An Automated Prior Robustness Analysis in Bayesian Model Comparison

Joshua C.C. Chan
Department of Economics
Purdue University and UTS

Liana Jacobi
Department of Economics
University of Melbourne

Dan Zhu*
Department of Business Statistics and Econometrics
Monash University

March 2021

Abstract

It is well-known that the marginal likelihood, the gold standard for Bayesian model comparison, can be sensitive to prior hyperparameter choices. However, most models require computationally intense simulation-based methods to evaluate the typically high-dimensional integral of the marginal likelihood expression. Hence, despite the recognition that prior sensitivity analysis is important in this context, it is rarely done in practice. We develop efficient and feasible methods to compute the sensitivities of the marginal likelihood, obtained via two common simulation-based methods, with respect to any prior hyperparameter, alongside the MCMC estimation algorithm. Our approach builds on Automatic Differentiation (AD), which has only recently been introduced to the more computationally intensive Markov chain Monte Carlo simulation setting. We illustrate our approach with two empirical applications in the context of widely used multivariate time series models.

Keywords: automatic differentiation, model comparison, vector autoregression, factor models

JEL classifications: C11, C53, E37

*Email: dan.zhu@monash.edu

1 Introduction

The marginal likelihood is central to Bayesian model comparison and Bayesian model averaging. Since analytical computation is only possible for a few simple models, most models require computationally intense simulation-based methods to evaluate the typically high-dimensional integral of the marginal likelihood expression. Consequently, there is a vast literature devoted to its estimation using Monte Carlo methods.[1] Despite its prominence, one well-known drawback of the marginal likelihood is that it is relatively sensitive to the choice of prior—a small change in the prior that keeps inference of the model parameters the same could have a large impact on the value of the marginal likelihood (see, e.g., Aitkin, 1991; O'Hagan, 1995).[2] As such, the importance of sensitivity analysis for marginal likelihood has long been recognized (e.g., Kass, 1993), but it is not routinely done in empirical work due to the computation complexity given the intensity of marginal likelihood estimation.[3]

In practice, when an informal prior sensitivity analysis is conducted, it is typically implemented with a limited scope due to computational costs. For example, researchers might assess a specific aspect of marginal likelihood sensitivities by re-computing its value using a different set of hyperparameters. However, this approach is ad hoc and cumbersome. We take the first step to address this issue by introducing a computationally feasible and systematic approach to assess marginal likelihood sensitivities with respect to prior hyperparameters. More specifically, we develop methods based on Automatic Differentiation (AD) to compute partial derivatives of the marginal likelihood—estimated using Monte Carlo—with respect to various hyperparameters. These partial derivatives are useful in a number of situations.

Firstly, one can use them to identify key hyperparameters that impact the marginal likelihood values. A further analysis focusing on those influential hyperparameters is then feasible, for example, eliciting those prior hyperparameters more carefully or choosing their values optimally. Secondly, these partial derivatives can be used to analyze whether

[1] Popular approaches include Gelfand and Dey (1994), Newton and Raftery (1994), Frühwirth-Schnatter (1995), Chib (1995), Gelman and Meng (1998), Chib and Jeliazkov (2001), Frühwirth-Schnatter and Wagner (2008) and Friel and Pettitt (2008).

[2] In the case of hypothesis testing, Lindley (1957) shows that a point null hypothesis will always be rejected if the variance of a conjugate prior goes to infinity. This observation can be traced back to Jeffreys (1939).

[3] Furthermore, to communicate the results of the analysis to other researchers who might have different prior information or to demonstrate the robustness of the results with respect to prior information, a prior sensitivity analysis becomes necessary (e.g., Poirier, 1988; Liu and Aitkin, 2008).

small changes in any hyperparameter would impact the ranking of a set of competing models. This gives the user a convenient way to assess prior sensitivities without re-estimating all the models and re-computing the corresponding marginal likelihoods.[4]

Finally, it is common in many applications to adopt an empirical Bayes approach that selects hyperparameters values by maximizing the marginal likelihood. For example, in the context of a standard vector autoregression with the natural conjugate prior, Del Negro and Schorfheide (2004), Schorfheide and Song (2015) and Carriero, Clark, and Marcellino (2015) obtain the optimal hyperparameters by maximizing the marginal likelihood over a grid of possible values. However, this grid-search approach is typically time-consuming and is only possible for models where the marginal likelihood is available analytically. For most models in which the marginal likelihood needs to be estimated using Monte Carlo methods, this brute-force approach is simply computationally infeasible. In those cases, the partial derivatives of the marginal likelihood become necessary in order to speed up the marginal likelihood maximization problem (e.g., using gradient-ascent methods).[5]

Due to the complexity of MCMC and the marginal likelihood computation, our approach is based on AD to obtain the complete set of prior hyperparameter sensitivities of the marginal likelihood alongside the model estimation. It is "automatic" in the sense that for an algorithm that maps inputs into any posterior output, there is an automatic way of deriving its complementary algorithm of computing the sensitivities. Importantly for our purpose, the AD-based approach would only require running the original algorithm once with derivatives computed alongside the estimation algorithm. While AD methods are now commonly used in classical simulation settings Financial Mathematics and in Machine Learning, the approach is yet to be widely adopted in Econometrics or Statistics. Jacobi, Joshi, and Zhu (2018) have developed the first AD-based approach for input sensitivity analysis of output from high-dimensional mappings based on Markov chain

[4]It is possible to generate a new MCMC sample given an alternative prior without re-running the MCMC. Specifically, one can implement an importance re-weighing of the old sample with respect to weights proportional to the new and old prior ratio. However, the quality of the sample depends on the tail behaviors of the two priors, and it is often infeasible to guarantee that the variance of the weights is finite (e.g., Robert and Casella, 2013, pp.94-99).

[5]An alternative approach to obtain partial derivatives is to approximate them using the finite-difference method. However, the finite differencing approach to robustness analysis has theoretical limitations as it generates biased derivative estimates. In addition, the choice of step size is crucial because it involves a critical bias-variance trade-off (e.g., decreasing the step size can reduce bias but increase variance); see, e.g., Section 7.1, Glasserman (2003). This issue is also illustrated in a simple example on page 6 in Berger, Insua, and Ruggeri (2000).

Monte Carlo (MCMC) methods. Chan, Jacobi, and Zhu (2019) extends this framework further to predictive simulation—in particular, to analyze the sensitivities of point and interval forecasts based on vector autoregressions on prior hyperparameters.

This paper further extends the AD-based approach to the more computationally intensive settings of computing the marginal likelihood using MCMC output and adaptive importance sampling. There is by now an extensive literature on marginal likelihood estimation using MCMC output; for a recent review, see Friel and Wyse (2012) and Ardia, Baştürk, Hoogerheide, and van Dijk (2012). Here we focus on two popular methods: Chib's method (Chib, 1995; Chib and Jeliazkov, 2001) and the improved cross-entropy method (Chan and Eisenstat, 2015, 2018). One of the paper's key innovations is to study the derivative of the cross-entropy parameter with respect to the prior hyperparameters, such that the cross-entropy parameters are obtained via a numerical search of optimum. One can readily apply a regular AD to differentiate the optimization algorithm. Unfortunately, if the number of steps required to reach convergence is high, the algorithm produces a large expression graph with substantive computation costs. Hence, instead of standard AD, we use its implicit derivatives and estimates based on the simulated samples.[6]

We illustrate our new methodology with two empirical applications in the context of multivariate time series analysis using vector autoregressions (VARs) and factor models. In each case, we use AD to compute the partial derivatives of each marginal likelihood estimator with respect to a variety of hyperparameters and assess various aspects of the marginal likelihood sensitivity. The first application compares two VARs for modeling a US macroeconomic dataset that involves GDP inflation and real output growth. In the second application, we fit daily returns on ten foreign exchange rates using factor models with a different number of latent factors. While the conclusion in the first application—that the VAR with t errors is more favored by the data over the benchmark Gaussian VAR—is robust over a wide range of hyperparameter values, the preferred number of factors in the second application is more uncertain—the weight of evidence can change noticeably if we alter some hyperparameter values. Our findings, therefore, not only illustrate the feasibility of the proposed methods but also highlight the importance of systematically performing a prior sensitivity analysis in Bayesian model comparison.

The rest of this paper is organized as follows. Section 2 first gives an overview of the

[6]To reduce the memory requirements, we further improve the implementation of the standard AD by storing only absolutely necessary quantities. We specify the exact quantities stored in the application section.

marginal likelihood and its estimation using Chib's and the cross-entropy methods. We then develop an AD-based framework to analyze the sensitivity of the two marginal likelihood estimators with respect to a set of prior hyperparameters in Section 3. It is followed by two empirical applications to illustrate the AD-based prior robustness analysis in Section 4. We further study the empirical computational complexity of the proposed method via a series of simulations in Section 5. Lastly, Section 6 concludes and briefly discusses some future research directions.

2 Marginal Likelihood Estimation

To set the stage, suppose we wish to compare the set of models $\{M_1, \ldots, M_K\}$, where each model M_k is formally defined by a likelihood function $p(\mathbf{y} \mid \boldsymbol{\psi}_k, M_k)$[7] and a prior on the model-specific parameter vector $\boldsymbol{\psi}_k$ denoted by $p(\boldsymbol{\psi}_k \mid M_k)$. The gold standard for Bayesian model comparison is the Bayes factor. Specifically, the *Bayes factor* in favor of M_i against M_j is defined as

$$\text{BF}_{ij} = \frac{p(\mathbf{y} \mid M_i)}{p(\mathbf{y} \mid M_j)},$$

where

$$p(\mathbf{y} \mid M_k) = \int p(\mathbf{y} \mid \boldsymbol{\psi}_k, M_k) p(\boldsymbol{\psi}_k \mid M_k) \mathrm{d}\boldsymbol{\psi}_k \tag{1}$$

is the *marginal likelihood* under model M_k, $k = i, j$. It therefore follows that if the Bayes factor BF_{ij} is larger than 1, observed data are more likely under model M_i than model M_j. This can be viewed as evidence in favor of M_i. For a more detailed discussion of the Bayes factor and its role in Bayesian model comparison, see Koop (2003), Kroese and Chan (2014) and Amisano and Giacomini (2007).

The marginal likelihood of a particular model can be interpreted as a joint density forecast from that model evaluated at the observed data $\mathbf{y}$—hence, if the observed data are likely under the model, the corresponding marginal likelihood would be "large" and vice versa. To see this, let $\mathbf{y}_{1:t} = (\mathbf{y}_1, \ldots, \mathbf{y}_t)$ denote all the data up to time t with $\mathbf{y}_{1:T} = \mathbf{y}$. Then, we can factor the marginal likelihood as follows:

$$p(\mathbf{y} \mid M_k) = p(\mathbf{y}_1 \mid M_k) \prod_{t=1}^{T-1} p(\mathbf{y}_{t+1} \mid \mathbf{y}_{1:t}, M_k), \tag{2}$$

[7] More precisely, for latent variable models, it is the observed-data or integrated likelihood function that is unconditional on the latent variables.

where $p(\mathbf{y}_{t+1} \mid \mathbf{y}_{1:t}, M_k)$ is the *predictive likelihood* under model M_k, which can be interpreted as a one-step-ahead density forecast for $\mathbf{y}_{t+1}$.

The factorization of the marginal likelihood in (2) also reveals that its value is likely to be sensitive to the choice of prior. For instance, the predictive likelihood $p(\mathbf{y}_1 \mid M_k)$ depends entirely on the prior distribution and not on the data. More generally, the component $p(\mathbf{y}_{t+1} \mid \mathbf{y}_{1:t}, M_k)$ is likely to be heavily influenced by the prior distribution when t is small. This highlights the relevance of performing sensitivity analysis when computing the marginal likelihood.

Analytical computation of the marginal likelihood in (1) is only possible for a few simple models. More complex models require simulation-based methods to evaluate the typically high-dimensional integral in (1). In what follows, we discuss two such methods. From here onwards we suppress the model indicator for clarity. For example, we denote the likelihood function simply by $p(\mathbf{y} \mid \boldsymbol{\psi})$. For the marginal likelihood estimators in this section, we are interested in their sensitivities with respect to the prior hyperparameters. More generally, let $\boldsymbol{\theta}_0$ denote the vector of all inputs that are of interest. We will then make the dependence on $\boldsymbol{\theta}_0$ explicit. For example, we write the prior density as $p(\boldsymbol{\psi}; \boldsymbol{\theta}_0)$ and the marginal likelihood as $p(\mathbf{y}; \boldsymbol{\theta}_0)$.

2.1 Chib's Method

Chib's method (Chib, 1995; Chib and Jeliazkov, 2001) is based on the observation that the marginal likelihood is the normalizing constant of the posterior distribution. By rearranging the definition of the posterior distribution, we have

$$p(\mathbf{y}; \boldsymbol{\theta}_0) = \frac{p(\mathbf{y} \mid \boldsymbol{\psi}) p(\boldsymbol{\psi}; \boldsymbol{\theta}_0)}{p(\boldsymbol{\psi} \mid \mathbf{y}; \boldsymbol{\theta}_0)}.$$

Hence, a natural estimator of $p(\mathbf{y})$ (written in log scale) is

$$\log \widehat{p(\mathbf{y}; \boldsymbol{\theta}_0)}_{\text{Chib}} = \log p(\mathbf{y} \mid \boldsymbol{\psi}^*) + \log p(\boldsymbol{\psi}^*; \boldsymbol{\theta}_0) - \log \widehat{p(\boldsymbol{\psi}^* \mid \mathbf{y}; \boldsymbol{\theta}_0)}. \tag{3}$$

The posterior ordinate $\boldsymbol{\psi}^*$ can, in principle, be any point in the posterior support, but for computational efficiency, it is typically chosen to be some "high density" point such as the posterior mean or mode.

In many situations, we can evaluate both the likelihood and the prior distribution ana-

lytically. The only unknown quantity is the posterior ordinate $p(\boldsymbol{\psi}^* \,|\, \mathbf{y}; \boldsymbol{\theta}_0)$, which can be estimated using Monte Carlo methods. In particular, if all the full conditional distributions are known, then $p(\boldsymbol{\psi}^* \,|\, \mathbf{y}; \boldsymbol{\theta}_0)$ can be estimated using posterior draws and additional draws from a series of suitably designed Gibbs samplers, the so-called reduced runs.

To give a concrete example, suppose we have a model with three parameter blocks $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \boldsymbol{\psi}_3)$ for which the posterior distribution $p(\boldsymbol{\psi} \,|\, \mathbf{y}; \boldsymbol{\theta}_0)$ can be obtained using a 3-block Gibbs sampler by drawing from the full-conditional posterior distributions of each parameter block. The posterior ordinate can be expressed as

$$p(\boldsymbol{\psi}^* \,|\, \mathbf{y}; \boldsymbol{\theta}_0) \equiv p(\boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^*, \boldsymbol{\psi}_3^* \,|\, \mathbf{y}; \boldsymbol{\theta}_0) = p(\boldsymbol{\psi}_1^* \,|\, \mathbf{y}; \boldsymbol{\theta}_0) p(\boldsymbol{\psi}_2^* \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*; \boldsymbol{\theta}_0) p(\boldsymbol{\psi}_3^* \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^*; \boldsymbol{\theta}_0).$$

The first quantity $p(\boldsymbol{\psi}_1^* \,|\, \mathbf{y}; \boldsymbol{\theta}_0)$ can be estimated using posterior draws from the Gibbs sampler, while the last quantity $p(\boldsymbol{\psi}_3^* \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^*)$ can be evaluated exactly at the posterior means of the draws. The middle term $p(\boldsymbol{\psi}_2^* \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*; \boldsymbol{\theta}_0)$ can be estimated using draws from two-block reduced Gibbs sampler from $p(\boldsymbol{\psi}_2 \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_3; \boldsymbol{\theta}_0)$ and $p(\boldsymbol{\psi}_3 \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2; \boldsymbol{\theta}_0)$ with $\boldsymbol{\psi}_1$ fixed at $\boldsymbol{\psi}_1^*$. The posterior ordinate of models with more blocks can be estimated similarly, albeit additional reduced runs are required.

2.2 The Cross-Entropy Method

The cross-entropy method was originally developed for rare-event simulation by Rubinstein (1997, 1999) using a multi-level procedure to construct the optimal importance sampling density. Chan and Kroese (2012) later show that the optimal importance sampling density can be obtained more accurately in one step using MCMC. This new variant is applied in Chan and Eisenstat (2015, 2018) for marginal likelihood estimation. Below we outline the main ideas.

For estimating the marginal likelihood in (1), the theoretical zero-variance importance sampling density is the posterior density $p(\boldsymbol{\psi} \,|\, \mathbf{y})$. Unfortunately, this density is only known up to a constant and cannot be used directly in practice. However, it provides a good benchmark to obtain a suitable importance sampling density. The key idea is to locate a density that is "close" to this ideal importance sampling density, denoted as $f^* = f^*(\boldsymbol{\psi}) = p(\boldsymbol{\psi} \,|\, \mathbf{y})$. Operationally, we consider a parametric family $\mathcal{F} = \{f(\boldsymbol{\psi}; \mathbf{v})\}$ indexed by the parameter vector $\mathbf{v} \in \mathbb{R}^{\dim_v}$, and then find the density $f(\boldsymbol{\psi}; \mathbf{v}^*) \in \mathcal{F}$ such that it is the "closest" to f^*.

One convenient measure of closeness between densities is the *Kullback-Leibler divergence* or the *cross-entropy distance*. Specifically, the cross-entropy distance from f_1 to f_2 is defined as: $\mathcal{D}(f_1, f_2) = \int f_1(\mathbf{x}) \log(f_1(\mathbf{x})/f_2(\mathbf{x}))\mathrm{d}\mathbf{x}$. Given this measure, we locate the density $f(\cdot; \mathbf{v}) \in \mathcal{F}$ such that $\mathcal{D}(f^*, f(\cdot; \mathbf{v}))$ is minimized. This minimization problem can be shown to be equivalent to finding

$$\mathbf{v}^*_{\text{ce}} = \underset{\mathbf{v}}{\text{argmax}} \int p(\mathbf{y} \mid \boldsymbol{\psi}) p(\boldsymbol{\psi}) \log f(\boldsymbol{\psi}; \mathbf{v}) \mathrm{d}\boldsymbol{\psi}.$$

This maximization problem is difficult to solve analytically, but $\mathbf{v}^*_{\text{ce}}$ can be estimated by

$$\widehat{\mathbf{v}}^*_{\text{ce}} = \underset{\mathbf{v}}{\text{argmax}} \frac{1}{R} \sum_{r=1}^{R} \log f(\boldsymbol{\psi}^r; \mathbf{v}), \tag{4}$$

where $\boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^R$ are posterior draws. This is analogous to finding the maximum likelihood estimate for $\mathbf{v}$ if we treat $f(\boldsymbol{\psi}; \mathbf{v})$ as the likelihood function with parameter vector $\mathbf{v}$ and $\boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^R$ as an observed sample. Since finding the maximum likelihood estimate is a standard problem, solving (4) is typically easy. For instance, analytical solutions are available for the exponential family (e.g., Rubinstein and Kroese, 2004, p. 70). Finally, once the optimal density $f(\cdot; \widehat{\mathbf{v}}^*_{\text{ce}})$ is obtained, it is used to construct the importance sampling estimator:

$$\widehat{p(\mathbf{y}; \boldsymbol{\theta}_0)}_{\text{ce}} = \frac{1}{N} \sum_{j=1}^{N} \frac{p(\mathbf{y} \mid \boldsymbol{\psi}^j) p(\boldsymbol{\psi}^j; \boldsymbol{\theta}_0)}{f(\boldsymbol{\psi}^j; \widehat{\mathbf{v}}^*_{\text{ce}})},$$

where $\boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^N$ are independent draws from the optimal importance sampling density $f(\boldsymbol{\psi}; \widehat{\mathbf{v}}^*_{\text{ce}})$.[8] One main advantage of this importance sampling approach is that it is easy to implement, and the numerical standard error of the estimator is readily available. We refer the readers to Chan and Eisenstat (2015) for a more thorough discussion.

3 Automatic Differentiation for Marginal Likelihood

In this section, we introduce a general framework to analyze the sensitivity of two marginal likelihood estimators with respect to a set of prior parameter hyperparameters, $\boldsymbol{\theta}_0 \in \mathbb{R}^p$. This builds on recent work by Jacobi, Joshi, and Zhu (2018) that has introduced prior sensitivity analysis for MCMC output based on Automatic Differentiation (AD), which

[8] See also Frühwirth-Schnatter (1995), which constructs a different importance sampling density by using a mixture of full conditional distributions given the latent states.

is designed to compute sensitivities with respect to the full set of input parameters.

3.1 AD Implementation

Bayesian MCMC algorithms are complicated high-dimensional mappings that take inputs such as hyperparameters of the prior distribution, the chain's starting values, and the data. For many applications, we are typically interested in the effect of a subset of these inputs, say the set of hyperparameters $\boldsymbol{\theta}_0$, on posterior outcomes. Formally, MCMC is a function that maps

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) \in \mathbb{R}^p \times \mathbb{R}^l \rightarrow \mathbf{G}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0),$$

where $\boldsymbol{\eta}_0$ refers to the set of inputs such as starting values that in combination with $\boldsymbol{\theta}_0$ are mapped via some MCMC algorithm $\mathbf{G}$ into posterior quantities.

AD is an efficient means of computing derivatives, i.e., the local sensitivity of the outputs with respect to the inputs of interest. In a nutshell, for a function $\mathbf{G}$, AD maps $\mathbf{G}$ into its vector of first-order partial derivatives automatically, $\frac{\partial}{\partial \boldsymbol{\theta}_0}\mathbf{G}$, i.e. a *function operator*

$$AD : \mathbf{G} \rightarrow \frac{\partial}{\partial \boldsymbol{\theta}_0}\mathbf{G}.$$

Technically, the complementary AD is able to compute the derivatives of the posterior output $\mathbf{G}$ with respect to the complete set of inputs, both $\boldsymbol{\theta}_0$ and $\boldsymbol{\eta}_0$. In practice, however, it is up to the analyst to choose which subset of inputs are included in $\boldsymbol{\theta}_0$.

Like symbolic differentiation, implemented in many widely used software packages, AD computes exact partial derivatives of the original mapping up to floating-point errors. However, while symbolic differentiation operates by first deriving and then evaluating the analytical expression of $\frac{\partial}{\partial \boldsymbol{\theta}_0}\mathbf{G}$, AD evaluates the derivatives *alongside* the algorithm $\mathbf{G}$ without any analytical derivations. This alleviates the issue of expression overloading and hence typically maintains a relative fast computation. More importantly, it avoids the infeasible derivation of symbolic expressions and focuses on the actual evaluation of derivative values.

AD is "automatic" in the sense that for an algorithm that maps the input vector $\boldsymbol{\theta}_0$ into the posterior output vector, there is an automatic way of evaluating its complementary sensitivities without manually deriving the symbolic formula of the derivatives. Instead, it is derived by first decomposing the original algorithm $\mathbf{G}$ into simpler opera-

tions $\mathbf{G}_1, \ldots, \mathbf{G}_k$:

$$\mathbf{G} = \mathbf{G}_k \circ \mathbf{G}_{k-1} \circ \cdots \circ \mathbf{G}_1,$$

where

$$\mathbf{G}_i : (\mathbf{x}_i, \boldsymbol{\theta}) \to \mathbf{x}_{i+1}$$

and $\mathbf{x}_i$ is the intermediary values at step i. Then, the derivative of $\mathbf{G}$ can be obtained via the chain-rule (that is implemented automatically in the computer program)

$$\frac{\partial \mathbf{G}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_0} = \sum_{i=1}^{k} \frac{\partial}{\partial \mathbf{x}_k} \mathbf{G}_k \frac{\partial}{\partial \mathbf{x}_{k-1}} \mathbf{G}_{k-1} \cdots \frac{\partial}{\partial \mathbf{x}_{i+1}} \mathbf{G}_{i+1} \frac{\partial}{\partial \boldsymbol{\theta}_0} \mathbf{G}_i,$$

where $\frac{\partial \mathbf{G}_i}{\partial \mathbf{x}_i}, i = 1, \ldots, k$ are the intermediate Jacobians of the simpler operations. While the end result $\frac{\partial \mathbf{G}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_0}$ is a dense matrix, the $\frac{\partial \mathbf{G}_i}{\partial \mathbf{x}_i}$'s are typically very sparse matrices because each operation $\mathbf{G}_i$ typically only updates one or two variables.

Since AD computes the derivatives of the algorithm $\mathbf{G}$, its computational requirements naturally depend on the computational complexity of $\mathbf{G}$. In fact, one can establish upper bounds for the computational complexity and memory requirement of AD relative to the original algorithm. These results are summarized in the following proposition.

Proposition 1. Consider computing the Jacobian of a function $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ with n input variables and m output variables. Suppose the base algorithm for evaluating $\mathbf{F}$ can be computed with L elementary operations. In that case, its gradient with respect to l of its inputs can be computed with $3lL$ elementary operations and with no more than twice the memory required for the original computation.

We refer the reader to Griewank and Walther (2008) for the proof and a more detailed exposition on AD. Note that in the classical simulation context, finite differencing offers an attractive alternative, however with well-document theoretical limitations, such as biased derivative estimates (Glasserman, 2003). In comparison, AD requires additional model analysis and programming, but this additional effort is justified by the improvement in the quality and computational efficiency expanding the scope of calculated local sensitivities. Due to the additional computation burden in MCMC settings, typically, only a very limited and mostly less formal numerical prior parameter sensitivity analysis is implemented.

While AD methods have been widely used to undertake input sensitivity analysis in the context of less computationally intensive classical simulation methods, particularly in

Financial Mathematics, it has only been recently introduced in the context of MCMC simulation by Jacobi, Joshi, and Zhu (2018). The paper develops an AD approach and AD based methods for a comprehensive prior parameters sensitivity analysis of MCMC output and shows how the forward mode of differentiation can be applied to compute Jacobian matrices of first order derivatives for MCMC based statistic based on prior input sensitivities of the parameter draws in Gibbs MCMC settings. In particular, sensitivities can often be derived using information about model dynamics in simulation,i.e., the dependence of the posterior distribution on the set of prior assumptions, as AD proceeds by differentiating the evolution of the underlying state variables along each path.

Since both Chib's and the cross-entropy methods require posterior draws of the model parameters, we apply the AD approach for MCMC Gibbs settings to obtain the first-order partial derivatives of the model parameters with respect to $\boldsymbol{\theta}_0$. In order to extend the approach to prior parameter sensitivities of marginal likelihood estimation, we introduce additional steps needed to compute the complete set of first-order derivatives of $\log \widehat{p(\mathbf{y};\boldsymbol{\theta}_0)}_{\text{Chib}}$ and $\log \widehat{p(\mathbf{y};\boldsymbol{\theta}_0)}_{\text{ce}}$ with respect to $\boldsymbol{\theta}_0$. We provide MATLAB code to implement the AD-based prior sensitivity analysis of marginal likelihood estimators.[9] It is important to stress that while most AD packages act as a black box supporting a particular statistical application, our AD implementation emphasizes transparency and flexibility to allow for extensions, such as computing the marginal likelihood as a post MCMC procedure considered in this paper. In what follows we focus on the key points in passing the AD operator through Chib's and the cross-entropy methods.

3.2 Gradient of Chib's Estimator

Chib's estimator for the log-marginal likelihood consists of three components: the log-likelihood, the log-prior and the log-posterior, all evaluated at some posterior ordinate $\boldsymbol{\psi}^*$, such as the posterior mean or mode. Furthermore, let $\boldsymbol{\psi}^r, r = 1, 2, \ldots, R$ denote the posterior draws obtained using MCMC.

Assuming that we have already obtained the Jacobian of the posterior ordinate $\frac{\partial \boldsymbol{\psi}^*}{\partial \boldsymbol{\theta}_0}$ as well as the draws of the model parameters $\frac{\partial \boldsymbol{\psi}^r}{\partial \boldsymbol{\theta}_0}$ for $r = 1, 2, \ldots, R$, the Jacobian of the

[9] The code associated with this paper can be downloaded at `http://joshuachan.org/research.html`.

first two components can be obtained via:

$$\frac{\partial \log p(\mathbf{y} \mid \boldsymbol{\psi}^*)}{\partial \boldsymbol{\theta}_0} = \frac{1}{p(\mathbf{y} \mid \boldsymbol{\psi}^*)} \frac{\partial p(\mathbf{y} \mid \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*} \frac{\partial \boldsymbol{\psi}^*}{\partial \boldsymbol{\theta}_0}$$
$$\frac{\partial \log p(\boldsymbol{\psi}^*; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} = \frac{1}{p(\mathbf{y} \mid \boldsymbol{\psi}^*)} \left[\frac{\partial p(\boldsymbol{\psi}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}\bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*} \frac{\partial \boldsymbol{\psi}^*}{\partial \boldsymbol{\theta}_0} + \frac{\partial p(\boldsymbol{\psi}^*; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} p(\boldsymbol{\psi}^*; \boldsymbol{\theta}_0) \right].$$

Here we assume that both the likelihood and the prior distribution functions are continuously differentiable in $\boldsymbol{\psi}$.[10]

For the log-posterior term estimated from reduced runs, the derivative operator needs to be applied through the additional Monte Carlo simulation as well. For example, for a three-block Gibbs sampler with $\boldsymbol{\psi} = (\boldsymbol{\psi}_1', \boldsymbol{\psi}_2', \boldsymbol{\psi}_3')'$, we have

$$\frac{\partial}{\partial \boldsymbol{\theta}_0} \log \widehat{p(\boldsymbol{\psi}^* \mid \mathbf{y}; \boldsymbol{\theta}_0)} = \frac{\frac{\partial p(\boldsymbol{\psi}_1^* \mid \mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}}{p(\boldsymbol{\psi}_1^* \mid \mathbf{y}; \boldsymbol{\theta}_0)} + \frac{\frac{\partial p(\boldsymbol{\psi}_2^* \mid \mathbf{y}, \boldsymbol{\psi}_1^*; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}}{p(\boldsymbol{\psi}_2^* \mid \mathbf{y}, \boldsymbol{\psi}_1^*; \boldsymbol{\theta}_0)} + \frac{\frac{\partial p(\boldsymbol{\psi}_3^* \mid \mathbf{y}, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^*; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}}{p(\boldsymbol{\psi}_3^* \mid \mathbf{y}, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^*; \boldsymbol{\theta}_0)}.$$

We can estimate the derivative of the first term via the original MCMC[11]

$$\frac{\partial p(\boldsymbol{\psi}_1^* \mid \mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} = \frac{1}{R} \sum_{r=1}^{R} \frac{\partial p(\boldsymbol{\psi}_1^* \mid \boldsymbol{\psi}_2^r, \boldsymbol{\psi}_3^r, \mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} + \frac{\partial p(\boldsymbol{\psi}_1^* | \boldsymbol{\psi}_2^r, \boldsymbol{\psi}_3^r, \mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}} \begin{bmatrix} \frac{\partial \boldsymbol{\psi}_1^*}{\partial \boldsymbol{\theta}_0} \\ \frac{\partial \boldsymbol{\psi}_2^r}{\partial \boldsymbol{\theta}_0} \\ \frac{\partial \boldsymbol{\psi}_3^r}{\partial \boldsymbol{\theta}_0} \end{bmatrix},$$

where $\dfrac{\partial p(\boldsymbol{\psi}_1^* \mid \boldsymbol{\psi}_2^r, \boldsymbol{\psi}_3^r, \mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}$ denotes the partial derivative of $p(\boldsymbol{\psi}_1 \mid \boldsymbol{\psi}_2, \boldsymbol{\psi}_3, \mathbf{y}; \boldsymbol{\theta}_0)$ with respect to $\boldsymbol{\psi}$ evaluated at $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^r, \boldsymbol{\psi}_3^r)'$.

The second term can be estimated via a reduced run of N sample by fixing $\boldsymbol{\psi}_1^*$

$$\frac{\partial p(\boldsymbol{\psi}_2^* \mid \mathbf{y}, \boldsymbol{\psi}_1^*; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} = \frac{1}{N} \sum_{n=1}^{N} \frac{\partial p(\boldsymbol{\psi}_2^* \mid \boldsymbol{\psi}_3^n, \boldsymbol{\psi}_1^*, \mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} + \frac{\partial p(\boldsymbol{\psi}_2^* \mid \boldsymbol{\psi}_3^n, \boldsymbol{\psi}_1^*, \mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}} \begin{bmatrix} \frac{\partial \boldsymbol{\psi}_1^*}{\partial \boldsymbol{\theta}_0} \\ \frac{\partial \boldsymbol{\psi}_2^*}{\partial \boldsymbol{\theta}_0} \\ \frac{\partial \boldsymbol{\psi}_3^n}{\partial \boldsymbol{\theta}_0} + \frac{\partial \boldsymbol{\psi}_3^n}{\partial \boldsymbol{\psi}_1^*} \frac{\partial \boldsymbol{\psi}_1^*}{\partial \boldsymbol{\theta}_0} \end{bmatrix}$$

[10] Since AD builds on an algorithm and evaluates its derivatives alongside the original algorithm, the applicability of our proposed method depends on the applicability of the original algorithm. In particular, since Chib's method requires the value of the likelihood at the posterior ordinate $\boldsymbol{\psi}^*$, the proposed method does so as well. This might limit the applicability of the proposed method as the evaluation of the likelihood can be time-consuming in some latent variable models. However, since Chib's method has been successfully applied to a wide range of high-dimensional latent variables models and is one of the dominant approaches to computing the marginal likelihood, we choose it as one of our examples.

[11] This amounts to changing the order of differentiation and integration, which is permissible if the posterior has continuous partial derivatives.

such that the sensitivities of $\boldsymbol{\psi}_3^n$ is obtained in the reduced run through its direct dependence on the hyperparameters and indirect dependence via $\boldsymbol{\psi}_1^*$.

Finally, the last term can be computed exactly as:

$$\frac{\partial p(\boldsymbol{\psi}_3^* \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^*; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} = \frac{\partial p(\boldsymbol{\psi}_3^* \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^*; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}} \begin{bmatrix} \frac{\partial \boldsymbol{\psi}_1^*}{\partial \boldsymbol{\theta}_0} \\ \frac{\partial \boldsymbol{\psi}_2^*}{\partial \boldsymbol{\theta}_0} \\ \frac{\partial \boldsymbol{\psi}_3^*}{\partial \boldsymbol{\theta}_0} \end{bmatrix},$$

where $\dfrac{\partial p(\boldsymbol{\psi}_3^* \,|\, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^*, \mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}$ denotes the partial derivative of $p(\boldsymbol{\psi}_3 \,|\, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \mathbf{y}; \boldsymbol{\theta}_0)$ with respect to $\boldsymbol{\psi}$ evaluated at $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^*, \boldsymbol{\psi}_3^*)'$.

In terms of memory budget of the original MCMC, the computer needs to store: 1) $\boldsymbol{\psi}^*$ and its associated derivatives; 2) $\boldsymbol{\psi}_2^r$ and $\boldsymbol{\psi}_3^r$ and their associated derivatives. To reduce the memory requirement, the blocking should be chosen in a way that $\boldsymbol{\psi}_1$ is the of the largest dimension.

Due to the construction of the marginal likelihood in Chib's method, in principle its derivative is independent of $\boldsymbol{\psi}$ as it is an integral over the prior space. More specifically, assuming $\boldsymbol{\psi}^*$, the ordinate at which to evaluate the densities, is fixed, we have

$$\frac{\partial}{\partial \boldsymbol{\theta}_0} \log p(\mathbf{y}; \boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}_0} \log p(\boldsymbol{\psi}^*; \boldsymbol{\theta}_0) - \frac{\partial}{\partial \boldsymbol{\theta}_0} \log p(\boldsymbol{\psi}^* | \mathbf{y}; \boldsymbol{\theta}_0)$$

as the likelihood drops out from the derivative expression above. This could reduce the algorithm's computational complexity, especially for models under which the likelihoods are expensive to evaluate. However, the above argument relies on fixing $\boldsymbol{\psi}^*$. In many practical situations where Chib's method is applied, the numerical stability of the estimator is sensitive to the choice of $\boldsymbol{\psi}^*$ (which is why $\boldsymbol{\psi}^*$ is often recommended to be chosen in some 'high-density' area such as the posterior mean). Hence, proceeding with the above expression might lead to an unstable estimator if $\boldsymbol{\psi}^*$ is sensitive to changes in $\boldsymbol{\theta}_0$.

3.3 Gradient of the Cross-Entropy Estimator

To calculate the gradient of the cross-entropy marginal likelihood estimator, we need to first obtain $\frac{\partial \mathbf{v}_{\text{ce}}^*}{\partial \boldsymbol{\theta}_0}$. For cases where analytical expressions of $\mathbf{v}_{\text{ce}}^*$ is available, e.g., for a Gaussian importance sampling density, the derivatives can be obtained directly via AD by passing through the analytical evaluation. The associated memory cost is then

negligible, i.e. $\mathbf{v}_{\text{ce}}^*$ is an analytical expression of $\boldsymbol{\psi}^r$'s, and the value and its derivative of $\mathbf{v}_{ce}^*$ are accumulated in the original MCMC algorithm.

When obtaining $\mathbf{v}_{\text{ce}}^*$ requires numerical search such the Newton-Raphson method, we can compute the derivative via the implicit function theorem.

Proposition 2. Assuming that the importance sampling density $f(\boldsymbol{\psi}; \mathbf{v})$ is twice continuously differentiable in both $\mathbf{v}$ and $\boldsymbol{\psi}$ with

$$\mathbb{E}_\pi \left[\left| \left| \frac{\partial^2 \log f(\boldsymbol{\psi}; \mathbf{v}_{\text{ce}}^*)}{\partial \mathbf{v}^2} \right| \right| \right] < \infty, \quad \mathbb{E}_\pi \left[\left| \left| \frac{\partial^2 \log f(\boldsymbol{\psi}; \mathbf{v}_{\text{ce}}^*)}{\partial \mathbf{v} \partial \boldsymbol{\psi}} \right| \right| \right] < \infty$$

and

$$\mathbb{E}_\pi \left[\frac{\partial^2 \log f(\boldsymbol{\psi}; \mathbf{v}_{\text{ce}}^*)}{\partial \mathbf{v}^2} \right]$$

is positive definite, then

$$\frac{\partial \mathbf{v}_{\text{ce}}^*}{\partial \boldsymbol{\theta}_0} = -\mathbb{E}_\pi \left[\frac{\partial^2 \log f(\boldsymbol{\psi}; \mathbf{v}_{\text{ce}}^*)}{\partial \mathbf{v}^2} \right]^{-1} \left(\mathbb{E}_\pi \left[\frac{\partial \log f(\boldsymbol{\psi}; \mathbf{v}_{\text{ce}}^*)}{\partial \mathbf{v}} \frac{\partial \log(\pi(\boldsymbol{\psi}; \boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta}_0'} \right] \right),$$

where the expectation $\mathbb{E}_\pi$ is taken with respect to the posterior measure.

Proof. Based on the first-order condition for $\mathbf{v}_{ce}$, we have

$$\int p(\mathbf{y}|\boldsymbol{\psi}) p(\boldsymbol{\psi}) \frac{\partial \log f(\boldsymbol{\psi}; \mathbf{v}_{ce}^*)}{\partial \mathbf{v}} \mathrm{d}\boldsymbol{\psi} = \mathbf{0}.$$

This is equivalent to

$$\int \pi(\boldsymbol{\psi}; \boldsymbol{\theta}_0) \frac{\partial \log f(\boldsymbol{\psi}; \mathbf{v}_{ce}^*)}{\partial \mathbf{v}} \mathrm{d}\boldsymbol{\psi} = \mathbb{E}_\pi \left[\frac{\partial \log f(\boldsymbol{\psi}; \mathbf{v}_{ce}^*)}{\partial \mathbf{v}} \right] = \mathbf{0}.$$

Given the regularity assumption, we can now apply derivative with respect to $\boldsymbol{\theta}_0$ to both sides

$$\mathbb{E}_\pi \left[\frac{\partial^2 \log f(\boldsymbol{\psi}; \mathbf{v}_{ce}^*)}{\partial \mathbf{v}^2} \right] \frac{\partial \mathbf{v}_{ce}^*}{\partial \boldsymbol{\theta}_0} + \mathbb{E}_\pi \left[\frac{\partial \log f(\boldsymbol{\psi}; \mathbf{v}_{ce}^*)}{\partial \mathbf{v}} \frac{\partial \log(\pi(\boldsymbol{\psi}; \boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta}_0'} \right] = \mathbf{0}.$$

The result is immediate by re-arranging the above expression. □

Let $\boldsymbol{\Psi} = \{\boldsymbol{\psi}^1, \boldsymbol{\psi}^2, ..., \boldsymbol{\psi}^R\}$ denote the collection of posterior draws, its consistent sample estimate is

$$\frac{\partial \widehat{\mathbf{v}_{\text{ce}}^*}}{\partial \boldsymbol{\theta}_0} = -\left[\sum_{r=1}^{R} \frac{\partial^2 \log f(\boldsymbol{\psi}^r; \mathbf{v}_{ce}^*)}{\partial \mathbf{v}^2} \right]^{-1} \left[\sum_{r=1}^{R} \frac{\partial^2 \log f(\boldsymbol{\psi}^r, \mathbf{v}_{\text{ce}}^*)}{\partial \mathbf{v} \partial \boldsymbol{\psi}^j} \frac{\partial \boldsymbol{\psi}^j}{\partial \boldsymbol{\theta}_0} \right].$$

This expression involves the storage of $\frac{\partial \boldsymbol{\psi}^r}{\partial \boldsymbol{\theta}_0}$'s from the original MCMC algorithm. Hence, it is operational if we choose the importance sampling density so that most of the parameters can be solved analytically, and the dimension of $\boldsymbol{\psi}$ that requires the above manipulation is small.

Finally, given the draws $\boldsymbol{\psi}^j, j = 1, \ldots, N$ from the importance sampling density $f(\boldsymbol{\psi}; \widehat{\mathbf{v}}^*_{\text{ce}})$, the derivative of the CE estimator is given by:

$$\frac{\partial \widehat{p(\mathbf{y}; \boldsymbol{\theta}_0)}_{\text{ce}}}{\partial \boldsymbol{\theta}_0} = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{\partial}{\partial \boldsymbol{\psi}} \Big(\frac{p(\mathbf{y} \mid \boldsymbol{\psi}^j) p(\boldsymbol{\psi}^j; \boldsymbol{\theta}_0)}{f(\boldsymbol{\psi}^j; \widehat{\mathbf{v}}^*_{\text{ce}})} \Big) \frac{\partial \boldsymbol{\psi}^j}{\partial \mathbf{v}^*_{\text{ce}}} - \frac{p(\mathbf{y} \mid \boldsymbol{\psi}^j) p(\boldsymbol{\psi}^j; \boldsymbol{\theta}_0)}{f(\boldsymbol{\psi}^j; \widehat{\mathbf{v}}^*_{\text{ce}})^2} \frac{\partial f(\boldsymbol{\psi}^j; \widehat{\mathbf{v}}^*_{\text{ce}})}{\partial \mathbf{v}} \right) \frac{\partial \widehat{\mathbf{v}^*_{\text{ce}}}}{\partial \boldsymbol{\theta}_0} + \frac{p(\mathbf{y} \mid \boldsymbol{\psi}^j)}{f(\boldsymbol{\psi}^j; \widehat{\mathbf{v}}^*_{\text{ce}})} \frac{\partial p(\boldsymbol{\psi}^j; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}.$$

In other words, the sensitivity of the cross-entropy estimator with respect to $\boldsymbol{\theta}_0$ is through its dependence on $\widehat{\mathbf{v}}^*_{\text{ce}}$. Depending on the complexity of obtaining $\mathbf{v}^*_{\text{ce}}$, the Jacobian $\frac{\partial \boldsymbol{\psi}^j}{\partial \mathbf{v}^*_{\text{ce}}}$ can be obtained either algorithmically or through the distributional derivative method in Jacobi, Joshi, and Zhu (2018).

3.4 Perturbation Analysis and Model Comparison

In most situations, it is not necessarily the sensitivity of the marginal likelihood that is of interest, but the sensitivity of the model ordering — e.g., one might worry that a small change in prior hyperparameter values would reverse the value of the Bayes factor from greater than one to less than one. Below we outline a procedure to extend our method of computing the gradient of the marginal likelihood to approximate the change in the Bayes factor from a small change of hyperparameter values.

To that end, consider the Bayes factor in favor of model i against model j:

$$\text{BF}_{i,j}(\boldsymbol{\eta}_0, \boldsymbol{\eta}_i, \boldsymbol{\eta}_j) = \frac{p_i(\boldsymbol{\eta}_0, \boldsymbol{\eta}_i)}{p_j(\boldsymbol{\eta}_0, \boldsymbol{\eta}_j)},$$

where p_i and p_j are respectively the marginal likelihoods of model i and model j, $\boldsymbol{\eta}_0$ denotes the vector of hyperparameters that are common in both models, and $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_j$ are model-specific hyperparameter vectors. Note that we suppress the dependence on the data $\mathbf{y}$ to simplify the notation. To approximate the change in Bayes factor when the hyperparameter values are changed from $(\boldsymbol{\eta}_0, \boldsymbol{\eta}_i, \boldsymbol{\eta}_j)$ to $(\boldsymbol{\eta}^*_0, \boldsymbol{\eta}^*_i, \boldsymbol{\eta}^*_j)$, one can use

perturbation analysis via AD as follows:

$$\mathrm{BF}_{i,j}(\boldsymbol{\eta}_0^*, \boldsymbol{\eta}_i^*, \boldsymbol{\eta}_j^*) \approx \mathrm{BF}_{i,j}(\boldsymbol{\eta}_0, \boldsymbol{\eta}_i, \boldsymbol{\eta}_j) + \left(\nabla \mathrm{BF}_{i,j}(\boldsymbol{\eta}_0, \boldsymbol{\eta}_i, \boldsymbol{\eta}_j)\right)' \begin{bmatrix} \boldsymbol{\eta}_0^* - \boldsymbol{\eta}_0 \\ \boldsymbol{\eta}_i^* - \boldsymbol{\eta}_i \\ \boldsymbol{\eta}_i^* - \boldsymbol{\eta}_i \end{bmatrix},$$

where the gradient $\nabla \mathrm{BF}_{i,j}(\boldsymbol{\eta}_0, \boldsymbol{\eta}_i, \boldsymbol{\eta}_j)$ can be computed using the AD-based methods described in previous sections. In particular, the three partial derivative vectors can be computed simultaneously via AD regardless of the perturbation size. Given the Bayes factor and its gradient at the baseline $(\boldsymbol{\eta}_0, \boldsymbol{\eta}_i, \boldsymbol{\eta}_j)$, this perturbation analysis allows an easy mechanism to assess the impact of small changes in hyperparameters on the Bayes factor and to check if the ranking of the models would change.

4 Empirical Applications

This section presents two empirical applications to illustrate the proposed automated prior sensitivity analysis based on Automatic Differentiation. The first application compares two vector autoregressions (VARs) for modeling a US macroeconomic dataset. In the second empirical example, we fit exchange rate data using factor models with different numbers of latent factors.

4.1 Vector Autoregressions for the US Economy

Since the seminal work of Sims (1980), vector autoregressions (VARs) have become a workhorse model for analyzing the evolving inter-relationships between multiple macroeconomic variables. VARs are widely used for structural analysis and macroeconomic forecasting. In particular, VARs combined with the Minnesota prior developed in Doan, Litterman, and Sims (1984) and Litterman (1986) are often used as benchmark models.

We perform a formal Bayesian model comparison exercise to compare two popular VARs for fitting a US macroeconomic dataset in the first application. We aim to identify salient model features that are useful in modeling the evolution and interdependence among the macroeconomic time series. To that end, let $\mathbf{y}_t$ be an $n \times 1$ vector of endogenous variables at time t with $t = 1, \ldots, T$. The first model we consider is the conventional VAR with

Gaussian innovations:

$$\mathbf{y}_t = \mathbf{b} + \mathbf{B}_1\mathbf{y}_{t-1} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\mathbf{b}$ is an $n \times 1$ vector of intercepts, $\mathbf{B}_1, \ldots, \mathbf{B}_p$ are $n \times n$ matrices of VAR coefficients, $\boldsymbol{\Sigma}$ is a covariance matrix, and $\mathcal{N}(\cdot, \cdot)$ denotes the normal distribution.

For estimation purpose, this VAR can be written in the seemingly unrelated regression (SUR) form as:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{5}$$

where $\mathbf{X}_t = \mathbf{I}_n \otimes (1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})$ and $\boldsymbol{\beta} = \text{vec}([\mathbf{b}, \mathbf{B}_1, \ldots, \mathbf{B}_p]')$ is the vector of intercepts and VAR coefficients stacked by rows. Note that the dimension of $\boldsymbol{\beta}$ is $k_\beta \times 1$ with $k_\beta = n(np + 1)$.

Despite the empirical success of the standard VAR with Gaussian innovations, recent research has found that macroeconomic variables are occasionally subject to large shocks (see, e.g. Cúrdia, Del Negro, and Greenwald, 2014). Hence, the second model we consider is a VAR with t innovations, which we denote as VAR-t. That is, instead of the Gaussian distribution for the innovations, we assume they follow a multivariate t distribution. For ease of estimation, we use the following latent variable representation: $(\boldsymbol{\varepsilon}_t \,|\, \boldsymbol{\Sigma}, \lambda_t) \sim \mathcal{N}(\mathbf{0}, \lambda_t\boldsymbol{\Sigma})$ with $(\lambda_t \,|\, \nu) \sim \mathcal{IG}(\nu/2, \nu/2)$, where $\mathcal{IG}(\cdot, \cdot)$ denote the inverse-gamma distribution. Then marginal of λ_t, $\boldsymbol{\varepsilon}_t$ has a multivariate t distribution with mean vector $\mathbf{0}$, scale matrix $\boldsymbol{\Sigma}$ and degree of freedom parameter ν (see, e.g., Geweke, 1993). Empirical work that uses VARs with t innovations include Clark and Ravazzolo (2015), Cross and Poon (2016) and Chiu, Mumtaz, and Pinter (2017).

4.1.1 Data, Priors and Estimation

For our first application, we use a US quarterly macroeconomic dataset that involves GDP deflator and real GDP from 1948:Q1 to 2019:Q4. Both variables are commonly used in structural analysis and forecasting (e.g., Banbura, Giannone, and Reichlin, 2010; Koop, 2013), and following standard practice, they are transformed to annualized growth rates. The data are sourced from the Federal Reserve Bank of St. Louis economic database, and they are plotted in Figure 1.

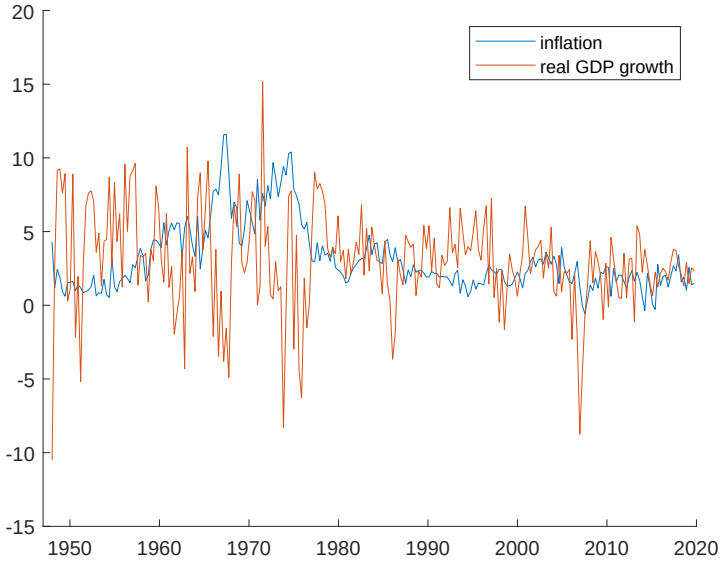

Figure 1: Plots of GDP deflator inflation and real GDP growth.

Next, we describe the priors for the two VARs. In general, we maintain the same priors for common parameters across models. For the Gaussian VAR, the parameters are $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. We assume a standard inverse-Wishart prior for $\boldsymbol{\Sigma}$ and a Minnesota-type prior for $\boldsymbol{\beta}$ that shrinks the VAR coefficients to zero:

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}}), \qquad \boldsymbol{\Sigma} \sim \mathcal{IW}(k_{0,\boldsymbol{\Sigma}}, \mathbf{S}_{0,\boldsymbol{\Sigma}}), \tag{6}$$

where $\mathcal{IW}(\cdot,\cdot)$ denotes the inverse-Wishart distribution. The prior covariance matrix $\mathbf{V}_{\boldsymbol{\beta}}$ is assumed to be diagonal with diagonal elements $v_{\boldsymbol{\beta},ii} = \kappa_1/(l^2\widehat{s}_r)$ for a coefficient associated to lag l of variable r and $v_{\boldsymbol{\beta},ii} = \kappa_2$ for an intercept, where $\widehat{s}_r$ is the sample variance of an AR(4) model for the variable r. To avoid theoretical issues of using the sample to calibrate the prior, we compute $\widehat{s}_r$ using a pre-sample. Specifically, we use the data from 1948:Q1 to 1957:Q4 as the pre-sample; the sample to compute the marginal likelihood then starts from 1958:Q1 and ends in 2019:Q4. We set $\kappa_1 = 0.4^2$ and $\kappa_2 = 10^2$. These values imply that the coefficient associated to a lag l variable is shrunk more heavily to zero as the lag length increases, but intercepts are not shrunk to zero. Further we set $k_{0,\boldsymbol{\Sigma}} = n + 3$, $\mathbf{S}_{0,\boldsymbol{\Sigma}} = \kappa_3 \mathbf{I}_n$ with $\kappa_3 = 1$. Similar hyperparameter values are widely used in the literature; see, e.g., Koop and Korobilis (2010) or Karlsson (2013). For the VAR with t innovations, the parameters are $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ (the degree of freedom parameter ν is fixed but we consider a range of values). We use exactly the same priors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ as

in the Gaussian VAR case given in (6).

Bayesian estimation of the two VARs is fairly standard. Estimation of the Gaussian VAR is done using a 2-block Gibbs sampler by sequentially drawing from $p(\boldsymbol{\beta} \,|\, \mathbf{y}, \boldsymbol{\Sigma})$ and $p(\boldsymbol{\Sigma} \,|\, \mathbf{y}, \boldsymbol{\beta})$. In particular, in the notations of Section 2.1, the two blocks of parameters are $\boldsymbol{\psi}_1 = \boldsymbol{\beta}$ and $\boldsymbol{\psi}_2 = \boldsymbol{\Sigma}$. For a textbook treatment on the estimation on the Gaussian VAR, see, e.g., Koop and Korobilis (2010) or Chan (2020b).

For estimating the VAR with t innovations, we implement a 3-block Gibbs sampler that sequentially draws from the conditional distributions: $p(\boldsymbol{\beta} \,|\, \mathbf{y}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$, $p(\boldsymbol{\Sigma} \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \nu)$ and $p(\boldsymbol{\lambda} \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)$, where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_T)'$ is the vector of latent variables. For more estimation details on a regression with t innovations, see Chan, Koop, Poirier, and Tobias (2019). To estimate the marginal likelihood, we integrate out analytically the latent variables $\boldsymbol{\lambda}$ and use $p(\mathbf{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)$ as the likelihood. Similar to the Gaussian VAR, the two blocks of parameters are $\boldsymbol{\psi}_1 = \boldsymbol{\beta}$ and $\boldsymbol{\psi}_2 = \boldsymbol{\Sigma}$.

Finally, in implementing the proposed method, we store all the posterior draws and their associated derivatives. In addition, to speed up the computations of the gradient of Chib's method, we also store the means and precision matrices of the full conditional distribution of the VAR coefficients.

4.1.2 Empirical Results

We fit the US quarterly dataset using VARs with Gaussian and t errors and lag length $p = 2$.[12] For each VAR, we first obtain 10,000 posterior draws after a burn-in period of 1,000. We then compute the marginal likelihood value using both Chib's method and the cross-entropy method. We set the simulation size of both the reduced-run for Chib's method and the importance sampling for the CE method to be 10,000. The results are reported in Table 1.

Both Chib's and the cross-entropy methods give exactly the same marginal likelihood estimates. Our results show that the data overwhelmingly prefer VARs with t errors to the benchmark with Gaussian errors. This is consistent with earlier empirical studies that show VARs with t errors generally forecast better than those with Gaussian errors (e.g., Cross and Poon, 2016; Chiu, Mumtaz, and Pinter, 2017; Chan, 2020a). Among the t VARs, the one with the heaviest tails ($\nu = 5$) receives the most support. Even the

[12]For all VARs, the choice of lag length $p = 2$ is best supported by the data. Additional model comparison results with various lag lengths are provided in Appendix B.

version with the thinnest tails ($\nu = 30$) is strongly favored by the data compared to the Gaussian VAR. At first glance, this finding might appear to be surprising; after all, the t distribution with degree of freedom $\nu = 30$ is very similar to the standard Gaussian distribution for most of their supports. To investigate this issue further, we compute the standardized residuals using the least squares (see Figure 5 in Appendix B). There are a few residuals with absolute magnitudes larger than 3 (and one is larger than 4), which are highly unlikely under the Gaussian assumption, but are much more probable under the t distribution.[13]

Table 1: Log marginal likelihood estimates of the VAR and VAR with t innovations using the cross-entropy method (CE) and Chib's method (Chib).

	VAR	VAR-t		
		$\nu = 5$	$\nu = 10$	$\nu = 30$
CE	-1039.6	-1013.0	-1019.7	-1031.3
Chib	-1039.6	-1013.0	-1019.7	-1031.3

Next, Table 2 reports the derivatives of the log marginal likelihood estimates with respect to the three key hyperparameters: κ_1, κ_2 and κ_3. Recall that κ_1 controls the overall shrinkage strength of the VAR coefficients; κ_2 is the prior variance for the intercepts; and κ_3 controls the prior mean of the covariance matrix $\mathbf{\Sigma}$.

Our results show that the marginal likelihood estimates are relatively sensitive to κ_1 and κ_3, but not to κ_2. For example, decreasing κ_1 from the baseline value of 0.4^2 to 0.3^2 would increase the log marginal likelihood value of the Gaussian VAR by about 1.54,[14] but decreasing κ_2 by the same proportion—from the baseline value of 100 to 56—has little impact on the marginal likelihood value.

[13]For instance, for a standard Gaussian random variable X, the probability that $|X| > 4$ is about 6.33×10^{-5}, which is an order of magnitude smaller than that for a t distributed random variable with degree of freedom parameter $\nu = 30$ (3.82×10^{-4}).

[14]To check this estimate, we redid the marginal likelihood estimation with $\kappa_1 = 0.3^2$, while keeping other hyperparameters exactly the same. The new marginal likelihood value increases by 2, which is similar to the original estimate.

Table 2: Derivatives of log marginal likelihood estimates of the VAR and VAR with t innovations with respect to the hyperparameters.

	VAR			VAR-t ($\nu = 5$)		
	κ_1	κ_2	κ_3	κ_1	κ_2	κ_3
CE	-22.0	-0.01	5.54	-12.8	-0.01	4.09
Chib	-22.0	-0.01	5.54	-12.9	-0.01	4.09

Interestingly, even though the three hyperparameters are common across the two VARs, their impacts on the marginal likelihood values differ across the two VARs. For example, decreasing κ_1, i.e., increasing the strength of shrinkage, helps the Gaussian VAR fit the data better relative to the t VAR. In view of the differential impact of the common hyperparameters, it would be of interest to assess if the ranking of the models would change over a range of reasonable hyperparameter values. For example, even if we halve the value of κ_1, the log marginal likelihood values of the Gaussian and t VARs would be about -1037.8 and -1012, respectively.[15] Since the difference between the two values remains large, the conclusion that the data strongly prefer the t VAR is reasonably robust.

To illustrate the computational requirements of the proposed method, Table 3 breaks down the runtime and memory requirements of the three components: obtaining posterior draws using MCMC and estimating the marginal likelihood using the CE method and Chib's method. Overall the proposed method is reasonably fast — even for the high-dimensional latent variable model VAR-t, obtaining 10,000 draws and all the necessary derivatives take only 2 minutes. The memory requirements to store all the necessary derivatives are also modest, which are well within a standard desktop computers' capacity.[16]

Table 3: Performance measures of the proposed AD algorithm for VAR and VAR-t.

	VAR			VAR-t ($\nu = 5$)		
	MCMC	CE	Chib's	MCMC	CE	Chib's
Runtime (s)	6	3	2	120	54	75
Memory used (GB)	2.9	< 0.01	< 0.01	2.9	< 0.01	< 0.01

[15] To verify these estimates, we redid the marginal likelihood estimation with $\kappa_1 = 0.4^2/2$ for the two VARs. The log marginal likelihood values of the Gaussian and t VARs are, respectively -1037.2 and -1012.

[16] The memory requirement is the total memory used by MATLAB. As a benchmark, an empty workspace in MATLAB requires about 2.2 GB of memory.

Next, to assess how the marginal likelihood estimates vary over a wider range of hyperparameter values, one can plot the estimates together with the corresponding partial derivatives against some hyperparameters of interest. In addition, this provides a visual way to select the values of some key shrinkage hyperparameters by maximizing the marginal likelihood, as is commonly done in applications (see, e.g., Del Negro and Schorfheide, 2004; Schorfheide and Song, 2015; Carriero, Clark, and Marcellino, 2015). As an example, Figure 2 plots the partial derivatives of the log marginal likelihood values with respect to the overall shrinkage parameter κ_1 on the VAR coefficients. As the figure shows, for values of κ_1 less than about 0.015, the derivatives are large and positive, indicating that a small increase in κ_1 would substantially increase the marginal likelihood value. However, for values of κ_1 greater than 0.015, the derivative values are small in magnitude and even negative. These results suggest that the maximizer is around 0.15. In addition, since the marginal likelihoods for a wide range of values of κ_1 are all less than -1033, one can confirm that the data favor the VAR with t errors.

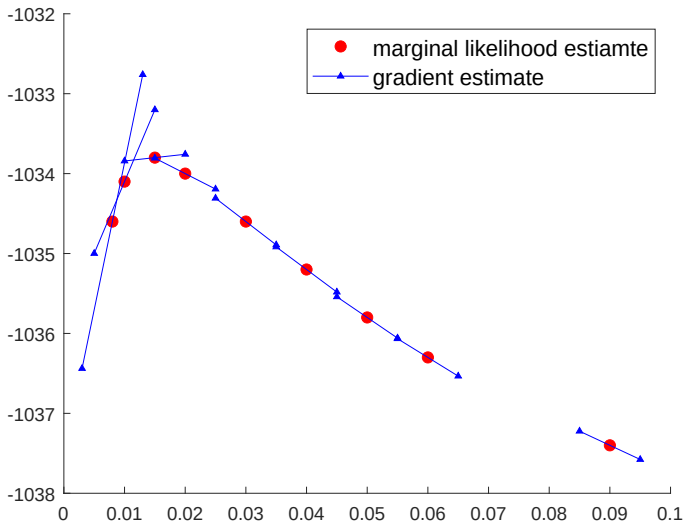

Figure 2: Marginal likelihood estimates (red dots) and the corresponding partial derivatives with respect to κ_1 (blue tangents) under the Gaussian VAR.

4.2 Factor Models for Exchange Rate Returns

Factor models have been widely used in many different areas, including psychology, bioinformatics, economics and finance. They are often used for modeling the dependence structure of high-dimensional data. One central interest in factor analysis is to determine

the number of latent factors. In the second application, we compare factor models with different number of factors for fitting a dataset of exchange rates.

More specifically, let $\mathbf{y}_t$ denote the $n \times 1$ vector of observations at time t with $t = 1, \ldots, T$, and let $\mathbf{f}_t$ represent a vector of k latent factors. Then, the k-factor model is specified as:

$$\mathbf{y}_t = \mathbf{A}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \tag{7}$$

where $\mathbf{A}$ is the $n \times k$ loading matrix. The factors and the innovations are assumed to be independent and normally distributed: $\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are diagonal. For the purpose of identification, we also require $n \geqslant 2k + 1$ and assume that $\mathbf{A}$ is lower triangular where the diagonal elements are set to be one (see, for example, the discussion in Geweke and Zhou, 1996).

4.2.1 Data, Priors and Estimation

In the second application, we analyze daily returns on ten international currency exchange rates relative to the US dollar beginning in January 2016 and ending in December 2020. Specifically, the exchange rate returns are computed as $y_{it} = 100 \log(p_{i,t}/p_{i,t-1})$, where p_{it} denotes the daily closing spot rate for currency i at time t. We then demean these returns so that their sample means are zero. The ten currencies are the Australian Dollar (AUD), Canadian Dollar (CAD), Swiss Franc (CHF), Chinese Yuan (CNY), Euro (EUR), British Pound (GBP), Japanese Yen (JPY), South Korean Won (KRW), New Zealand Dollar (NZD) and New Taiwan Dollar (TWD). These represent some of the most heavily traded currencies over the period. The dataset is sourced from the Federal Reserve Bank of St. Louis economic database.

To specify the priors, first let $\mathbf{a}$ denote the vector of free elements in the factor loadings $\mathbf{A}$ stacked by row. Note that the dimension of $\mathbf{a}$ is $k_a = kn - k(k+1)/2$. Now, the parameters for the k-factor model are $\mathbf{a}, \boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, and $\boldsymbol{\Omega} = \text{diag}(\omega_1^2, \ldots, \omega_k^2)$. We consider the following independent priors:

$$\mathbf{a} \sim \mathcal{N}(\mathbf{a}_0, \mathbf{V}_\mathbf{a}), \quad \sigma_i^2 \sim \mathcal{IG}(\nu_{\sigma_i^2}, S_{\sigma_i^2}), \quad \omega_j^2 \sim \mathcal{IG}(\nu_{\omega_j^2}, S_{\omega_j^2}) \tag{8}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, k$. We parameterize the priors so that they depend on 4 key hyperparameters $\kappa_4, \kappa_5, \kappa_6$ and κ_7. More specifically, we set $\mathbf{a}_0 = \kappa_4 \mathbf{1}_{k_a}$, $\mathbf{V}_\mathbf{a} = \kappa_5 \mathbf{I}_{k_a}$, $\nu_{\sigma_i^2} = \kappa_6$, $S_{\sigma_i^2} = \kappa_6 - 1$, $\nu_{\omega_j^2} = \kappa_7$ and $S_{\omega_j^2} = \kappa_7 - 1$, where $\kappa_4 = 0$ and $\kappa_5 = 1$, $\kappa_6 = \kappa_7 = 3$.

Using these hyperparameter values, the prior means of σ_i^2 and ω_j^2 are both 1, and their prior variances are, respectively, $1/(\kappa_6 - 2)$ and $1/(\kappa_7 - 2)$.

The factor model is estimated using a 4-block Gibbs sampler by sequentially drawing from $p(\mathbf{A} \,|\, \mathbf{y}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}, \mathbf{f})$, $p(\boldsymbol{\Sigma} \,|\, \mathbf{y}, \mathbf{A}, \boldsymbol{\Omega}, \mathbf{f})$, $p(\boldsymbol{\Omega} \,|\, \mathbf{y}, \mathbf{A}, \boldsymbol{\Sigma}, \mathbf{f})$ and $p(\mathbf{f} \,|\, \mathbf{y}, \mathbf{A}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$, where $\mathbf{f} = (\mathbf{f}_1', \ldots, \mathbf{f}_T')'$ is the vector of latent factors. Estimation details can be found in standard sources such as Geweke and Zhou (1996) and Lopes and West (2004). We note that while this approach is easy to implement, in some instances mixing could be slow. In those cases, one might consider a more efficient sampler of jointly sampling the factors and the factors loadings, as proposed in Chib, Nardari, and Shephard (2006). Lastly, to estimate the marginal likelihood, we integrate out analytically the latent factors $\mathbf{f}$ and use $p(\mathbf{y} \,|\, \mathbf{A}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ as the likelihood; its analytical expression is given in Appendix A. In the notations of Section 2.1, the three blocks of parameters are $\boldsymbol{\psi}_1 = \mathbf{A}$, $\boldsymbol{\psi}_2 = \boldsymbol{\Sigma}$ and $\boldsymbol{\psi}_2 = \boldsymbol{\Omega}$.

4.2.2 Results

We fit the exchange rate returns data using the factor models with $k = 1$ to $k = 4$ factors. For each factor model, we first obtain 10,000 posterior draws after a burn-in period of 1,000. We then compute the marginal likelihood value using both Chib's and the cross-entropy methods. We set the simulation size for both methods to 10,000. For computing the gradients of Chib's and CE methods, we store all the posterior draws and their associated derivatives. We report the log marginal likelihood estimates in Table 4.

The two marginal likelihood estimators give slightly different estimates, but they are within simulation errors.[17] In addition, both methods are consistent in terms of the ranking of the models — both indicate that the 3-factor model is most preferred by the data, whereas the 4-factor model comes in second. Furthermore, both methods show a substantial initial increase in log marginal likelihood—e.g. log marginal likelihood

[17]It is not uncommon for different estimators of the marginal likelihood of a factor model to give somewhat different estimates. For example, Lopes and West (2004) computes the log marginal likelihood of a 2-factor model using a variety of estimators, and the estimates range from -871.0 from the harmonic mean estimator to -935.3 from the Chib's method. However, these differences are likely due to the overestimation in some inaccurate estimators, such as the harmonic mean estimator. Finally, we also compute the derivatives of the log marginal likelihood estimate from Chib's method with respect to the posterior ordinates, and the results are reported in Appendix B. In principle, these derivatives should all be zero. In practice, however, Chib's estimates are sometimes found to be sensitive to the posterior ordinates' choice. Our results indicate that while Chib's estimates are insensitive to the selected values of the factor loadings, but they are relatively sensitive to those of the variances.

increases by about 422 from $k = 1$ to $k = 2$ factors and 272 from $k = 2$ to $k = 3$ factors—but it starts to decrease when we move from $k = 3$ to $k = 4$ factors.

Table 4: Log marginal likelihood estimates of the factor model with k factors using the cross-entropy method (CE) and Chib's method (Chib).

	$k = 1$	$k = 2$	$k = 3$	$k = 4$
CE	-6106.7	-5684.5	-5411.2	-5432.4
Chib	-6104.9	-5683.1	-5411.9	-5434.4

Next, Table 5 reports the derivatives of the log marginal likelihood estimates for the 3- and 4-factor models with respect to the four key hyperparameters: $\kappa_4, \kappa_5, \kappa_6$ and κ_7. Recall that κ_4 and κ_5 control, respectively, the prior means and the prior variances of the factor loadings $\mathbf{a}$. The prior means of σ_i^2 and ω_i^2 are both fixed at 1, but their priors variances are inversely proportional to κ_6 and κ_7, respectively.

Table 5: Derivatives of the log marginal likelihood estimates of the 3- and 4-factor models with respect to the hyperparameters.

	$k = 3$				$k = 4$			
	κ_4	κ_5	κ_6	κ_7	κ_4	κ_5	κ_6	κ_7
CE	12.9	-5.7	-124	-35.9	11.1	-8.0	-129	-63.2
Chib	12.8	-5.8	-125	-35.6	11.1	-8.0	-128	-64.2

The results suggest that the marginal likelihood values are more sensitive to κ_4, which controls the prior means of the factor loadings, relative to κ_5, which controls the prior variances. For instance, increasing κ_4 by 1 would increase the marginal likelihood of the 3-factor model by about 13, whereas the same increase in value for κ_5 would decrease the marginal likelihood by about 6. Next, between κ_6 and κ_7, the former has a much larger impact on the marginal likelihood than the latter. For example, increasing κ_6 by 1 would decrease the marginal likelihood of the 3-factor model by about 124-125. Since κ_6 controls the shrinkage strength on $\sigma_1^2, \ldots, \sigma_n^2$ to unity, these results suggest that stronger shrinkage of the idiosyncratic variances to unity has a large negative impact on the model-fit, and this negative impact is much larger than similar shrinkage of the factor variances.

It is also interesting to note that except for κ_7, the hyperparameters seem to have a similar impact on both factor models. Since κ_7 controls the shrinkage strength on the

factor variances $\omega_1^2, \ldots, \omega_k^2$, it is not surprising that its impact on the marginal likelihood is noticeably larger for the 4-factor model than the 3-factor model. Due to the differential impact of κ_7, the marginal likelihood of the 4-factor model increases more than that of the 3-factor model as κ_t decreases. Given that these two models have rather similar marginal likelihood values at the baseline setting, their marginal likelihood values would become even closer if we reduce the value of κ_7. We thus conclude that the 3- and 4-factor models receive similar support from the data. This sensitivity analysis highlights the value of having gradient estimates of the log marginal likelihood with respect to some key hyperparameters.

Finally, we report in Table 6 the runtimes and memory requirements of the proposed method. As is clear from the table, the proposed method is fast and requires only a modest amount of memory to store all the necessary derivatives. For example, the runtime of the main MCMC for a 3-factor model takes only 1.5 minutes and requires about 3.2 GB of memory.

Table 6: Performance measures of the proposed AD algorithm for the factor models.

	$k = 3$			$k = 4$		
	MCMC	CE	Chib's	MCMC	CE	Chib's
Runtime (s)	90	15	69	106	15	80
Memory used (GB)	3.2	0.02	< 0.01	3.2	0.02	< 0.01

5 A Simulation Study: Empirical Computational Complexity

In this section, we conduct a series of simulations to illustrate the computational requirements of the proposed method in a variety of high-dimensional settings. More specifically, we report the runtimes of the different components of the proposed method and the associated memory requirements as a function of various inputs (e.g., the dimensionality of the model and sample size) in the context of a factor model. For all the simulations, we set the main MCMC run, reduced-run in Chib's method, and the Monte Carlo sample size in the CE method to be 10,000.

We first investigate the computational requirements of the proposed method as a function of the sample size T, ranging from 200 to 3,200. For this experiment, we set the number

of factors r to be 3 and the number of variables n to be 10. Figure 3 reports the runtimes of the main MCMC run, the CE method and Chib's method, as well as the memory requirements of the main MCMC run (the memory requirements of the CE method and Chib's method are negligible). As is clear from the figure, the runtimes increase linearly in the sample size T, whereas the memory required appears to be insensitive to T. Overall, even for a large dataset with $T = 3,200$ observations, the proposed method is fast and requires only a modest amount of memory.

Next, we report in Figure 4 the computational requirements of the proposed method as a function of the number of variables n, ranging from 5 to 50. For this experiment, we set the number of factors r to be 3 and the number of observations T to be 500. The runtimes of the main MCMC run and the Chib's reduced-run appear to be linear in n, whereas that of the CE method rises more rapidly than linear. Consequently, the CE method is faster than Chib's for lower dimensions ($n < 30$) but slower for higher dimensions. Here the memory required increases slowly with the model's dimension, from about 3 GB for $n = 5$ to about 3.1 GB for $n = 50$ (as a benchmark, an empty workspace in MATLAB uses about 2.2 GB of memory). All in all, these results show that the proposed method remains computationally feasible for high-dimensional settings.

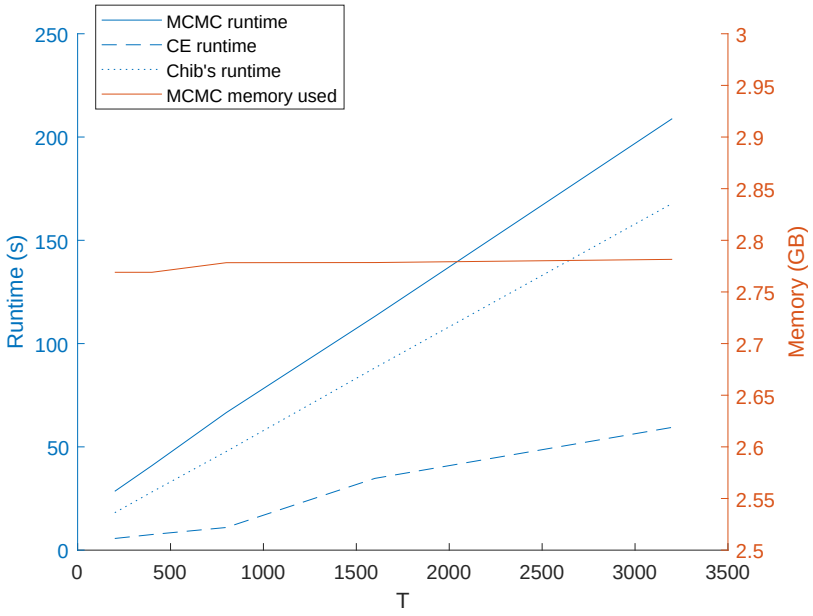

Figure 3: Runtimes (in second) and memory requirements (in GB) of the proposed method for a 3-factor model with $n = 10$.

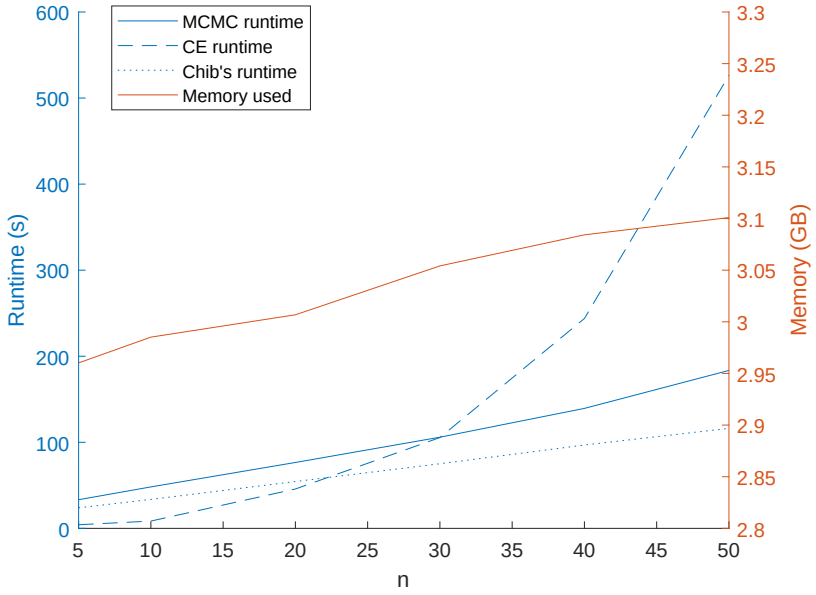

Figure 4: Runtimes (in second) and memory requirements (in GB) of the proposed method for a 3-factor model with $T = 500$.

6 Concluding Remarks and Future Research

We have developed a general method based on Automatic Differentiation to compute the sensitivities of two common marginal likelihood estimators with respect to a set of prior hyperparameters. We have illustrated the methodology using two empirical applications. While in the VAR example the conclusion is robust over a wide range of hyperparameter values, the most preferred number of factors in the factor model application is more uncertain. Our findings, therefore, highlight the importance to conduct a prior sensitivity analysis in Bayesian model comparison routinely.

In future work, it would be useful to develop similar automated prior sensitivity analysis for time-varying models. This is motivated by recent findings that models that allow for time-varying parameters and stochastic volatility, such as those developed in Cogley and Sargent (2001, 2005) and Primiceri (2005), tend to forecast substantially better, as demonstrated in Clark (2011), D'Agostino, Gambetti, and Giannone (2013) and Cross and Poon (2016). Furthermore, AD-based prior sensitivity analysis is instrumental when strong prior information is used, such as estimating dynamic stochastic general equilibrium models.

Appendix A: Integrated Likelihood of the Factor Model

In this appendix we provide an explicit expression for the integrated likelihood factor model in (7). Recall that $\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. By integrating out the factors $\mathbf{f}_t$, we have

$$(\mathbf{y}_t \,|\, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}) \sim \mathcal{N}(\mathbf{X}_t\boldsymbol{\beta}, \mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma}).$$

Evaluating this Gaussian distribution in the conventional way would involve computing the $n \times n$ inverse $(\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma})^{-1}$, which is a time-consuming operation when n is large. As pointed out in Geweke and Zhou (1996), one can avoid this computation problem by using the Woodbury matrix identity:

$$(\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma})^{-1} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{A}(\boldsymbol{\Omega}^{-1} + \mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\Sigma}^{-1}, \tag{9}$$

which only requires computing the $k \times k$ inverse $(\boldsymbol{\Omega}^{-1} + \mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}$.[18] In typical situations where n is much larger than k, the computation saving is substantial. We further improve the efficiency of this approach by vectorizing the operations and by implementing sparse matrix routines.

To that end, we stack the observations over t and write (7) as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_T \otimes \mathbf{A})\mathbf{f} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_T')'$, $\mathbf{f} = (\mathbf{f}_1', \ldots, \mathbf{f}_T')'$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_T')'$ and $\mathbf{X}$ is similarly defined. It follows that unconditional on $\mathbf{f}$, $\mathbf{y}$ is jointly distributed as:

$$(\mathbf{y} \,|\, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_T \otimes (\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma})).$$

Hence, the integrated likelihood (in log) of this model is given by

$$\begin{aligned} \log f(\mathbf{y} \,|\, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}) = &-\frac{Tn}{2}\log(2\pi) - \frac{T}{2}\log|\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma}| \\ &- \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\left(\mathbf{I}_T \otimes (\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma})^{-1}\right)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{10}$$

[18] Note that $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are both diagonal matrices and their inverses are fast to compute.

Appendix B: Additional Results

In this appendix we provide additional empirical results. Table 6 presents the log marginal likelihood estimates of the VAR and VAR-t models with various lag lengths. For all models, the choice of $p = 2$ is best supported by the data.

Table 7: Log marginal likelihood estimates of the VAR and VAR-t models with lag length $p = 1, \ldots, 4$.

	$p = 1$	$p = 2$	$p = 3$	$p = 4$
VAR	-1044.1	-1039.6	-1042.0	-1043.3
VAR-t ($\nu = 5$)	-1020.1	-1013.0	-1016.9	-1020.4
VAR-t ($\nu = 10$)	-1025.8	-1019.7	-1023.3	-1026.0
VAR-t ($\nu = 30$)	-1036.6	-1031.3	-1034.7	-1036.6

Next, Figure 5 plots the standardized residuals of the Gaussian VAR of GDP deflator inflation and real GDP growth. There are a few residuals with absolute values larger than 3, which are highly unlikely under the Gaussian assumption.

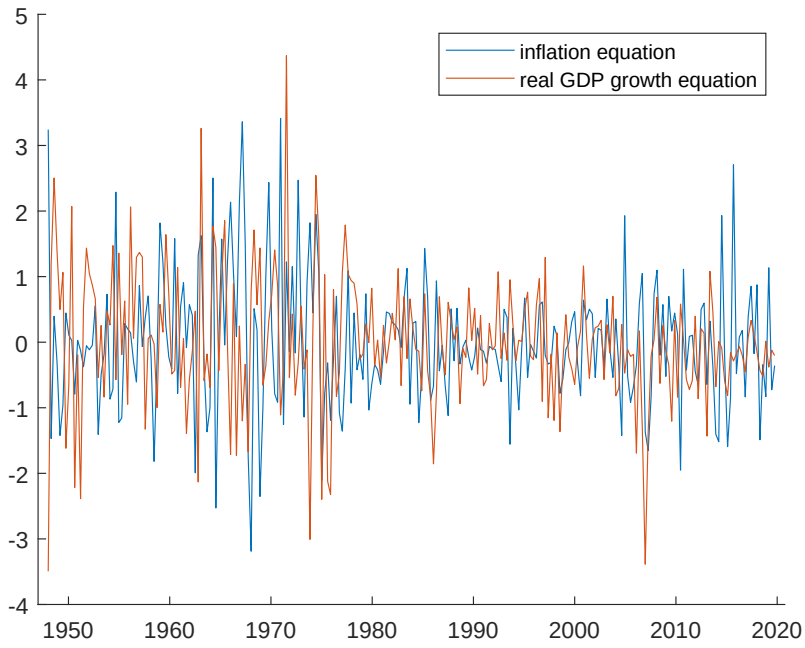

Figure 5: Standardized residuals of the Gaussian VAR of GDP deflator inflation and real GDP growth.

We next report the boxplots of the derivatives of the log marginal likelihood estimate from Chib's method with respect to the posterior ordinates of 3 groups of parameters: the factor loadings $\mathbf{a}$, the idiosyncratic variances $\sigma_1^2, \ldots, \sigma_n^2$ and the factor variances $\omega_1^2, \ldots, \omega_r^2$. As is clear from Figure 6, while the derivatives with respect to the factor loadings $\mathbf{a}$ are small in magnitude, those with respective to the variances are relatively large.

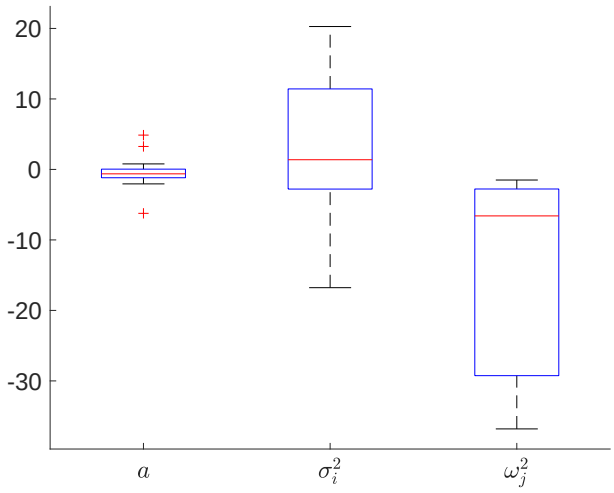

Figure 6: Boxplots of the derivatives of the log marginal likelihood estimate from Chib's method with respect to the posterior ordinates.

References

Aitkin, M. (1991): "Posterior Bayes Factors," *Journal of the Royal Statistical Society Series B*, 53(1), 111–142.

Amisano, G., and R. Giacomini (2007): "Comparing density forecasts via weighted likelihood ratio tests," *Journal of Business and Economic Statistics*, 25(2), 177–190.

Ardia, D., N. Baştürk, L. Hoogerheide, and H. K. van Dijk (2012): "A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood," *Computational Statistics and Data Analysis*, 56(11), 3398–3414.

Banbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25(1), 71–92.

Berger, J. O., D. R. Insua, and F. Ruggeri (2000): "Bayesian robustness," *Robust Bayesian Analysis*, pp. 1–32.

Carriero, A., T. E. Clark, and M. Marcellino (2015): "Bayesian VARs: Specification Choices and Forecast Accuracy," *Journal of Applied Econometrics*, 30(1), 46–73.

Chan, J. C. C. (2020a): "Large Bayesian VARs: A Flexible Kronecker Error Covariance Structure," *Journal of Business and Economic Statistics*, 38(1), 68–79.

——— (2020b): "Large Bayesian Vector Autoregressions," in *Macroeconomic Forecasting in the Era of Big Data*, ed. by P. Fuleky, pp. 95–125. Springer.

Chan, J. C. C., and E. Eisenstat (2015): "Marginal Likelihood Estimation with the Cross-Entropy Method," *Econometric Reviews*, 34(3), 256–285.

——— (2018): "Bayesian Model Comparison for Time-Varying Parameter VARs with Stochastic Volatility," *Journal of Applied Econometrics*, 33(4), 509–532.

Chan, J. C. C., L. Jacobi, and D. Zhu (2019): "How Sensitive Are VAR Forecasts to Prior Hyperparameters? An Automated Sensitivity Analysis," *Advance in Econometrics*, 40(A), 229–248.

Chan, J. C. C., G. Koop, D. J. Poirier, and J. L. Tobias (2019): *Bayesian Econometric Methods*. Cambridge University Press, 2 edn.

Chan, J. C. C., and D. P. Kroese (2012): "Improved Cross-Entropy Method for Estimation," *Statistics and Computing*, 22(5), 1031–1040.

Chib, S. (1995): "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Chib, S., and I. Jeliazkov (2001): "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.

Chib, S., F. Nardari, and N. Shephard (2006): "Analysis of high dimensional multivariate stochastic volatility models," *Journal of Econometrics*, 134(2), 341–371.

Chiu, C. J., H. Mumtaz, and G. Pinter (2017): "Forecasting with VAR models: Fat tails and stochastic volatility," *International Journal of Forecasting*, 33(4), 1124–1143.

Clark, T. E. (2011): "Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility," *Journal of Business and Economic Statistics*, 29(3), 327–341.

Clark, T. E., and F. Ravazzolo (2015): "Macroeconomic Forecasting Performance under alternative specifications of time-varying volatility," *Journal of Applied Econometrics*, 30(4), 551–575.

Cogley, T., and T. J. Sargent (2001): "Evolving post-world war II US inflation dynamics," *NBER Macroeconomics Annual*, 16, 331–388.

——— (2005): "Drifts and volatilities: Monetary policies and outcomes in the post WWII US," *Review of Economic Dynamics*, 8(2), 262–302.

Cross, J., and A. Poon (2016): "Forecasting structural change and fat-tailed events in Australian macroeconomic variables," *Economic Modelling*, 58, 34–51.

Cúrdia, V., M. Del Negro, and D. L. Greenwald (2014): "Rare shocks, great recessions," *Journal of Applied Econometrics*, 29(7), 1031–1052.

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

Del Negro, M., and F. Schorfheide (2004): "Priors from General Equilibrium Models for VARs," *International Economic Review*, 45, 643–673.

Doan, T., R. Litterman, and C. Sims (1984): "Forecasting and conditional projection using realistic prior distributions," *Econometric reviews*, 3(1), 1–100.

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

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

Frühwirth-Schnatter, S. (1995): "Bayesian model discrimination and Bayes factors for linear Gaussian state space models," *Journal of the Royal Statistical Society Series B*, 57(1), 237–246.

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

Gelfand, A. E., and D. K. Dey (1994): "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society Series B*, 56(3), 501–514.

Gelman, A., and X. Meng (1998): "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

Geweke, J. (1993): "Bayesian Treatment of the Independent Student-t Linear Model," *Journal of Applied Econometrics*, 8(S1), S19–S40.

Geweke, J., and G. Zhou (1996): "Measuring the Pricing Error of the Arbitrage Pricing Theory," *The Review of Financial Studies*, 9, 557–587.

Glasserman, P. (2003): *Monte Carlo Methods in Financial Engineering*, vol. 53. Springer Science & Business Media.

Griewank, A., and A. Walther (2008): *Evaluating derivatives: principles and techniques of algorithmic differentiation*, vol. 105. Siam.

Jacobi, L., M. S. Joshi, and D. Zhu (2018): "Automated Sensitivity Analysis for Bayesian Inference via Markov Chain Monte Carlo: Applications to Gibbs Sampling," Available at SSRN: http://dx.doi.org/10.2139/ssrn.2984054.

Jeffreys, H. (1939): *Theory of Probability*. Oxford University Press.

Karlsson, S. (2013): "Forecasting with Bayesian vector autoregressions," in *Handbook of Economic Forecasting*, ed. by G. Elliott, and A. Timmermann, vol. 2 of *Handbook of Economic Forecasting*, pp. 791–897. Elsevier.

Kass, R. E. (1993): "Bayes Factors in Practice," *The Statistician*, 42(5), 551–560.

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

——— (2013): "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics*, 28(2), 177–203.

Koop, G., and D. Korobilis (2010): "Bayesian Multivariate Time Series Methods for Empirical Macroeconomics," *Foundations and Trends in Econometrics*, 3(4), 267–358.

Kroese, D. P., and J. C. C. Chan (2014): *Statistical Modeling and Computation*. Springer, New York.

Lindley, D. V. (1957): "A statistical paradox," *Biometrika*, 44, 187–192.

Litterman, R. (1986): "Forecasting With Bayesian Vector Autoregressions — Five Years of Experience," *Journal of Business and Economic Statistics*, 4, 25–38.

Liu, C. C., and M. Aitkin (2008): "Bayes factors: Prior sensitivity and model generalizability," *Journal of Mathematical Psychology*, 52(6), 362–375.

Lopes, H. F., and M. West (2004): "Bayesian model assessment in factor analysis," *Statistica Sinica*, 14(1), 41–67.

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

O'Hagan, A. (1995): "Fractional Bayes Factors for Model Comparison," *Journal of the Royal Statistical Society Series B*, 57(1), 99–138.

Poirier, D. J. (1988): "Frequentist and subjectivist perspectives on the problems of model building in economics," *Journal of Economic Perspectives*, 2(1), 121–144.

Primiceri, G. E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *Review of Economic Studies*, 72(3), 821–852.

Robert, C., and G. Casella (2013): *Monte Carlo Statistical Methods*. Springer Science & Business Media.

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

Rubinstein, R. Y. (1999): "The cross-entropy method for combinatorial and continuous optimization," *Methodology and Computing in Applied Probability*, 2, 127–190.

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

Schorfheide, F., and D. Song (2015): "Real-Time Forecasting With a Mixed-Frequency VAR," *Journal of Business and Economic Statistics*, 33(3), 366–380.

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