,T}$$

$$\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Phi}_t = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Phi}_0$$

Next for $t = T-1$, collect the items involving $\boldsymbol{\theta}_{T-1}$ in $p(\boldsymbol{\theta}_{T-1}|\boldsymbol{\theta}_{T-2})\exp(g_{T-1}(\boldsymbol{\theta}))$ as well as the items involving $\boldsymbol{\theta}_{T-1}$ from $p(\boldsymbol{\theta}_T|\boldsymbol{\theta}_{T-1})\exp(g_T(\boldsymbol{\theta}))$. Do the same for $q(\boldsymbol{\theta}_{T-1}|\boldsymbol{\theta}_{T-2})$. By matching the items, one obtains:

$$\boldsymbol{\Sigma}_{T-1}^{-1} + \boldsymbol{\Phi}_T'\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Phi}_T = \boldsymbol{\Sigma}_0^{-1} - 2\boldsymbol{A}_{T-1} + \boldsymbol{\Phi}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Phi}_0$$

$$\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\mu}_{T-1} - \boldsymbol{\Phi}_T'\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\mu}_T = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{a}_{2,T-1} - \boldsymbol{\Phi}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$$

$$\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Phi}_{T-1} = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Phi}_0$$

Repeating these steps backward until $t = 1$ gives the formulas in Theorem 2.2.

□

C The Gelfand-Dey Method

To compute the marginal likelihood, the GD method uses posterior draws of model parameters that are often simulated by a Markov chain Monte Carlo sampler and are serially correlated. Thus, the suitable measure of the variability of the Monte Carlo estimate of $1/p(\boldsymbol{y})$ is its long-run variance that takes into account the serial correlations between posterior draws of model parameters and is complicated to operate. To simplify the task, we take the position that, by adequately thinning the posterior draws of model parameters (e.g. keeping 1 in every 10 draws), the serial correlations in the retained posterior draws could be sufficiently small such that, when the thinned posterior draws are used to apply the GD method, one can meaningfully focus on minimizing the variance term of the long-run variance.

Specifically, assume that the serial correlations of the posterior draws of model parameters are negligible by sufficient thinning, we have:

$$V(1/\hat{p}(\boldsymbol{y})) \approx \frac{1}{m}\int\left(\frac{q(\boldsymbol{\theta},\boldsymbol{\delta})}{p(\boldsymbol{\theta},\boldsymbol{\delta})p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\delta})}-\frac{1}{p(\boldsymbol{y})}\right)^2 p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y})d\boldsymbol{\theta}d\boldsymbol{\delta}$$
$$= \frac{1}{mp(\boldsymbol{y})^2}\int\left(\frac{q(\boldsymbol{\theta},\boldsymbol{\delta})}{p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y})}-1\right)^2 p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y})d\boldsymbol{\theta}d\boldsymbol{\delta}$$

Let $f(\boldsymbol{\theta},\boldsymbol{\delta}) = \log(p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y}))-\log(q(\boldsymbol{\theta},\boldsymbol{\delta}))$. Taking a first-order Taylor expansion of $\exp(-f(\boldsymbol{\theta},\boldsymbol{\delta}))$ at $f(\boldsymbol{\theta},\boldsymbol{\delta}) = 0$ and inserting this approximation into the expression of $V(1/\hat{p}(\boldsymbol{y}))$ leads to the following simplification:

$$V(1/\hat{p}(\boldsymbol{y})) \approx \frac{1}{mp(\boldsymbol{y})^2}\int\left(\log(p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y}))-\log(q(\boldsymbol{\theta},\boldsymbol{\delta}))\right)^2 p(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{y})d\boldsymbol{\theta}d\boldsymbol{\delta}$$

where the integral is the same as in the Monte Carlo variance of the IS estimate of the marginal likelihood (see Equation (15) of the main text). Thus, one can apply the same approach discussed in Section 2.2 to construct $q(\boldsymbol{\theta},\boldsymbol{\delta})$ in order to minimize the variance $V(1/\hat{p}(\boldsymbol{y}))$.

D Regressors of the Probit Regression Model

The regressors of the probit regression model in Section 4.1 are listed in Table D1.

E Regressors of the Linear Regression Models

The regressors of the linear regression models in Section 4.2 and 4.3 are listed in Table E1.

Table D1: List of Predictors for Economic Recessions

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

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

Table E1: List of Predictors for Equity Premium

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

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

References

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

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

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

Chan, J. and E. Eisenstat (2015). Marginal likelihood estimation with the cross-entropy method. *Econometric Reviews 34*(3), 256–285.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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